MERCon 2026
MORATUWA ENGINEERING RESEARCH CONFERENCE

Fault of Our Stars : Behavioral Drivers of Rating - Sentiment Incongruence

Ramanaish Abaiyan!, Ruththiragayan Sutharsan, Kusal Amantha, Anusan Krishnathas, Sriyathursan Kovindharaja, Asma Rauff , Dr. Nisansa De Silva, Dr. Sandareka Wickramanayake

Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

Abstract

Online reviews carry two signals: a star rating and written text. Sentiment analysis widely treats the rating as a proxy for the text's sentiment, but whether the two actually agree is rarely tested. This study measures **sentiment–rating incongruence** – disagreement between text sentiment and the sentiment implied by the star rating – in Sri Lankan tourism reviews.

Using **16,156 reviews** (2010–2023) and a transformer-based sentiment pipeline that infers text sentiment independently of the rating, we find incongruence in **18.6%** of reviews, organized into six directional patterns dominated by *Conservative Rater* and *Obligatory 5-Star* behavior.

Statistical testing, Logistic Regression, Random Forest, and SHAP identify venue type, reviewer expertise, review length, and time as drivers of this divergence, with effects that hold after multiple-testing correction and multicollinearity checks. Star ratings are not interchangeable with textual sentiment, and should be validated before use as ground-truth NLP labels.

Project Objectives

- Infer text sentiment **independently of ratings**, avoiding the circularity of using ratings to validate ratings.
- Classify mismatches into a **six-pattern directional taxonomy** rather than a binary match/mismatch flag.
- Identify which reviewer, venue, and temporal factors predict incongruence, and quantify effect sizes.
- Determine whether star ratings are safe to use as ground-truth sentiment labels in NLP pipelines.

Related Work

- Ratings are widely assumed to track text sentiment [1], a convention still applied without validation in tourism NLP, concentrated on hotels and restaurants [2, 3]. Empirical evidence on rating–text alignment is **mixed** [4, 5].
- Prior Sri Lankan work [6, 7] uses lexicon-based methods and treats mismatch as an error to *correct*, without testing which factors predict it.
- **Gap:** we use transformer-based inference, a directional taxonomy, and multivariable modeling to frame incongruence as a **weak-label reliability** problem in NLP, not a correction task.

Methodology

1. **Dataset:** 16,156 English reviews, 11 attraction types, 2010–2023 [8]. Source columns in Table 1.
2. **Preprocessing:** ratings grouped as Negative (1–2★) / Neutral (3★) / Positive (4–5★); review length, review delay, province/district, and reviewer tier engineered.
3. **Sentiment inference:** four transformer models evaluated on 1,000 hand-labeled reviews (700/300 split); see Table 2 in Experiments.
4. **Incongruence flag:** disagreement between predicted *Sentiment* and *Rating_Class* → six-category *Pattern* variable.
5. **Explanatory modeling:** bivariate tests (BH-FDR corrected for 5 predictors) → Logistic Regression + Logit → Random Forest → SHAP. Predictors used to build *Incongruent* were excluded from the model to avoid leakage.

Table 1. Source columns used in analysis

Feature	Description
Location	Used to derive province and district
Location_Type	Type of attraction (museum, beach etc.)
User_Contributions	Total reviews posted by the reviewer
Travel_Date	Date of visit; used to derive travel year
Published_Date	Date of review posting; used to derive review delay
Rating	Star rating used to create Rating_Class
Title	Review title; combined with text for sentiment inference
Text	Review body; used for sentiment and review length

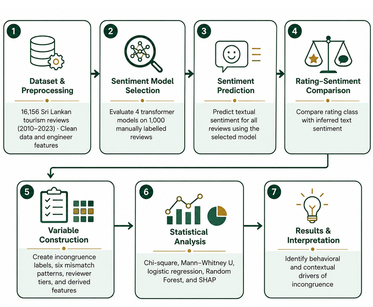


Figure 1. Overview of the proposed methodology.

Experiments

Sentiment Model Selection (test set, $n = 300$)

Table 2. Sentiment model performance (test set, $n = 300$)

Model	Acc.	Mac. F1	Wt. F1	Decision
gosorio/robertaSentimentFT	0.730	0.281	0.616	Rejected
cardiffnlp/twitter-roberta (pretrained) [9]	0.830	**0.692**	0.805	**Selected**
Fine-tuned Cardiff RoBERTa	0.733	0.677	0.758	Lower overall fit
Fine-tuned RoBERTa-base	0.777	0.698	0.780	Higher Mac. F1 only

Pretrained CardiffNLP selected on Macro F1 (class-imbalance-robust), not accuracy alone – the fine-tuned RoBERTa-base scored marginally higher on Macro F1 but was rejected due to overfitting risk on only 700 training rows and weaker weighted F1.

Predictive Modeling (predictors: venue, province, reviewer tier, review length/delay, travel year; max VIF = 3.663, low multicollinearity)

- **1A – Logistic Regression:** balanced classes, 5-fold CV; mean AUC 0.589 ± 0.009, test AUC 0.584 – low variance across folds indicates a stable, not overfit, baseline.
- **1B – Explanatory Logit:** 19 predictors significant at 95% CI (bivariate screen in Table 3).
- **2 – Random Forest:** GridSearchCV; test AUC 0.6095, confirming mild nonlinear structure beyond the linear baseline.
- **SHAP:** TreeExplainer on best RF, post-hoc only. Top features by mean $|\phi|$: review length, reviewer expertise, travel year, review delay, venue type – consistent with the logit coefficients, cross-validating the two modeling approaches.

Results

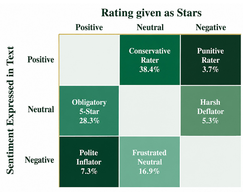


Figure 2. Distribution of the six directional incongruence patterns (n=3,005 incongruent reviews).

Prevalence & Taxonomy

- **18.6%** of all 16,156 reviews (3,005) are incongruent – roughly 1 in 5.
- *Conservative Rater* (38.4%) + *Obligatory 5-Star* (28.3%) = 66.7% of mismatches: structured, not random.
- *Frustrated Neutral* (16.9%) + *Polite Inflator* (7.3%): negative text paired with non-negative ratings in 24.2% of mismatches – the direction that matters most for downstream label noise, since it silently inflates apparent satisfaction.
- *Harsh Deflator* (5.3%) and *Punitive Rater* (3.7%) are rare: reviewers almost never rate harsher than their text reads, so the noise this dataset produces is directionally biased toward inflated sentiment, not symmetric.

Table 3. Bivariate screening with BH-FDR correction

Predictor	Test	Raw p	q-value	Retained
Reviewer Tier	$\chi^2(3)$=85.99	1.59×10^{-18}	7.94×10^{-18}	✓
Province	$\chi^2(7)$=49.10	2.17×10^{-8}	3.62×10^{-8}	✓
Travel Year	Mann–Whitney	7.73×10^{-11}	1.93×10^{-10}	✓
Review Length	Mann–Whitney	0.0011	0.0014	✓
Review Delay	Mann–Whitney	0.7503	0.7503	✗

Venue & Reviewer Effects

- Venue matters: $\chi^2(10) = 125.85$, $p < 0.001$, but effect size is modest (Cramer's V = 0.088) – venue shifts the odds, it does not determine the outcome. Museums highest (26.3%), National Parks lowest (12.8%); Museums vs. Parks Adjusted OR = 2.386, with similarly elevated odds for beaches, inland waterbodies, zoological gardens, and waterfalls.
- Experts are **1.97×** more likely than novices to produce incongruent reviews; Conservative Rater rises from 27.4% (novice) to 40.4% (expert) – experienced reviewers may reserve top ratings for exceptional cases while still writing positively.
- Incongruent reviews run longer at the median (296 vs. 279 characters), suggesting more descriptive text is where nuance gets lost in the numeric score.
- Review delay is **not** significant in bivariate screening ($q = 0.7503$), yet shows a weak negative association in the full logit model (OR = 0.960) – a reminder that bivariate and multivariable results can diverge once other predictors are held constant, and that timing plays a secondary role at best.

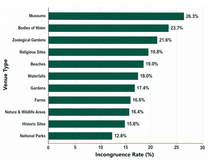


Figure 3. Incongruence rate by venue type.

Temporal Trend

- Linear travel-year term OR = 0.946 (CI < 1): incongruence declines modestly over time; quadratic term not significant (OR = 0.995) – trend is linear, ruling out a U-shaped or accelerating pattern.
- SHAP corroborates this direction (mean $|\phi|$ for Travel_Year_c = 0.0219), so the temporal decline is not an artifact of the logit model's functional-form assumptions alone.

Conclusions

- Incongruence is **systematic, not noise**: 18.6% prevalence, concentrated in two dominant patterns, and directionally biased toward inflated rather than deflated sentiment.
- Drivers are structural and behavioral – venue, reviewer expertise, review length, and time (Table 3) – and remain significant after FDR correction and VIF screening for multicollinearity.
- Predictive power is real but modest (AUC 0.58–0.61): the observed features explain part of the pattern, but most of the variance in any single review's incongruence remains unexplained by venue, reviewer tier, or timing alone.
- **Practical implication:** star ratings should not be used as ground-truth sentiment labels without venue-/reviewer-aware validation, calling into question prior low-resource sentiment work that derives labels from ratings alone [10, 11].
- **Robustness:** SHAP and logit rankings agree on the top predictors, giving cross-method convergence rather than a single-model artifact; the six-pattern taxonomy itself is a reusable framework independent of which predictive model is applied to it.
- **Limitation:** sentiment labels rely on a single pretrained model evaluated on only 1,000 hand-labeled reviews, with model selection performed directly on the 300-review test set and no held-out validation split – some flagged "incongruence" may reflect model error rather than reviewer behavior. Future work: alternative sentiment baselines (VADER, TextBlob) and a dedicated validation split.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] A. Alaei, S. Becken, and B. Stantic, "Sentiment analysis in tourism: Capitalising on big data," *Journal of Travel Research*, vol. 58, no. 2, pp. 175–191, 2019.

[3] A. Ameur, S. Hamdi, and S. B. Yahia, "Sentiment analysis for hotel reviews: A systematic literature review," *ACM Computing Surveys*, vol. 56, no. 2, p. Article 51, Sep. 2023.

[4] E. Bigne, C. Ruiz, C. Perez-Cabanero, and A. Cuenca, "Are customer star ratings and sentiments aligned? a deep learning study of the customer service experience in tourism destinations," *Service Business*, vol. 17, pp. 281–314, 2023.

[5] B. Kwon, J. Lee *et al.*, "Beyond the stars: The impact of rating–text inconsistency on perceived review usefulness," *Asia Pacific Journal of Information Systems*, vol. 35, no. 1, pp. 49–72, 2025.

[6] H. P. P. M. Abeysinghe and C. K. Walgampaya, "Sentiment analysis in user reviews: A study of incompatibility in hotel reviews in city of anuradhapura, sri lanka," in *Proceedings of iPURSE*, vol. 23, Nov. 2021.

[7] P. Abeysinghe and T. Bandara, "A novel self-learning approach to overcome incompatibility on tripadvisor reviews," *Data Science and Management*, vol. 5, pp. 1–10, 2022.

[8] T. Sewwandi, "Tourism and travel reviews: Sri lankan destinations," Mendeley Data, V1, 2023.

[9] F. Barbieri, J. Camacho-Collados *et al.*, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Nov. 2020, pp. 1644–1650.

[10] V. Jayawickrama, G. Weerapraneshwara *et al.*, "Seeking sinhala sentiment: Predicting facebook reactions of sinhala posts," in *International Conference on Advances in ICT for Emerging Regions*, 2021, pp. 177–182.

[11] ——, "Facebook for sentiment analysis: Baseline models to predict facebook reactions of sinhala posts," *The International Journal on Advances in ICT for Emerging Regions*, vol. 15, no. 2, 2022.

Abaiyan-27/Group-J---Research-Paper

polyglots/tourism-sentiment-incongruence

ORGANIZED BY

Engineering Research Unit, University of Moratuwa, Sri Lanka

University of Moratuwa IEEE Student Branch

FINANCIAL SPONSOR

IEEE Sri Lanka Section

TECHNICAL CO-SPONSORS

GCASCEDA — Circuits and Systems Society, Geoscience and Remote Sensing Society, Electronics Design Architecture — IEEE Sri Lanka Section Joint Chapter

IEEE COMPUTER SOCIETY Sri Lanka Chapter

IEEE COMMUNICATIONS SOCIETY Sri Lanka

IEEE EMBS Engineering in Medicine & Biology Society — Sri Lanka Chapter

GRSS Geoscience and Remote Sensing Society

IAS Industry Applications Society — Sri Lanka Section

pels IEEE POWER ELECTRONICS SOCIETY — Powering a Sustainable Future

IEEE Industrial Electronics Society — Ice — Sri Lanka Chapter

IEEE PES Power & Energy Society — Sri Lanka Chapter

IEEE Robotics & Automation Society — Sri Lanka Chapter

IEEE Signal Processing Society — SRI LANKA CHAPTER

MTT-S — Sri Lanka