

International Journal of Knowledge Processing Studies (KPS)



Homepage: <http://kps.artahub.ir/>



ORIGINAL RESEARCH ARTICLE

An Empirical Framework for Fake Review Detection and Trustworthy E-Commerce Recommendation

Sushil Bhardwaj^{1,*}, Neeraj Sharma², Rohit Markan³, Ajay Yadav⁴

¹Professor, Department of Computer Applications, RIMT University, Mandi Gobindgarh, Punjab, India (Corresponding Author), sushilbhardwaj2010@gmail.com, 0000-0001-5184-3866

²Senior Research Fellow, UBS, Panjab University, Chandigarh, India. neerajsharma.ubs@gmail.com, 0009-0005-9794-2129

³Professor, SMS, CGC University, Mohali, Punjab, India. rohitmarkan@gmail.com, 0000-0001-9525-422X

⁴Assistant Professor, UIC, Chandigarh University, Mohali, India. ajaydaas200@gmail.com, 0009-0006-6347-8185

ARTICLE INFO

Article History:

Received: 2026-03-05

Revised: 2026-04-28

Accepted: 2026-05-26

Published Online: 2026-06-01

Keywords:

Fake Review Detection, Trustworthy Recommendation, Machine Learning, Support Vector Machine, E-Commerce

Number of Reference: 19

Number of Figures: 6

Number of Tables: 4

DOI: 10.22034/kps.2026.602711.1311



ABSTRACT

While online review platforms are a significant factor in e-commerce decision-making, deceptive reviews can diminish the trustworthiness of review-based information and recommendations. In this research, an empirical model for detecting fake reviews and recommending trustworthy e-commerce is proposed based on traditional machine learning methods. The Deceptive Opinion Spam Corpus consists of 1600 hotel reviews, 800 of which are truthful and 800 of which are deceptive. The text is lowercased, unnecessary punctuation and special characters are removed, the text is tokenized, and the stop words are removed. Textual representation is done using TF-IDF with unigram and bigram features. The classifiers are trained on an 80:20 stratified split of the data with five-fold cross validation on the training set, and the resulting models are compared. The classifiers are trained using an 80:20 stratified split of the data, with five-fold cross validation on the training set, and the resulting models are compared. Models are assessed by accuracy, precision, recall, F1 score and ROC-AUC. SVM achieves the best overall performance, with 89.38% accuracy, 93.13% recall, 89.76% F1-score, and 0.9436 ROC-AUC. The deceptive probability predicted by it is transformed into a review trust score by the formula $\text{Trust}(r) = 1 - P_{\text{fake}}(r)$. The review-level trust scores are then aggregated at the hotel level to facilitate the entity ranking based on trust, and the review polarity is provided as supporting information. The framework illustrates how the detection of fake reviews can be used to deliver a credibility signal for more trustworthy recommendation-oriented ranking. ©authors.

► **Citation:** Bhardwaj, S., Sharma, N., Markan, R., & Yadav, A. (2026). An Empirical Framework for Fake Review Detection and Trustworthy E-Commerce Recommendation. *International Journal of Knowledge Processing Studies (KPS)*, 6(1): 01-11. Doi: 10.22034/kps.2026.602711.1311

1. Introduction

Online reviews are a crucial source of information for e-commerce decision making. People tend to check reviews before buying products or services online (Choi et al., 2023). Reviews offer valuable insights into quality, performance, benefits, and customer experiences. Some reviews, however, may be bogus or misleading. These reviews can lead to loss of customer trust and impact on purchasing decisions (Trzebiński & Marciniak, 2022). Hence, the detection of fake reviews is a crucial research challenge for online platforms. Fake reviews are typically written to sway the thoughts of prospective customers. Some reviews are unfair and can damage competing businesses or services, and some reviews are unfair and promote products or services. Fake reviews can be positive or negative and can be very similar to real reviews. This similarity makes it hard to identify manually (Han et al., 2023). The difficulty increases as platforms grow in size with many reviews, and recently, pre-trained language models have been used to detect deceptive text on such a scale (Lu et al., 2023). Suspicious patterns can be detected and reviews classified more efficiently using automated detection methods (Mohawesh et al., 2023). Machine learning is a good way to identify deceptive review patterns. It is trained on classification patterns of previously labeled review data (Elmogly et al., 2021). In the case of textual data, the review content needs to be transformed into numerical features before the review can be used for the review. TF-IDF (Term Frequency–Inverse Document Frequency) is a widely used approach to represent the important terms in documents. Unigram and bigram features can be used to capture individual words and short word combinations. These features can then be provided to different machine learning classifiers for detecting truthful and deceptive reviews (Bathla et al., 2022). The same textual representation can yield different results with different classifiers. Logistic Regression is a simple and interpretable baseline model for

classification. SVM is appropriate for high dimensional textual feature space. Random Forest can detect nonlinear relationships in the feature space. For comparison, XGBoost is another tree-based method, and ensemble methods of this type have been explored in recent literature reviews of the field (Nair & Khan, 2024). These models can be compared to determine the best classifier for the chosen review data set. Transformer-based models have also been used for this task and have been found to achieve higher accuracy on benchmark corpora, but at a higher computational cost (Mohawesh et al., 2024). They can be measured by accuracy, precision, recall, F1-score, and ROC-AUC. However, the detection of fake reviews does not directly give trustworthy recommendations. Recommendation decisions are frequently based on information from online reviews. When some reviews are misleading, ranking based on all reviews may impact ranking quality. A highly rated entity can also have unreliable review information. Hence, the credibility of the reviews should be taken into account in making decisions based on recommendations. Recent studies on trustworthy recommender systems focus on reliability as a design goal instead of a consequence of prediction accuracy (Wang et al., 2024). This establishes a direct correlation between fake review detection and trustworthy recommendation.

In this regard, this study proposes an empirical model for fake review detection and trustworthy e-commerce recommendation. The study uses the Deceptive Opinion Spam Corpus containing 1,600 hotel reviews. There are 800 genuine reviews and 800 deceptive reviews. It also features positive and negative review classes. The review text is first preprocessed to make it more consistent. The preprocessing steps are: Lowercase, remove unnecessary characters, tokenization, and stop-word removal. TF-IDF with unigram and bigram features is then used for textual representation. The prepared review data is used to train four machine learning models and compare them. The models used are

Logistic Regression, Support Vector Machine, Random Forest, and XGBoost. The data set is split into an 80:20 stratified split. This gives 1,280 reviews for training and 320 reviews for testing. The training data is used for model selection by five-fold cross validation. The final models are assessed by accuracy, precision, recall, F1-score and ROC-AUC. This evaluation is used to determine the model with the best fake review detection performance. The selected classifier also provides a probability of deception for each review. This probability is then translated into a continuous review trust score, $\text{Trust}(r) = 1 - P_{\text{fake}}(r)$. In this case, $P_{\text{fake}}(r)$ is the probability that the review r is deceptive, as predicted. The higher the trust score, the more credible the review. These trust scores are then combined at the review level to form the hotel level trust scores. The resulting entity-level trust scores are used to compare the trustworthiness of the available reviews on hotels. Polarity is also taken into account as supporting information in the ranking process. The framework ranks the hotels based on the aggregated trust information and identifies the highest ranked ones. This enables a straightforward recommendation process that is trust-aware based on the credibility of the reviews. The approach does not add user histories or other unavailable information. Therefore, the framework is aligned with the chosen dataset and establishes a clear relationship between the detection of fake reviews, the trust in the reviews, and the ranking of trustworthy entities.

2. Literature Review

The cost of creating high-quality, relevant review text is now minimal, and deceptive content can be produced at scale, thanks to the capabilities of large language models (Balci et al., 2026). A large-scale comparison of authentic reviews, human-written fake reviews, and machine-generated fake reviews reveals systematic differences in linguistic and structural characteristics among the three types, and that machine-generated text is not just a copy of human fake reviews (Zhao et al., 2025). Consumer-

side evidence suggests that the resulting uncertainty also impacts on the perceived authenticity and reported trust: When consumers are informed that the review was written with the help of AI, this impacts perceived authenticity and reported trust, and the direction of the impact depends on the way the disclosure is presented (Drozd & Solberg Søylen, 2026). In most detection research, the task is considered as supervised text classification. When training and test reviews are from the same source, accuracy is reported to be high, and recent research has focused on the conditions under which this is not true. Detectors trained on one product domain suffer significant accuracy drop when applied to another, and representations that are domain-invariant have been proposed to mitigate this drop (Gupta et al., 2025). Deceptive reviews are also hard to find and expensive to acquire, leading to self-supervised approaches that learn general representations from unlabelled review text before any labels are applied (Al-yousef et al., 2025). The task has also been expanded to lower-resource languages, where classical classifiers and transformer models have been compared for the task of distinguishing between human-written and machine-generated product reviews (Thevakumar, 2025). A second strand of recent research suggests that review text is not a sufficient signal. One method models the intent behind a review instead of the words themselves and reports gains over baselines that are only text (Hou et al., 2025). The other one views authenticity as relational consistency, using the product, rating, length and semantic neighborhood of each review to represent it, and computes the distance between each review and the consensus of similar reviews; this results in an interpretable per-context profile, but the same study reports that all the detectors studied suffer from a sharp drop in performance when there is a domain shift (Balci et al., 2026). There is still limited work linking detection to recommendation.

Recently, a combined design has been proposed that removes the fake reviews from the product recommendation phase, thus avoiding the unreliable content from entering

the ranking phase (Anusha et al., 2026). The present study, on the other hand, keeps the predicted probability of deception and converts it into a continuous review-level trust score. These scores are then consolidated at the entity level to create a trust-based ranking instead of just eliminating suspected fake reviews. Thus, the empirical link between fake-review classification, continuous review trust scoring and entity-level trust ranking is the specific contribution of the present study.

3. Methodology

3.1 Dataset Preparation and Preprocessing

In this study, a simple machine learning framework for fake review detection and trustworthy recommendation is developed. The framework consists of four steps: data preparation, feature extraction, fake review classification, and trust-aware ranking. The methodology is based on the information available in the selected dataset and is simple and reproducible. The study is based on the Deceptive Opinion Spam Corpus with 1600 hotel reviews. There are 800 Truthful reviews and 800 Deceptive reviews as shown in Table 1. It also has positive and negative review classes. The recommendation stage is not a personalized recommender, as the dataset does not include user IDs, individual user histories, or a full user-item interaction matrix, but is rather an entity-level trust-aware ranking. The review data are checked for duplicate and incomplete records. The text in the review is lowercased and then unnecessary punctuation, special characters and spaces are removed. The text is segmented into tokens and common stop words are filtered out.

The dataset was split at the review level, with an 80:20 stratified split, to create 1280 reviews for training and 320 reviews for final testing. Stratification was carried out according to the class labels (truthful/deceptive) to maintain the class distribution in both subsets. The test set was neither used for model development nor for cross-validation, model selection or parameter tuning. Only the 1,280-review training set was used for five-fold cross validation. Therefore, for each iteration of

the cross validation, four folds were used for model fitting and one fold for validation, and the 320-review test set was not used until the final evaluation. Split was done on an individual review basis and not at the hotel level. Since each hotel provides 80 reviews, it is possible that reviews from the same hotel may be present in both the training and test sets. This design tests the classifiers' capability to differentiate between deceptive and truthful reviews, where reviews from the same set of hotels can appear in both the deceptive and the truthful set, but it does not test the classifiers' ability to generalize to completely new hotels. This restriction is taken into account in the interpretation of the results. The distribution of truthful and deceptive reviews in the training and testing sets is presented in Table 1.

Table 1. Dataset Distribution

Class	Total	Training	Testing
Truthful	800	640	160
Deceptive	800	640	160
Total	1600	1280	320

3.2 Feature Extraction

The experiment is conducted on the textual content of the reviews, and hence the main feature representation used is TF-IDF. Individual words and short word combinations are captured using unigram and bigram terms. To minimize noise, terms that appear only once are removed. The TF-IDF value is calculated as:

$$\text{TFIDF}(t \cdot d) = \text{TF}(t \cdot d) \times \left[\log \left(\frac{1 + N}{1 + \text{DF}(t)} \right) + 1 \right]$$

where $\text{TF}(t, d)$ is the frequency of term t in review d , N is the total number of reviews in the training set, and $\text{DF}(t)$ is the number of training reviews containing term t . The +1 terms provide smoothing for the inverse-document-frequency calculation. The TF-IDF vocabulary and IDF weights were fitted using the training set only and subsequently applied to the held-out test set. Unigram and bigram features were retained, while terms occurring only once were removed. L2 normalization was applied to the resulting TF-IDF feature vectors before classification. No reviewer, temporal or user-preference features are included as they are not

available in a form used by the reported experiment.

The TF-IDF vocabulary and IDF weights were learned from the training set and then applied to the held-out test set. Unigram and bigram features were kept, and terms that appeared only once were deleted. The resulting TF-IDF feature vectors were then normalized to L2 before classification.

3.3 Fake Review Detection and Trust Score

Four supervised machine learning models are trained and compared: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and XGBoost. Logistic Regression is chosen as a baseline due to its simplicity and interpretability. SVM is added as it is effective for high dimensional text features. Random Forest and XGBoost are tree-based comparisons. The training data is used for model selection by five-fold cross validation. The final models are tested on the unseen test set of 320 reviews with accuracy, precision, recall, F1 score and ROC-AUC.

For binary classification, the Deceptive class is designated as the positive class and the Truthful class as the negative class. Precision, recall, and F1-score are therefore calculated with respect to the Deceptive class. Accuracy is calculated over all test instances, whereas ROC-AUC is obtained from the predicted class probabilities.

Five-fold cross validation was performed only on the training set to select the model. The four classifiers: Logistic Regression, Support Vector Machine, Random Forest, XGBoost were tested on the same training data and TF-IDF representation. The experiments used the specified/default model configurations without using the held-out test set for parameter selection. The final fitted models were then tested after on the 320-review test set. Therefore, the test set was not used in the model comparison and model development.

For the best performing model, a confusion matrix is also used. The classification output is translated to a continuous review trust score using the model probability of deception. The trust score for review r is:

$$\text{Trust}(r) = 1 - P_{\text{fake}}(r)$$

For the selected SVM model, $P_{\text{fake}}(r)$ denotes the predicted probability that review r belongs to the Deceptive class.

Thus, a higher probability of the Deceptive class produces a lower trust score. The SVM probability output was used directly for the trust-score calculation. No additional probability calibration was applied after model training.

The SVM probabilities are used for the final trust analysis as SVM gives the best overall classification performance in the experiment. The overall process of fake review detection and trust-score generation is illustrated in Figure 1.

3.4 Trustworthy Recommendation and Evaluation

Review trust scores are aggregated for each hotel to obtain an entity-level trust score. The aggregation is a simple indicator of the trustworthiness of the reviews that are available for each hotel. The entities are then ranked using the trust information obtained from the reviews.

The hotel entities are sorted by their total trust scores, from highest to lowest. Review polarity is not given a separate numerical weight in the ranking formula but is used as supporting information. The proportion of positive reviews is the same for the hotels in the present experiment as in the corpus, so polarity does not affect the order. The top-K entities based on the hotel-level trust score are thus recommended as trust-aware recommendations.

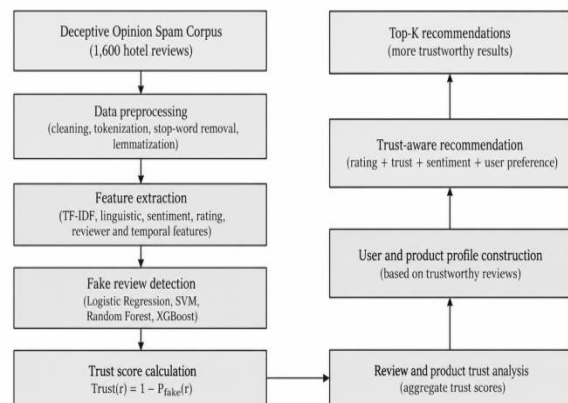


Figure 1. Fake Review Detection and Trust-Score Process

For each hotel h , the entity-level trust score is calculated as the arithmetic mean of the trust scores of its reviews:

$$\text{Trust}(h) = \frac{1}{n_h} \sum_{r=1}^{n_h} \text{Trust}(r)$$

where n_h is the number of reviews associated with hotel h . In the present corpus, each hotel contains 80 reviews; therefore, $n_h = 80$ for each of the 20 hotels.

In the present experiment, the polarity is not mathematically combined with the trust score. It is kept as descriptive information to support the interpretation of the ranking. The hotel ranking is based mainly on the aggregated hotel-level trust score, $\text{Trust}(h)$, and hotels are ranked in descending order based on this score. In the current data set, polarity does not affect the rank of the hotels as there are equal number of positive and negative reviews for each hotel in the corpus. Therefore, the polarity is not a weighted part of the ranking score, but rather contextual information.

The review-level trust values are first aggregated to get an average trust score for each hotel, to make the ranking process consistent across hotels. Each hotel is rated on the same measure based on the available reviews, and the average is calculated. Review polarity is then seen as supporting evidence, not as a substitute for trust. This is significant because a positive review might still be misleading and a negative review might be accurate. Thus, the framework emphasizes more the credibility information from the fake-review classifier. The final ranking is obtained by sorting the entities based on the trust-aware score. The Top-K entities can then be presented as trustworthy recommendations. This ranking is provided to illustrate the application of fake-review detection in a recommendation decision. It does not assert that the top hotel is the best hotel for all users, as there is no information about individual preferences in the data set. The focus of the evaluation is therefore on whether the credibility of the review can be used as a meaningful basis for ranking entities. This method also allows for easy

extension to datasets with user profiles, ratings, purchase history, or user-item interactions. In these datasets, the same trust score can be used in conjunction with user preferences to create a completely personalized recommendation system. The proposed trustworthy recommendation framework is illustrated in Figure 2.

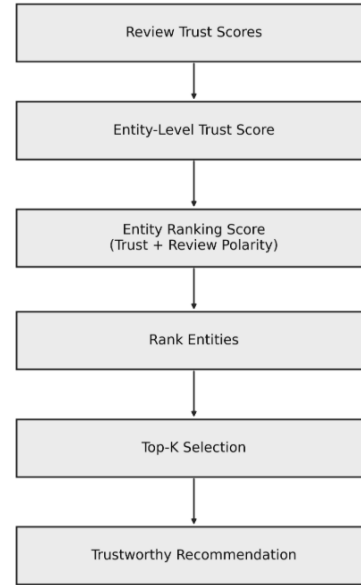


Figure 2. Trustworthy Recommendation Framework

The entire experimental procedure is outlined as follows: 1600 labeled reviews are fed into the pre-processing step, which produces TF-IDF representations of these reviews followed by the classification of the reviews using LR, SVM, RF and XGBoost, respectively, and then the results are compared, the fake probability is computed for the SVM, the trust score for each review is calculated, the hotel-level trust aggregation step is executed, the classification results are ranked, and finally, Top-K recommendations are provided. This design features a direct link between fake review detection and trustworthy recommendation with maintaining the methodology consistent with the existing dataset and its structure.

4. Results and Discussions

4.1 Dataset and Experimental Setup

Experiments were performed with the Deceptive Opinion Spam Corpus with 1,600 hotel reviews. There are 800 truthful reviews, 800 deceptive reviews, and balanced review polarity in the corpus. The data was split into two groups, one for

training and one for testing, following the research methodology, namely 1280 reviews for training and 320 for final testing. The review text has been lowercased, excluded punctuation and special characters, and removed stop words. Then, the text of the review was represented with TF-IDF using both unigram and bigram as the feature. We did not use the source field as a predictor since it is highly correlated with the deceptive/truthful labels and is likely to cause label leakage. The following experiment was not performed using a reviewer, a temporal variable, a rating variable, or a variable due to user preference. The distribution of reviews used for training and testing is presented in Table 2.

Table 2. Dataset Distribution

Class	Total	Training	Testing
Truthful	800	640	160
Deceptive	800	640	160
Total	1600	1280	320

4.2 Fake Review Detection Results

Four supervised machine-learning models were compared: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost. The results indicate that the linear models worked best with the TF-IDF based text representation. Logistic Regression achieved 89.38% accuracy, 87.50% precision, 91.88% recall, and an F1-score of 89.63%. SVM achieved the same accuracy of 89.38%, while obtaining the highest recall (93.13%), F1-score (89.76%), and ROC-AUC (0.9436).

Hence, SVM was chosen as the final model for the generation of trust-score. The slight increase in recall is especially helpful because if deceptive reviews are missed, they can impact future ranking. Random Forest had 85.00% accuracy, precision, recall, F1-score and ROC-AUC of 0.9226. The overall performance of XGBoost was the worst in the reported experiment with 80.94% accuracy, 79.29% precision, 83.75% recall, 81.46% F1-score, and an ROC-AUC of 0.9000. The lower performance of the tree-based approaches is in line with the challenge of modelling high dimensional sparse TF-IDF representations with tree-based decision boundaries. Among the evaluated models, SVM achieved the highest recall, F1-score, and ROC-AUC, while its accuracy was equal to that of Logistic Regression. Based on these evaluation measures and the requirement to identify deceptive reviews, SVM was selected for subsequent trust-score generation. The comparative performance of the four classification models is presented in Table 3.

For the reported precision, recall, and F1-score in Table 3, the metrics are calculated with the Deceptive class as the positive class rather than using macro averaging across the two classes. Accuracy is calculated as the proportion of correctly classified instances among all test instances. ROC-AUC is calculated from the model's class-probability scores across the test samples.

Table 3. Performance of Fake Review Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
Logistic Regression	89.38	87.50	91.88	89.63	0.9429
SVM	89.38	86.63	93.13	89.76	0.9436
Random Forest	85.00	85.00	85.00	85.00	0.9226
XGBoost	80.94	79.29	83.75	81.46	0.9000

Note: Precision, recall, and F1-score are reported for the Deceptive class, designated as the positive class. Accuracy is calculated over all test samples, while ROC-AUC is calculated using the model probability scores.

The performance differences among the four models are further illustrated in Figure 3.

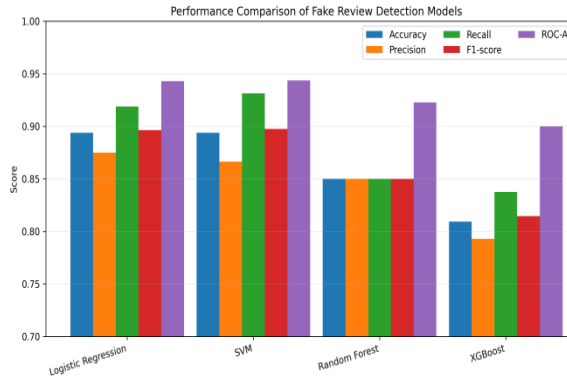


Figure 3. Performance comparison of fake review detection models

4.3 ROC and Classification Analysis

The ROC analysis also shows the superiority of the SVM classifier. SVM achieved the best ROC-AUC value of 0.9436, closely followed by Logistic Regression with 0.9429. Random Forest and XGBoost achieved 0.9226 and 0.9000, respectively. The close performance of SVM and Logistic Regression suggests that the TF-IDF representation of the text is useful for capturing linear patterns to differentiate truthful and deceptive reviews. The ROC curves of the four models are shown in Figure 4.

The SVM confusion matrix is shown in Figure 5, where the rows indicate the true classes and the columns indicate the predicted classes. The "Deceptive" class is considered as the positive class and the "Truthful" class is considered as the negative class in this study. Accordingly, the confusion matrix gives TP = 149, FN = 11, FP = 23, and TN = 137. Thus, the recall for the deceptive class is $149/(149 + 11) = 93.13\%$, while its precision is $149/(149 + 23) = 86.63\%$. The values are similar to the SVM precision and recall reported in Table 3. The classifier correctly classified 149 of the 160 deceptive reviews and 11 of the deceptive reviews were classified as truthful. Of the 160 truthful reviews, 137 were correctly classified as truthful and 23 were incorrectly classified as deceptive.

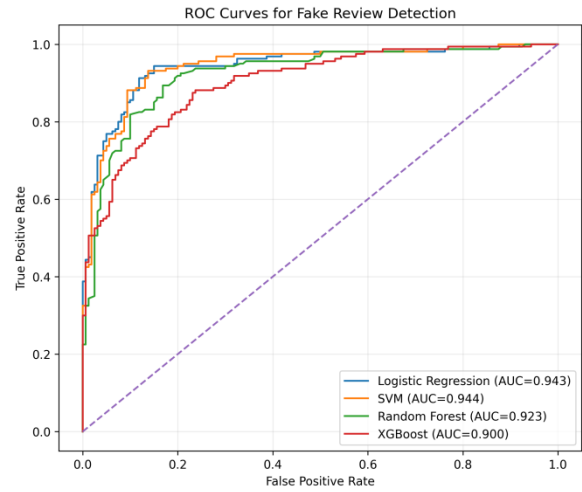


Figure 4. ROC curves for fake review detection

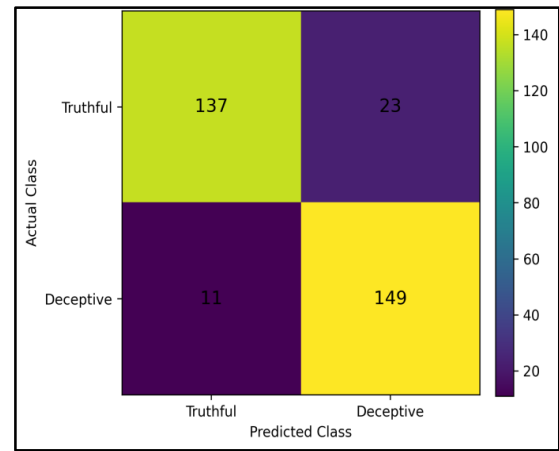


Figure 5. Confusion matrix of the SVM mode (rows: actual class; columns: predicted class; Deceptive = positive class)

4.4 Trust Score Analysis

The corrected methodology was then directly translated to a review trust score, where $\text{Trust}(r) = 1 - P_{\text{fake}}(r)$. The score is thus between 0 and 1, with a higher score signifying a higher level of confidence that a review is truthful. On the test set, truthful reviews had a mean trust score of approximately 0.797, whereas deceptive reviews had a mean trust score of approximately 0.161. The difference between the two groups is quite large, which shows that the classifier probability is a useful continuous credibility measure. There is some overlap between the two distributions, meaning that the classifier is not 100% certain for all reviews. This helps to facilitate the use of a continuous trust score, not just a binary label of truthful/deceptive. The distribution of trust

scores for truthful and deceptive reviews is shown in Figure 6.



Figure 6. Trust score distribution by review class

4.5 Entity-Level Trust and Trustworthy Recommendation

The last step combines the hotel level review trust scores based on the SVM model. There are 20 hotels, each with 80 reviews, giving 80 observations for each hotel for comparison at the entity level. The entity-level trust score is the average of the 80 reviews of each hotel. Review polarity is kept as descriptive information and does not get another numerical weight in the ranking. In the present experiment, the corpus consists of an equal number of positive and negative reviews for each hotel, so the polarity does not affect the hotel ranking. So, the hotels are listed mainly based on their review trust scores. The top ten hotels ranked according to average review trust are presented in Table 4. This ranking represents the relative credibility of the available review content for each hotel, rather than a measure of hotel quality, user-specific preference, or recommendation accuracy.

The results at the entity level show how the fake review detection process can be used to enable trustworthy ranking. The hotels with the higher average trust values were awarded: Affinia, Hyatt and Monaco, and hotels with lower average trust values were awarded below. All hotels get the same percentage of positive reviews, which means that the difference in ordering cannot be due to differences in polarity (either positive or negative) but rather to the differences in the estimated credibility of the review text. The result demonstrates the practical connection

between the prediction of fake reviews and trustworthy recommendations: False-revelation judgements on reviews are associated with a higher trust grade on the entity level.

Table 4. Top 10 Entities by Average Review Trust

Rank	Hotel	Reviews	Average Trust	Positive Reviews
1	affinia	80	0.534	50.0%
2	hyatt	80	0.532	50.0%
3	monaco	80	0.527	50.0%
4	intercontinental	80	0.527	50.0%
5	talbott	80	0.523	50.0%
6	knickerbocker	80	0.521	50.0%
7	palmer	80	0.518	50.0%
8	swissotel	80	0.510	50.0%
9	homewood	80	0.506	50.0%
10	allegro	80	0.489	50.0%

These results are generalizable to the characteristics of the selected dataset. The experiments were performed on the Deceptive Opinion Spam Corpus, which consists of Chicago hotel reviews with balanced classes of truthful and deceptive reviews, and balanced polarity of positive and negative reviews. The reported classification and trust-ranking results should not be taken as a direct generalization to other product or service domains, review platforms, languages, or datasets with different class distributions. Specifically, newer generative approaches to reviews might produce reviews with linguistic features that are different from those found in the current corpus. The proposed framework should therefore be generalized to other domains, platforms, languages and sources of deceptive content through external validation.

5. Conclusion

The results of the study indicate the potential of machine learning in determining unreliable online reviews. The same text for review and TF-IDF as features were used to test four classification models. Of these, SVM showed to be the best overall performer. It was able to correctly detect most of the deceptive reviews and the accuracy and the F1 score obtained were 89.38% and 89.76% respectively. The study also utilised the model output to give a trust score to every review. This score provides

for a quick estimate of the reliability of a review as judged by the trained model. The more trustworthy reviews were given higher trust scores. The scores were then aggregated at the hotel level to determine the trustworthiness of the overall reviews for each hotel. The result indicates that credibility of reviews is helpful for developing recommendation-oriented rankings. The proposed approach goes beyond relying on review opinions and also takes into account the trustworthiness of the underlying reviews. This offers another layer of information to guide decision making. In this overall, the proposed framework provides a simple and practical way to link fake review classification to the trust-based ranking. It could also be a starting point for further analysis with more data and more recommendation information.

Future Work

The framework can be tested in the future with larger and more varied datasets of reviews. Other linguistic, behavioural, temporal and rating-based features are also available for exploration. Comparisons can be made between the advanced deep learning and the transformer-based models with the traditional classifiers. Trust-score calibration for more reliable reliability assessment can be further explored in future research. The framework can be extended by including user preferences and user-item interactions. Experiments across domains could give a better understanding of model robustness. Finally, the proposed approach can be tested with real world e-commerce data sets for further validation.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Reference

- Al-yousef, A., Mohawesh, R., Saifan, A. A., & Al-Shannaq, M. A. (2025). A self-supervised transformer framework for cross-domain fraudulent review detection (CROSS-FRD). *The Journal of Supercomputing*, 81, 1575. <https://doi.org/10.1007/s11227-025-08061-5>
- Anusha, P., Sahas, B. S., & Surekha, T. L. (2026). Enhancing e-commerce trust: An integrated product recommendation and fake review detection system. In S. Namasudra, N. Kar, S. Kumar Patra, & B.-G. Kim (Eds.), *Data Science and Network Engineering (ICDSNE 2025)*, *Lecture Notes in Networks and Systems*, Vol. 1668. Springer. https://doi.org/10.1007/978-3-032-07735-6_3
- Bacı, M. A., Erdoğan, D. R., Akgüller, Ö., & Gaban, L. (2026). Detecting AI-generated e-commerce reviews with sheaf-hypergraph consensus residuals: Interpretable signals and cross-domain asymmetry. *Applied Sciences*, 16(14), 7071. <https://doi.org/10.3390/app16147071>
- Bathla, G., Singh, P., Singh, R. K., Cambria, E., & Tiwari, R. (2022). Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. *Neural Computing and Applications*, 34(22), 20213–20229. <https://doi.org/10.1007/s00521-022-07531-8>
- Choi, W., Nam, K., Park, M., Yang, S., Hwang, S., & Oh, H. (2023). Fake review identification and utility evaluation model using machine learning. *Frontiers in Artificial Intelligence*, 5, 1064371. <https://doi.org/10.3389/frai.2022.1064371>
- Droz, D., & Solberg Søilen, K. (2026). AI labels, perceived authenticity, and consumer trust in user-generated reviews. *Journal of Theoretical and Applied Electronic Commerce Research*, 21(5), 154. <https://doi.org/10.3390/jtaer21050154>
- Elmogy, A. M., Tariq, U., Mohammed, A., & Ibrahim, A. (2021). Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications*, 12(1). <https://doi.org/10.14569/IJACSA.2021.0120169>
- Gupta, R., Jindal, V., & Kashyap, I. (2025). Cross-domain fake review detection via

- orthogonal counterfactual representations. *IEEE Access*, 13, 119004–119022. <https://doi.org/10.1109/ACCESS.2025.3584731>
- Han, S., Wang, H., Li, W., Zhang, H., & Zhuang, L. (2023). Explainable knowledge integrated sequence model for detecting fake online reviews. *Applied Intelligence*, 53(6), 6953–6965. <https://doi.org/10.1007/s10489-022-03822-8>
- Hou, J., Tan, Z., Zhang, S., Hu, Q., & Wang, P. (2025). Detecting fake review intentions in the review context: A multimodal deep learning approach. *Electronic Commerce Research and Applications*, 70, 101485. <https://doi.org/10.1016/j.elerap.2025.101485>
- Lu, J., Zhan, X., Liu, G., Zhan, X., & Deng, X. (2023). BSTC: A fake review detection model based on a pre-trained language model and convolutional neural network. *Electronics*, 12(10), 2165. <https://doi.org/10.3390/electronics12102165>
- Mohawesh, R., Salameh, H. B., Jararweh, Y., Alkhalaileh, M., & Maqsood, S. (2024). Fake review detection using transformer-based enhanced LSTM and RoBERTa. *International Journal of Cognitive Computing in Engineering*, 5, 250–258. <https://doi.org/10.1016/j.ijcce.2024.06.001>
- Mohawesh, R., Xu, S., Springer, M., Jararweh, Y., & Al-Hawawreh, M. (2023). An explainable ensemble of multi-view deep learning model for fake review detection. *Journal of King Saud University – Computer and Information Sciences*, 35(8), 101644. <https://doi.org/10.1016/j.jksuci.2023.101644>
- Nair, R. T., & Khan, M. A. (2024). Metaheuristic-based selective ensembles for fake review detection: A review. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-024-08620-x>
- Thevakumar, J., & Thevakumar, L. (2025). RATHAN@DravidianLangTech 2025: Annaparavai — Separate the authentic human reviews from AI-generated one. *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, 449–453. <https://doi.org/10.18653/v1/2025.dravidianlangtech-1.66>
- Trzebiński, W., & Marciniak, B. (2022). Recommender system information trustworthiness: The role of perceived ability to learn, self-extension, and intelligence cues. *Computers in Human Behavior Reports*, 6, 100193. <https://doi.org/10.1016/j.chbr.2022.100193>
- Wang, S., Zhang, X., Wang, Y., & Ricci, F. (2024). Trustworthy recommender systems. *ACM Transactions on Intelligent Systems and Technology*, 15(4), 1–20. <https://doi.org/10.1145/3627826>
- Zhao, Y., Tang, S., Zhang, H., & Lyu, L. (2025). AI vs. human: A large-scale analysis of AI-generated fake reviews, human-generated fake reviews and authentic reviews. *Journal of Retailing and Consumer Services*, 87, 104400. <https://doi.org/10.1016/j.jretconser.2025.104400>