engineering proceedings

MDPI

Proceeding Paper

Explainable Reputation Estimation from Web Service Reviews †

Elmira Saeedi Taleghani, Ronald Iván Maldonado Valencia, Ana Lucila Sandoval Orozco and Luis Javier García Villalba *

Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Facultad de Informática, Universidad Complutense de Madrid (UCM), 28040 Madrid, Spain; elmirasa@ucm.es (E.S.T.); ronaldim@ucm.es (R.I.M.V.); asandov@ucm.es (A.L.S.O.)

* Correspondence: javierv@ucm.es

† Presented at the First Summer School on Artificial Intelligence in Cybersecurity, Cancun, Mexico, 3–7 November 2025.

Abstract

Star ratings alone are noisy, manipulable, and ignore aspect-level sentiment. We present Scrape2Repute, a compact and reproducible pipeline that ingests Yelp reviews under policy constraints; cleans and normalises text/metadata; learns a calibrated text sentiment per review; fuses stars and text via a tunable hybrid label; downweights suspicious reviews with unsupervised anomaly scoring; and aggregates evidence into a time-decayed business reputation with uncertainty bounds. The system is explainable (top-k rationales, aspect summaries), runs on commodity hardware, and ships with CLI/GUI. On the Yelp Open Dataset, we show strong predictive validity for forecasting future ratings and stable behaviour under sensitivity sweeps. We release implementation and an ethics checklist for compliant use.

Keywords: reputation modelling; open dataset; sentiment analysis; anomaly detection; calibration; explainability

1. Introduction

User reviews on *Yelp* influence consumer decisions, yet naive star averaging ignores textual nuance, temporal drift, and manipulation risk. An auditable solution should (1) integrate textual sentiment with stars; (2) reduce the impact of stale or suspicious evidence; (3) expose explanations; and (4) be reproducible and policy-compliant.

Contributions: *Scrape2Repute* provides (i) an end-to-end pipeline with minimal dependencies; (ii) a hybrid per-review label combining normalised stars with a calibrated text score; (iii) a lightweight anomaly screen; (iv) time-decayed aggregation with confidence intervals; and (v) explainability via n-gram/token attributions and aspect summaries. We release CLI/GUI entry points and an ethics and compliance checklist tailored to data. An overview of the end-to-end architecture is shown in Figure 1.

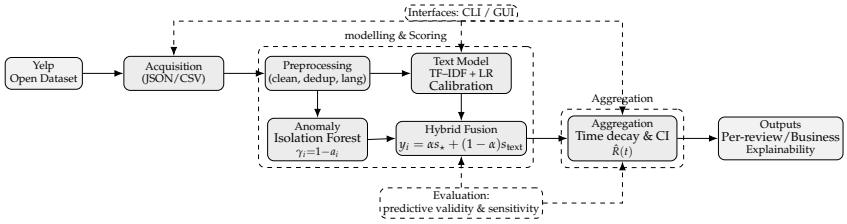


Figure 1. End-to-end architecture of Scrape2Repute.

Check for updates

Academic Editors: Héctor Manuel Pérez Meana and Gabriel Sánchez Pérez

Published: 5 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Paper organisation. Section 3 details the pipeline; Section 4 reports predictive validity and sensitivity; Section 5 states compliance; and Section 6 concludes.

2. Related Work

Work on online reputation blends ratings, text, and interactions. Collaborative filtering and trust modelling improve robustness and mitigate bias across contexts [1,2]. Reputation dynamics, such as inertia, motivate temporal and adaptive treatment [3]. For web services, QoS-aware models and matrix factorisation fuse performance indicators with reputation signals [4,5]; integrity is improved by graph-based filtering of dishonest raters and credibility modelling [6,7]. Personalised reputation adapts scores to user history [8]. Our pipeline operationalises these ideas end-to-end on Yelp: calibrated text+stars fusion, unsupervised anomaly dampening, and time-decayed aggregation with uncertainty.

3. Methodology

Our pipeline comprises five stages: Acquisition (Yelp-compliant loading), Preprocessing (normalisation, dedup), Modelling (text classifier + calibration; hybrid fusion), Screening (unsupervised anomalies), and Aggregation (time decay with uncertainty).

3.1. Data Access and Normalisation

We use the *Open Dataset* [9], focusing on `review` and `business`. For each review we keep `(review_id, business_id, user_id, stars, date, text)`. Stars are normalised as $s_\star = (\text{stars} - 1)/4 \in [0,1]$, with timestamps t_i. We remove empty/near-empty texts, strip HTML, lowercase, normalise whitespace/emoji, optionally filter to English, and deduplicate by hashing normalised text.

3.2. Text Modelling and Calibration

Let x_i denote a vectorised review (TF–IDF). A logistic classifier produces a calibrated probability $s_{\text{text}}(x_i) = \sigma(w^\top x_i + b) \in [0,1]$ (Platt/isotonic). Sentence embeddings with a non-linear classifier can replace TF–IDF where resources permit.

3.3. Hybrid, Anomaly, and Aggregation

$$y_i(\alpha) = \alpha\, s_{\star,i} + (1-\alpha)\, s_{\text{text},i}, \quad \alpha \in [0,1], \tag{1}$$

$$w_i(t) = \exp\big(-\lambda\,[t - t_i]_+\big)\,\gamma_i, \quad \gamma_i = (1 - a_i), \tag{2}$$

$$\hat{R}(t) = \frac{\sum_i w_i(t) y_i}{\sum_i w_i(t)}, \quad \hat{R} \pm 1.96\sqrt{\hat{R}(1-\hat{R})/\sum_i w_i}. \tag{3}$$

An Isolation Forest estimates soft anomaly $a_i \in [0,1]$ from length, repetition, burstiness, and (where legal) account cues [10]. The robustness weight γ_i attenuates suspicious reviews. We select α via cross-validation or stability criteria.

4. Experiments and Results

4.1. Setup

We train a *TF-IDF + Logistic* classifier with weak supervision from extreme stars (normalised ≥ 0.7 positive, ≤ 0.3 negative) and Platt calibration (3-fold; 50% calibration split). Inference runs in chunks on the cleaned review file, followed by Bayesian reputation aggregation and anomaly estimation.

Data Scale and Footprint: We processed $\approx$ 6.99 M reviews; the 90-day split retained 7059 businesses with min_early ≥ 5 and min_future ≥ 5. Training (TF-IDF + LR + Platt) took $\sim$88 min; the sensitivity sweep $\sim$30 min; anomaly scoring $\sim$15 min; and other

https://doi.org/10.3390/engproc2026123020

stages < 2 min each on a commodity workstation. These figures suggest that the pipeline is practical for routine, offline recalibration at scale.

4.2. Text Model Quality

On a held-out sample (n = 7846), accuracy is 0.875; per-class metrics are in Table 1. The positive class shows strong precision/recall; neutral remains challenging under weak supervision.

Table 1. Calibrated text classifier (TF-IDF + LR).

Class	Precision	Recall	F1
−1 (neg)	0.832	0.885	0.857
0 (neu)	0.543	0.256	0.348
+1 (pos)	0.913	0.965	0.938
Accuracy	0.875 (macro F1 0.715)		

4.3. Predictive Validity (Early → Future)

We test whether early hybrid scores forecast future stars at the business level (90-day horizon; min_early ≥ 5, min_future ≥ 5; 7059 businesses; see Figures 2 and 3a). Weighted mean future star01 = 0.7239; Pearson r = 0.7553 (weighted 0.7652); Spearman ρ = 0.6737 (weighted 0.6873); and top-decile mean = 0.9069 (+25.27% vs. overall).

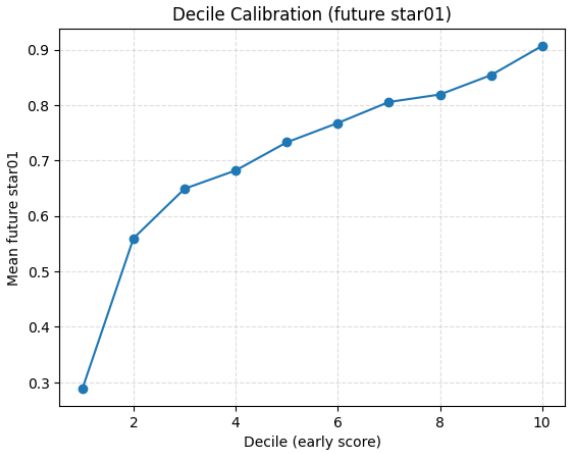
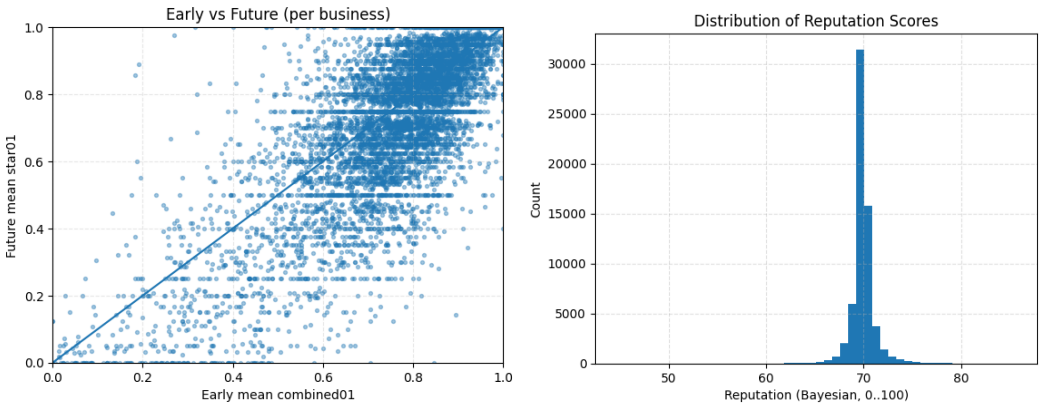


Figure 2. Decile calibration: early combined score vs. future star01.



(**a**) Early vs. future ($r = 0.755$, $\rho = 0.674$). (**b**) Distribution of Bayesian reputation.

Figure 3. Evaluation on Yelp: (**a**) predictive validity; (**b**) reputation distribution.

https://doi.org/10.3390/engproc2026123020

4.4. Distribution and Sensitivity

Figure 3 summarises predictive validity and the distribution of Bayesian reputation (mode ∼70); see panels (a, b).

We sweep half-life $\in \{180, 365, 540, 720\}$, suspicious-weight $\in \{0.1, 0.2, 0.3, 0.5, 0.8, 1.0\}$, and $\alpha \in \{10, 20, 30, 50, 80\}$; Table 2 contrasts the baseline (HL = 365, SW = 0.3, $\alpha = 20$) with best settings and—practically—recommends HL 540–720, SW 0.5–1.0, and $\alpha \in \{10, 20\}$.

Baseline contrast: A stars-only aggregator is included as a baseline; the hybrid shows better decile calibration (Figure 2) and higher predictive validity across the sweep (Table 2).

Reproducibility: Code, configs, fixed seeds, and artefacts enable end-to-end replication.

Practical impact: The pipeline supports periodic offline recalibration and lightweight deployment for marketplace ranking, vendor monitoring, and early-warning alerts.

Table 2. Sensitivity summary (90-day horizon).

Config	HL	SW	α	Metric
Baseline	365	0.3	20	$r = 0.7457$, lift = 21.66%
Best-r	540	1.0	10	$r = 0.7664$, lift = 24.41%
Best-lift	720	1.0	10	$r = 0.7663$, lift = 24.49%

5. Ethics, Compliance

We use the Yelp Open Dataset under its research terms [9]; we avoid PII, apply rate limiting, and honour robots and policies.

6. Conclusions

Scrape2Repute offers an auditable pathway from Yelp reviews to a time-aware, anomaly-robust reputation with explanations. Predictive validity is strong (Pearson ≈ 0.76 at 90-day horizon) and stable across decay/anomaly settings. Future work includes multilingual aspect extraction, category-specific calibration, and user studies on explanation usefulness.

Author Contributions: Conceptualisation, E.S.T., R.I.M.V., A.L.S.O. and L.J.G.V.; methodology, E.S.T., A.L.S.O. and L.J.G.V.; validation, E.S.T., A.L.S.O. and L.J.G.V.; investigation, E.S.T., R.I.M.V., A.L.S.O. and L.J.G.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the PRIVATEER EU project, Grant agreement N 101096110, by the Programme UNICO-5G I+D of the Spanish Ministerio de Asuntos Económicos y Transformación Digital, the European Union—NextGeneration EU in the framework of the ”Plan de Recuperacion, Transformación y Resiliencia” under reference “TRAZA5G (TSI-063000-2021-0050)”, by the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation EU), through the Chair “Cybersecurity for Innovation and Digital Protection” INCIBE-UCM and by Comunidad Autonoma de Madrid, CIRMA-CM Project (TEC-2024/COM-404). The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

Data Availability Statement: The dataset analyzed in this study is Yelp Open Datasets [9].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Moreno, N.; Pérez-Vereda, A.; Vallecillo, A. Managing reputation in collaborative social computing applications. *J. Object Technol.* **2022**, *21*, 3. [CrossRef]
2. Muslim, H.S.M.; Rubab, S.; Khan, M.M.; Iltaf, N.; Bashir, A.K.; Javed, K. S-RAP: relevance-aware QoS prediction in web-services and user contexts. *Knowl. Inf. Syst.* **2022**, *64*, 1997–2022. [CrossRef]

3. Duradoni, M.; Gronchi, G.; Bocchi, L.; Guazzini, A. Reputation matters the most: The reputation inertia effect. *Hum. Behav. Emerg. Technol.* **2020**, *2*, 71–81. [CrossRef]
4. Xu, J.; Chen, Y.; Zhu, C. A QoS-based User Reputation Measurement Method for Web Services. In *Proceedings of the 2018 International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018), Wuhan, China, 7–8 February 2018*; Advances in Computer Science Research Series; Atlantis Press: Paris, France, 2018; pp. 470–473. [CrossRef]
5. Ghafouri, S.H.; Hashemi, S.M.; Razzazi, M.R.; Movaghar, A. Web service quality of service prediction via regional reputation-based matrix factorization. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e6318. [CrossRef]
6. Tibermacine, O.; Tibermacine, C.; Kerdoudi, M.L. Reputation evaluation with malicious feedback prevention using a HITS-based model. In Proceedings of the 2019 IEEE International Conference on Web Services (ICWS), Milan, Italy, 8–13 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 180–187.
7. Paul, A.; Dhar, S.; Roy, S. Applying interrater reliability measure for user credibility assessment in reputation-oriented service discovery. In Proceedings of the Web Intelligence, Venice, Italy, 26–29 October 2023; SAGE Publications Sage: London, UK, 2023; Volume 21, pp. 167–180.
8. Du, X.; Xu, J.; Cai, W.; Zhu, C.; Chen, Y. Oprc: An online personalized reputation calculation model in service-oriented computing environments. *IEEE Access* **2019**, *7*, 87760–87768. [CrossRef]
9. Yelp Open Dataset. Available online: https://www.yelp.com/dataset (accessed on 3 February 2026).
10. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining(ICDM), Washington, DC, USA, 15–19 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 413–422.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.