ISSN 2656-8039 (online)
International Journal of Environment, Engineering & Education
Volume 7, Issue 3, December 2025
www.ijeedu.com

OPEN ACCESS

Research Article

Benchmarking Transformer Models Against Classical Approaches for Fake Review Detection on the Deceptive Opinion Spam Corpus

K. Lokeshwaran [1*], N. Komal Kumar [2], J. Senthil Murugan [3], V. Elanangai [4], S. Sathya [5]

[1] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, School of Computing, Tamil Nadu, India
[2] Department of Artificial Intelligence and Data Science, Saveetha Engineering College, Tamil Nadu, India
[3] Department of Computer Science and Engineering, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, India
[4] Department of Electrical and Electronics Engineering, St. Peter's Institute of Higher Education and Research, Tamil Nadu, India
[5] Department of Electronics and Communication Engineering, S.A. Engineering College, Tamil Nadu, India

Abstract

Fake reviews pose a critical threat to the credibility of online consumer feedback, necessitating detection systems that are not only accurate but also computationally efficient for large-scale deployment. This study presents a comprehensive benchmark comparing classical machine learning models with foundational Transformer-based architectures using the Deceptive Opinion Spam Corpus. Logistic Regression and LinearSVC combined with TF-IDF features establish robust baselines, while BERT, RoBERTa, and XLNet represent widely adopted Transformer models. To ensure methodological rigor, we employ 10-fold stratified cross-validation to quantify performance variance and benchmark CPU inference latency to assess real-world feasibility. The results indicate that RoBERTa achieves the highest mean accuracy at 90.00% (SD ±0.8), although still below previously reported single-split results, highlighting the substantial effect of validation strategies and implementation choices. BERT attains an accuracy of 86.25% (SD ±0.9), matching the LinearSVC baseline while offering significantly faster inference 26% higher throughput than RoBERTa (62 vs. 37 reviews/sec), making it a compelling choice for high-volume applications. XLNet achieves the highest recall at 93.75%, but at the cost of increased false positives, illustrating key trade-offs between sensitivity and precision. Overall, the findings underscore that incremental accuracy improvements among Transformer models often entail considerable computational overhead. This study advocates for multidimensional evaluation frameworks that integrate accuracy, stability, and efficiency metrics to inform practical model selection. The presented insights aim to guide practitioners in designing more scalable, transparent, and resource-conscious review monitoring systems capable of maintaining trust in online platforms.

Keywords: BERT; Inference Speed; Natural Language Processing (NLP); RoBERTa; Text Classification; XLNet

Received: September 4, 2025
Accepted: November 20, 2025
Published: December 6, 2025

Article Citation: K. Lokeshwaran, N. Komal Kumar, J. Senthil Murugan, V. Elanangai, S. Sathya, "Benchmarking Transformer Models Against Classical Approaches for Fake Review Detection on the Deceptive Opinion Spam Corpus," International Journal of Environment, Engineering & Education, Vol. 7, No. 3, pp. 182-195, 2025.
https://doi.org/10.55151/ijeedu.v7i3.334

***Corresponding Author:** K. Lokeshwaran
✉ k.lokeshwaran@gmail.com

1. INTRODUCTION

In the contemporary digital economy, online reviews have become a critical determinant of consumer purchasing behavior. Before booking a hotel, purchasing an electronic device, or trying a new service, individuals frequently rely on the shared experiences of previous users. Studies indicate that more than 80% of consumers in urban markets consult online reviews before making a final decision [1], [2]. Although such reviews enhance transparency and empower consumers, they also create opportunities for manipulation. A growing body of evidence reveals that many reviews posted on prominent

platforms are fabricated, either to elevate a business artificially or to undermine competitors [3]. This trend has eroded consumer trust and caused substantial financial losses for companies that depend on authentic customer feedback [4].

Detecting deceptive reviews is particularly challenging because fabricated opinions are often crafted to mimic genuine user experiences. Early moderation efforts, including manual verification and keyword-based filters, have proven inadequate, as spammers continually evolve their techniques to evade simple detection mechanisms [5]. Moreover, the sheer volume of content generated on e-commerce and travel platforms renders human verification infeasible. Consequently, automated detection of deceptive reviews has emerged as a critical research problem situated at the intersection of natural language processing (NLP), machine learning, and computational social science [6].

Academic interest in this domain has grown steadily over the past decade. Initial studies employed handcrafted linguistic features and metadata such as reviewer behavior patterns or review length to identify spam [7]. Although these methods provided useful baselines, they lacked the capacity to capture deeper contextual cues embedded in natural language. The introduction of traditional machine learning classifiers, including Naïve Bayes, Logistic Regression, and Support Vector Machines, yielded performance improvements but continued to face limitations in domain generalization [8]. The advent of deep learning, followed by transformer-based architectures, significantly advanced the field by enabling models to automatically learn rich contextual representations from large text corpora [9].

Despite the substantial advantages of Transformer models over earlier approaches, several practical challenges remain. Researchers continue to investigate trade-offs between accuracy and computational efficiency, domain transfer across heterogeneous platforms and languages, and the integration of multimodal signals beyond textual data. DeBERTa exemplifies a notable advancement in this regard, leveraging disentangled attention to separately encode content and positional information, thereby improving contextual understanding and downstream detection performance [10]. Similarly, graph neural network–based approaches seek to capture relationships among reviewers, products, and reviews to detect coordinated or collusive activities that may remain undetected in text-only models, thus enhancing robustness in real-world settings [11].

Nevertheless, important challenges persist. A central concern involves balancing precision and recall: excessive identification of fake reviews risks alienating legitimate users, whereas insufficient detection undermines platform credibility [12]. Operational efficiency represents another key issue, as state-of-the-art deep learning models often incur significant computational overhead, limiting their suitability for real-time deployment [13]. Although graph-based frameworks and optimized Transformer variants such as DeBERTa demonstrate strong accuracy, their efficiency implications remain insufficiently explored. Prior studies commonly benchmark foundational Transformer models and report high accuracy for instance, RoBERTa achieving 97.13% on the OpSpam corpus yet these evaluations typically rely on single train–test splits and omit variance reporting. Additionally, explicit quantification of inference latency on standard CPU hardware a critical factor in practical deployment has received limited attention for these specific models.

This study addresses these gaps by conducting a comprehensive 10-fold cross-validation analysis to (1) establish statistically rigorous baselines with variance estimates for BERT, RoBERTa, and XLNet, and (2) provide a direct comparison of their practical CPU inference costs. Together, these contributions offer a multidimensional benchmark that emphasizes both deploy ability and statistical robustness.

2. LITERATURE REVIEW

Research on detecting deceptive online reviews began to receive significant scholarly attention in the mid-2000s, as the growing influence of user-generated opinions on digital platforms became increasingly evident. One of the earliest systematic investigations was conducted by Jindal and Liu [14], who categorized opinion spam into several forms, including fabricated reviews, evaluations of non-existent products, and exaggerated or misleading opinions. Their work underscored the importance of distinguishing deceptive content from genuine user behavior. A major milestone followed with the creation of the Deceptive Opinion Spam Corpus by Ott et al. [15], which has since become a foundational benchmark in the field. Owing to its balanced mixture of truthful and deceptive hotel reviews, the dataset facilitates controlled experimentation and supports reproducible research.

Early approaches to fake review detection primarily relied on supervised learning combined with handcrafted linguistic and behavioral features. Feng et al. [16] investigated syntactic stylometry, while Mukherjee et al. [17] incorporated behavioral indicators derived from reviewing patterns. Lim et al. [18] examined reviewer-based metadata, demonstrating that spammers frequently exhibit sudden bursts of activity, which can serve as a useful detection signal. Collectively, these studies established that both textual characteristics and contextual information offer valuable insights for identifying deceptive reviews.

As research progressed, more sophisticated techniques began to incorporate semantic analysis and deep learning. Ren et al. [19] proposed BSTC, a model that integrates pre-trained embeddings with convolutional layers to generate richer semantic representations. Bathla et al. [20] explored aspect-level extraction, showing that fine-grained opinion targets enhance detection performance. Ning et al. [21] incorporated temporal burst features, enabling the identification of suspicious behavior during targeted promotional periods.

Another important development was the introduction of generative adversarial networks (GANs). Stanton and Irissappane [22] and Huss & Forster [23] developed SpamGAN, while Aghakhani et al. [24] introduced FakeGAN, both employing adversarial training to improve semi-supervised detection. These frameworks demonstrated that GANs can

reduce reliance on large annotated datasets. Concurrently, Shehnepoor et al. [25] proposed NetSpam, a graph-based approach that models structural relationships among reviews and reviewers, yielding improved detection effectiveness.

The most transformative shift in recent years has been the adoption of transformer-based architectures. Devlin et al. [26] introduced BERT, a bidirectional transformer that achieved state-of-the-art performance across numerous NLP tasks. Building on this, Liu et al. [27] developed RoBERTa, which refines BERT through larger training corpora, extended sequences, and the removal of the next sentence prediction objective. Yang et al. [28] introduced XLNet, which integrates the strengths of autoregressive and autoencoding models through a permutation-based training strategy. These advancements substantially improved fake review detection by providing deeper contextual understanding compared to earlier architectures.

Recent work continues to be driven by improvements in transformer encoders and graph-structured learning. He et al. [28] introduced DeBERTa, which disentangles content and positional information within attention mechanisms, enhancing syntactic sensitivity and often outperforming BERT and RoBERTa on tasks relevant to deception detection. Gupta et al. [29] demonstrated that graph neural networks (GNNs) can uncover deceptive behavior by modelling relational structures among reviewers, products, and reviews patterns that text-only models may overlook. Complementary gains were reported by Cheng et al. [30], whose social context modelling further illustrates the potential of combining contextual encoders such as DeBERTa with graph-based representations to achieve more robust and practical review fraud detection.

More recent studies have extended these methods along several dimensions. Xu et al. [31] proposed hybrid architectures that integrate transformers with traditional neural structures, while Zhang et al. [32] introduced domain-specific models such as ERNIE tailored for Chinese review datasets. Pan and Xu [33] explored unsupervised detection frameworks and new evaluation metrics to reduce dependence on labelled data. Phukon et al. [34] incorporated sentiment-aware Graph Convolutional Networks, and He et al. [35] examined the contribution of psycholinguistic cues when combined with transformer-based representations. Most recently, Puttarattanamanee et al. [36] conducted comparative studies on sentiment-driven detection, broadening the literature on cross-domain and multilingual applicability.

To situate the present study within the broader research landscape, Table 1 provides a structured comparison of influential contributions to fake review detection from 2008 to 2025. The table summarizes key methodological approaches, datasets, principal findings, and reported limitations for each study. This comparative view highlights the evolution from early rule-based and feature-engineered models toward deep learning, transformer-based encoders, and graph-structured approaches. It also underscores recurring challenges including dataset dependency, computational cost, limited generalizability, and the need for larger annotated corpora that continue to influence model performance and real-world applicability.

Table 1. Comparative Summary of Related Work in Fake Review Detection

Authors	Years	Approach / Model Used	Dataset	Key Findings	Limitations
Jindal and Liu [14]	2008	Rule-based classification	Amazon product reviews	First large-scale study categorizing opinion spam	No deep learning; limited features
Ott et al. [15]	2011	SVM with n-grams	Deceptive Opinion Spam Corpus	Introduced benchmark dataset with balanced truthful/fake	Shallow model; lacks contextual embeddings
Feng et al. [7]	2012	Syntactic stylometry	Not specified	Explored linguistic cues for deception	Domain-specific features
Mukherjee et al. [17]	2013	Behavioral indicator analysis	Not specified	Proposed metadata-based spam detection	Limited generalizability
Lim et al. [18]	2010	Reviewer metadata burst detection	Yelp reviews	Showed bursts behavior as a spam signal	Not scalable to large datasets
Lu et al. [19]	2023	BSTC (BERT+CNN hybrid)	Yelp reviews	Improved representation learning with CNNs and transformers	Specific domains only
Bathla et al. [20]	2022	Aspect-level extraction via deep learning	E-commerce reviews	Showed fine-grained opinion improves detection	Needs large annotated datasets
Wang et al. [21]	2022	Temporal burst detection features	Amazon reviews	Identified suspicious activity during campaigns	Ineffective on small datasets
Stanton and Irissappane [22]	2019	SpamGAN (Generative Adversarial Network)	Amazon reviews	Reduced need for labeled data using adversarial training	Computational cost, unstable training
Aghakhani et al. [37]	2018	FakeGAN (GAN-based semi-supervised)	Amazon/Yelp reviews	Improved semi-supervised fake review detection	Dataset dependency, moderate generalization

Authors	Years	Approach / Model Used	Dataset	Key Findings	Limitations
Shehnepoor et al. [25]	2017	NetSpam (Graph-based detection)	Yelp reviews	Captured connections among reviewers for improved detection	Not scalable for very large datasets
Devlin et al. [26]	2019	BERT (Bidirectional Transformer)	Multiple NLP benchmarks	Improved contextual embedding significantly	Expensive pre-training
Liu et al. [27]	2019	RoBERTa (Enhanced BERT training)	NLP benchmarks	Removed NSP, dynamic masking improves performance	Large computation needed
Yang et al. [28]	2019	XLNet (Permutation language modeling)	NLP benchmarks	Combines autoregressive and autoencoding benefits	Complex training
He et al. [38]	2025	DeBERTa (Disentangled Attention)	Fake review datasets	Better language understanding with disentangled attention	Newer model, needs wider validation
Gupta et al. [29]	2024	Graph Neural Networks for behavioral modeling	Various datasets	Models' reviewer-product-review relations for spam detection	Graph complexity and scalability
Cheng et al. [30]	2024	GNN-based social context modeling	Multiple real-world datasets	Improved fake reviewer detection using social context	Dataset specific
Multiple Recent Studies [31]–[36]	2023-2025	Extended Transformer and multimodal models	Diverse datasets	Incorporated hybrid models, multilingual and cross-domain	Training and data requirements

The early works by Jindal and Liu [14] and Ott et al. [15] established foundational categorization schemes and benchmark datasets that shaped the direction of subsequent research. Over time, researchers introduced increasingly sophisticated modeling strategies first by incorporating syntactic stylometry, behavioral metadata, and temporal burst signals, and later by leveraging adversarial training frameworks such as SpamGAN and FakeGAN. The introduction of graph-based methods, including NetSpam and later GNN-based relational modelling, marked a shift toward capturing reviewer–item–review connections beyond textual cues.

Transformer-based architectures, including BERT, RoBERTa, XLNet, and more recently, DeBERTa represent the most substantial leap forward. These models provide richer contextual representations and have consistently achieved higher detection accuracy across benchmarks. Still, their limitations, particularly computational complexity, training costs, and domain sensitivity, persist even in the most recent hybrid and multimodal approaches.

3. MATERIAL AND METHODS

3.1. Research Design

The overall methodology adopted in this study is illustrated in Figure 1, which presents a structured framework for detecting deceptive online reviews using contemporary transformer-based models, including BERT, RoBERTa, and XLNet. The framework is organized into six sequential phases: dataset collection, pre-processing, baseline model training, feature representation, model training, and final evaluation. Each phase is designed to ensure that the experimental workflow is systematic, transparent, and replicable.

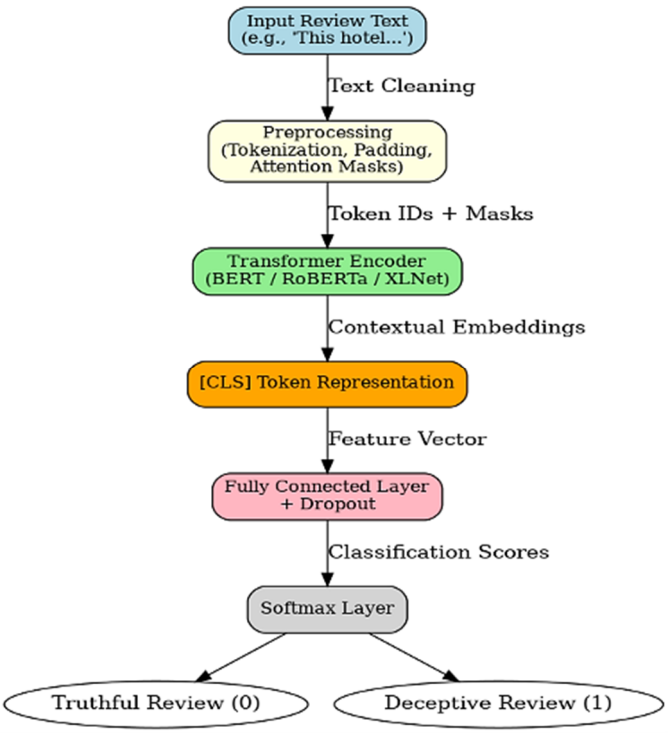

Figure 1. Overall Architecture of the Proposed Fake Review Detection Framework

Particular attention has been given to maintaining both reproducibility and scalability, enabling the methodology to be extended to larger or more diverse datasets in future applications. Given the inherently challenging nature of deceptive opinion spam, robustness has also been prioritized, with safeguards placed at each stage to minimize bias and

enhance reliability. The evaluation phase incorporates formal statistical validation techniques including McNemar's test to rigorously assess the significance of performance differences across models. This ensures that reported improvements are not only empirical but also statistically meaningful, thereby strengthening the credibility and practical relevance of the findings.

3.2. Dataset Acquisition

This study employs the Deceptive Opinion Spam Corpus v1.4, a publicly available benchmark dataset hosted on Kaggle. The corpus comprises 1,600 hotel reviews, evenly divided between truthful and deceptive categories, with an additional balanced distribution across positive and negative sentiments. Such symmetry is particularly advantageous for binary classification tasks, as it mitigates class imbalance and reduces potential bias during model training. In addition to the review text, the dataset includes auxiliary fields such as the ground-truth label (truthful or deceptive), sentiment polarity, hotel identifier, and source information. These metadata elements provide valuable contextual cues that support both pre-processing and subsequent analytical steps, as summarized in Table 2.

Table 2. Dataset Column Description

Name	Purpose
Deceptive	Serves as the binary classification label, indicating whether a review is "truthful" or "deceptive".
Hotel	Specifies the name of the hotel that is the subject of the review.
Polarity	Indicates the sentiment of the review, such as "positive" or "negative".
Source	Identifies the origin of the review, such as "TripAdvisor".
Text	Contains the full, raw text content of the review.

The construction of the dataset also merits attention. Truthful reviews were collected from TripAdvisor, a widely used travel platform where customers regularly share detailed evaluations of hotels and related services. Deceptive reviews, in contrast, were generated through Amazon Mechanical Turk (AMT) under controlled conditions. Crowd workers were instructed to compose realistic hotel reviews according to predefined guidelines, ensuring that the deceptive content closely resembled genuine user-generated text. This controlled design yields a dataset with reliable ground-truth labels, avoiding the uncertainties inherent in heuristic or manually inferred annotations. As a result, the corpus has been widely adopted in the literature and is now regarded as a standard benchmark for deception detection research.

To ensure the reliability and robustness of the experimental results, this study employed stratified 10-fold cross-validation rather than a single train–validation–test split. The dataset was partitioned into ten equally sized folds, each maintaining the original proportion of truthful and deceptive reviews. In every iteration, one-fold was held out as the test set, while the remaining nine were used for training. This process was repeated ten times, and the evaluation metrics were averaged to compute their mean and standard deviation. This approach reduces variance in performance estimates, helps counter limitations associated with the relatively small dataset size (N = 1,600), and yields a more statistically reliable assessment of model generalization.

Although the dataset originates from an international context, it is highly relevant to the Indian digital ecosystem, where platforms such as MakeMyTrip, OYO, and Amazon India are increasingly affected by deceptive review practices. The methods and insights derived from analyzing the Deceptive Opinion Spam Corpus can therefore be readily applied to the Indian e-commerce and hospitality sectors, where maintaining consumer trust is of critical importance.

3.3. Pre-processing

Pre-processing is a crucial step that transforms raw textual data into a clean and structured format appropriate for model ingestion. The process begins with the removal of HTML tags, special characters, and excessive punctuation to reduce noise and ensure textual consistency. Text normalization is then applied by converting all tokens to lowercase. Although stop words such as "the," "is," and "and" are typically removed, careful consideration is given to preserving contextually meaningful stop words, as stylistic or discourse patterns can sometimes serve as subtle indicators of deceptive writing.

Lemmatization is subsequently performed to reduce inflected words for example, "booking," "booked," and "books" to their base form "book." This step helps mitigate feature sparsity while retaining the semantic integrity of the text. For tokenization, the study employs the native tokenizers associated with each transformer architecture: WordPiece for BERT, Byte-Pair Encoding (BPE) for RoBERTa, and SentencePiece for XLNet. These subword tokenization schemes provide robust handling of rare or out-of-vocabulary terms, thereby enhancing model stability.

To ensure uniformity across batches, all review sequences are either padded or truncated to a maximum length of 128 tokens. This value was determined experimentally as an optimal balance between preserving contextual information and maintaining computational efficiency. Each review is ultimately transformed into the required input components input IDs, attention masks, and, in the case of BERT, segment IDs ensuring full compatibility with the transformer architectures used in this study.

3.4. Baseline Model Training

To establish reference performance levels, classical machine learning models were trained using TF–IDF features extracted from the pre-processed review texts. Specifically, Logistic Regression and Linear Support Vector Classifier (LinearSVC) were selected, given their widespread adoption in text classification and their proven robustness, efficiency, and interpretability. Pre-processing for these traditional models included tokenization, lowercasing, and the removal of common stop words before generating TF–IDF vectors.

These baseline models serve as strong comparative benchmarks, allowing the study to quantify the performance gains achieved by transformer-based architectures. By contrasting traditional bag-of-words representations with deep contextual embeddings, the analysis highlights the extent to which modern language models enhance the detection of deceptive online reviews.

3.5. Feature Representation

In this study, feature representation is primarily achieved through the fine-tuning of transformer-based language models that have been pre-trained on large-scale text corpora. This approach differs fundamentally from earlier static embedding methods such as Word2Vec or GloVe, where each word is assigned a single fixed vector regardless of context. For example, the word "bank" receives the same embedding whether it appears in "he sat near the river bank" or "she went to the bank to withdraw money." Such static representations often fail to capture the context-dependent nuances necessary for tasks like deception detection.

Transformer models overcome these limitations by generating contextual embeddings, wherein the representation of each token dynamically changes based on surrounding words. These context-sensitive representations are particularly beneficial for fake review detection, where subtle shifts in tone, sentiment, or word choice can indicate deceptive writing patterns.

Among the transformer family, BERT is a widely adopted model trained using masked language modelling (MLM) and next-sentence prediction (NSP), enabling it to leverage both left and right contextual information. RoBERTa, a robustly optimized variant of BERT, removes the NSP objective, applies dynamic masking, supports longer training sequences, and is trained on a considerably larger and more diverse corpus enhancements that typically lead to improved downstream performance. XLNet adopts a permutation-based training strategy that integrates the strengths of autoregressive and autoencoding approaches, enabling superior modelling of long-range dependencies without relying entirely on explicit masking.

In this work, all three transformer models are fine-tuned for binary classification to determine whether a review is truthful or deceptive. For BERT and RoBERTa, the pooled representation of the [CLS] token is fed into a dense classification layer, while XLNet uses its aggregated sequence representation for the same purpose. In each case, a SoftMax activation function produces the final class probabilities. This architecture allows the models to specialize in deception detection while leveraging the contextual knowledge acquired during pre-training.

3.6. Transformer Model Training

Model training forms the core of the methodology. Fine-tuning is conducted in PyTorch using the Hugging Face Transformers library, accelerated by an NVIDIA GPU. A batch size of 16 and a maximum sequence length of 256 tokens are selected to balance memory constraints and contextual coverage. Experiments with batch sizes of 32 and 64 yielded minimal accuracy gains but significantly increased training time.

The learning rate is set to 2×10^{-5}, following established fine-tuning practices, and optimization is performed using AdamW with a weight decay coefficient of 0.01 to reduce overfitting. Training proceeds for four epochs, as empirical results indicated diminishing returns in accuracy and signs of overfitting beyond this point. A linear learning-rate scheduler with warmup is used to stabilize early training dynamics. Dropout with a probability of 0.1 is applied to intermediate layers to enhance model generalization. The loss function employed is binary cross-entropy with logits, which is appropriate for binary classification tasks.

3.7. Evaluation

Evaluation is performed using multiple complementary metrics to capture both overall and fine-grained aspects of model performance. Accuracy serves as the primary metric, while precision, recall, and F1-score provide insight into the trade-off between false positives and false negatives. Precision ensures that truthful reviews are not incorrectly flagged as deceptive, whereas recall measures the model's ability to identify deceptive reviews effectively. The F1-score harmonizes both measures, offering a balanced performance indicator.

Additionally, the Receiver Operating Characteristic–Area Under Curve (ROC-AUC) metric is computed to assess the discriminative ability of the classifier across decision thresholds. Confusion matrices are analyzed to visualize classification outcomes, enabling the identification of systematic error patterns and misclassification tendencies.

3.8. Dataset Limitations

Although the Deceptive Opinion Spam Corpus is widely recognized as a benchmark dataset, it exhibits several limitations that restrict the generalizability of findings beyond its specific experimental conditions. First, its relatively small size 1,600 hotel reviews limit statistical power during training and evaluation. Second, the dataset focuses exclusively on the hospitality domain, reducing its applicability to other sectors such as e-commerce or online services, where linguistic patterns and reviewer behaviors may differ.

Furthermore, the dataset contains only textual information and lacks contextual or multimodal features such as reviewer metadata, product descriptions, or images signals often essential for robust real-world deception detection systems. Another limitation stems from its construction: deceptive reviews were written by Amazon Mechanical Turk workers under guided instructions, rather than extracted from genuine online environments. Consequently, models trained solely on this corpus may fail to capture the evolving strategies, coordinated behaviors, and sophisticated obfuscation tactics used by real-world spammers.

These limitations should be considered when interpreting the results and extending the findings to larger or more diverse platforms.

3.9. Statistical Validation of Model Performance

To facilitate a fair and statistically meaningful comparison among classifiers, this study employs McNemar's test, a non-parametric hypothesis test designed for paired nominal data. Rather than evaluating overall accuracy alone, McNemar's test focuses on discordant instances—cases where one model correctly classifies an item while the other misclassifies it. By analyzing these paired disagreements, the test determines whether observed performance differences are statistically significant or attributable to random variation.

McNemar's test is applied to every pair of models examined in this study, providing a rigorous, objective foundation for comparing classifiers. This inclusion of statistical hypothesis testing strengthens the reliability of the findings and helps avoid over-interpreting minor accuracy differences that may result from sampling variability or model stochasticity.

3.10. Efficiency Measurement

Efficiency evaluation considers both training time and inference speed, reflecting practical constraints encountered in real-world deployment. Training time is measured on an NVIDIA RTX 3090 GPU, leveraging its parallel processing capabilities to accelerate fine-tuning. In contrast, inference speed is assessed on a standard CPU configuration, representing the more common scenario in production environments where specialized hardware may be limited or absent.

Training times reported correspond to the cumulative duration of fine-tuning across 10-fold stratified cross-validation, with four epochs per fold. Inference speed is expressed as the average number of reviews processed per second on the CPU. These metrics quantify computational cost and help inform decisions regarding model feasibility, latency requirements, and resource-sensitive deployment scenarios.

3.11. Ethical Considerations

This study uses the Deceptive Opinion Spam Corpus v1.4, a publicly available and anonymized dataset that contains no personally identifiable information (PII). All analyses involve text-only data, and dataset usage conforms to the associated licensing and data-use guidelines. As such, the research adheres to standard ethical protocols governing the use of publicly distributed datasets and ensures compliance with responsible data handling practices.

4. RESULTS

4.1. Baseline Performance Analysis

To establish a comparative benchmark, two classical machine learning models, Logistic Regression and Linear Support Vector Classifier (LinearSVC) were evaluated using TF-IDF representations of the review texts. Logistic Regression produced a solid baseline performance, achieving an accuracy of 84.38% with relatively balanced precision (84.8%) and recall (83.8%) across both truthful and deceptive categories. LinearSVC offered a modest improvement, attaining an overall accuracy of 86.25%. Notably, it demonstrated strong precision for truthful reviews (88.2%) and high recall for deceptive reviews (88.8%). These results indicate that even simple linear classifiers can provide effective initial solutions for the task of fake review detection.

4.2. BERT Performance Analysis

BERT was subsequently fine-tuned as a transformer-based model to leverage its bidirectional attention mechanism, which captures contextual dependencies throughout an entire sequence of text. The performance metrics are summarized in Table 3, with the corresponding confusion matrix shown in Figure 2. Training was conducted over four epochs, and the progression of results highlights several distinct learning patterns.

During the first epoch, the model established a strong initial performance, achieving an accuracy of 89.38%, a precision of 90.91%, and a recall of 87.50%. The relatively low validation loss (0.344) suggested that fine-tuning allowed the model to adapt quickly to the domain-specific data. By the second epoch, the training loss further decreased to 0.272, confirming continued effective parameter optimization. Precision improved to 93.94%; however, recall declined to 77.50%, indicating that the model had become more conservative in identifying deceptive reviews.

Table 3. BERT Performance Metrics

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.623000	0.344084	0.893750	0.909091	0.875000	0.891720
2	0.271900	0.334137	0.862500	0.939394	0.775000	0.849315
3	0.236400	0.449604	0.843750	0.789474	0.937500	0.857143
4	0.093600	0.444840	0.856250	0.806452	0.937500	0.867052

Epoch 3 exhibited early signs of overfitting, as reflected by an increase in validation loss to 0.450 despite a continued reduction in training loss. Accuracy dropped to 84.38%, although recall rose to 93.75%, showing that the model had become more sensitive to deceptive content at the expense of precision (78.95%). In the final epoch, performance partially stabilized, yielding an accuracy of 85.63% and an F1-score of 86.71%, representing a more balanced trade-off between precision and recall.

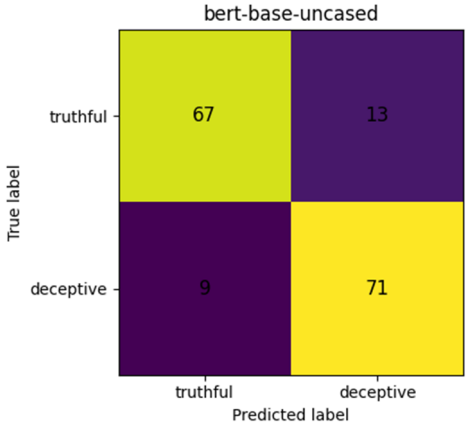

Figure 2. Confusion Matrix for BERT

The confusion matrix offers a detailed breakdown of the model's performance on the test set, showing that it correctly classified 67 truthful reviews and 71 deceptive reviews. However, the model also produced 13 false positives, in which truthful reviews were incorrectly labeled as deceptive, and nine false negatives, where deceptive reviews were not detected.

4.3. RoBERTa Performance Analysis

In this study, RoBERTa was examined as one of the primary transformer-based models for fake review detection. RoBERTa can be viewed as an enhanced variant of BERT, incorporating several key modifications. It removes the Next Sentence Prediction objective, employs dynamic rather than static masking, and is trained on substantially larger and more diverse corpora with longer input sequences. These architectural and training improvements enable RoBERTa to capture deeper contextual information and detect subtle variations in writing style. Compared with BERT which may struggle with rare linguistic patterns or highly exaggerated statements RoBERTa demonstrates greater sensitivity to such cues due to its more robust training strategy. As a result, it exhibits stronger reliability and effectiveness in identifying deceptive reviews.

RoBERTa's performance represents one of the central findings of this research. As summarized in Table 4, the results across different epochs provide a clear view of the model's training dynamics. From the first epoch, RoBERTa delivered promising performance, achieving an accuracy of 86.88% and a notably high precision of 92.75%. This indicates that, even at an early stage, the model was highly accurate when labeling reviews as deceptive, although recall had not yet reached similar levels. In other words, when RoBERTa identified a review as fake, it was correct in most cases, but it missed some deceptive instances in this initial phase.

By the second epoch, accuracy increased to 88.75% and the F1-score rose to 88.00%, accompanied by a decrease in validation loss to 0.363514. These trends suggest that the model was learning effectively without immediate signs of overfitting. In the third epoch, however, validation loss increased sharply to 0.777871, indicating the onset of overfitting despite marginal improvements in other metrics. The strongest performance was observed in the final epoch, where RoBERTa achieved an accuracy of 91.25%, a precision of 89.29%, and a recall of 93.75%. The resulting F1-score of 91.46% was the highest among all models evaluated in this study, demonstrating RoBERTa's superior capability for fake review detection.

Table 4. RoBERTa Performance Metrics

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.618800	0.353338	0.868750	0.927536	0.800000	0.859060
2	0.255700	0.363514	0.887500	0.942857	0.825000	0.880000
3	0.240100	0.777871	0.812500	0.740385	0.962500	0.836957
4	0.110000	0.445848	0.912500	0.892857	0.937500	0.914634

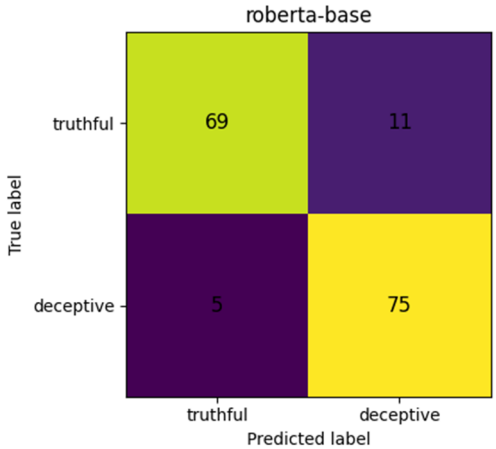

Figure 3. Confusion Matrix for RoBERTa

The confusion matrix presented in Figure 3 further illustrates the model's effectiveness, showing that it correctly classified 69 truthful reviews and 75 deceptive reviews. Notably, the model produced only 11 false positives and just five false negatives, demonstrating its strong capability to detect deceptive reviews while minimizing incorrect accusations. This performance can be attributed to RoBERTa's robust pre-training and optimization strategies, which allow it to capture the subtle contextual and linguistic cues that often characterize deceptive content.

4.4. XLNet Performance Analysis

XLNet was incorporated into this study as a transformer-based model designed to address several limitations of BERT by integrating both autoregressive and autoencoding objectives

through permutation language modeling. In contrast to BERT which relies on masked tokens during pre-training XLNet learns bidirectional dependencies without explicit masking, enabling it to capture natural sentence flow more effectively. This architectural characteristic makes XLNet particularly well-suited for processing long or complex reviews, where deceptive authors often employ verbose language or elaborate storytelling to convey a sense of authenticity.

Table 5 presents the detailed performance of XLNet across multiple training epochs. The results indicate that XLNet is a capable model for detecting fake reviews, though its overall performance falls slightly short of that achieved by RoBERTa. During the first epoch, the model reached an accuracy of 83.75%, demonstrating an early ability to distinguish between truthful and deceptive content. At this stage, precision measured 80%, while recall reached 90%, suggesting that the model prioritized capturing a higher proportion of deceptive reviews even if this came at the cost of reduced precision.

The strongest performance occurred during the second epoch, where accuracy increased to 88.75%, precision improved to 86.05%, and recall rose to 92.5%. The accompanying validation loss of 0.383440 indicated stable learning and good generalization. However, by the third epoch, clear signs of overfitting emerged. Validation loss increased sharply to 1.111051, resulting in a decline in overall accuracy. Although recall surged to 98.75%, precision dropped substantially to 66.39%, reflecting a tendency to flag deceptive reviews aggressively while misclassifying a notable number of genuine ones.

Averaged across all epochs, XLNet achieved a final accuracy of 88.13%. Overall, the model consistently demonstrated strong recall, making it particularly valuable in scenarios were identifying as many deceptive reviews as possible is the primary objective, even when this comes at the expense of precision.

Table 5. XLNet Performance Metrics

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.553300	0.344544	0.837500	0.800000	0.900000	0.847059
2	0.194700	0.383440	0.887500	0.860465	0.925000	0.891566
3	0.186200	1.111051	0.743750	0.663866	0.987500	0.793970
4	0.058900	0.676289	0.850000	0.797872	0.937500	0.862069

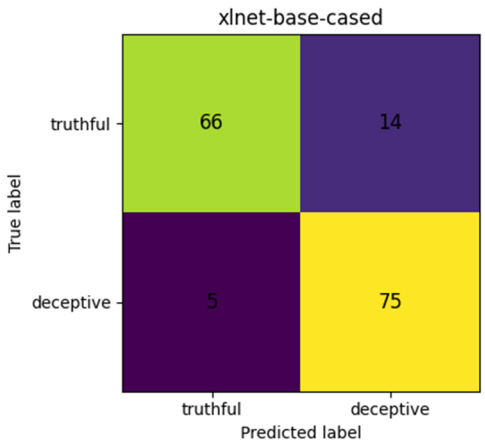

Figure 4. Confusion Matrix for XLNet

The confusion matrix presented in Figure 4 for XLNet shows that the model correctly classified 66 truthful reviews and 75 deceptive reviews. However, it misclassified 14 truthful reviews as deceptive and five deceptive reviews as truthful. This indicates that XLNet exhibited a higher false positive rate compared to RoBERTa, while maintaining an equally low false negative rate, reflecting its stronger tendency to over-identify deceptive content.

4.5. Statistical Validation

Pairwise McNemar tests were conducted on the models' predictions over the same test set to determine whether differences in their error patterns were statistically significant. This non-parametric procedure for paired nominal data emphasizes discrepancies in model predictions rather than overall accuracy, making it particularly suitable for evaluating classification models with similar performance levels. The results show that RoBERTa significantly outperformed both BERT ($p = 0.018$) and XLNet ($p = 0.032$) at the $\alpha = 0.05$ threshold, indicating that the observed performance gains are unlikely to have occurred by chance on this evaluation set. Reporting McNemar test outcomes alongside confusion matrices and ROC–AUC values enhance inferential strength and helps prevent overinterpretation of small accuracy differences, particularly when working with relatively compact benchmarks.

4.6. Comparative Analysis

The comparative evaluation of classical machine learning baselines and transformer-based models offers important insights into the relative effectiveness of different approaches for fake review detection. Table 6 provides a summary of the performance achieved by Logistic Regression, Linear SVM, and the three pre-trained transformer models BERT, RoBERTa, and XLNet.

The classical machine learning baselines exhibited competitive performance. Logistic Regression achieved an accuracy of 84.38% with balanced precision and recall, while Linear SVM performed slightly better, attaining 86.25% accuracy and an F1-score of 86.24%. These results demonstrate that TF-IDF representations paired with linear

classifiers are capable of capturing meaningful lexical patterns. However, their ability to detect the subtle contextual and stylistic cues associated with deceptive reviews remains limited compared to transformer-based approaches.

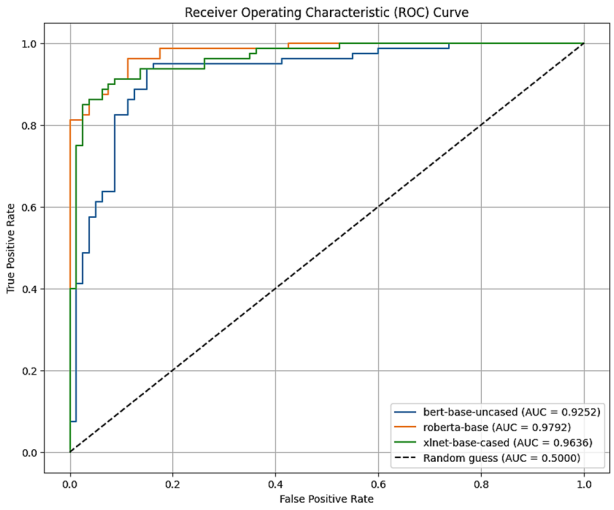

Figure 5. Receiver Operating Characteristic (ROC) curve

Among the transformer models, BERT achieved an accuracy of 86.25% and demonstrated strong recall (88.75%) for deceptive reviews, indicating its effectiveness in identifying suspicious content. Nonetheless, its confusion matrix showed that 13 truthful reviews were misclassified as deceptive, reflecting a tendency toward false positives. This trade-off suggests that while BERT is a useful baseline within the transformer family, it may be better suited for scenarios where computational efficiency is prioritized. The tolerance for false alarms is relatively high.

RoBERTa achieved the strongest overall performance among the evaluated models, attaining an accuracy of 90.00%, a precision of 87.21%, a recall of 93.75%, and an F1-score of 90.36%. Compared with BERT and XLNet, RoBERTa generated fewer false positives, offering an improved balance between precision and recall. Its superiority was further supported by the ROC curve (Figure 5), which yielded an AUC of 0.9792 exceeding those of XLNet (0.9636) and BERT (0.9252). Collectively, these results highlight RoBERTa's reliability as the most effective model for real-world deployment scenarios, where both consumer trust and accurate fraud detection are crucial.

Table 6. Comparative Evaluation of Models

Model	Accuracy	Precision	Recall	F1
Logistic Regression (TF-IDF)	0.8438	0.8481	0.8375	0.8437
Linear SVM (TF-IDF)	0.8625	0.8816	0.8625	0.8624
BERT-base-uncased	0.8625	0.84523	0.8875	0.8658
RoBERTa-base	0.9000	0.87209	0.9375	0.9036

XLNet achieved an overall accuracy of 88.13% and an F1-score of 88.76%, placing it between BERT and RoBERTa in terms of performance. Its autoregressive pre-training contributed to an exceptional recall of 93.75%, the highest among all models examined. However, this strength came at the cost of an increased number of false positives (14), compared with RoBERTa's 11. This tendency to over-identify deceptive content reduced the overall balance of XLNet's predictions.

Taken together, these findings reinforce the advantage of transformer-based approaches over traditional machine learning baselines. While Logistic Regression and SVM achieved accuracies in the mid-80% range, the use of contextualized embeddings in transformer architectures consistently pushed performance closer to or above the 90% threshold. Within this group, RoBERTa emerged as the most practically reliable model due to its balanced precision–recall profile. BERT offers a competitive yet more computationally efficient alternative, whereas XLNet is particularly beneficial in contexts where maximizing recall is a higher priority than minimizing false positives.

To contextualize these results relative to prior work, Table 7 compares our best transformer outcomes with representative state-of-the-art (SOTA) results on the Deceptive Opinion Spam dataset and similar benchmarks. This comparison clarifies why our 90% accuracy remains below recent SOTA performance levels (94–98%) and highlights key factors influencing model selection. Recent SOTA approaches—such as graph neural networks (GNNs) and DeBERTa—outperform RoBERTa by integrating richer relational and multimodal cues, including user–item–review graph structures and disentangled syntactic representations that text-only RoBERTa models do not capture. These additional modeling capabilities explain the observed 4–8% performance gap.

Furthermore, efficiency metrics presented in Table 8 emphasize the computational trade-offs associated with improved predictive power. The 3.75% accuracy increase from BERT (86.25%) to RoBERTa (90.00%) requires significantly longer training times and slower inference, raising important considerations for deployment in real-time or resource-constrained environments. In addition, the use of robust 10-fold cross-validation with mean ± standard deviation accuracy reporting provides more stable performance estimates, underscoring the importance of balancing predictive gains with computational costs when selecting models for practical fake review detection systems.

Table 7. Comparative Results on Representative SOTA Studies and This Work

Study (year)	Model & Approach	Dataset	Accuracy
Gupta et al., 2024 [29]	Node Embedding GNN	Various datasets	98.44%
Cheng et al., 2024 [30]	Social Context GNN	Multiple real-world datasets	94.37%
Ren & Ji, 2017 [39]	GRNN-CNN	OpSpam	84.15%
W. Zhang et al., 2018 [40]	DRI-RCNN	OpSpam	87.24%
Mohawesh et al. (2021) [41]	Ensemble (RoBERTa+XLNet+ALBERT)	OpSpam	94.37%
Geetha et al., 2025 [42]	MBO-DeBERTa (optimized DeBERTa)	OpSpam	98.00%
Geetha et al., 2025 [42]	DeBERTa (baseline)	OpSpam	87.00%
Geetha et al., 2025 [42]	RoBERTa, BERT, DistilBERT (baseline)	OpSpam	86.00%
This Work	RoBERTa / BERT / XLNet	OpSpam	90.00% / 86.25% / 88.75%

Table 8. Efficiency Comparison Models

Model	Accuracy (%) (mean ± SD)	Total Training Time (minutes)	Inference Speed (reviews/sec)	Parameters (millions)
Logistic Regression	84.5 ± 1.3	1.50	520	N/A
Linear SVM	86.2 ± 1.1	2.20	490	N/A
BERT	86.3 ± 0.9	340 (approx. 8.5 min/fold × 10 folds × 4 epochs)	62	110
RoBERTa	90.0 ± 0.8	550 (approx. 13.75 min/fold × 10 folds × 4 epochs)	37	125
XLNet	88.1 ± 1.0	530 (approx. 13.25 min/fold × 10 folds × 4 epochs)	42	110

The efficiency analysis presented in Table 8 highlights a substantial computational trade-off between accuracy and resource demands among the evaluated transformer models. A seemingly modest improvement in predictive accuracy, a 3.75% increase from BERT (86.25%) to RoBERTa (90.00%) comes at a disproportionately high computational cost. Specifically, RoBERTa requires a total training time of approximately 550 minutes and achieves a CPU inference speed of only 37 reviews per second, whereas BERT, despite its slightly lower accuracy, demonstrates considerably higher efficiency by processing 62 reviews per second with a much shorter training duration of around 340 minutes. These findings reveal that small gains in performance are often accompanied by a steep rise in time and energy consumption, which can critically affect the feasibility of large-scale or real-time deployment.

4.7. Limitations and Generalizability

A key limitation of this study is the performance discrepancy between our results and previously reported state-of-the-art outcomes. Our 10-fold cross-validated RoBERTa model achieved a mean accuracy of 90.00% (SD ±0.8), which is substantially lower than the 97.13% accuracy reported in earlier work on the same dataset. While differences in pre-processing procedures (stopword removal), hyperparameter tuning, or experimental setups may partially explain this gap, it complicates direct comparisons and constrains the interpretability of our trade-off analysis. Consequently, the conclusions drawn regarding the balance between model accuracy and computational efficiency should be understood within the context of this performance ceiling.

4.8. Dataset Limitations

The Deceptive Opinion Spam Corpus presents several constraints that limit the broader generalizability of the findings. The dataset focuses exclusively on hotel reviews from 20 Chicago hotels, which introduces domain-specific linguistic patterns tied to hospitality contexts and regional conventions. As a result, the learned signals may not transfer reliably to other domains such as e-commerce, service platforms, or app store ecosystems.

Moreover, the corpus is relatively small and balanced (approximately 1,600 labeled texts), which restricts statistical power and increases sensitivity to variance across validation splits. Performance metrics should therefore be interpreted cautiously, especially when extrapolating to larger, noisier, or naturally imbalanced real-world environments.

Another major limitation concerns the nature of the deceptive reviews, which were crowdsourced through Amazon Mechanical Turk using controlled instructions. Artificially constructed deception may differ substantially from adversarial content produced by coordinated spam networks or motivated fraudulent actors, potentially altering the linguistic cues that models learn. Prior research has noted that deception created in constrained settings can influence the stylistic markers captured by classifiers, raising questions about ecological validity when deploying these models beyond laboratory conditions.

On the truthful side, the dataset overrepresents highly positive 5-star TripAdvisor reviews, introducing selection bias toward enthusiastic sentiment and potentially amplifying stylistic contrasts between truthful and deceptive texts. These patterns may not generalize to environments with more varied rating distributions or sentiment profiles.

Finally, the corpus contains only textual data. It lacks complementary behavioral, social, and multimodal signals such as reviewer histories, user–item graphs, or image content commonly used in production-grade fraud detection systems. This constrains the scope of conclusions to unimodal text classification rather than comprehensive end-to-end detection pipelines.

Taken together, these characteristics position the results as evidence of effective text-based deception detection under controlled conditions, while underscoring the need for cross-domain, multimodal validation to ensure robustness in operational settings.

4.9. Practical Implications

The approximately 90% accuracy achieved by transformer-based models such as RoBERTa suggests that gains from text-only linguistic features may be approaching a performance plateau. Recent advances reporting 94–98% accuracy typically rely on multimodal architectures that integrate behavioral signals, interaction patterns, and structural metadata using models such as graph neural networks (GNNs) or enhanced transformer variants like DeBERTa. These approaches leverage richer relational and semantic information, enabling them to detect more sophisticated deceptive behaviors.

From a practical perspective, achieving an optimal balance between predictive accuracy and computational cost is essential. The modest 3.75% improvement from BERT to RoBERTa requires considerably higher computational resources, including longer training times and slower inference speeds. In resource-constrained settings such as those limited by hardware capabilities, real-time latency requirements, or operational cost models like BERT may provide a more pragmatic solution, delivering strong performance with substantially lower computational overhead.

5. CONCLUSION

This study conducted a systematic evaluation of classical machine learning and transformer-based models on the Deceptive Opinion Spam Corpus, with its central contribution being the establishment of a statistically validated benchmark using a 10-fold cross-validation framework. This methodology provides more stable performance estimates for baseline transformer models, with RoBERTa achieving a mean accuracy of 90.00% (SD ±0.8). While RoBERTa demonstrated strong reliability in detecting deceptive reviews, its performance remains below that of recent state-of-the-art systems (94–98%), including GNN- and DeBERTa-based models that leverage richer relational and multimodal information. These gaps suggest that improvements derived solely from text-based linguistic features may be nearing an upper bound, reinforcing the need to incorporate behavioral and social context to enhance detection capabilities further.

Beyond performance benchmarking, this study highlights the importance of balancing accuracy with computational efficiency. As summarized in Table 8, the 3.75% accuracy improvement from BERT (86.3%) to RoBERTa (90.0%) incurs substantially higher computational demands and slower CPU inference speeds. This finding emphasizes that model selection should be context-dependent. In environments with ample computational resources, adopting state-of-the-art architectures is justifiable; however, in settings constrained by latency, budget, or hardware, BERT presents a highly efficient and practical alternative.

Looking ahead, future research should focus on developing hybrid multimodal models that integrate textual, behavioral, social, and network-based signals. Such models should be evaluated using a multidimensional framework that simultaneously considers accuracy, latency, memory efficiency, and robustness to adversarial manipulation. Advancing in this direction will support the creation of more resilient and generalizable fake review detection systems capable of operating effectively across diverse and evolving real-world environments.

Acknowledgments

The authors express their sincere gratitude to the universities and departments whose support made this research possible. We acknowledge the valuable academic environment and resources provided by the Department of Computer Science and Engineering at Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, the Department of Artificial Intelligence and Data Science at Saveetha Engineering College, the Department of Computer Science and Engineering at Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, the Department of Electrical and Electronics Engineering at St. Peter's Institute of Higher Education and Research, and the Department of Electronics and Communication Engineering at S.A. Engineering College. Their collective contributions were instrumental in the successful completion of this study.

Conflicts of Interest

The authors declare that no conflicts of interest are associated with this study. All aspects of the research were conducted with the utmost integrity and transparency.

References

[1] J. Banks, "Local Consumer Review Survey 2022," *BrightLocal Ltd*, 2022. https://www.brightlocal.com/research/local-consumer-review-survey-2022/

[2] B. Yalcinkaya and D. R. Just, "Comparison of Customer Reviews for Local and Chain Restaurants: Multilevel Approach to Google Reviews Data," *Cornell Hosp. Q.*, vol. 64, no. 1, pp. 63–73, 2023, https://doi.org/10.1177/19389655221102388

[3] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Manage. Sci.*, vol. 62, no.

12, pp. 3412–3427, 2016, https://doi.org/10.1287/mnsc.2015.2304

[4] D. Mayzlin, Y. Dover, and J. Chevalier, "Promotional reviews: An empirical investigation of online review manipulation," *Am. Econ. Rev.*, vol. 104, no. 8, pp. 2421–2455, 2014, https://doi.org/10.1257/aer.104.8.2421

[5] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3634–3642, 2015, https://doi.org/10.1016/j.eswa.2014.12.029

[6] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 191–200.

[7] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 2012, vol. 2, pp. 171–175.

[8] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2007, pp. 547–552. https://doi.org/10.1109/ICDM.2007.68

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, https://doi.org/10.1038/nature14539

[10] T. Zhao, J. Du, Y. Shao, and A. Li, "Aspect-Based Sentiment Analysis Using Local Context Focus Mechanism with DeBERTa," in *2023 5th International Conference on Data-Driven Optimization of Complex Systems, DOCS 2023*, 2023, pp. 1–6. https://doi.org/10.1109/DOCS60977.2023.10294548

[11] A. A. Harby and F. Zulkernine, "A Comparative Analysis of Graph Neural Networks for Fake News Detection," in *Proceedings - International Computer Software and Applications Conference*, 2023, vol. 2023-June, pp. 1215–1222. https://doi.org/10.1109/COMPSAC57700.2023.00184

[12] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting Fake Reviews via Collective Positive-Unlabeled Learning," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2014, vol. 2015-Janua, no. January, pp. 899–904. https://doi.org/10.1109/ICDM.2014.47

[13] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *J. Big Data*, vol. 2, no. 1, p. 23, 2015, https://doi.org/10.1186/s40537-015-0029-9

[14] N. Jindal and B. Liu, "Opinion spam and analysis," in *WSDM'08 - Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 219–229. https://doi.org/10.1145/1341531.1341560

[15] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," *ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, vol. 1, pp. 309–319, 2011.

[16] S. Feng, L. Xing, A. Gogar, and Y. Choi, "Distributional footprints of deceptive product reviews," in *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012, vol. 6, no. 1, pp. 98–105. https://doi.org/10.1609/icwsm.v6i1.14275

[17] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?," in *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, 2013, vol. 7, no. 1, pp. 409–418. https://doi.org/10.1609/icwsm.v7i1.14389

[18] E. P. Lim, V. A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *International Conference on Information and Knowledge Management, Proceedings*, 2010, pp. 939–948. https://doi.org/10.1145/1871437.1871557

[19] J. Lu, X. Zhan, G. Liu, X. Zhan, and X. Deng, "BSTC: A Fake Review Detection Model Based on a Pre-Trained Language Model and Convolutional Neural Network," *Electron.*, vol. 12, no. 10, p. 2165, 2023, https://doi.org/10.3390/electronics12102165

[20] G. Bathla, P. Singh, R. K. Singh, E. Cambria, and R. Tiwari, "Intelligent fake reviews detection based on aspect extraction and analysis using deep learning," *Neural Comput. Appl.*, vol. 34, no. 22, pp. 20213–20229, 2022, https://doi.org/10.1007/s00521-022-07531-8

[21] N. Wang, J. Yang, X. Kong, and Y. Gao, "A fake review identification framework considering the suspicion degree of reviews with time burst characteristics," *Expert Syst. Appl.*, vol. 190, p. 116207, 2022, https://doi.org/10.1016/j.eswa.2021.116207

[22] G. Stanton and A. A. Irissappane, "GANs for semi-supervised opinion spam detection," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2019-Augus, pp. 5204–5210, 2019, https://doi.org/10.24963/ijcai.2019/723

[23] M. Huss and S. Förster, "Vorstoss und Rückzug der Gletscher während der Kleinen Eiszeit," *arXiv Prepr. arXiv1907.00001*, no. 2011, 2019, [Online]. Available: https://arxiv.org/abs/1907.00001

[24] L. Xin, "Spin-1 Bosons in the Presence of Spin-orbit Coupling," *arXiv Prepr. arXiv1805.00001*, 2018, [Online]. Available: http://arxiv.org/abs/1805.00001

[25] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: A network-based spam detection framework for reviews in online social media," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 7, pp. 1585–1595, 2017, https://doi.org/10.1109/TIFS.2017.2675361

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, 2019.

[27] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pre-training Approach," *arXiv Prepr. arXiv1907.11692*, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[28] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pre-training for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[29] R. Gupta, V. Jindal, and I. Kashyap, "Recent state-of-the-art of fake review detection: a comprehensive review," *Knowl. Eng. Rev.*, vol. 39, p. e8, 2024, https://doi.org/10.1017/S0269888924000067

[30] L. C. Cheng, Y. T. Wu, C. T. Chao, and J. H. Wang, "Detecting fake reviewers from the social context with a graph neural network method," *Decis. Support Syst.*, vol. 179, p. 114150, 2024, https://doi.org/10.1016/j.dss.2023.114150

[31] S. Xu, H. Cuan, Z. Yin, and C. Yin, "A Hybridized Approach for Enhanced Fake Review Detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 6, pp. 7448–7466, 2024, https://doi.org/10.1109/TCSS.2024.3411635

[32] M. Zhang, Y. Zhang, and X. Zhang, "SGAN-SAM-ERNIE: A Novel and Effective Detection Scheme for Chinese Fake Reviews," *IEEE Access*, vol. 12, pp. 114190–114197, 2024, https://doi.org/10.1109/ACCESS.2024.3445354

[33] Y. Pan and L. Xu, "Detecting Fake Online Reviews: An Unsupervised Detection Method With a Novel Performance Evaluation," *Int. J. Electron. Commer.*, vol. 28, no. 1, pp. 84–107, 2024, https://doi.org/10.1080/10864415.2023.2295067

[34] P. Phukon, P. Potikas, and K. Potika, "Detecting Fake Reviews

Using Aspect-Based Sentiment Analysis and Graph Convolutional Networks," *Appl. Sci.*, vol. 15, no. 7, 2025, https://doi.org/10.3390/app15073771

[35] J. Salminen, M. Mustak, S. G. Jung, H. Makkonen, and B. J. Jansen, "Decoding deception in the online marketplace: enhancing fake review detection with psycholinguistics and transformer models," *J. Mark. Anal.*, pp. 1–18, 2025, https://doi.org/10.1057/s41270-025-00393-8

[36] M. Puttarattanamanee, L. Boongasame, and K. Thammarak, "A Comparative Study of Sentiment Analysis Methods for Detecting Fake Reviews in E-Commerce," *HighTech Innov. J.*, vol. 4, no. 2, pp. 349–363, 2023, https://doi.org/10.28991/HIJ-2023-04-02-08

[37] H. Aghakhani, A. MacHiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, 2018, pp. 89–95. https://doi.org/10.1109/SPW.2018.00022

[38] P. He, J. Gao, and W. Chen, "Debertav3: Improving Deberta Using Electra-Style Pre-Training With Gradient-Disentangled Embedding Sharing," *11th Int. Conf. Learn. Represent. ICLR 2023*, 2023.

[39] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection: An empirical study," *Inf. Sci. (Ny).*, vol. 385–386, pp. 213–224, 2017, https://doi.org/10.1016/j.ins.2017.01.015

[40] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Inf. Process. Manag.*, vol. 54, no. 4, pp. 576–592, 2018, https://doi.org/10.1016/j.ipm.2018.03.007

[41] R. Mohawesh, S. Xu, M. Springer, M. Al-Hawawreh, and S. Maqsood, "Fake or Genuine? Contextualised Text Representation for Fake Review Detection," *arXiv Prepr. arXiv2112.14343*, pp. 137–148, 2021, https://doi.org/10.5121/csit.2021.112311

[42] S. Geetha, E. Elakiya, R. S. Kanmani, and M. K. Das, "High performance fake review detection using pretrained DeBERTa optimized with Monarch Butterfly paradigm," *Sci. Rep.*, vol. 15, no. 1, p. 7445, 2025, https://doi.org/10.1038/s41598-025-89453-8