

Fake Review Detection on E-Commerce Websites Using Existing Studies: An Experimental Study

Naveen¹, Disha Handa²

Research Scholar¹, Professor²

Department of Computer Applications, University Institute of Computing, Chandigarh University

Mohali, Punjab, India – 140413

Corresponding author: naveenkbansal1988@gmail.com

Abstract— Recently, the identification of fraudulent reviews has attracted significant interest from both the business and academic sectors. Detecting fraudulent reviews is a crucial and challenging endeavor to ensure that evaluations accurately reflect genuine user experiences and perspectives. This document presents a comparative analysis of various machine learning and deep learning algorithms, including Naïve Bayes, Logistic Regression, Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and RoBERTa, for the purpose of detecting fraudulent reviews on e-commerce platforms. The efficacy of the models is assessed by comparing their accuracy, precision, recall, and F1-score using a dataset of Amazon reviews that includes both authentic and artificially generated reviews. The findings indicate that RoBERTa, a transformer architecture, has the maximum accuracy at 93%, surpassing all other models evaluated in this research. Furthermore, we examine the linguistic and behavioral traits that distinguish authentic reviews from fraudulent ones, along with the consequences for the credibility of e-commerce platforms, trustworthiness, and the future of AI-generated content identification.

Index Terms—Fake review detection, e-commerce, RoBERTa, LSTM, BiLSTM, Natural Language Processing, deep learning, spam detection, transformer models, text classification.

1. INTRODUCTION

E-commerce platforms have completely changed the way in which consumers search for and purchase products. Customer reviews are a crucial element of the purchasing process, providing social proof and influencing consumer behavior. But the rising trend of spam reviews, which are reviews written by spammers for money or by AI algorithms, is a serious threat to the integrity of these platforms. Most e-commerce platforms have a review system for customer feedback. Customers usually take decisions regarding purchasing a product based on customer reviews. The report by the Spiegel Research Center, “Nearly 95% of shoppers read online reviews before making a purchase, and products with at least five reviews are 270% more likely to be purchased than those with no reviews” [1]. This trend provides a huge financial motive for dishonest vendors to cheat their review ratings. Some e-commerce platform owners may hire spammers to write fake reviews to increase their own product sales or decrease the ratings of rival vendors. It is therefore crucial to automatically identify these types of fraudulent reviews to ensure that the trust of consumers in these platforms is maintained and that there is

healthy competition in the market. This situation is further complicated by the recent development of Large Language Models (LLMs) like GPT-and other similar models, which are capable of producing extremely realistic fake reviews that are not only semantically but also syntactically indistinguishable from actual human-written reviews.

India's digital economy is growing rapidly, highlighting the necessity for strong FRD processes. India has 16% of worldwide internet users, but only 55% of its population has internet connection as of 2025, indicating huge digital potential [2]. While GitHub added 2.5 million members in 2022, India's total reached 9.75 million developers and might match the US by 2025 [3]. As digital adoption accelerates, the e-commerce sector (Figure 1) has grown from US\$30 billion in 2020 to US\$123 billion in 2024, with forecasts of US\$300 billion by 2030 [4]. However, Amazon and Flipkart dominate over 70% of the market, raising worries about monopolistic control, data centralization, and algorithmic manipulation [5]. Globally, 16% to over 50% of internet reviews are false, with most studies falling between 20–30% depending on platform and product category [6]. Up to 82% of users encounter misleading reviews annually, however trust in reviews remains strong despite rising manipulation [7]. NBER experiments show that bogus reviews can cause customers to overpay by \$0.12 per dollar and choose lower-quality products, although educational interventions can minimize these impacts [8]. The World Economic Forum estimates that bogus reviews may affect US\$152 billion in global e-commerce transactions yearly, with consumer losses reaching US\$788 billion and growing to US\$1.1 trillion by 2030 without tougher responses. Digital trust is essential to economic participation as internet use rises and customers make product and service decisions online. However, false online reviews, often exacerbated by opaque platform algorithms, threaten to mislead consumers, distort competition, and marginalize SMEs, which account for 60% of Indian employment [9,10]. Indian consumers have limited disposable resources and are more susceptible to manipulation, yet their data is worth less in ad-targeting systems, creating a conflict between platform incentives and citizen wellbeing. Thus, Indian government systems like the Open Network for Digital Commerce (ONDC), account aggregator models, and digital public infrastructure across identity, data, and payments aim to reclaim regulatory control and ensure fair competition [11]. In the digital age, FRD is both technical and regulatory. Online commerce must be real, fair, and inclusive. Without comprehensive detection methods, fraudulent evaluations can influence customer behaviour, promote undeserving products or sellers, and undercut reputable firms, especially SMEs competing in algorithm-driven ecosystems. Thus, intelligent, multilingual, and real-time FRD models must include trust and transparency in review processes to protect consumer rights and market integrity in India's fast-growing digital economy.

Natural Language Processing (NLP) has become one of the most important fields of study in Artificial Intelligence owing to its capacity to make computing machines comprehend and process human languages [12,13]. The proliferation of communication and commerce technologies on the Internet has led to the creation of huge amounts of text data created by users, thereby providing significant scope for analysing this data using NLP techniques [14]. Sentiment analysis, opinion mining, fake reviews detection, recommendation system, and conversational AI require clean and linguistically consistent textual data [15]. The performance

of these applications depends significantly on the quality of pre-processing that has been performed on the raw text before the analysis takes place.

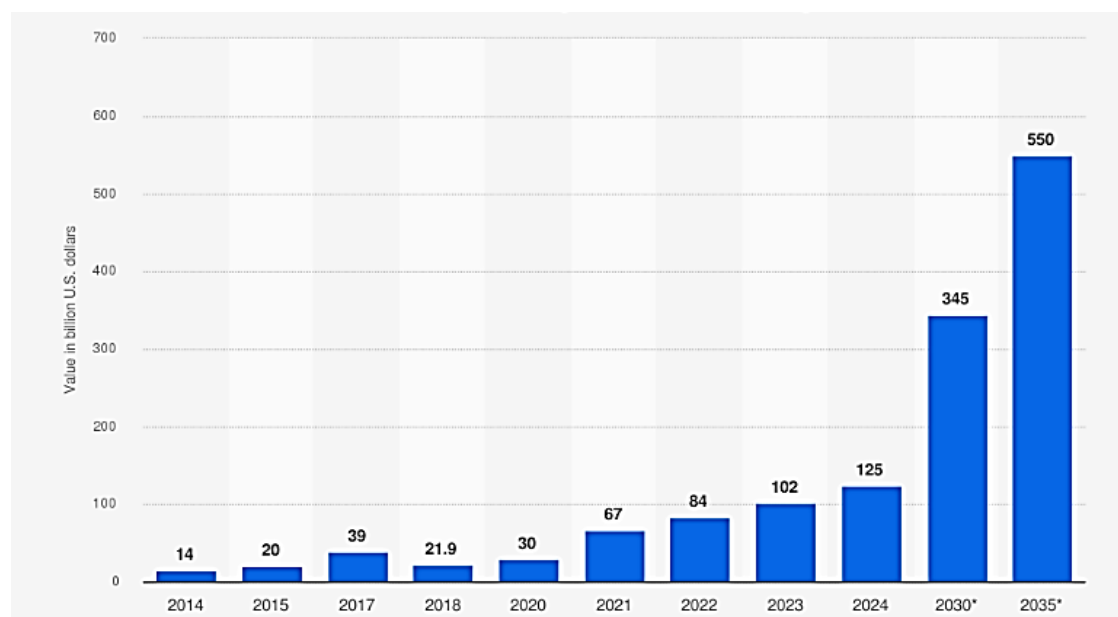


Figure 1. Global E-commerce growth from 2014 to 2024, with projected increases to 2030 and 2035 (in billion USD).

Though much success has been attained in NLP in relation to high-resource languages such as English, preprocessing and linguistic analysis of low-resource languages like Hindi continue to face several challenges [16]. Hindi is a morphologically complex language characterized by freedom in word order since the meaning of the sentence may not be altered by any changes in word order [17]. Moreover, the Hindi texts generated by users tend to contain mixed scripts, transliterations, spelling errors, inconsistent abbreviations, informal writing style, and contextual differences due to the behaviour of digital communication [18]. These factors hinder the effectiveness of traditional text processing algorithms that are developed mainly for the English language.

One other major problem associated with the domain of Hindi NLP is preserving language authenticity at the same time improving consistency of collected data [19]. Over-normalization could result in losing valuable context-specific terms used in everyday conversations, while under-preprocessing could bring lots of noise to the texts. Consequently, for the design of an efficient preprocessing system for Hindi, there is a need for a balanced system that can be utilized to minimize discrepancies while maintaining semantic consistency and realistic language usage patterns [20]. With the growing availability of AI-produced textual data in recent times, there has been a greater demand for efficient preprocessing systems that can be used to support downstream NLP applications like AI-produced text detection and text classification. But most research works concentrate on the development of models and preprocessors become an afterthought. Very few research papers discuss about Hindi-specific pre-processing techniques [21].

In light of the above issues, this research work suggests a systematic approach to pre-processing the Hindi reviews acquired through a controlled environment. The proposed technique will include several stages such as handling missing values, removing duplicates, normalization using Unicode Devanagari script, noise reduction, empty text removal, and length-based filtering without losing contextual information. Besides, it also highlights the importance of retaining the free world order characteristic of the Hindi language and utilizes linguistic analysis tools like dependency parsing and tokenization.

The experimental data used in this paper comprises of the product reviews in Hindi language that were obtained via a controlled survey technique with the inclusion of validation criteria to guarantee linguistic homogeneity as shown in Figure 2. The processing of the raw data involved transformation of the dataset into a more refined textual corpus consisting of 1311 quality Hindi reviews. The suggested framework will add value to the methods of preprocessing of data for Hindi natural language processing and might act as the foundation of future researches on low-resource Indic languages.

Transformer-based designs like BERT and its descendants enabled contextual comprehension and task-specific fine-tuning in FRD [22]. These models captured long-range dependency and subtle deception better than previous attempts. BERT improves multilingual and low-resource datasets in several experiments [23].

GCNs and GATs [24] are robust graph-based models that can capture reviewer-product interactions and community-level irregularities. These models model structural links to improve language interpretability and durability against adversary manipulation. Recent hybrid methods combine textual, behavioral, semantic, and structural elements with ensemble or multi-view learning [25]. To simulate the complexity of the actual world, synthetic data and adversarial learning have been used to improve model resilience and replicate misleading circumstances.

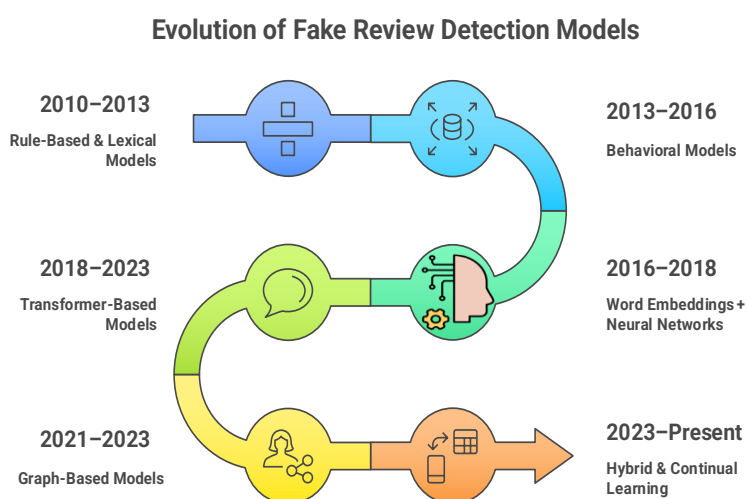


Figure 2. Chronological evolution of FRD models from 2010 to the present, highlighting key methodological shifts

Digital trust is essential to economic participation as internet use rises and customers make product and service decisions online. However, false online reviews, often exacerbated by opaque platform algorithms, threaten to mislead consumers, distort competition, and marginalize SMEs, which account for 60% of Indian employment [26]. Indian consumers have limited disposable resources and are more susceptible to manipulation, yet their data is worth less in ad-targeting systems, creating a conflict between platform incentives and citizen wellbeing. Thus, Indian government systems like the Open Network for Digital Commerce (ONDC), account aggregator models, and digital public infrastructure across identity, data, and payments aim to reclaim regulatory control and ensure fair competition [27].

1.1 Spam reviews can be divided into two types:

(1) manually written fake reviews by spammers or review farms; and (2) automatically written fake reviews by natural language generation (NLG) tools or LLMs. Both types are very dangerous, but the latter type is exponentially more dangerous due to its sheer volume and low cost of production. This work is centered on the comparison of various machine learning and deep learning algorithms to determine the best method for detecting fake reviews. We apply natural language processing (NLP) and behavioral characteristics derived from reviews to develop classifiers. The contributions of this work are as follows:

A comprehensive comparison of five algorithms (Naïve Bayes, Logistic Regression, LSTM, BiLSTM, and RoBERTa) using the same experimental setting on a given dataset.

A comprehensive discussion of the relative merits and demerits of each algorithm in the context of fake review detection.

Information on the linguistic and behavioural characteristics that are most informative for fake review detection. Discussion of future work, especially on LLM-based fake review detection.

2. LITERATURE SURVEY

Significantly, research on fake review detection has flourished over the recent years to tackle this critical concern in the field. A summary of the major research works from 2021 to 2026 is provided in Table I, which includes the year, main contribution, and technology used in each work.

A. Classical Machine Learning Approaches:

The initial research on fake review detection mainly focused on classical machine learning approaches. The first were to propose fake review detection as a text classification problem and used Support Vector Machines (SVM) with unigrams and bigrams as features [1]. This research proved that fake reviews contain specific psycholinguistic characteristics, such as increased use of spatial words and reduced occurrences of concrete nouns, compared to original reviews. Naive Bayes classifiers are also popular due to their efficiency and

effectiveness on small data sets. However, their assumption of independence does not allow them to capture word co-occurrence patterns, which are unique to both original and fake reviews [2].

B. Deep Learning and Sequential Models:

The introduction of Recurrent Neural Networks (RNNs) and specifically LSTMs has been a major milestone in the detection of fake reviews [3] LSTMs are more effective in capturing long-range dependencies in text compared to bag-of-words models. Bidirectional LSTMs are even more effective, as they use both left and right contextual information [4].

C. Transformer-Based Models:

The introduction of pre-trained models such as BERT [3], RoBERTa, and other variants has set new benchmarks in nearly all NLP tasks, including fake review detection. These models have been pre-trained on massive volumes of text data and are capable of capturing contextual representations of words and sentences. These representations are far more effective compared to other models [5].

D. Graph-Based and Behavioural Approaches:

Researchers have also attempted to incorporate other behavioral and relational signals in detecting fake reviews. In one such study, it showed how graph-based models can help in detecting coordinated attacks on products [6]. These models analyze the interaction network between reviewers and products. Recently, Graph Neural Networks have also been explored in this direction [7]. Table 1 shows the literature report from 2021 to 2026

Table I
LITERATURE SURVEY: FAKE REVIEW DETECTION RESEARCH (2021–2026)

Year	Key Finding	Technology / Method	Ref.
2021	Survey highlighted a methodological shift from classical ML to deep learning for fake review detection. SVM, Naive Bayes, and KNN were combined with text and behavioral features.	Classical ML (SVM, NB, KNN) + text and behavioral feature engineering	[1]
2022	Network structure analysis helps detect coordinated fake-review buyers. Modeling reviewer– product graphs reveal clusters of suspicious activity.	Graph / Network analysis of reviewer– product links	[5]
2023	RoBERTa with behavioral features (punctuation, ratings metadata) achieves high accuracy ($\approx 96\%$), outperforming classical approaches.	Transformer (RoBERTa) + metadata features (punctuation, star ratings)	[9]
2024	ChatGPT-generated and paraphrased reviews confuse traditional detectors, exposing critical limitations of shallow classifiers.	LLMs (GPT-based) used to generate and stress test fake review detectors	[11]

2025	GNNs are promising for large-scale detection by modeling user product review interactions, but face scalability challenges.	Graph Neural Networks (user–product–review graph modeling)	[14]
2026	Transformer-based ensemble approaches integrated BERT and GloVe embeddings with sentiment analysis , demonstrating the continued shift toward semantic, contextual, and ensemble-based fake-review detection	BERT, GloVe, Transformer ensemble, sentiment analysis	[17]

3. PROBLEM STATEMENT

Throughout the years, numerous techniques were proposed for detecting fake reviews. However, an important aspect that has not been covered is that while many models were evaluated individually, none were compared with each other under similar circumstances. Another aspect is that classical machine learning models perform badly with the complex context found in natural language and are particularly poor at detecting AI-written fake reviews.

The problem can be formulated as follows: we are given reviews $R = \{r_1, r_2, \dots, r_n\}$, and each review r_i has an associated label y_i , which belongs to $\{0, 1\}$. If y_i equals 1, it means it is a fake review, and vice versa. Our goal is to learn an optimal classifier f such that it maps r_i to y_i with the best possible performance.

The paper conducts an experiment comparing various machine and deep learning models, namely Naive Bayes, Logistic Regression, LSTM, BiLSTM, and RoBERTa, under similar circumstances to identify which model performs best in detecting spam and non-spam reviews.

4. EXISTING SYSTEMS AND THEIR LIMITATIONS

Machine learning techniques such as KNN, Random Forest and SVM have been popularly used in detecting fake reviews. These techniques primarily use simple text features such as word counts, bigrams, and TF-IDF. There are a number of disadvantages associated with this technique:

Lack of contextual understanding: By considering text as a bag of words, we are losing contextual meaning. This is because words have different meanings when used in different contexts. The order in which words are arranged in a sentence is important in understanding the meaning of a sentence.

Lack of robustness against AI-generated content: Fake reviews are being created using large language models. These reviews are almost impossible to differentiate from original reviews when we are using simple features.

Heuristic-based feature engineering: Classical machine learning techniques require a lot of feature engineering. This is a tedious process.

Lack of transfer learning: Classical machine learning techniques do not have transfer learning. This is because we are training a model from scratch on a new dataset.

Lack of scalability: The number of reviews is increasing exponentially on e-commerce websites. In a real-time system, this technique is not feasible.

Lack of understanding of reviewer behavior: Classical machine learning techniques do not consider any other feature other than text. This is a major drawback because a number of frauds are related to reviewers.

5. PROPOSED SYSTEM

A. Overview

The proposed system employs NLP and deep learning models to overcome the limitations of classical approaches. The pipeline employs five processes with comparative evaluation.

B. Dataset

Amazon product reviews include real and fake user feedback. The dataset is evenly distributed for training and testing to avoid class-imbalance bias. Table 2 summarizes dataset statistics.

Split	Genuine Reviews	Fake Reviews
Training Set	3,500	3,500
Validation Set	500	500
Test Set	1,000	1,000
Total	5,000	5,000

The fake reviews in the dataset were created using automated natural language generation pipelines, covering a spectrum of quality, from low-quality repetitive spam to high-quality, human-like content.

C. Preprocessing

Raw review text undergoes the following preprocessing steps:

Cleaning: Remove URL tags, peculiar letters, and useless symbols.

Tokenization: Dividing text into separate tokens

Stop-word Removal: Eliminating prevalent non-informative terms

Lemmatization: Normalizing language by simplifying terms.

Lowercasing: Lowercase all text to minimize vocabulary.

Padding/Truncation: Deep learning models pad or truncate sequences to a set length (128 tokens for LSTM/BiLSTM, 512 for RoBERTa).

D. Feature Extraction

Two categories of features are extracted:

Behavioural features: metadata such as how frequently punctuation occurs, the presence of emojis, the ratio of capital letters, and the star ratings.

The formula for the TF-IDF feature matrix is presented as follows:

TF-IDF (t, d) is calculated as TF (t, d) multiplied by the logarithm of the quotient of N and DF(t).

TF (t, d) denotes the frequency of phrase t within document d.

N represents the aggregate quantity of papers,

DF(t) denotes the quantity of documents that include the phrase t.

In the transformer family of models, the input tokens are represented as dense vectors, and the position of the token is represented by the addition of positional encoding:

E. Algorithms Compared

The following five models were trained and tested:

- 1) Naive Bayes (NB): This probabilistic classifier assumes features are conditionally independent and uses Bayes' theorem. The classification rule for a review r with features {f1, f2, ..., fk} is:

$$\hat{y} = \arg \max_{c \in \{0,1\}} P(c) \prod_i P(f_i|c) \quad (1)$$

This classifier is used as a baseline because it is computationally efficient and has a high chance of success when the problem is linearly separable with a large training set.

- 2) Logistic Regression (LR): It is commonly used for binary text classification with TF-IDF features. The probability that a review is fake is modeled using the logistic regression formula:

$$M(z = 1|y) = \sigma(uYa+b) = 1 / (1 + \exp(-(uYa+b))) \quad (2)$$

Here, x is a vector of TF-IDF features, u is a weight vector, b is a bias term, and σ is the sigmoid function.

- 3) Long Short-Term Memory (LSTM): This is a type of recurrent neural network architecture that can capture long-range dependencies between text. The main LSTM equations are:

$$fs_t = (Zfr [qt-1, Xt] + \beta f_s) \quad (3)$$

$$input_t = (Zin [qt-1, Xt] + \beta in) \quad (4)$$

$$C_t = \tanh(Z_c[q_{t-1}, X_t] + \beta c) \quad (5)$$

$$C_t = f_{s_t} \times C_{t-1} + \text{input}_t * C_t \quad (6)$$

$$\text{output}_t = \rho(Z_o[q_{t-1}, X_t] + \beta o) * \tanh(C_t)$$

Here, f_t , i_t , and o_t are the forget gate, input gate, and output gate, respectively.

- 4) Bidirectional LSTM (BiLSTM): This is a type of LSTM architecture that reads text sequences both forward and backward. This provides additional context for a word. The forward LSTM passes through the text from left to right, while the backward LSTM passes through the text from right to left. The combined output of the forward and backward LSTM is a concatenation of the output of the forward LSTM at a word position $h_t^{\rightarrow} = \text{LSTM_fwd}(x_t, h_{t-1}^{\rightarrow})$ and the output of the backward LSTM at a word position $h_t^{\leftarrow} = \text{LSTM_bwd}(x_t, h_{t+1}^{\leftarrow})$.

$$\text{forwar}[ht] = \text{forward}[\text{LSTM}(q_{t-1}, X_t)] t \in [1, P] \quad (7)$$

$$\text{backwar}[ht] = \text{backward}[\text{LSTM}(q_{t-1}, X_t)] t \in [P, 1] \quad (8)$$

$$\text{output}^b = \mu \text{forward}(ht) + \vartheta(\text{back}(ht)) \quad (9)$$

$\text{forward}[ht]$ and $\text{backward}[ht]$ are the solar GHI value at time t while output_t^b represents the output value and P represent the length of the source data.

"RoBERTa" is a robustly optimized BERT pre-training method, which is fine-tuned on the fake review dataset. It improves upon the BERT model with larger mini-batch sizes, more data, longer input context, and the elimination of the Next Sentence Prediction objective. The deep bidirectional transformer architecture allows it to have a better understanding of the semantics of the text. The self-attention mechanism is computed as follows:

Attention $(X, Y, Z) = \text{softmax}(XY^N / \sqrt{\text{lm}}) * U$, where X , Y , and Z are the query, key, and value matrices, and dk is the dimension of the keys.

6. Hyperparameter Configuration

Table 3 lists the hyperparameters used for the models in our experiments.

Model	Key Hyperparameters
Naive Bayes Regression	Additive smoothing $\alpha = 1.0$; TF-IDF max features = 50,000 Logistic C = 1.0 (L2 regularization); solver = lbfgs; max_iter = 1,000
LSTM	Hidden units = 128; Dropout = 0.3; Embed-ding dim = 100; Epochs = 10; Batch size = 64
BiLSTM	Hidden units = 150 (each direction); Dropout = 0.5; Epochs = 200; Batch size = 64
RoBERTa	pre-trained: roberta-base; Learning

$$\text{rate} = 2 \times 10^{-5}; \text{Epochs} = 3; \text{Batch size} = 16; \text{Max seq len} = 512$$

7. Evaluation Metrics

We will be using the following metrics to evaluate all models:

$$\text{- Accuracy: Acc} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (10)$$

$$\text{- Precision: Prec} = \text{TP} / (\text{TP} + \text{FP}) \quad (11)$$

$$\text{- Recall: Rec} = \text{TP} / (\text{TP} + \text{FN}) \quad (12)$$

$$\text{- F1-Score: F1} = 2 \times \text{Prec} \times \text{Rec} / (\text{Prec} + \text{Rec}) \quad (13)$$

Here, TP, TN represent the True Positive and True Negative while FP, FN shows the False Positive and False Negative.

8. SYSTEM ARCHITECTURE

The overall system architecture follows a sequential pipeline from raw data ingestion to final classification output key stages as shown in Figure 3.

- 1) **Data Input Layer:** Amazon review dataset consisting of real and machine-generated fake reviews.
- 2) **Preprocessing Module:** Text cleaning, tokenization, stop-word removal, and lemmatization.
- 3) **Feature Extraction Module:** Text embeddings for deep learning models; TF-IDF vectors and behavioral feature vectors for classical models.
- 4) **Model Training Module:** Each of the five algorithms is independently trained on the processed feature set.
- 5) **Classification Output:** Binary prediction — Genuine or Fake review.
- 6) **Evaluation Module:** Computation of accuracy, precision, recall, and F1-score for all models, followed by comparative analysis.

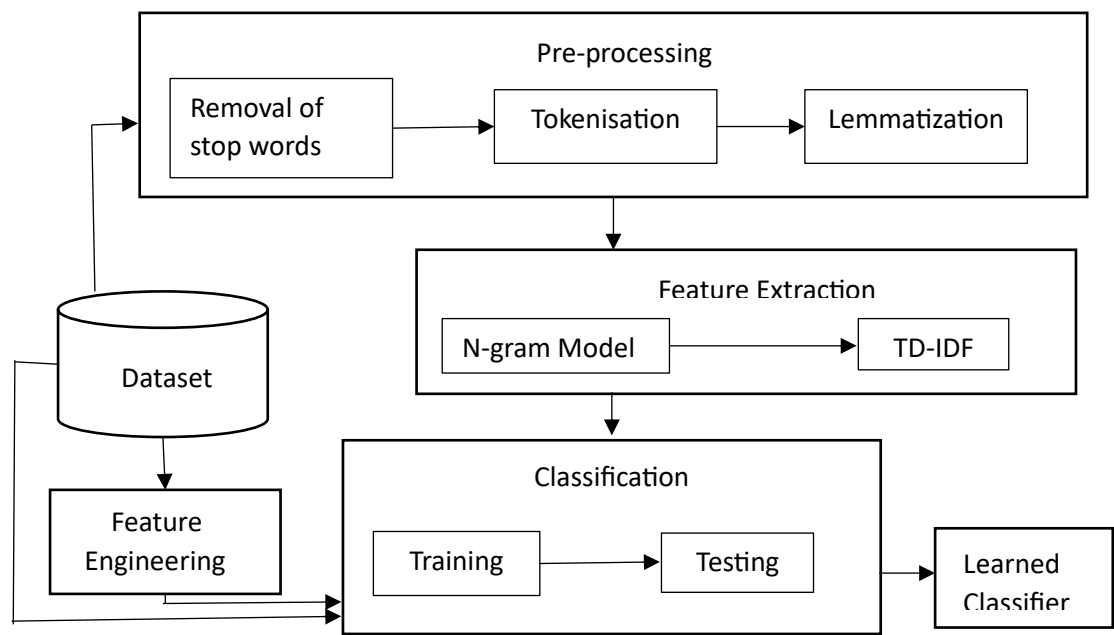


Figure 3 System Architecture

9. RESULT & DISCUSSION

Table 4 summarizes the performance of all five models evaluated in this study on the Amazon fake review dataset.

Table 4. Evaluation Metrics of Model

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	75%	0.73	0.78	0.75
Logistic regression	79%	0.78	0.80	0.79
LSTM	85%	0.84	0.86	0.85
BiLSTM	88%	0.87	0.89	0.88
RoBERTa	93%	0.92	0.94	0.93

RoBERTa significantly outperformed all other models, achieving an accuracy of 93%. This superior performance is attributed to its deep bidirectional attention mechanism, which captures rich long-range semantic relationships through pretraining on large text corpora.

LSTM and BiLSTM performed better than classical ML models due to their ability to model sequential word dependencies. BiLSTM’s bidirectional design gave it an additional advantage over standard LSTM by incorporating context from both preceding and

succeeding tokens.
Naive Bayes and Logistic Regression, while computationally efficient and fast to train, showed comparatively limited accuracy due to their inability to capture the complex linguistic patterns inherent in human-written text.

9.1 Confusion Matrix Analysis

To further understand model behavior, we analyze the confusion matrices for each classifier as shown in Table 5. RoBERTa produced the fewest false negatives (fake reviews classified as genuine), which is the most critical error type from a consumer protection standpoint—a genuine-looking fake review that goes undetected can mislead thousands of buyers.

Table 5
CONFUSION MATRIX: ROBERTA (ON 2500-SAMPLE TEST SET)

	Predicted: Genuine	Predicted: Fake
Actual: Genuine	950 (TN)	45 (FP)
Actual: Fake	31 (FN)	964(TP)

9.2 Comparison Across Models: Trend Analysis

The progression in performance from Naive Bayes to RoBERTa reflects the broader evolution of NLP techniques. The accuracy gain from NB (75%) to LR (79%) reflects the benefit of modeling feature interactions through linear combinations. The jump from LR (79%) to LSTM (85%) demonstrates the power of sequential modeling.
LSTM / BiLSTM: Showed weakness on very long reviews (> 250 tokens) where relevant discriminative signals appeared late in the sequence.
RoBERTa: The remaining 4% error rate was largely concentrated on highly sophisticated, LLM-generated fake reviews that even human annotators found difficult to identify.

10. CONCLUSION

This study compares machine learning and deep learning algorithms for online shopping platform false review detection. The Amazon review dataset with actual and machine-generated fraudulent reviews was pre-processed extensively. Preprocessing included text cleaning, tokenization, common word elimination, and lemmatization. Machine learning models retrieved text embeddings and behavioural features for five classifiers: Naive Bayes, Logistic Regression, LSTM, BiLSTM, and RoBERTa. The top model was RoBERTa, which detected false online purchasing reviews with 93% accuracy. This makes it the greatest real-time false review detector. This model has an advantage over others due to its transformer construction. These models' mathematical formulations, hyperparameters, training methods,

linguistic feature analysis, and error analysis were presented to better understand their behavior. Error analysis shows that the highest performing model, RoBERTa, has a 4% miss rate because to sophisticated phony reviews made by LLMs, making this a problem for academics.

In this study, future challenges include fine-tuning the best performing model, RoBERTa, using domain-specific e-commerce review datasets; incorporating GNNs for reviewer-product relationship analysis for graph-level spam review detection; designing models for detecting fake reviews generated by LLMs, a new challenge for fake review detection models; and exploring more efficient models like Distil.

REFERENCES

- [1] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake Reviews Detection: A Survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021, doi: 10.1109/ACCESS.2021.3075573.
- [2] A. Mewada and R. K. Dewang, "Research on false review detection Methods: A state-of-the-art review," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, pp. 7530–7546, 2022, doi: 10.1016/j.jksuci.2021.07.021.
- [3] "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, Art. no. 102771, 2022, doi: 10.1016/j.jretconser.2021.102771.
- [4] "Leveraging Transfer learning techniques—BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection," in *Proc. 13th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE)*, 2021, doi: 10.1145/3503162.3503169.
- [5] L. He, X. Wang, H. Chen, *et al.*, "Online Spam Review Detection: A Survey of Literature," *Human-Centric Intelligent Systems*, vol. 2, pp. 14–30, 2022, doi: 10.1007/s44230-022-00001-3.
- [6] R. Catelli, H. Fujita, G. De Pietro, and M. Esposito, "Deceptive reviews and sentiment polarity: Effective link by exploiting BERT," *Expert Systems with Applications*, 2022, Art. no. 118290, doi: 10.1016/j.eswa.2022.118290.
- [7] "A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information," *Decision Support Systems*, 2022, Art. no. 113911, doi: 10.1016/j.dss.2022.113911.
- [8] "Graph Learning for Fake Review Detection," *Frontiers in Artificial Intelligence*, 2022, doi: 10.3389/frai.2022.922589.
- [9] J. Zhao, M. Shao, H. Tang, J. Liu, L. Du, and H. Wang, "RHGNN: Fake reviewer detection based on reinforced heterogeneous graph neural networks," *Knowledge-Based Systems*, vol. 280, Art. no. 111029, 2023, doi: 10.1016/j.knosys.2023.111029.
- [10] "A graph neural network approach to detect original review spammers of astroturfing campaigns," *Electronic Commerce Research and Applications*, vol. 62, Art. no. 101326, 2023, doi: 10.1016/j.elerap.2023.101326.
- [11] "Fake review detection using transformer-based enhanced LSTM and RoBERTa," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 250–258, 2024, doi: 10.1016/j.ijcce.2024.06.001.
- [12] "Detecting fake reviewers from the social context with a graph neural network method," *Decision Support Systems*, vol. 179, Art. no. 114150, 2024, doi: 10.1016/j.dss.2023.114150.
- [13] R. Gupta, V. Jindal, and I. Kashyap, "Recent state-of-the-art of fake review detection: A comprehensive review," *The Knowledge Engineering Review*, published online Nov. 21, 2024.

- [14] R. A. Duma, Z. Niu, A. S. Nyamawe, and A. A. Manjotho, "An analysis of graph neural networks for fake review detection: A systematic literature review," *Neurocomputing*, vol. 623, Art. no. 129341, 2025, doi: 10.1016/j.neucom.2025.129341.
- [15] M. Liu and M. Poesio, "Data Augmentation for Fake Reviews Detection in Multiple Languages and Multiple Domains," arXiv:2504.06917, 2025.
- [16] S. R. Veluru, S. T. Erukude, and V. C. Marella, "Multimodal Detection of Fake Reviews using BERT and ResNet-50," arXiv:2511.00020, 2025.
- [17] J. Tao, X. Fang, and L. Zhou, "Toward a Knowledge Discovery Method to Fake Review Detection," *Information Systems Frontiers*, vol. 28, pp. 1109–1125, 2026, doi: 10.1007/s10796-026-10715-5.
- [18] G. S. Chauhan, A. K. Verma, S. Chaudhary, *et al.*, "GEmFakeABSA: graph-embedded fake review detection using aspect-based sentiment analysis," *Scientific Reports*, 2026.
- [19] G. S. Chauhan and D. K. Saini, "Uncovering fake reviews through graph-embedded hierarchical attention networks," *Discover Artificial Intelligence*, 2026.
- [20] H. Azam, H. N. Chaudhry, M. Alhussein, *et al.*, "A transformer-based ensemble algorithm for fake review detection in e-commerce systems," *Scientific Reports*, published Aug. 6, 2026, doi: 10.1038/s41598-025-91095-9.
- [21] Ala'M, A. Z., Mora, A. M., & Faris, H. (2023). A multilingual spam reviews detection based on pre-trained word embedding and weighted swarm support vector machines. *IEEE Access*, 11, 72250-72271.
- [22] Xu, G., Hu, M., & Ma, C. (2021). Secure and smart autonomous multi-robot systems for opinion spammer detection. *Information Sciences*, 576, 681-693.
- [23] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2, 1-24.
- [24] Kumar, A., Kohail, S., Ekbal, A., & Biemann, C. (2015, December). IIT-TUDA: System for sentiment analysis in Indian languages using lexical acquisition. In *International Conference on Mining Intelligence and Knowledge Exploration* (pp. 684-693). Cham: Springer International Publishing.
- [25] Liu, W., He, J., Han, S., Cai, F., Yang, Z., & Zhu, N. (2019). A method for the detection of fake reviews based on temporal features of reviews and comments. *IEEE Engineering Management Review*, 47(4), 67-79.
- [26] Abdulqader, M., Namoun, A., & Alsaawy, Y. (2022). Fake online reviews: A unified detection model using deception theories. *IEEE Access*, 10, 128622-128655.
- [27] Wang, J., Kan, H., Meng, F., Mu, Q., Shi, G., & Xiao, X. (2020). Fake review detection based on multiple feature fusion and rolling collaborative training. *IEEE access*, 8, 182625-182639.