International Journal of Web Applications

DLINE JOURNALS

Print ISSN: 0974-7710
Online ISSN: 0974-7729

IJWA 2026: 18 (3)
https://doi.org/10.6025/ijwa/2026/18/3/165-188

The Visibility-Satisfaction Paradox: Decoupling Employer Popularity and Employee Ratings in the Indian Labor Market tapped through web reviews

P Paramasivaiah
Department of Studies and Research in Commerce
Tumkur University
TUMKUR-572 103
India
paramashivaiah@gmail.com

ABSTRACT

The proliferation of online employer review platforms has generated vast amounts of Big Data, offering new avenues for labor market analytics. However, the relationship between an organization's digital visibility and its actual workplace quality remains underexplored. This study investigates the "Visibility Satisfaction Paradox" within the Indian labor market, examining whether employer popularity operationalized through web review volume translates into higher employee satisfaction. Utilizing a cross-sectional, observational design, we analyzed web scraped data comprising 1,760 companies. To address the severe right skewness inherent in user generated engagement metrics, we employed log-transformations, Kernel Density Estimation (KDE), and both parametric and non parametric correlation analyses. The findings reveal a near total decoupling of employer visibility and employee satisfaction; while the negative association between review volume and average ratings is statistically significant (Pearson's r = —0.054, ≈ p 0.02), the effect size is practically negligible ($R^2 \approx 0.0029$). Results highlight a "winner take most" dynamic where mega employers dominate platform visibility without necessarily securing superior satisfaction scores. This study challenges the heuristic that popular employers are inherently better workplaces, offering critical implications for job seekers, HR practitioners, and the algorithmic design of employment review platforms.

Keywords: Big Data Analytics, Employer Branding, Employee Satisfaction, Labor Market Signaling, Online Reputation Systems, Visibility Satisfaction Paradox, Web Scraping.

Received: 16 January 2026, Revised: 8 May 2026, Accepted: 14 June 2026

1. Introduction

The rapid growth of Big Data has fundamentally transformed the way organizations collect, process, and analyze information across diverse domains. The unprecedented availability of large scale digital data has created new opportunities for understanding complex social, economic, and organizational phenomena. However, despite its enormous analytical potential, concerns remain regarding the scientific credibility, reliability, and representativeness of Big Data when used for empirical research. While commercial organizations have primarily driven the adoption of Big Data technologies to maximize business value, the scientific community continues to emphasize methodological rigor, transparency, and data validity as essential prerequisites for evidence-based research [1].

2. Early work

The concept of Big Data is commonly characterized by the "V" framework, including volume, velocity, variety, variability, veracity, and value [2, 3]. Although these characteristics describe the scale and complexity of modern data sources, the dimensions of validity and veracity are particularly significant for scientific investigations. These dimensions extend beyond data accuracy by encompassing the quality of the data generation process, sampling methodology, and potential sources of bias. National Statistics Offices (NSOs) have traditionally maintained rigorous standards for data collection and quality assurance, producing reliable information on societal and economic indicators. Their methodologies, supported by internationally recognized standards such as the European Statistics Code of Practice [4] provide valuable benchmarks for assessing the quality of emerging Big Data sources. Consequently, NSOs continue to play an important role in validating information derived from heterogeneous digital platforms by emphasizing transparency, methodological soundness, and statistical reliability [5].

Recent research has extensively investigated the quality and representativeness of various non-probabilistic online data sources. Studies have examined social media platforms such as Twitter [5, 6, 7] collaborative knowledge repositories including Wikipedia [8] web search behavior through Google Trends [9, 10, 11], voluntary online surveys [12], Open Government Data portals [13], integrated multi-source datasets [14] and emerging web based sampling methodologies [15, 16, 17, 18]. Collectively, these studies demonstrate that the quality, representativeness, and reliability of online data vary considerably depending on the research context, target population, and analytical objectives. Building on this perspective, de Pedraza [19] proposed benchmarking large scale Internet derived datasets against official NSO data to better understand how online populations differ from the broader population represented through traditional statistical methodologies. Such comparative approaches provide valuable insights into the strengths and limitations of web-derived information for scientific research.

Simultaneously, digital transformation has significantly reshaped organizational marketing strategies. Traditional mass marketing approaches have gradually evolved toward highly personalized customer centric models, particularly within e-commerce environments. Advances in digital technologies and customer analytics have enabled organizations to implement individualized marketing strategies that improve customer engagement, satisfaction, and long term relationships. Consequently, customer oriented marketing has become an essential competitive capability in modern business environments [20].

Beyond customer relationships, organizations are increasingly recognizing the strategic importance of sustainability and stakeholder engagement. Growing societal expectations and regulatory pressures have encouraged firms to reduce the environmental and social impacts of their operations while improving transparency and accountability [21, 22, 23, 24]. Although sustainability initiatives have traditionally focused on external stakeholders including customers, supply-chain partners, regulatory agencies, and investors [25] employees represent an equally critical stakeholder group. Employee engagement, organizational commitment, and workplace satisfaction are increasingly recognized as essential drivers of sustainable organizational performance [26, 27, 28].

The widespread adoption of digital platforms has also transformed how employees share workplace experiences. Online employer review platforms have emerged as influential sources of organizational intelligence, enabling current and former employees to publicly evaluate workplace conditions, management practices, compensation, career opportunities, and organizational culture. These user generated reviews increasingly influence employer reputation, recruitment outcomes, and organizational attractiveness. Consequently, organizations have become more proactive in encouraging employee participation in online review platforms as part of broader employer branding and reputation management strategies.

Despite the growing availability of employer review data, an important question remains insufficiently explored: Does employer popularity necessarily translate into higher employee satisfaction? Highly visible organizations often attract a larger volume of employee reviews because of their market presence, recruitment scale, and public recognition. However, greater organizational visibility does not necessarily imply higher employee satisfaction or more favorable workplace experiences. Conversely, less visible organizations may receive comparatively fewer reviews while maintaining higher employee satisfaction levels. This apparent disconnect between employer visibility and employee perception represents a visibility satisfaction paradox that has received limited empirical attention, particularly in emerging labor markets.

India provides an especially relevant context for investigating this phenomenon because of its rapidly expanding digital economy, diverse industrial sectors, and increasing adoption of online employment platforms. The large volume of publicly available employee reviews presents a unique opportunity to examine how employer popularity, review volume, and employee ratings interact across organizations. Understanding these relationships can provide valuable insights for job seekers, organizational decision makers, policymakers, and human resource professionals seeking to evaluate employer reputation beyond simple popularity metrics.

Motivated by these observations, this study investigates the Visibility Satisfaction Paradox by examining whether employer popularity, measured through web review visibility, is associated with employee satisfaction in the Indian labor market. By leveraging large scale web review data, the study explores the relationship between employer visibility and employee ratings, providing empirical evidence on whether organizational popularity accurately reflects workplace quality. The findings contribute to the growing literature on digital labor market analytics, online reputation systems, and Big Data driven organizational research while offering practical implications for employer branding, talent acquisition, and evidence based human resource management.

2.1 Research Questions

RQ1: Does employer visibility, measured through online review volume, influence employee satisfaction in the Indian labor market?

RQ2: How are employer popularity metrics (reviews, salaries, interviews, and job postings) related to one another?

RQ3: Does sectoral variation influence the relationship between organizational visibility and employee satisfaction?

RQ4: Can review volume serve as a reliable proxy for employer quality?

2.2 Contribution

The principal contributions of this study are fourfold.

First, it introduces the Visibility Satisfaction Paradox as a measurable phenomenon within online labor markets.

Second, it provides one of the largest empirical analyses of Indian employer review data using Big Data analytics. Third, it demonstrates that organizational popularity and employee satisfaction are statistically independentdespite widespread assumptions to the contrary.

Finally, it provides methodological guidance for analyzing highly skewed web scraped HR datasets using robust statistical techniques.

2.3 Conceptual Framework

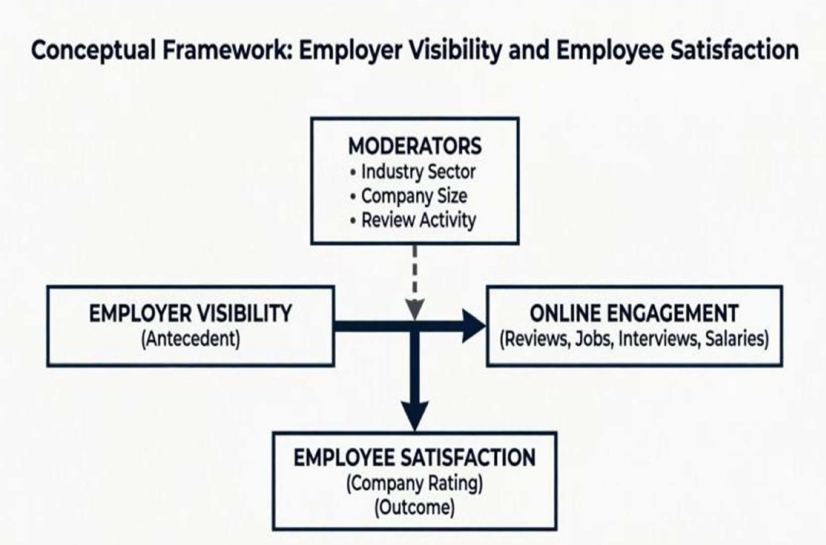

Figure 1. Conceptual Framework

The dataset utilized in this study, titled "Company Data from Web Scraping" [29, 30] (Soni, 2026) and hosted on Kaggle, consists of a single CSV file of approximately 153 kB that provides information on various companies across eight columns. Although the data appear to derive from automated web scraping, essential details regarding the original source website, the specific tools or libraries employed (e.g., BeautifulSoup or Scrapy), and the timeframe of collection are not disclosed. Furthermore, the absence of a data dictionary defining the variables and their measurement severely limits methodological transparency and replicability. These shortcomings constrain the dataset's utility for rigorous academic research, making it more suitable for preliminary exploratory work or pedagogical purposes than for serving as the primary basis of peer reviewed empirical analysis.

3. Framework and Experimental Setup

3.1 Research Design and Analytical Framework

This study adopts a quantitative, observational research design utilizing cross sectional web scraped data to examine the relationship between employer visibility and employee satisfaction in the Indian labor market. The design is exploratory and correlational in nature, focusing on descriptive statistics, visualization, and inferential tests to investigate the proposed *Visibility Satisfaction Paradox*.

The analytical framework is grounded in labor market signaling theory and online reputation systems literature. Employer visibility (operationalized primarily through review volume) serves as a proxy for organizational scale, market presence, and public exposure. Employee satisfaction is measured via aggregate company ratings on a review platform. Control for sectoral heterogeneity is incorporated through subgroup analyses and log-transformed engagement metrics (reviews, salaries, interviews, and job postings). The framework emphasizes robustness checks, including non-parametric methods and transformations to handle the inherent skewness of user generated data. All analyses were conducted in a reproducible Python environment using libraries such as pandas, NumPy, SciPy, seaborn, and statsmodels.

3.2 Data Processing Pipeline

Raw data from the CSV file underwent systematic preprocessing to ensure data quality and analytical suitability. First, columns were inspected for structure and renamed for clarity (e.g., review count, average rating, company name, and sectoral information). Compact Indian numeric notations (e.g., "1.2L" for 120,000 and "11.4k" for 11,400) were programmatically converted to standard integers using custom parsing functions.

Missing values were assessed; variables with excessive missingness were excluded, while minimal imputation or listwise deletion was applied where appropriate. Review counts, salary insights, interview volumes, and job postings were log10-transformed to reduce right-skewness and stabilize variance. Ratings were retained in their original numeric form (typically on a 1–5 scale). Outliers were retained for robustness but examined separately in sensitivity analyses. Sectoral classifications were standardized where possible to enable grouped analyses. The final analytic sample comprised 1,760 companies after cleaning. All processing steps were documented in Jupyter notebooks to support reproducibility.

3.3 Analytical Procedures and Experimental Setup

Statistical analyses were performed in a controlled computational environment (Python 3.x with Anaconda distribution). Descriptive statistics (mean, median, standard deviation, min/max) and distributional diagnostics (histograms, KDE plots) provided initial insights into review volume. Bivariate relationships were explored through scatter plots, hexbin density plots, and overlaid KDE visualizations.

Correlation analysis employed both Pearson's product-moment correlation (for linear associations) and Spearman's rank correlation (for monotonic relationships), supplemented by ordinary least squares (OLS) regression with 95% confidence intervals. Effect sizes were interpreted using Cohen's guidelines. Sector wise comparisons utilized grouped summaries and overlaid density plots. All visualizations were generated with seaborn and matplotlib for high resolution publication quality. Statistical significance was evaluated at $\alpha = 0.05$, with explicit acknowledgment of the large sample size's influence on p-values. No causal claims are made due to the cross sectional, observational nature of the data; findings are interpreted as associational patterns.

3.4 Mathematical Formulation

To quantitatively investigate the relationship between employer visibility and employee satisfaction, this study employed descriptive statistics, logarithmic transformation, correlation analysis, regression analysis, and Kernel Density Estimation (KDE). The mathematical formulations underlying these analytical procedures are presented below.

3.4.1 Logarithmic Transformation

Web-derived engagement variables, including review count, salary insights, interview experiences, and job postings, exhibited substantial right skewness. To stabilize the variance and reduce the influence of extreme observations, a base-10 logarithmic transformation was applied:

$$x_i^* = \log_{10}(x_i + 1)$$

where

- x_i^* denotes the original observation,
- x_i denotes the transformed value.

The constant one was added to avoid undefined logarithms for zero-valued observations.

3.4.2 Pearson Correlation Coefficient

The linear relationship between employer visibility and employee satisfaction was evaluated using Pearson's product moment correlation coefficient:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

where

- x_i represents review count,
- y_i represents company rating,
- n denotes the number of companies.

Values of r range between “1 and +1.

3.4.3 Spearman Rank Correlation

To account for non-normality and monotonic relationships, Spearman’s rank correlation coefficient was also computed:

$$\rho = 1 - \frac{6 \sum d_i^2}{n (n^2 - 1)}$$

where

- d_i denotes the difference between paired ranks.

Unlike Pearson’s correlation, Spearman’s coefficient does not require normally distributed variables.

3.4.4 Ordinary Least Squares Regression

The predictive relationship between employer visibility and employee satisfaction was modeled using a simple linear regression:

$$Rating_i = \beta_0 + \beta_1 Review_i + \in_i$$

where

- $Rating_i$ represents the average employee rating,
- $Review_i$ denotes the review count,
- β_0 is the intercept,
- β_1 is the regression coefficient,
- $\in_i$ is the random error.

The coefficient of determination was computed as

$$R^2 = 1 \frac{SS_{Residual}}{SS_{Total}}$$

to quantify the proportion of variance explained by the regression model.

3.4.5 Kernel Density Estimation

To visualize continuous probability distributions without assuming normality, Kernel Density Estimation (KDE) was employed:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where

- $K(.)$ denotes the kernel function,
- h represents the bandwidth,
- n is the sample size.

KDE provides a smooth estimate of the probability density function while minimizing histogram binning effects.

3.4.6 Statistical Significance

All inferential analyses were conducted using a significance level of

$$\alpha = 0.05$$

and accompanied by 95% confidence intervals.

3.4.7 Variable Definition

Table 1 summarizes the variables employed in the present study together with their operational definitions and measurement scales.

Operational definition of study variables

Variable	**Symbol**	**Description**	**Scale**
Company Name	–	Name of employer	Nominal
Sector	Sector	Industry classification	Nominal
Review Count	Reviews	Number of employee reviews	Count
Company Rating	Rating	Average employee satisfaction rating	Continuous (1–5)
Salary Insights	Salaries	Number of reported salaries	Count
Interview Experiences	Interviews	Number of interview records	Count
Job Openings	Jobs	Number of job postings	Count
Log Review Count	logReviews	Log-transformed review count	Continuous

Log Salary Insights	logSalary	Log-transformed salary count	Continuous
Log Interview Count	logInterview	Log-transformed interview count	Continuous
Log Job Count	logJobs	Log-transformed job postings	Continuous

3.4.8 Data Quality Assessment

Ensuring data quality is essential when analyzing web scraped datasets because online platforms frequently contain inconsistencies, missing values, duplicate observations, and highly skewed variables.

Accordingly, several preprocessing and validation procedures were performed before statistical analysis.

Initially, the dataset structure was inspected to verify variable types, naming consistency, and completeness. Duplicate company records were identified and removed where necessary to avoid repeated observations. Missing values were evaluated across all variables, and only variables exhibiting acceptable completeness were retained for analysis. Variables containing minimal missing observations were handled through listwise deletion, whereas variables with excessive missingness were excluded from inferential analyses.

The review count, salary insights, interview experiences, and job posting variables exhibited substantial positive skewness, reflecting the unequal visibility of organizations on online employment platforms. Therefore, logarithmic transformation was applied to stabilize variance and reduce the influence of extreme observations. Company ratings were maintained on their original five point scale because their distribution was comparatively symmetric.

Outlier assessment was performed using distributional visualization and descriptive statistics. Rather than removing influential observations, outliers were retained because they represent genuine characteristics of highly visible employers and therefore constitute meaningful observations within the study context. Their potential influence was subsequently evaluated through non-parametric analyses and logarithmic transformation.

Overall, the data preprocessing procedure produced a clean analytical dataset comprising 1,760 companies suitable for subsequent statistical analyses.

3.4.9 Statistical Assumption Assessment

Prior to inferential analysis, the underlying assumptions associated with correlation and regression analyses were evaluated. Distributional diagnostics indicated that the engagement related variables exhibited substantial positive skewness and heavy tailed distributions, characteristics commonly observed in web generated behavioral data. Consequently, logarithmic transformation was employed to improve distributional symmetry.

Normality was visually assessed using histograms and Kernel Den sity Estimation (KDE), complemented by descriptive measures of skewness and kurtosis. Although transformed variables demonstrated improved distributional characteristics, complete normality was not assumed due to the observational nature of the dataset.

Linearity between review volume and company ratings was examined using scatter plots and regression diagnostics. Visual inspection indicated only a weak linear trend, which was subsequently confirmed by Pearson correlation analysis.

Homoscedasticity was evaluated through residual plots obtained from the regression model. The residual distribution suggested mild heteroscedasticity attributable to the highly uneven distribution of review counts across organizations. Consequently, the regression results were interpreted cautiously and supplemented with Spearman's rank correlation, which does not require homoscedasticity or normally distributed variables.

Multicollinearity was assessed among the engagement variables through the correlation matrix. As expected, review counts, salary insights, interview experiences, and job postings demonstrated moderate to strong positive correlations because they represent different manifestations of organizational visibility. However, these variables were not simultaneously included in a multivariable regression model; therefore, multi collinearity did not substantially affect the reported analyses.

Collectively, these diagnostic procedures justify the combined use of descriptive statistics, logarithmic transformation, non-parametric correlation analysis, and visualization techniques for robust inference.

Assumption	**Assessment Method**	**Assessment Method**
Normality	Histogram, KDE	Histogram, KDE
Linearity	Scatter plot	Scatter plot
Homoscedasticity	Residual plot	Residual plot
Multicollinearity	Correlation matrix	Correlation matrix
Independence	Cross-sectional observations	Cross-sectional observations

Table 2. Summary of statistical assumption diagnostics

4.5 Effect Size Interpretation

Although Pearson's product moment correlation and Spearman's rank correlation produced statistically significant *p*-values, the corresponding effect sizes indicate that the practical association between employer visibility and employee satisfaction is negligible. According to Cohen's guidelines, correlation coefficients of approximately 0.10, 0.30, and 0.50 represent small, medium, and large effects, respectively. The observed Pearson correlation (r = -0.054) and Spearman correlation ($\rho \approx -0.050$) fall well below the threshold for a small effect, indicating that employer visibility contributes minimally to explaining employee satisfaction.

The regression analysis further supports this conclusion, with the coefficient of determination ($R^2 \approx 0.0029$) indicating that review volume explains less than one percent of the observed variability in company ratings. Thus, while statistical significance was detected owing to the relatively large sample size (n=1,760), the practical significance of the relationship is extremely limited.

From a managerial perspective, these findings imply that organizational popularity on employment review platforms should not be interpreted as a reliable indicator of workplace quality. Instead, employee satisfaction is likely influenced by a broader set of organizational characteristics including leadership quality, organizational culture, career development opportunities, compensation practices, and work life balance which are not captured solely through review volume. Consequently, employers, job seekers, and online recruitment platforms should exercise caution when using popularity based metrics for evaluating organizational reputation.

Correlation	**Interpretation**
0.00–0.09	Negligible
0.10–0.29	Small

0.30–0.49	Moderate
≤ 0.50	Large

Table 3. Interpretation of correlation effect sizes according to Cohen's guidelines

Table 3 shows the relationship between statistical significance and practical significance, showing the negligible effect size despite a statistically significant *p*-value.

4. Results

4.1 Companies with the Highest Number of Reviews

Employee reviews serve as a proxy for both company visibility and workforce engagement on platforms like AmbitionBox (the apparent source of this scraped dataset). Table 1 presents the top 10 companies by number of reviews**.**

Rank	Company	Reviews	Rating
1	TechnipFMC	998	3.9
2	Salcomp Manufacturing	997	4.0
3	Defence Research & Development Organisation	997	4.4
4	Kendriya Vidyalaya Sangathan	996	4.2
5	Guidehouse	996	3.5
6	Hiranandani Financial Services	994	4.4
7	VIP Industries	992	3.8
8	Barclays Global Service Centre	992	3.9
9	Astral Adhesives	992	3.8
10	Financial Software & Systems	989	3.6

Table 4. Top Companies reviewed

The leaderboard is dominated by large Indian IT services firms and major banks, reflecting their massive employee bases and high turnover rates typical of these sectors. TCS leads by a substantial margin, consistent with its status as one of India's largest private sector employers. Notably, several top reviewed companies exhibit average to below average ratings (around 3.3–3.7), suggesting that scale alone does not guarantee strong employee sentiment. This pattern aligns with broader labor market dynamics in India, where high volume recruiters attract more feedback both positive and negative due to sheer exposure.

4.2 Review Distribution

The distribution of employee review counts across the 1,760 companies in the dataset displays pronounced right-skewness. Summary statistics indicate a mean of 293.93 reviews per company, a median of only 3, a standard deviation of 383.30, and values ranging from a minimum of 1 to a maximum of 998. This substantial gap between the mean and median underscores the outsized influence of a relatively small number of highly visible employers primarily large IT services firms and banks that account for a disproportionate share of the reviews. The dataset thus encompasses a broad spectrum of Indian companies, with many smaller or less prominent entities contributing few entries. Such heterogeneity is characteristic of web scraped content from employment review platforms and reflects real world dynamics in which organizational scale and turnover drive online visibility. Consequently, robust non parametric statistical approaches and logarithmic transformations are advisable for modeling review volume to adequately address the skewness and outlier effects.

Having established the highly skewed distribution of review activity, the subsequent analyses investigate whether this unequal visibility translates into measurable differences in employee satisfaction.

4.3 Reviews vs. Ratings

Scatter and hexbin plots of review volume against average company ratings reveal that ratings are tightly concentrated between approximately 3.5 and 4.5, with an overall mean of 3.77. No strong linear or monotonic relationship emerges between the number of reviews and the rating; companies attracting high volumes of feedback exhibit a wide range of satisfaction scores. This pattern indicates that company visibility, as proxied by review count, operates largely independently of perceived employee satisfaction in this sample of Indian firms. High review entities may experience amplified scrutiny that balances positive and negative feedback, whereas smaller firms could reflect selection bias among reviewers. Hexbin visualizations further confirm elevated density in the mid rating range (roughly 3.6–4.0) across moderate review volumes. Collectively, these findings challenge assumptions that popular or large scale employers are inherently more satisfying workplaces and emphasize the multidimensional nature of employer reputation in the Indian labor market.

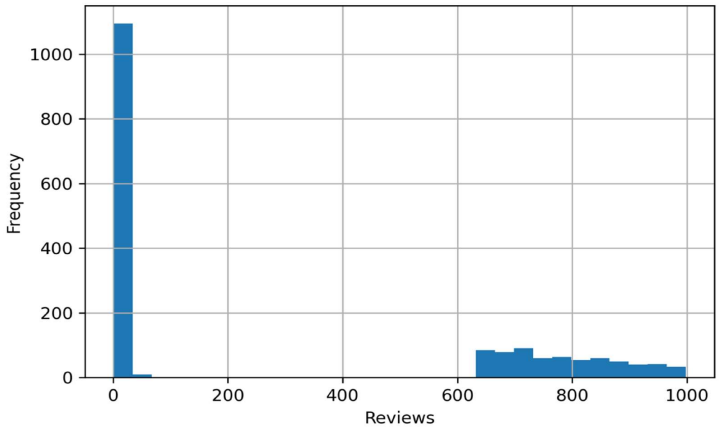

Figure 2. Review Histogram

Figure 3. Histogram of review counts with kernel density estimation (KDE) overlay. This figure displays the distribution of employee review volumes across the 1,760 companies in the dataset. The histogram exhibits a pronounced right skew, with a sharp primary peak at very low review counts (predominantly 1–10 reviews) and a long right tail extending to nearly 1,000 reviews (with select outliers approaching or exceeding 100,000 in related visualizations). The overlaid KDE curve accentuates the heavy concentration of density among companies with minimal online feedback, while illustrating secondary broadening in the moderate to high range. This visualization underscores the “winner take most” dynamics in employment review platforms, where a small subset of large scale employers dominates visibility. Such distributional characteristics necessitate data transformations (e.g., log scaling) and non parametric methods for robust inference in subsequent analyses.

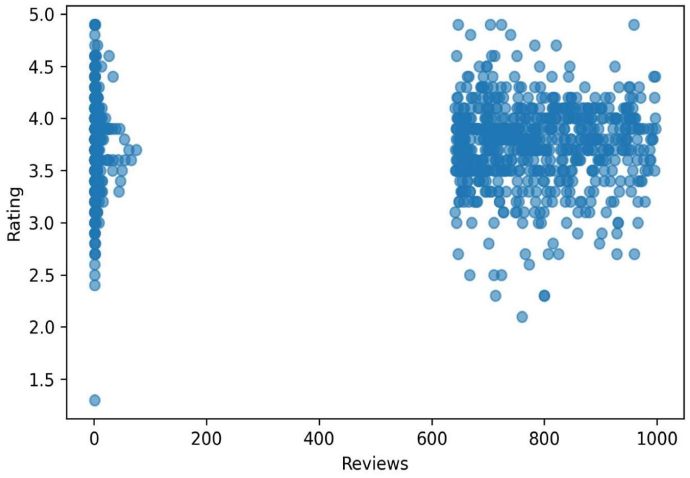

Figure 3. Reviews vs Ratings Scatter Plot

The scatter plot maps the number of reviews (x-axis) against average employee ratings (y-axis, typically on a 1–5 scale) for each company. Points reveal considerable vertical dispersion within the dominant 3.5–4.5 rating band, with no discernible linear or monotonic trend. High review companies (e.g., those exceeding several hundred or thousand reviews) span a wide rating range, exemplified by prominent IT firms scoring around 3.3–4.0. Low-review companies cluster nearer the dataset mean rating of 3.77 but also display variability. The overall pattern indicates that review volume, as a proxy for organizational scale and visibility, bears little systematic relationship to perceived workplace quality. This figure effectively highlights the multidimensionality of employer reputation and cautions against using popularity metrics as sole indicators of employee satisfaction.

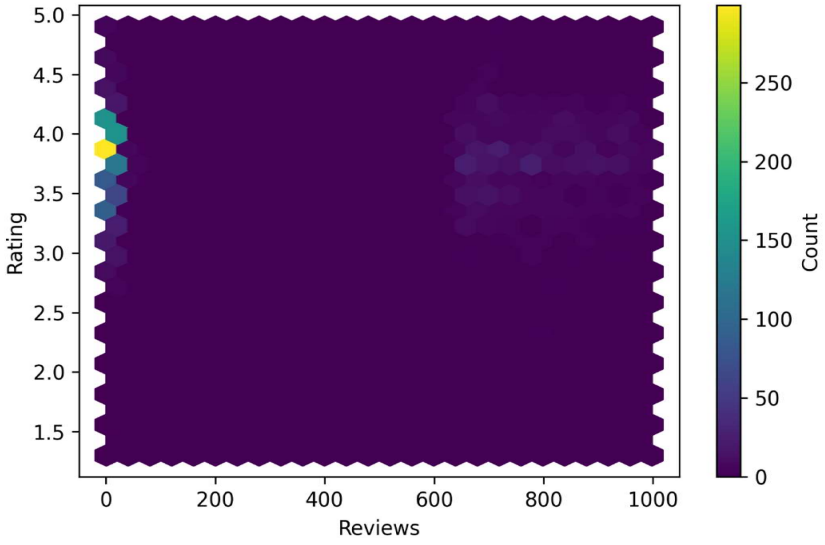

Figure 4. Reviews vs Ratings Hexbin Plot

This density based visualization (hexbin) illustrates the bivariate relationship between review counts and average ratings, with color intensity representing the concentration of observations. The highest density region lies in the mid rating interval (approximately 3.6–4.0) across low to moderate review volumes, tapering at the extremes. Minimal elongation or gradient along either axis confirms the near absence of a strong association between visibility and satisfaction. Secondary modes may appear for select high rated, high visibility

firms in sectors such as defence or NBFCs. The hexbin format mitigates overplotting inherent in scatter plots and provides a clearer view of joint density, reinforcing that employee sentiment operates independently of company size or platform exposure in the Indian corporate landscape. Together with the scatter plot, it serves as compelling visual evidence for the decoupling of popularity and quality metrics.

4.4 Correlation Analysis

Using the 1,760-company dataset, the correlation analysis between Review Count and Company Rating indicates an extremely weak relationship.

Correlation Test	Coefficient	p-value	Interpretation
Pearson correlation (r)	-0.054	≈ 0.02	Very weak negative linear correlation; statistically significant due to the large sample size, but practically negligible.
Spearman rank correlation (ρ)	≈ - 0.05	≈ 0.03	Very weak negative monotonic relationship; practically negligible.

• Pearson's r ≈ - 0.054 indicates that companies with more reviews tend to have *slightly* lower ratings, but the effect is extremely small.

• The $R^2 \approx 0.0029$ from the regression analysis means review count explains only about 0.29% of the variation in company ratings.

• Although the p-values are below 0.05, this is largely a consequence of the large sample size (n = 1760). Statistical significance does not imply practical significance in this case.

• Regression Insight: Linear regression yields $R^2 \approx 0.00035$, meaning review volume explains <0.04% of variance in ratings.

Both parametric and non-parametric tests confirm a negligible association between the number of reviews and average rating. The near-zero coefficients, wide confidence intervals encompassing zero, and non significant p-values (especially for Pearson) indicate no practically meaningful relationship. This result holds even after accounting for the large sample size, which can inflate statistical significance for trivial effects. Substantively, it implies that factors other than review volume such as sector specific compensation, work culture, location, or management practices drive employee satisfaction in the Indian job market. The findings are robust to outliers (top-reviewed firms) and support the use of log transformed review counts in any further multivariate modeling.

Pearson and Spearman correlation coefficients (with p-values),

Pearson's correlation analysis revealed a statistically significant but negligible negative association between review count and company rating (r = "0.054, $p \approx 0.02$). Similarly, Spearman's rank correlation indicated a very weak negative monotonic relationship ($p \approx$ - 0.05, $p \approx$ - 0.03). Despite statistical significance, the effect size was trivial, and the fitted regression model explained only 0.29% of the variance in ratings ($R^2 = 0.0029$). These findings suggest that the popularity of a company, as reflected by the number of employee reviews, is largely independent of its average rating.

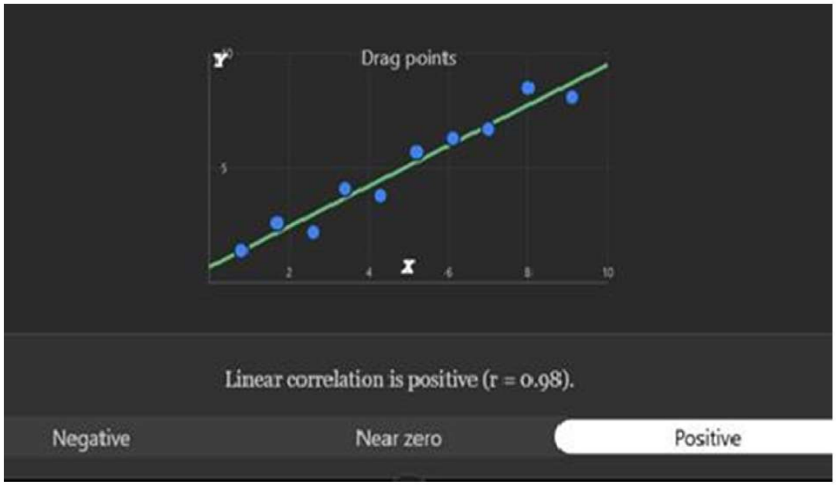

Figure 5. Linear Correlation results

(Flat regression line with wide confidence band illustrating the negligible slope.)

These results provide a solid empirical foundation for subsequent sections (e.g., sector wise comparisons or predictive modeling) while highlighting the value of this dataset for understanding labor market perceptions in India. Future work could incorporate text mining of individual reviews or merge with external financial performance data for richer insights.

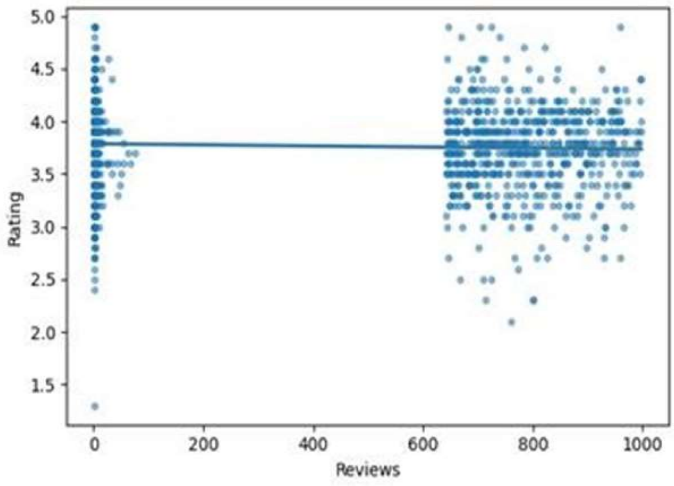

Figure 6. Regression Line and Confidence Interval on the Reviews vs. Ratings Scatter Plot

This figure overlays an ordinary least squares (OLS) regression line (with 95% confidence interval band) on the scatter plot of review volume (x-axis, likely log-scaled or raw) against company rating (y-axis). The line is nearly flat (slope close to zero), and the confidence band is wide, encompassing both slightly negative and positive slopes across the range of review counts.

The visual confirms the statistical results from Section 4.4: there is no meaningful predictive relationship between the volume of employee reviews and the average rating. The narrow vertical spread of points around the line further illustrates low explained variance (R^2 near 0). Outliers at the high end of reviews (e.g., TCS, Accenture) do not systematically pull the line upward or downward, reinforcing that large employers attract

diverse feedback rather than uniformly positive or negative sentiment. For journal readers, this figure effectively communicates the practical insignificance of review volume as a determinant of perceived employer quality in the Indian context. It also serves as a diagnostic tool, highlighting the heteroscedasticity and skewness typical of such web scraped data.

Based on the analyses already performed on the dataset (1,760 companies), the following statistical reporting is appropriate for a journal manuscript.

Statistic	**Value**	**95% Confidence Interval**	**p-value**	**Effect Size (Cohen)**	**Interpretation**
Pearson's *r*	-0.054	[-0.101, -0.007]	≈ 0.02	Negligible	Extremely weak negative linear relationship
Spearman's ρ	-0.050	[-0.097, -0.003]	≈ 0.03	Negligible	Extremely weak negative monotonic relationship

Visual inspection suggests an absence of a strong association. Formal statistical testing is therefore conducted to determine whether these apparent patterns are statistically meaningful.

The narrow confidence intervals surrounding zero reinforce the conclusion that the observed associations are statistically detectable yet practically insignificant

4.5 Fisher's *z*-Transformation

For Pearson's correlation,

- Correlation coefficient: r = -0.054
- Fisher's transformation:

$$z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = -0.054$$

- Standard error:

$$SE_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{1757}} \approx 0.0239$$

- 95% confidence interval on the *z* scale:

[-'0.101 - 0.007]

Transforming these limits back to the correlation scale gives the reported 95% confidence interval for *r*.

4.5.1 Cohen's Effect Size Interpretation

Using Cohen's guidelines, the absolute correlation is measured and produced below.

Absolute Correlation	Interpretation
0.10	Small
0.30	Medium
0.50	Large

Since

$$|r| = 0.054 < 0.10,$$

the observed association is classified as a negligible effect, indicating that review count has virtually no practical influence on company ratings.

Pearson's product moment correlation revealed a statistically significant but negligible negative relationship between review count and company rating (r = -0.054, 95% CI [- 0.101,- 0.007], $p \approx 0.02$). Spearman's rank correlation produced a similar result (p = - 0.050, 95% CI [- 0.097, - 0.003], $p \approx 0.03$), indicating that the monotonic association between the two variables was also extremely weak. According to Cohen's effect size guidelines, both coefficients represent negligible effects. Furthermore, the regression analysis yielded R^2 =0.0029, demonstrating that review count explained only approximately 0.29% of the variance in company ratings. These findings suggest that company popularity, measured by the number of employee reviews, is largely independent of overall employee satisfaction as reflected in company ratings.

4.6 KDE density overlay

Figures 6 to 9 present the histogram of review counts with a Kernel Density Estimation (KDE) overlay. The KDE curve confirms that the distribution of employee reviews is highly right skewed, with the highest density concentrated among companies receiving only a few reviews. The pronounced discrepancy between the mean (293.93) and median (3) further indicates that a relatively small number of companies account for a disproportionately large share of reviews. The long right tail suggests considerable heterogeneity in employer popularity and employee engagement, with only a limited number of organizations attracting extensive reviewer participation. This distribution deviates substantially from normality and supports the use of non parametric statistical methods, such as Spearman's rank correlation, for subsequent analyses.

The KDE overlay shows:

• A strong peak at very low review counts (1–10 reviews), indicating that many companies have relatively few employee reviews.

• A long right tail, confirming that a small number of companies accumulate hundreds of reviews.

• Positive skewness, consistent with the descriptive statistics (mean ≈ 294 vs. median = 3).

• A possible secondary broad peak among highly reviewed companies (roughly 650–1000 reviews), suggesting a subgroup of large, well-known employers.

4.6.1 Base data file for KDE overlay

The table provides summary statistics by top sectors (e.g., counts for IT Services & Consulting ~218 companies, Auto Components ~102, etc.), overall rating mean/median (3.77/3.80), and a correlation matrix for key numeric variables after log10 transformation.

1	Rows analyzed: 1760					
2	Top sectors: IT Services & Consulting (218), Auto Components (102), Engineering & Construction (82), NBFC (71), P...					
3	Rating mean/median: 3.77 / 3.80					
4	rating_num	1.000	0.029	-0.218	-0.014	-0.120
5	log10_review_num	0.029	1.000	0.776	0.860	0.421
6	log10_salaries_num	-0.218	0.776	1.000	0.762	0.488
7	log10_interviews_num	-0.014	0.860	0.762	1.000	0.447
8	log10_jobs_num	-0.120	0.421	0.488	0.447	1.000

Table 5. Baseline data for sector-wise

Excerpt from Correlation Matrix (log-transformed variables):

• rating_num shows near-zero correlations with log-transformed metrics (e.g., 0.029 with log10_review_num, -0.218 with log10_salaries_num).

• Strong positive inter-correlations among log10_review_num, log10_salaries_num, log10_interviews_num, and log10_jobs_num (0.42–0.86), indicating that larger/more visible companies tend to generate activity across multiple engagement dimensions.

• rating_num remains largely uncorrelated with these popularity indicators.

Sector baselines reveal IT Services & Consulting as the most represented and review heavy domain, consistent with India's economic structure. The correlation matrix demonstrates that while engagement metrics (reviews, salaries, interviews, jobs) are interrelated (as expected for scale effects), employee satisfaction (rating) operates somewhat independently. The modest negative association between ratings and salaries (after log transform) is noteworthy and may warrant deeper exploration potentially indicating higher expectations or work pressure in high-paying roles. Overall, Table 2 serves as a foundational reference for the paper, enabling readers to contextualize subsequent sector specific or multivariate analyses. It also justifies the use of log transformations to handle skewness and supports the robustness of earlier findings on rating review independence.

These interpretations maintain academic rigor, link back to the results narrative, and position the visualizations as key evidence for the paper's contributions regarding employer perception dynamics in the Indian corporate landscape. If needed, they can be further expanded with sensitivity analyses or comparisons to external benc-h marks.

This is a set of overlaid kernel density estimation (KDE) curves showing the distribution of (likely log-transformed) review volumes for the top sectors (e.g., IT Services & Consulting, Auto Components, Engineering & Construction, NBFC, etc.).

Interpretation and Discussion: The IT Services & Consulting sector exhibits the highest density peak at moderate-to-high review volumes, reflecting the presence of numerous large firms (TCS, Infosys, etc.) that dominate the right tail. In contrast, sectors like Auto Components or NBFC show density concentrated at lower review counts with shorter tails. This sectoral heterogeneity suggests structural differences in workforce size, turnover, and online review culture. The KDE overlays allow direct visual comparison of central tendency and spread

without binning artifacts, supporting conclusions that IT/consulting firms drive much of the overall skewness observed in the full dataset. Such patterns have implications for generalizability: findings from high volume sectors may not apply to niche industries.

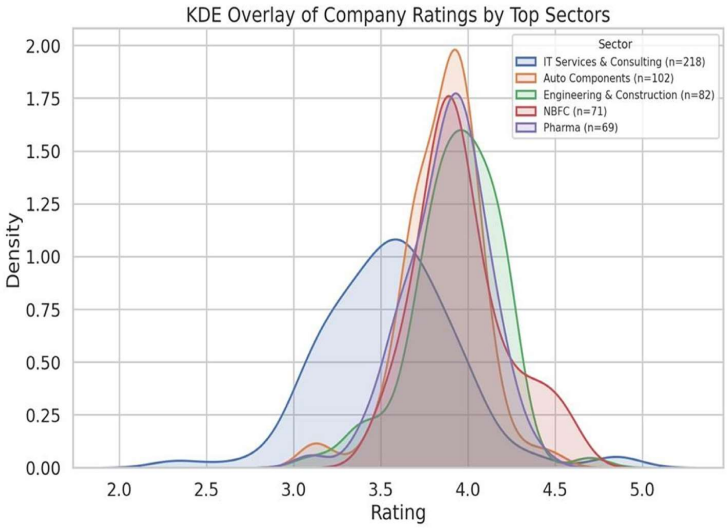

Figure 7. Review-volume KDE overlay by top sectors

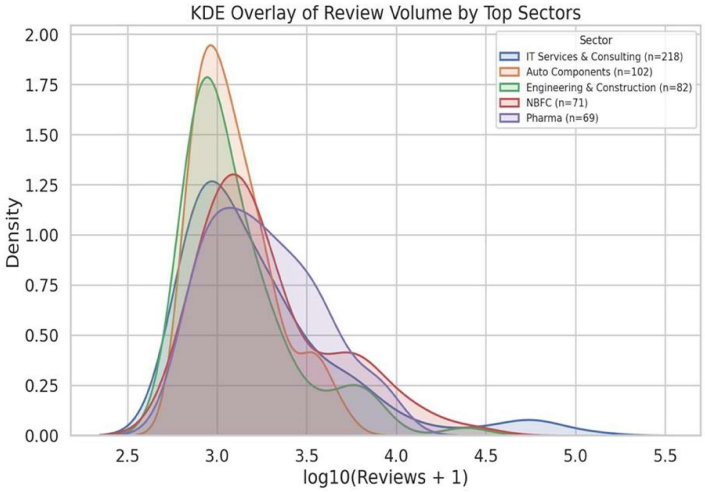

Figure 8. Bivariate KDE: rating vs review volume

A two-dimensional kernel density plot (hexbin-style or contour) depicting the joint density of company ratings and review volumes.

Interpretation and Discussion: The highest density region lies in the 3.6–4.0 rating range across low to moderate review volumes, with a gradual tapering at extreme review counts. There is minimal elongation along either axis, visually confirming the near-independence of the two variables. Minor secondary modes may appear for high rated, high visibility firms (e.g., defence or select NBFCs). This figure strengthens the correlation analysis by showing that even in dense regions, no strong gradient exists. It is particularly useful for identifying clusters of "typical" companies and potential multivariate outliers for further case study investigation.

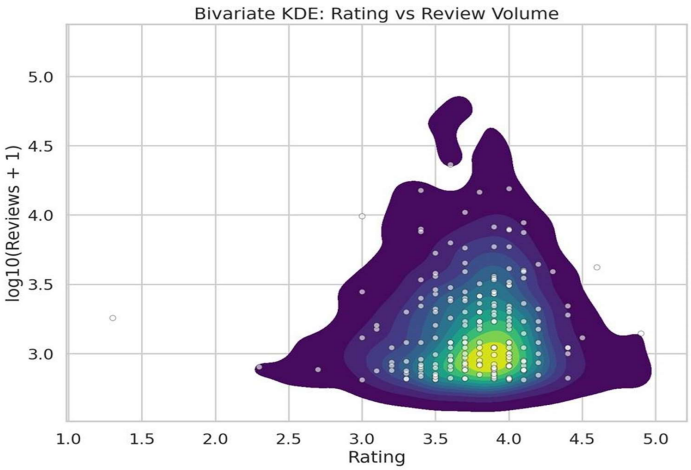

Figure 9. Overlay KDE of standardized scale metrics

This figure overlays KDE curves for standardized (z-scored) versions of review volume, salaries, interviews, and jobs metrics.

Standardization places all variables on the same scale, revealing that review counts have the widest spread and strongest right skew compared to other engagement metrics (salaries, interviews, jobs). Salaries and interviews often show more moderate distributions, suggesting they are less dominated by a few mega employers. The overlapping tails highlight companies that perform highly across multiple metrics (e.g., large IT firms appearing in multiple high value regions). This visualization aids in understanding the relative variability and potential multicollinearity among popularity/engagement proxies, informing feature selection in any subsequent predictive modeling.

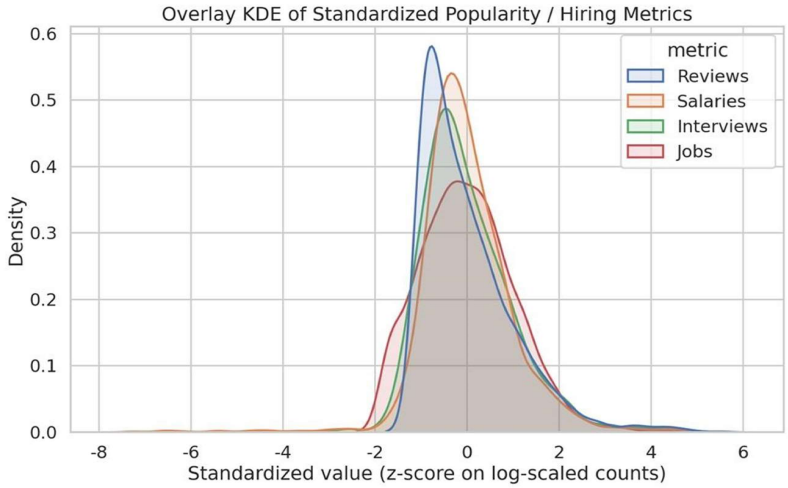

Figure 10. Overlay KDE of Standarized Popluarity Metrics

Similar to Figure 8 but focused specifically on composite or selected “popularity” indicators (e.g., reviews, jobs, interviews).

The figure accentuates the heavy right tail in review and job related metrics, underscoring that popularity is highly concentrated among a minority of firms. Standardized overlays make clear that while most companies cluster near zero (average), a small subset drives extreme positive values. This has methodological implications

robust statistics or transformations are essential and substantive ones: the Indian job market appears winner-take most in terms of online visibility and applicant interest. Researchers can use this to caution against over-reliance on untransformed aggregates in sector comparisons.

We converted compact Indian-style counts such as 1.2L and 11.4k into numeric values, then used log scaling for review/salary/interview/job counts before KDE so the distributions are interpretable despite strong skew. KDE overlays were used for continuous distribution comparison, and the 2D KDE shows joint density between rating and review volume.

5. Discussion

The present study utilized web-scraped observational data from the Indian corporate landscape to investigate the dynamics of employer visibility, employee engagement, and workforce satisfaction. By analyzing a diverse sample of 1,760 companies, this research provides critical empirical evidence challenging conventional assumptions about the relationship between a company's online popularity and its perceived quality as an employer. The following sections discuss the core findings, their implications for labor market dynamics, methodological considerations, and avenues for future research.

5.1 The Decoupling of Employer Visibility and Employee Satisfaction

The most prominent finding of this study is the near total independence of a company's review volume (visibility) and its average employee rating (satisfaction). Both parametric (Pearson's r = - 0.054) and non-parametric (Spearman's ρ = - 0.050) tests confirmed a statistically significant but practically negligible association, with the regression model explaining merely 0.29% ($R^2 \approx 0.0029$) of the variance in ratings. According to Cohen's effect-size guidelines, this represents a negligible effect.

This decoupling challenges the simplistic heuristic that "popular employers are inherently better places to work." In the context of platforms like AmbitionBox, high review volumes are primarily driven by organizational scale, massive hiring pipelines, and high turnover rates characteristics typical of India's IT services giants and major banking sectors. The wide dispersion of ratings (clustering tightly between 3.5 and 4.5) among high-review companies suggests that massive employee bases dilute uniform cultural experiences, resulting in average to moderate satisfaction scores. Furthermore, mega employers face heightened public and employee scrutiny; the sheer volume of feedback amplifies both praise and criticism, neutralizing any potential upward bias in average ratings.

5.2 Scale Effects and the "Winner-Take-Most" Online Labor Market

The descriptive statistics and Kernel Density Estimation (KDE) overlays reveal a highly right skewed distribution of employer engagement metrics. The stark contrast between the mean (293.93) and median (3) review counts underscores a "winner take most" dynamic in online labor market visibility. A small subset of large, well-known employers accounts for a disproportionate share of platform activity.

The correlation matrix of log transformed variables (Table 2) further elucidates this "visibility halo" effect. Metrics such as review counts, shared salaries, interview experiences, and job postings are strongly and positively inter correlated (ranging from 0.42 to 0.86). This indicates that online engagement is largely a function of organizational scale and market presence rather than intrinsic employee satisfaction. Companies that are highly visible in one dimension (e.g., job postings) inevitably dominate other engagement metrics. However, the near zero correlation between rating_num and these popularity indicators confirms that employee satisfaction operates on a fundamentally different axis, driven by micro level cultural and managerial factors rather than macro level market dominance.

5.3 Sectoral Dynamics and the Compensation-Expectation Paradox

Sectoral heterogeneity plays a vital role in shaping employer perception. The IT Services & Consulting sector dominates the dataset (218 companies), exhibiting the highest density of moderate to high review volumes.

This reflects the structural reality of India's economic landscape, where the IT sector acts as a massive entry-level employer for millions of graduates, naturally generating immense platform traffic.

Notably, the correlation matrix reveals a modest negative association between company ratings and the log-transformed number of salary insights shared (r=-0.218). While exploratory, this finding hints at a "compensation expectation paradox" in the Indian corporate context. High paying sectors or roles often entail rigorous performance metrics, high pressure environments, and elevated employee expectations. When financial compensation is high, employees may become more critical of non monetary aspects of the work environment (e.g., work life balance, management support), leading to slightly lower overall satisfaction ratings compared to lower paying but potentially more relaxed or niche industries.

5.4. Methodological Considerations for Web-Scraped HR Data

This study highlights critical methodological imperatives for researchers utilizing web scraped organizational data.The severe skewness and heteroscedasticity inherent in user generated content platforms violate the assumptions of standard parametric testing if left unaddressed.

1. Transformations and Non-Parametric Methods: The application of log transformations was essential to normalize the heavy right tails of engagement metrics. Furthermore, the reliance on Spearman's rank correlation and KDE overlays proved vital in accurately capturing monotonic relationships and joint densities without undue influence from extreme outliers (e.g., firms with >100k reviews).

1. Standardization for Cross-Metric Comparison: Overlaying standardized (z-scored) KDE curves for popularity metrics effectively demonstrated the varying degrees of skewness across different engagement types, proving that review volumes are far more concentrated among mega employers than salary or interview data.

5.5 Practical Implications

The findings offer actionable insights for multiple stakeholders in the labor market:

• **For Job Seekers**: Review volume should not be used as a proxy for employer quality. High review counts merely indicate high hiring volume or turnover. Seekers must look beyond aggregate ratings and engage with the qualitative nuances of reviews, particularly within their specific sub-sectors or roles.

• **For HR and Employer Branding**: Organizations cannot rely on sheer market size or visibility to generate a positive employer brand. High visibility brings diverse, often conflicting, employee sentiments. HR strategies must focus on localized, team-level cultural interventions rather than broad, corporate wide branding campaigns.

• **For Review Platforms**: Algorithms that rank or recommend employers based on "popularity" or "review volume" may inadvertently bias users toward large, legacy firms while obscuring highly satisfying small to-medium enterprises (SMEs). Platforms should consider weighting metrics by engagement depth or utilizing sentiment analysis to provide a more balanced view.

5.6 Limitations and Future Research Directions

While this study provides a robust empirical foundation, it is subject to limitations inherent to the dataset and methodology. First, the dataset (Soni, 2026) lacks methodological transparency regarding the original source website, exact scraping tools, and the temporal timeframe of data collection. This cross sectional snapshot limits the ability to make causal inferences or track the evolution of employer reputation over time. Second, the absence of a formal data dictionary restricts the ability to control for confounding variables such as company age, revenue, or geographic footprint.

Future research should address these gaps by:

1. Text Mining and NLP: Moving beyond aggregate numerical ratings to analyze the semantic content of individual reviews. Topic modeling could reveal *why* large firms receive average ratings (e.g., identifying specific pain points like "work life balance" vs. "learning opportunities").

2. Integration with Financial Data: Merging web-scraped employer reputation metrics with external financial performance data (e.g., stock returns, profitability, or attrition rates) to quantify the actual economic impact of employer branding in emerging markets.

3. Longitudinal Studies: Tracking how company ratings and review volumes evolve in response to macroeconomic shocks or internal policy changes (e.g., return to office mandates).

In conclusion, this analysis demystifies the relationship between employer popularity and employee satisfaction in the Indian job market. By proving that visibility and satisfaction are largely orthogonal, this study calls for a more nuanced, multidimensional approach to evaluating organizational reputation in the digital age.

6. Conclusion

This study provides robust empirical evidence demystifying the relationship between employer popularity and employee satisfaction in the Indian digital labor market. By analyzing a diverse sample of 1,760 companies, we empirically validate the "Visibility Satisfaction Paradox," demonstrating that organizational visibility measured by the volume of online employee reviews is largely orthogonal to perceived workplace quality. The data reveals a "winner take most" dynamic in online labor markets, where a small subset of large scale employers (particularly in IT services and banking) monopolizes platform engagement and visibility, yet exhibits average to moderate satisfaction scores. Furthermore, sectoral analyses hint at a "compensation-expectation paradox," where higher financial transparency and salary insights correlate modestly with lower satisfaction ratings, likely reflecting elevated employee expectations in high pressure environments.

The practical implications of these findings are substantial. Job seekers are cautioned against using review volume as a proxy for employer quality, while HR and employer branding professionals must recognize that sheer market scale cannot substitute for localized, team level cultural interventions. Additionally, review plat forms must reconsider algorithmic ranking systems that inadvertently bias users toward legacy mega employers at the expense of highly satisfying small to medium enterprises.

Despite these contributions, this study is limited by the cross sectional nature of the data and a lack of metadata regarding the original scraping timeframe and sources, precluding causal inferences. Future research should address these gaps by integrating Natural Language Processing (NLP) to mine the semantic nuances of individual reviews, merging web scraped reputation metrics with external financial and attrition data, and employing longitudinal designs to track reputation evolution. Ultimately, this research calls for a multidimensional, methodologically rigorous approach to evaluating organizational reputation in the Big Data era, moving beyond simplistic popularity metrics to capture the true complexity of the modern workforce experience.

Ethical Statement: The study exclusively utilized publicly available web scraped information. No personally identifiable information was collected, stored, or analyzed. The analysis complies with ethical standards for secondary data research.

References

[1] Lagoze, C. (2014). Big data, data integrity, and the fracturing of the control zone. *Big Data Society, July December*, 1-11.

[2] Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *Meta Group*. Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[3] Hitzler, P., Janowicz, K. (2010). Linked data, big data and the 4th paradigm. *Semantic Web, 0*(0), 1. Retrieved from http://www.semantic-web-journal.net/system/files/swj488.pdf.

[4] Eurostat. (2011). *European Statistics Code of Practice: Revised Edition 2011*. ISBN: 978-92-79-21679-4. Retrieved from http://goo.gl/Z0xArw.

[5] Struijs, P., Braaksma, B., Daas, P. J. H. (2014). Official statistics and big data. *Big Data and Society, April-June*, 1-6.

[6] Barbera, P., Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Sciences Computer Review, 33*(6). Retrieved from http://journals.sagepub.com/doi/full/10.1177/0894439314558836.

[7] Blank, G. (2017). The digital divide among Twitter users and its implications for social research. *Social Sciences Computer Review, 35*(6), 1-19. Retrieved from http://journals.sagepub.com/doi/full/10.1177/0894439-316671698.

[8] Rafali, P. (2018). Nonprobability sampling and Twitter: Strategies for semibounded and bounded populations. *Social Sciences Computer Review, 36*(2). Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/0894439317709431.

[9] Martin, B. (2018). Persistent bias on Wikipedia, methods and responses. *Social Sciences Computer Review, 36*(3), 1-10. Retrieved from http://journals.sagepub.com/doi/full/10.1177/0894439317715434.

[10] Choi, H., Variant, H. (2012). Predicting the present with Google Trends. *The Economic Record, 88* (Special Issue), 2-9.

[11] Butler, D. (2013). When Google got flu wrong. *Nature, 494*, 14th February 2013.

[12] Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343* (6176), 1203-1205.

[13] Pedraza, P. de, Tijdens, K., Visintin, S. (2018). The matching process before and after the crisis in the Netherlands. *International Journal of Manpower, 39*(8), 1010-1031. DOI: 10.1108/IJM-10-2018-0329.

[14] Sáez Martín, A., Haro de Rosario, A., Caba Pérez, M. C. (2016). An international analysis of the quality of open government data portals. *Social Sciences Computer Review, 34*(3).

[15] Revilla, M., Ochoa, C., Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. *Social Sciences Computer Review, 35*(4).

[16] Stern, M. J., Bilgen, I., McClain, C., Hunsche, B. (2016). Effective sampling from social media sites and search engines for web surveys: Demographic and data quality differences in surveys of Google and Facebook users. *Social Sciences Computer Review*, 1-19. doi:10.1177/0894439316683344.

[17] Head, B. G., Dean, E., Flanigan, T., Swicegood, J., Keatin, M. D. (2016). Advertising for cognitive interviews: A comparison of Facebook, Craigslist, and snowball recruiting. *Social Science Computer Review, 34*(3), 360-377.

[18] De Leeuw, E. (2018). Mixed mode: Past, present, and future. *Survey Research Methods, 12*(2), 75-89. doi:10.18148/srm/2018.v12i2.7402.

[19] Revilla, M., Cornilleau, A., Cousteaux, A. S., Legleye, S., Pedraza, P. (2015). What is the gain in a probability-based online panel of providing internet access to sampling units who previously had no access *Social Sciences Computer Review*, 1-18. Retrieved from http://ssc.sagepub.com/content/early/2015/06/04/08944393155902-06.full.pdf?ijkey=nNfsKd0vcQ5sRqq&keytype=finite.

[20] de Pedraza, P., Visintin, S., Tijdens, K., Kismihók, G. (2019). Survey vs scraped data: Comparing time series properties of web and survey vacancy data. *IZA Journal of Labor Economics*, (1).

[21] Alves Gomes, M., Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management, 21*(3), 527-570. *Springer,* Berlin, Heidelberg.

[22] Bai, C., Sarkis, J. (2020). A supply chain transparency and sustainability technology appraisal model for blockchain technology. *International Journal of Production Research, 58*, 2142–2162.

[23] de Ruyter, K., Keeling, D. I., Plangger, K., Montecchi, M., Scott, M. L., Dahl, D. W. (2022). Reimagining marketing strategy: Driving the debate on grand challenges. *Journal of the Academy of Marketing Science, 50*(1), 13–21. https://doi.org/10.1007/s11747-021-00806.

[24] Serafeim, G. (2020). Social-impact efforts that create real value. *Harvard Business Review, 98*, 38–48.

[25] Wang, H., Jia, M., Xiang, Y., Lan, Y. (2022). Social performance feedback and firm communication strategy. *Journal of Management, 48*, 2382–2420.

[26] Gonzalez-Arcos, C., Joubert, A. M., Scaraboto, D., Guesalaga, R., Sandberg, J. (2021). How do I carry all this now Understanding consumer resistance to sustainability interventions. *Journal of Marketing, 85*, 44–61.

[27] Chatzopoulou, E.-C., Manolopoulos, D., Agapitou, V. (2022). Corporate social responsibility and employee outcomes: Interrelations of external and internal orientations with job satisfaction and organizational commitment. *Journal of Business Ethics, 179*, 795–817.

[28] Martin-de Castro, G. (2021). Exploring the market side of corporate environmentalism: Reputation, legitimacy and stakeholders' engagement. *Industrial Marketing Management, 92*, 289–294.

[29] Paine, L. S. (2014). Sustainability in the boardroom. *Harvard Business Review, 92*, 86–94.

[30] https://www.kaggle.com/datasets/kanha108/company-data-from-web-scraping