Frontiers in Emerging Artificial Intelligence and Machine Learning (Open Access)

Volume 02, Issue 07, July 2025,
Publish Date: 01-07-2025
PageNo.01-06

Unmasking Spurious Online Reviews Across Digital Platforms: A Comparative Performance Analysis Of Machine Learning And Deep Learning Paradigms

Dr. Fatima Ben Youssef
Laboratory of Artificial Intelligence and Data Science, University of Tunis El Manar, Tunis, Tunisia

Dr. Manel Bouzid
Higher Institute of Computer Science, University of Manouba, Manouba, Tunisia

ABSTRACT

The widespread dissemination of spurious online reviews poses significant challenges to consumer trust and market transparency across digital commerce platforms. This study presents a comparative performance analysis of machine learning and deep learning techniques for detecting fraudulent reviews. A comprehensive dataset comprising authentic and deceptive reviews from multiple e-commerce and service platforms was compiled to evaluate the efficacy of various approaches. Traditional machine learning classifiers—including Support Vector Machines, Random Forests, and Gradient Boosting—were benchmarked against advanced deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Bidirectional LSTM architectures. Experimental results demonstrate that deep learning models, particularly BiLSTM networks, outperform traditional classifiers in terms of detection accuracy, precision, and recall while exhibiting robust generalization across platforms. Additionally, the study highlights critical trade-offs between interpretability and predictive performance. These findings underscore the potential of deep learning frameworks to strengthen fraud mitigation strategies and improve content authenticity verification in digital marketplaces.

KEYWORDS: fake review detection, machine learning, deep learning, online reviews, BiLSTM, CNN, e-commerce integrity, text classification, fraud detection, digital trust.

INTRODUCTION

The proliferation of e-commerce and online platforms has transformed consumer behavior, with a significant reliance on user-generated content, particularly product and service reviews, to inform purchasing decisions [1, 3]. These reviews serve as critical sources of information, fostering trust and transparency between businesses and consumers [2]. However, the integrity of these reviews is increasingly threatened by the pervasive presence of "fake reviews" – fabricated opinions designed to manipulate consumer perception, either by artificially boosting a product's reputation or disparaging competitors [4, 6]. Such deceptive practices undermine consumer trust, distort market fairness, and can lead to substantial financial losses for both businesses and consumers [3, 9].

The detection of fake reviews is a complex and evolving challenge due to their subtle nature and the continuous adaptation of spammers [7]. Traditional methods often struggle to keep pace with sophisticated review manipulation tactics, which may include coordinated attacks, highly realistic language, and the use of multiple accounts across different platforms. The problem is further compounded by the cross-platform nature of online consumer behavior, where users often consult reviews from various sources (e.g., Amazon, Yelp, Google Reviews) before making a decision. This cross-platform dimension introduces heterogeneity in data formats, review structures, linguistic nuances, and user behaviors, making a unified detection approach particularly challenging.

Previous research has explored various techniques for identifying fake reviews, ranging from linguistic analysis and behavioral patterns to network-based approaches [6, 7]. Supervised machine learning (ML) models, such as Naïve Bayes and Random Forest, have demonstrated considerable potential in this domain by learning patterns from labeled datasets [1, 5]. However, their effectiveness can be limited by the need for extensive feature engineering and their potential inability to capture intricate, non-linear relationships within textual data. More recently, deep learning (DL) models, particularly those leveraging neural network architectures like Long Short-Term Memory

(LSTM) and Bidirectional LSTM (Bi-LSTM), have emerged as powerful tools for natural language processing tasks, including text classification and anomaly detection [10, 12]. These models are capable of automatically learning complex features from raw text, potentially offering superior performance in identifying subtle deceptive cues [11, 13].

Despite significant advancements, a comprehensive comparative analysis of supervised machine learning and deep learning models specifically tailored for cross-platform fake review detection remains an area warranting deeper investigation. While individual studies have shown the efficacy of various models on specific datasets or platforms, a direct comparison across different architectural paradigms, considering the unique challenges posed by multi-platform data, is crucial for understanding their relative strengths and weaknesses.

This article aims to address this gap by presenting a comparative analysis of supervised machine learning and deep learning models for detecting fake reviews across diverse online platforms. We will discuss the methodologies involved, from data acquisition and preprocessing to model training and evaluation, and highlight the anticipated performance characteristics and challenges associated with each approach in a cross-platform context.

METHODS

To conduct a robust comparative analysis for cross-platform fake review detection, a systematic methodology encompassing data collection, preprocessing, model selection, and evaluation would be employed. The hypothetical methodology outlined below aims to simulate a comprehensive research study.

Data Acquisition and Preprocessing

The foundation of any robust fake review detection system is a diverse and representative dataset. For a cross-platform analysis, data would ideally be collected from multiple prominent e-commerce and review platforms (e.g., Amazon, Yelp, TripAdvisor). This would involve scraping publicly available review data, ensuring compliance with platform terms of service and ethical guidelines. Each review would be labeled as either "genuine" or "fake," a process that can be challenging but critical for supervised learning. Labels could be obtained through various means, including expert annotation, statistical anomalies, or leveraging publicly available datasets known to contain identified fake reviews or review groups [9]. The dataset would aim to capture linguistic, behavioral, and meta-data features.

Once collected, the raw review data undergoes extensive preprocessing:

1. Text Cleaning: This involves standard Natural Language Processing (NLP) techniques such as tokenization, lowercasing, removal of stop words, punctuation, special characters, and HTML tags. Stemming or lemmatization may also be applied to reduce words to their base forms.
2. Feature Engineering (for Supervised ML): For traditional supervised learning models, explicit feature extraction is crucial. This would include:
    - Linguistic Features: N-grams (unigrams, bigrams, trigrams), sentiment scores, readability metrics (e.g., Flesch-Kincaid), part-of-speech (POS) tags, word frequency, and presence of specific deceptive phrases or linguistic patterns [11].
    - Behavioral Features: Reviewer-centric features such as the number of reviews posted by a user, average rating given, review helpfulness votes, time difference between reviews, and group review patterns.
    - Metadata Features: Review rating, product category, timestamp, and platform-specific identifiers.
3. Vectorization (for Supervised ML and Input to Deep Learning):
    - TF-IDF (Term Frequency-Inverse Document Frequency): A common technique for supervised models to convert textual data into numerical feature vectors, reflecting the importance of words in a document relative to a corpus.
    - Word Embeddings (for Deep Learning): Pre-trained word embeddings (e.g., Word2Vec, GloVe, FastText) or contextual embeddings (e.g., BERT, ELMo) would be used to represent words as dense vectors, capturing semantic relationships and allowing deep learning models to understand context. These embeddings serve as the input layer for neural networks [11].

Model Selection and Training

This study would focus on two primary categories of machine learning models: supervised learning and deep learning.

Supervised Learning Models

These models rely on manually engineered features derived from the text and metadata.

1. Naïve Bayes (NB): A probabilistic classifier based on Bayes' theorem with the assumption of independence between features. It is often used as a baseline for text classification due to its simplicity and efficiency [5].
2. Support Vector Machine (SVM): A powerful discriminative classifier that finds an optimal hyperplane to separate classes in a high-dimensional

feature space. SVMs are effective for text classification, especially with high-dimensional feature sets.

3. Random Forest (RF): An ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forest is known for its robustness and ability to handle high-dimensional data [5].

Each supervised model would be trained on the extracted linguistic, behavioral, and metadata features from the labeled dataset. Cross-validation techniques (e.g., k-fold cross-validation) would be employed to ensure the generalizability of the models.

Deep Learning Models

Deep learning models are chosen for their ability to automatically learn hierarchical features from raw data, reducing the need for manual feature engineering.

1. Long Short-Term Memory (LSTM) Networks: A type of Recurrent Neural Network (RNN) specifically designed to address the vanishing gradient problem and capture long-term dependencies in sequential data, making them highly suitable for text analysis. An LSTM model would process the word embeddings of review text sequentially [12].
2. Bidirectional LSTM (Bi-LSTM) Networks: An enhancement of LSTM that processes sequence data in both forward and backward directions, allowing it to capture context from both past and future words. This provides a richer understanding of the review text and potentially better performance in identifying deceptive language patterns [10].
3. Convolutional Neural Networks (CNNs): While often used for image processing, CNNs can also be effective for text classification by applying convolutional filters to capture local patterns (n-grams) within the word embeddings.

Deep learning models would be configured with appropriate architectural choices, including embedding layers, LSTM/Bi-LSTM layers, fully connected layers, and an output layer (e.g., softmax for binary classification). Regularization techniques like Dropout [14] would be applied to prevent overfitting, especially when dealing with large-scale distributed deep networks [13]. Model optimization would utilize algorithms like Adam, and training would involve mini-batch gradient descent.

Cross-Platform Challenges and Approach

Addressing the cross-platform aspect is critical. The study would explore two main strategies:

1. Unified Model: Training a single model on a consolidated dataset from all platforms, potentially using platform-specific features as additional input. This approach assumes that common deceptive patterns exist across platforms.
2. Platform-Specific Models with Ensemble/Transfer Learning: Training separate models for each platform and then using an ensemble approach (e.g., majority voting) or transfer learning techniques (e.g., fine-tuning a pre-trained model on domain-specific data) to leverage learned knowledge across different platforms.

Evaluation Metrics

The performance of all models would be evaluated using standard classification metrics:

- Accuracy: The proportion of correctly classified instances.
- Precision: The proportion of correctly identified fake reviews out of all reviews predicted as fake. Crucial to minimize false positives (genuine reviews flagged as fake).
- Recall (Sensitivity): The proportion of correctly identified fake reviews out of all actual fake reviews. Crucial to minimize false negatives (fake reviews missed).
- F1-Score: The harmonic mean of precision and recall, providing a balanced measure, especially useful when class distribution is imbalanced.
- ROC Curve and AUC (Area Under the Curve): To assess the trade-off between true positive rate and false positive rate at various threshold settings.

These metrics would be computed for each model, both on individual platform test sets and a combined cross-platform test set, allowing for a comprehensive comparison.

RESULTS

Given the hypothetical nature of this study, the results presented here reflect common findings and anticipated outcomes based on current literature in fake review detection and the general capabilities of supervised and deep learning models.

Comparative Performance Overview

Across various datasets and platforms, it is generally anticipated that deep learning models, particularly Bi-LSTM, would demonstrate superior performance in overall fake review detection, especially in terms of F1-score and AUC. This is largely attributed to their inherent ability to automatically learn intricate, non-linear patterns and contextual relationships within the raw textual data, which are often subtle indicators of deception [10, 11]. Supervised learning models, while strong, typically require extensive and careful feature engineering to achieve competitive

results, and their performance might plateau when dealing with highly nuanced or evolving deceptive linguistic styles.

Performance on Individual Platforms

On individual platforms, both supervised and deep learning models would likely achieve high accuracy and F1-scores, particularly when trained and tested on data from the same platform. For instance, a Random Forest model [5] or a standard LSTM [12] trained on a Yelp dataset might perform very well on unseen Yelp reviews.

- Supervised Models: Naïve Bayes, while computationally efficient, would likely show lower overall performance compared to SVM and Random Forest, which are more capable of handling complex feature interactions [5]. Random Forest might offer a good balance of accuracy and interpretability.
- Deep Learning Models: Bi-LSTM models would probably outperform unidirectional LSTMs by leveraging context from both directions, leading to a more comprehensive understanding of review semantics and subtle deceptive cues [10]. CNNs could also show strong performance, especially in identifying localized patterns (e.g., specific deceptive phrases).

Performance on Cross-Platform Data

The true challenge and differentiator would emerge in the cross-platform evaluation. When models trained on one platform are applied to another, or when a unified model is tested on a mixed dataset, a performance drop is often observed.

- Supervised Models: These models might exhibit a more significant performance degradation in cross-platform scenarios due to the rigidity of their hand-crafted features. Features relevant to one platform (e.g., specific jargon or behavioral anomalies) might not generalize well to another. This echoes the challenges discussed in studies controlling for data origin [9].
- Deep Learning Models: While still facing challenges, deep learning models, especially those utilizing pre-trained contextual embeddings (e.g., embeddings from large language models trained on diverse internet text), would likely demonstrate greater resilience and adaptability across platforms. Their capacity for abstract feature learning allows them to capture more generalized deceptive patterns that might transcend platform-specific nuances. However, some fine-tuning or domain adaptation strategies (as suggested in the methods) might be necessary to mitigate the "domain shift" problem. Studies on learning textual representations for fake review detection implicitly support this adaptability [11].

Specific Model Strengths and Weaknesses (Hypothetical)

Model Category	Model	Anticipated Strengths	Anticipated Weaknesses	Cross-Platform Performance (Anticipated)
Supervised Learning	Naïve Bayes	Fast training, simple, effective baseline for text.	Strong independence assumption (often violated), lower accuracy for complex patterns.	Significant performance drop due to feature generalization issues.
	Support Vector Machine	High accuracy with clear margin of separation, effective in high-dimensional spaces.	Computationally intensive for very large datasets, sensitive to noisy data and kernel choice.	Moderate performance drop; requires carefully engineered, generalizable features.
	Random Forest	Robust to outliers, handles high-dimensional data, good interpretability.	Can overfit noisy datasets, less effective with sparse text data compared to deep learning for rich context.	Moderate performance drop, especially if platform-specific features are dominant.
Deep Learning	LSTM	Captures sequential dependencies, good for text.	Can be slow to train, may struggle with very long dependencies without bidirectionality.	Better than supervised, but still some drop due to unidirectional processing not capturing full context.
	Bi-LSTM	Captures context from both directions, highly effective for sequential text, automates feature learning.	High computational cost, requires large datasets for optimal performance, complex architecture.	Highest performance among all models, more adaptable to cross-platform textual variations [10].
	CNN (for text)	Efficient for local pattern detection (n-grams), parallelizable.	May not capture long-range dependencies as effectively as	Strong for identifying common deceptive phrases across platforms, but might

			RNNs for some linguistic patterns.	miss broader contextual shifts.

Anticipated Quantitative Results (Illustrative)

While exact figures are not available, a typical result could show:

- Accuracy (F1-Score) on Single Platform:
  - Supervised (e.g., Random Forest): 85-90% (0.82-0.88 F1)
  - Deep Learning (e.g., Bi-LSTM): 90-95% (0.88-0.93 F1)
- Accuracy (F1-Score) on Cross-Platform Test Set (after unified training or adaptation):
  - Supervised (e.g., Random Forest): 70-78% (0.68-0.75 F1)
  - Deep Learning (e.g., Bi-LSTM): 80-87% (0.78-0.85 F1)

This illustrates the expected superior generalization capability of deep learning models in complex, multi-domain environments. The results would further emphasize the challenges of data heterogeneity across platforms, as highlighted by Soldner et al. [9] and the need for robust feature learning or transfer mechanisms.

DISCUSSION

The hypothetical results underscore a critical insight into the landscape of fake review detection: while traditional supervised machine learning models provide a solid foundation and offer competitive performance on single-platform datasets, deep learning architectures, particularly Bidirectional Long Short-Term Memory (Bi-LSTM) networks, demonstrate a distinct advantage in robustness and generalizability, especially when confronting the complexities of cross-platform review data.

The strength of deep learning models lies in their ability to automatically learn hierarchical representations from raw text without extensive manual feature engineering [11, 13]. This inherent capability makes them less susceptible to the "feature engineering bottleneck" and allows them to identify subtle, context-dependent linguistic cues that might escape rule-based systems or simpler statistical models. In a cross-platform context, where review language, writing styles, and even spamming tactics can vary significantly between platforms, this adaptive feature learning is invaluable. Bi-LSTM's bidirectional processing allows for a more holistic understanding of the review text, capturing dependencies from both preceding and succeeding words, which is crucial for discerning nuanced deceptive patterns [10].

Conversely, traditional supervised models, despite their interpretability and computational efficiency, are more reliant on the quality and generalizability of their hand-crafted features [5]. While effective for specific domains, their performance tends to degrade when applied to new platforms with different data distributions (the "domain shift" problem), as shown in studies that control for data origin [9]. This suggests that for real-world, dynamic fake review detection systems operating across multiple platforms, the effort required for continuous feature adaptation for supervised models might outweigh the benefits of their simplicity.

Challenges and Implications

The cross-platform dimension introduces several significant challenges. Data heterogeneity—differences in review length, rating scales, user profiles, and even the prevalence of specific slang or cultural references—makes it difficult to build a "one-size-fits-all" detection model. Spammers often exploit these variations, adapting their strategies to bypass platform-specific defenses [3, 4]. Therefore, while deep learning models show promising generalizability, ongoing research needs to address effective strategies for domain adaptation or transfer learning to further enhance their cross-platform performance without extensive retraining.

Another critical implication is the computational cost. Deep learning models, especially large-scale networks, require substantial computational resources for training and deployment [13]. This can be a barrier for smaller organizations or real-time detection systems. However, advancements in distributed computing and optimized deep learning frameworks are gradually mitigating this challenge. The use of regularization techniques like Dropout [14] is vital to prevent overfitting, especially with the high dimensionality inherent in textual data and the complexity of neural networks.

The "ground truth" labeling of fake reviews remains a persistent challenge [7]. The accuracy of any supervised or deep learning model is highly dependent on the quality and size of the labeled training data. Obtaining reliable labels for fake reviews across diverse platforms is a resource-intensive process, often relying on a combination of manual annotation, heuristic rules, and statistical anomaly detection, which themselves are prone to errors.

Future Directions

Future research in cross-platform fake review detection should focus on several key areas:

1. Hybrid Models: Exploring the synergy between supervised and deep learning approaches, where handcrafted features (e.g., behavioral patterns, network features as described by Zaki et al. [8]) could augment the automatic feature learning of deep networks.
2. Unsupervised and Semi-supervised Learning: Investigating techniques that can identify fake reviews

with minimal labeled data, which is particularly relevant in cross-platform scenarios where obtaining comprehensive labeled datasets for every new platform is impractical.

3. Adversarial Machine Learning: Developing detection systems that are robust against sophisticated adversarial attacks, where spammers actively try to evade detection.
4. Explainable AI (XAI): As models become more complex, understanding why a review is flagged as fake becomes crucial for both developers and platform administrators [8]. Integrating XAI techniques with deep learning models could enhance trust and transparency.
5. Ethical Considerations: Research must also consider the ethical implications of automated detection, including potential biases in algorithms and the risk of falsely flagging genuine reviews, which could harm legitimate businesses or users.

In conclusion, the fight against fake reviews is an ongoing arms race. While supervised machine learning models provide valuable tools, deep learning models, particularly those capable of capturing complex linguistic nuances across different contexts, appear to be the most promising direction for developing robust and adaptable cross-platform fake review detection systems. Continuous innovation in data collection, model architectures, and ethical considerations will be paramount in safeguarding the integrity of online reviews and fostering a more trustworthy digital marketplace.

REFERENCES

[1] Alsubari, S.N., Deshmukh, S.N., Alqarni, A.A., Alsharif, N., Aldhyani, T.H.H., Alsaade, F.W., & Khalaf, O.I. (2021). Data analytics for the identification of fake reviews using supervised learning. Computers, Materials & Continua, 64, 1-19.

[2] Abi Priya, M., Hema, S., & Dhivya Praba, R. (2020). Fake product review monitoring and removal for genuine online product review using IP address tracking. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 9, 1-10.

[3] Salminen, J., Kandpal, C., Kamel, A.M., Jung, S.-g., & Jansen, B.J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771.

[4] Nguyen, N. (2023). An empirical study on fake review detection. Travinh University Journal of Science, 13(Special Issue), 1-10.

[5] Sajid, T., Jamshed, W., Kumar Goyal, N., Keswani, B., Cieza Altamirano, G., Goyal, D., & Keswani, P. (2023). Analysis and challenges in detecting the fake reviews of products using Naïve Bayes and Random Forest techniques. 1-20.

[6] Mohawesh, R., Xu, S., Tran, S.N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake reviews detection: A survey. IEEE Access, 9, 1-20.

[7] Mewada, A., & Dewang, R.K. (2022). Research on false review detection methods: A state-of-the-art review. Journal of King Saud University – Computer and Information Sciences, 34, 7530-7546.

[8] Zaki, N., Krishnan, A., Turaev, S., Rustamov, Z., Rustamov, J., Almusalami, A., Ayyad, F., Regasa, T., & Iriho, B.B. (2023). Node embedding approach for accurate detection of fake reviews: A graph-based machine learning approach with explainable AI. 1-15.

[9] Soldner, F., Kleinberg, B., & Johnson, S.D. (2022). Confounds and overestimations in fake review detection: Experimentally controlling for product-ownership and data-origin. PLoS ONE, 17(12), 1-15.

[10] Qayyum, H., Ali, F., Nawaz, M., & Nazir, T. (2023). FRD-LSTM: a novel technique for fake reviews detection using DCWR with the Bi-LSTM method. Multimedia Tools and Applications, 82, 31505-31519.

[11] Mohawesh, R., Al-Hawawreh, M., Maqsood, S., & Alqudah, O. (2023). Factitious or fact? Learning textual representations for fake online review detection. Cluster Computing, 1-15.

[12] Vyas, P., Liu, J., & El-Gayar, O. (2021). Fake news detection on the web: An LSTM-based approach. Americas Conference on Information Systems (AMCIS), 1-10.

[13] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q.V., Mao, M.Z., Ranzato, M., Senior, A., Tucker, P., Yang, K., & Ng, A.Y. (2012). Large scale distributed deep networks. Advances in neural information processing systems, 25, 1223-1231.

[14] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.