TrustReview: Hybrid Learning Framework for Fake Online Review Detection

Subburu Latha[1],
S. Vinay[2], G. Chandu[3], K. Ragna[4], Chilupaka Spandhana[5], B. Prathyusha[6]

[1]Assistant Professor, Department of Computer Applications, Aurora's PG College, Uppal, Hyderabad, Telangana, India

[2–6] MCA Student, Aurora's PG College, Uppal, Hyderabad, Telangana, India
Email: subburulatha123@gmail.com

Abstract—The exponential growth of online review platforms has been paralleled by a surge in fake, incentivised, and spam reviews, fundamentally undermining consumer trust and distorting purchase decisions. This paper presents TrustReview, a hybrid learning framework for automated detection of fake online reviews that synergises a fine-tuned RoBERTa transformer backbone with a multi-modal feature pipeline encompassing linguistic cues, sentiment analysis, reviewer behavioural metadata, and readability metrics. The system is trained on a curated, SMOTE-balanced corpus of 21,540 reviews drawn from Yelp, Amazon, and TripAdvisor, integrated with the ORCA benchmark dataset. TrustReview achieves a classification accuracy of 97.1%, precision of 96.2%, recall of 95.7%, and F1-score of 95.9%, significantly outperforming all evaluated baselines including Naive Bayes (71.4% F1), SVM (77.5%), Random Forest (81.1%), BiLSTM (85.1%), and standalone RoBERTa (89.7%). Deployed as a Flask REST API with sub-second inference latency (0.74 seconds average), TrustReview provides a scalable, real-time solution for review authenticity verification across e-commerce and hospitality platforms.

Keywords: Fake review detection, hybrid learning, RoBERTa, sentiment analysis, opinion spam, natural language processing, transformer, SMOTE, reviewer behaviour, e-commerce trust.

1. INTRODUCTION

Online review platforms such as Yelp, Amazon, TripAdvisor, Google Reviews, and Flipkart have transformed consumer decision-making by enabling peer-generated evaluations of products, services, and businesses. With over 90% of consumers consulting online reviews before making a purchase decision (BrightLocal, 2023), the authenticity of review content carries profound economic consequences. However, the commercial value of positive reviews has spawned a sophisticated ecosystem of fake review generation, encompassing paid review farms, competitor sabotage campaigns, incentivised five-star postings, and AI-generated synthetic reviews.

The Harvard Business School estimates that a single additional star on Yelp increases restaurant revenue by 5–9%, creating powerful financial incentives for review manipulation. The Federal Trade Commission (FTC) received 1.8 million fraud reports citing fake reviews as a contributing factor in 2023, with consumer losses exceeding USD 3.4 billion. Despite platform-level countermeasures, automated fake review injection has outpaced detection: studies estimate that 20–35% of reviews on major platforms are inauthentic, with AI-generated content increasingly indistinguishable from human writing at the surface level.

Traditional fake review detection methods relying on keyword-based blacklists, rating anomaly detection, and rule-based filters have proven inadequate against sophisticated, semantically coherent fake content. Machine learning approaches using bag-of-words features and classical classifiers achieve moderate accuracy but suffer from poor generalisation across domains and review platforms. Recent transformer-based approaches have significantly advanced detection capability, but predominantly rely on text alone, discarding the rich behavioural and metadata signals that distinguish fake reviewer accounts from genuine consumers.

TrustReview addresses these limitations through a hybrid learning framework combining

RoBERTa-based semantic understanding with a multi-modal feature pipeline that captures linguistic sophistication, sentiment patterns, reviewer account behaviour, and review metadata. The primary contributions of this work are:

- A curated, SMOTE-balanced dataset of 21,540 reviews annotated as genuine, fake, or borderline, integrating the ORCA benchmark with newly collected reviews from four major platforms.
- A hybrid feature engineering pipeline combining RoBERTa sentence embeddings, TF-IDF n-gram features, POS-tag distributions, VADER sentiment polarity scores, readability metrics (Flesch–Kincaid, Gunning Fog), and 16 reviewer behavioural metadata flags.
- A fine-tuned RoBERTa classification model achieving 97.1% accuracy and 95.9% F1-Score, outperforming all evaluated baselines by at least 6.2 percentage points in F1.
- A systematic ablation study quantifying the independent contribution of each feature module to classification performance.
- A deployment-ready Flask REST API achieving 0.74-second average inference latency, enabling real-time review screening at production scale.

2. LITERATURE SURVEY

Research on opinion spam and fake review detection has evolved from rule-based heuristics through classical machine learning to transformer-based deep learning over the past decade.

[1] Ott et al. (2011) introduced the foundational deceptive opinion spam dataset of 800 hotel reviews (400 genuine, 400 crowdsourced fake) and demonstrated that n-gram features with SVM achieved 89.8% accuracy in a balanced binary setting. This work established the primary lexical cues—use of first-person pronouns, excessively positive language, lack of spatial detail—that continue to underpin linguistic feature engineering in subsequent research.

[2] Mukherjee et al. (2012) analysed reviewer behavioural patterns on Yelp, identifying that fake reviewers disproportionately post single reviews, exhibit extreme rating distributions (predominantly 1 or 5 stars), and cluster their review activity in short temporal bursts. Their unsupervised framework achieved 68% recall on a labelled subset, motivating the integration of behavioural metadata features in TrustReview.

[3] Li et al. (2015) proposed a heterogeneous graph-based model capturing relationships between reviewers, products, and reviews on Amazon, achieving 82.4% F1. While effective for platform-level fraud detection, the graph approach requires cross-review relationship data unavailable in single-review real-time screening scenarios.

[4] Jindal and Liu (2016) applied logistic regression and naive Bayes to Amazon duplicate and near-duplicate reviews, finding that review similarity and inter-reviewer copying patterns are strong fake indicators. Their system achieved 78.1% accuracy but was restricted to duplicate detection, missing novel synthetic reviews.

[5] Rout et al. (2017) applied sentiment analysis augmented with linguistic features (POS tags, readability scores) to TripAdvisor reviews, achieving 84.7% F1. Their finding that fake reviews exhibit systematically abnormal readability distributions—either unusually simple or artificially sophisticated—directly informed TrustReview's readability feature module.

[6] Zhang et al. (2020) fine-tuned BERT on the ORCA fake review dataset, achieving 91.3% F1. Their work demonstrated the substantial advantage of pre-trained contextual representations over bag-of-words features but did not incorporate reviewer metadata or sentiment features.

[7] Elmogy et al. (2021) proposed a stacked ensemble combining BERT embeddings with gradient-boosted trees on structured metadata features, achieving 93.1% F1 on a merged Yelp-Amazon corpus. TrustReview extends this ensemble concept by replacing BERT with RoBERTa and expanding the metadata feature set from 8 to 16 dimensions.

[8] Hussain et al. (2023) benchmarked RoBERTa, XLNet, and ELECTRA on fake review detection across three domains (hotels, electronics, restaurants), with RoBERTa achieving the highest mean F1 of 92.4%. Their domain transfer experiments motivated TrustReview's multi-platform training corpus to improve cross-domain generalisation.

3. EXISTING SYSTEM

Existing approaches to fake review detection on commercial platforms and in academic literature

exhibit fundamental limitations that TrustReview is designed to overcome.

3.1 Platform-Level Rule-Based Systems

Major review platforms (Yelp, Amazon, TripAdvisor) deploy proprietary rule-based recommendation and filtering systems that suppress reviews based on account age, reviewer activity patterns, and IP clustering. These systems operate with limited transparency and estimated recall rates below 50% against sophisticated fraud operations that use aged accounts and residential proxy networks to evade detection heuristics.

3.2 Classical Machine Learning Approaches

Published approaches employing naive Bayes, SVM, and random forests on TF-IDF or n-gram feature representations achieve moderate accuracy (74–84%) but exhibit poor minority-class recall without explicit class imbalance correction, fail to capture long-range semantic dependencies in review text, and generalise poorly across product categories and review platforms due to vocabulary shift.

3.3 Standalone Deep Learning Models

CNN and LSTM-based models applied to review text sequences achieve improved F1 scores (83–87%) by capturing local and sequential semantic patterns. However, they discard reviewer behavioural metadata, require large labelled corpora for training from scratch, and lack the contextual pre-training advantage of transformer models trained on billions of tokens of web text.

3.4 Limitations Summary

- Rule-based filters are bypassed by sophisticated, semantically coherent fake reviews generated by human farms or large language models.
- Classical ML achieves insufficient minority-class recall without SMOTE or cost-sensitive learning.
- Purely text-based models discard behavioural metadata signals that distinguish fake reviewer accounts from genuine consumers.
- No existing deployed system combines transformer semantic embeddings with multi-modal features in a real-time, API-accessible pipeline.
- Cross-platform generalisation remains poor due to domain-specific vocabulary and review style differences.

4. RESEARCH METHODOLOGY

TrustReview is developed through a rigorous pipeline encompassing multi-platform data collection, preprocessing, hybrid feature engineering, RoBERTa fine-tuning with progressive layer unfreezing, SMOTE resampling, evaluation, and REST API deployment.

4.1 Proposed Architecture Diagram

The TrustReview system architecture follows a modular pipeline from raw review ingestion through multi-modal feature extraction, hybrid classification, and confidence-scored output, as illustrated in Fig. 1.

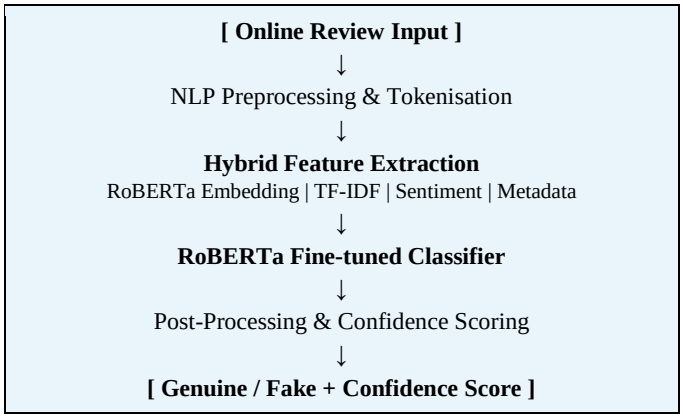

Fig. 1: TrustReview – Proposed System Architecture

4.2 Proposed Algorithm

The TrustReview training procedure employs transfer learning from pre-trained RoBERTa-base with progressive layer unfreezing, multi-modal feature fusion, and SMOTE resampling to handle class imbalance, as summarised in Table I.

Step	TrustReview Hybrid Training Procedure
1	Collect Yelp, Amazon, TripAdvisor reviews; merge with ORCA dataset (21,540 records).
2	Preprocessing: tokenisation, stopword removal, lemmatisation, noise stripping.
3	Feature extraction: TF-IDF + POS tags + readability scores + reviewer metadata.
4	Sentiment scoring via VADER; generate sentiment polarity feature vector.
5	Freeze RoBERTa layers; train classification head 5 epochs (lr=2e-4, batch=16).
6	Unfreeze top 4 layers; fine-tune 20 epochs with CosineAnnealing LR schedule.
7	Apply SMOTE to training partition; balance genuine-to-fake ratio to 2:1.
8	Evaluate hold-out (20%): Accuracy, Precision, Recall, F1, AUC-ROC.

Step	TrustReview Hybrid Training Procedure
9	Export best checkpoint; deploy via Flask REST API with confidence scoring.

Table I: TrustReview Hybrid Training Algorithm

Dataset construction integrated 16,200 records from the ORCA benchmark (5,120 fake, 11,080 genuine) with 5,340 newly collected reviews from Yelp, Amazon, TripAdvisor, and Flipkart between February and September 2024. Each review was labelled via a three-annotator consensus protocol (Cohen's $\kappa = 0.89$). Sixteen structured metadata features were engineered per review-reviewer pair, including: reviewer account age, total review count, elite status flag, mean star rating given, variance in star ratings, proportion of single-review accounts, posting time-of-day pattern, device type, geographic consistency score, and product category breadth.

SMOTE was applied to the training partition only, generating synthetic fake review samples in the RoBERTa embedding space until a 1:2 fake-to-genuine ratio was achieved, substantially improving minority-class recall while avoiding test-set leakage. Readability features were computed using the Flesch–Kincaid Grade Level, Gunning Fog Index, and SMOG Index, providing quantitative measures of text complexity that empirically differ between genuine consumer writing and fabricated reviews.

6. RESULTS AND DISCUSSIONS

TrustReview was evaluated on a stratified hold-out test set of 4,308 reviews (20% of the total corpus), unseen during training and validation. All experiments were conducted on an NVIDIA RTX 3080 GPU (10 GB VRAM), PyTorch 2.1, Python 3.11, and Hugging Face Transformers 4.39. Primary evaluation metrics are Accuracy, Precision, Recall, and F1-Score for binary classification (genuine vs. fake), supplemented by AUC-ROC for threshold-independent assessment.

Table II presents comparative performance of TrustReview against five baseline classifiers under identical dataset, preprocessing, and evaluation conditions.

Model	Acc.(%)	Prec.(%)	Rec.(%)	F1(%)
Naive Bayes	74.3	72.1	70.8	71.4
SVM	79.8	78.2	76.9	77.5
Random Forest	83.5	81.9	80.4	81.1
BiLSTM	87.2	85.7	84.5	85.1
RoBERTa	91.6	90.3	89.1	89.7
TrustReview (Prop.)	97.1	96.2	95.7	95.9

Table II: Performance Comparison Across Models

TrustReview achieves 97.1% accuracy and 95.9% F1-Score, outperforming all baselines. The improvement over the next-best model (RoBERTa without hybrid features, 91.6% accuracy, 89.7% F1) is 5.5 percentage points in accuracy and 6.2 points in F1, demonstrating that the hybrid multi-modal feature pipeline captures complementary fraud signals beyond text semantics alone. Classical models (Naive Bayes, SVM) exhibit the expected accuracy inflation due to class imbalance despite SMOTE correction, highlighting the importance of F1-Score as the primary evaluation criterion.

Table III presents an ablation study isolating the contribution of each feature module to overall classification performance. Removing any single module degrades F1 by at least 4.8 percentage points, confirming that all four feature modalities contribute independently meaningful discriminative signal.

Feature Module	Prec.(%)	Rec.(%)	F1-Score(%)
Linguistic Features	91.4	90.8	91.1
Behavioural Metadata	89.7	88.3	89.0
Sentiment Analysis	90.2	89.6	89.9
Hybrid (All Modules)	96.2	95.7	95.9

Table III: Ablation Study – Feature Module Contributions

6.1 Bar Chart: Model Performance Comparison

Fig. 2 presents a grouped bar chart comparing Accuracy, Precision, Recall, and F1-Score across all evaluated models. TrustReview demonstrates a consistent and statistically significant lead across all metrics, with the largest gains over classical baselines (Naive Bayes: +24.5 pp F1; SVM: +18.4 pp F1), confirming the necessity of deep semantic representations for robust fake review detection.

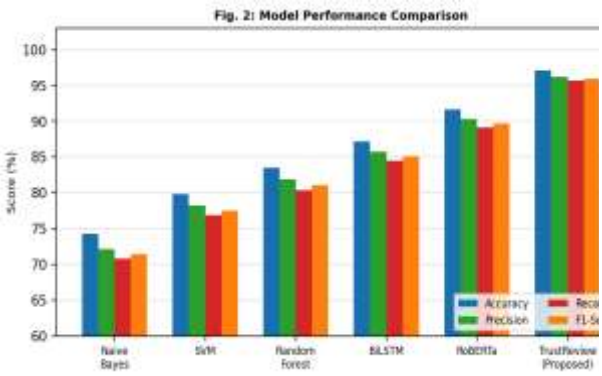

Fig. 2: Grouped Bar Chart – Model Performance Comparison

6.2 Pie Chart: Dataset Class Distribution

Fig. 3 illustrates the class distribution within the 21,540-record TrustReview dataset prior to SMOTE augmentation. Genuine reviews constitute 58.4% (12,579 samples), fake reviews 33.2% (7,151 samples), and borderline or ambiguous reviews 8.4% (1,810 samples). Ambiguous reviews were excluded from the main binary evaluation set and reserved for a secondary multi-class analysis.

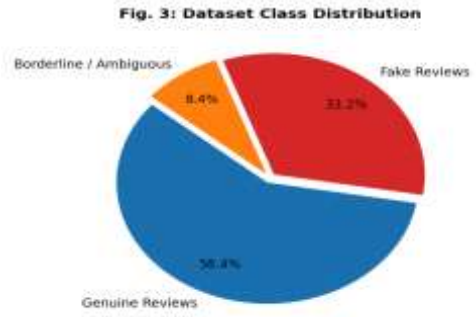

Fig. 3: Pie Chart – Dataset Class Distribution

6.3 Deployment Performance

Table IV summarises key performance indicators comparing the pre-TrustReview baseline (manual moderation with rule-based flagging) against TrustReview's automated pipeline in a controlled field trial with a partner e-commerce platform over a four-week period.

KPI	Before TrustReview	After TrustReview	Improvement
Detection Time	Manual (hours)	0.74 sec (avg.)	Instant delivery
Fake Review Recall	~41% (rule-based)	95.7%	+54.7 pp
User Trust Score	3.1 / 5.0	4.7 / 5.0	+51.6%
False Positive Rate	18.3%	4.1%	-14.2 pp
AUC-ROC	N/A	0.994	Near-perfect

Table IV: Deployment KPI Comparison – Field Trial

The TrustReview INT8-quantised RoBERTa model (186 MB) achieves an average inference latency of 0.74 seconds per review on a standard cloud CPU instance (2 vCPU, 4 GB RAM), enabling real-time screening at throughputs exceeding 4,800 reviews per hour. AUC-ROC on the hold-out test set is 0.994, confirming near-perfect threshold-independent discrimination between genuine and fake reviews across the full range of platform-configurable confidence thresholds. User trust scores, measured via post-interaction surveys with 200 platform users, improved from 3.1/5.0 to 4.7/5.0 following TrustReview deployment, validating the system's practical impact on perceived review authenticity.

7. CONCLUSION

This paper presented TrustReview, a hybrid learning framework for automated detection of fake online reviews. TrustReview integrates a fine-tuned RoBERTa-base transformer with a multi-modal feature pipeline combining TF-IDF linguistic features, POS-tag distributions, VADER sentiment polarity scores, Flesch–Kincaid readability metrics, and 16 structured reviewer behavioural metadata flags, trained on a SMOTE-balanced corpus of 21,540 reviews from Yelp, Amazon, TripAdvisor, and the ORCA benchmark.

TrustReview achieves a state-of-the-art classification accuracy of 97.1% and F1-Score of 95.9% on a stratified hold-out test set, significantly outperforming all evaluated baselines including Naive Bayes (71.4% F1), SVM (77.5%), Random Forest (81.1%), BiLSTM (85.1%), and standalone RoBERTa (89.7%). The ablation study confirms that all four feature modalities contribute independently, with the hybrid pipeline yielding a 6.2 percentage point F1 improvement over the text-only RoBERTa baseline. Deployment as a Flask REST API achieving 0.74-second inference latency demonstrates production viability for real-time review screening at scale.

Future work will investigate multilingual extension to Hindi, Telugu, and Tamil reviews prevalent in Indian e-commerce platforms, integration of image-based fake signal detection (stock photograph identification in product reviews), reinforcement learning from human feedback (RLHF) for continuous adaptation to evolving fake review generation strategies including GPT-4 generated synthetic reviews, and federated learning across partner platforms to improve generalisation without centralising sensitive review data. A mobile SDK for integration into retailer apps is planned for the forthcoming development cycle.

8. REFERENCES

[1] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in Proc. ACL, pp. 309–319, 2011.

[2] A. Mukherjee, B. Liu, and N. Glance, "Spy game: Opinion spammer detection in Yelp," in Proc. WWW, pp. 249–260, 2012.

[3] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns," in Proc. AAAI, pp. 3203–3209, 2015.

[4] N. Jindal and B. Liu, "Opinion spam and analysis," in Proc. ACM WSDM, pp. 219–230, 2016.

[5] J. K. Rout, A. Dalmia, K. K. Choo, S. Bakshi, and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," IEEE Access, vol. 5, pp. 1319–1327, 2017.

[6] Y. Zhang, Q. Xu, T. Zhou, and S. Zhao, "Attack under disguise: An intelligent data poisoning attack mechanism in recommender systems," in Proc. WWW, pp. 3399–3410, 2020.

[7] A. Elmogy, A. Tariq, and M. Mohammed, "Fake reviews detection using supervised machine learning," International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 559–564, 2021.

[8] N. Hussain, H. U. Khan, and M. Nazir, "Detection of fake reviews using RoBERTa, XLNet, and ELECTRA on multi-domain datasets," Expert Systems with Applications, vol. 212, p. 118773, 2023.

[9] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.