A Hybrid Machine Learning Approach for Fake Review Detection in E-Commerce

Divya Gupta, Aman Shukla, Tarun Maini

Department of Computer Science and Engineering
Galgotias University, Greater Noida –203201, India.
Corresponding Author: divyagupta8131@gmail.com

ABSTRACT

Fake review detection is a significant issue in e-commerce where product reviews are significant in consumer decision-making. This paper suggests a hybrid machine learning method that involves textual TF-IDF features and sentiment-based and behavioral Meta features like rating-sentiment gap and review length. The reason behind the use of a Linear Support Vector Machine (SVM) classifier is its performance in the high-dimensional feature spaces. Although text-only models have high in-dataset accuracy (99%), their accuracy decreases dramatically in cross-dataset tests. The hybrid model proposed has 92% accuracy and thus it has good generalization. The findings demonstrate the significance of incorporating linguistic and behavioral aspects to make fake review detection effective.

Keywords: Fake Review Detection, Opinion Spam, TF-IDF, Sentiment Analysis, Hybrid Features, Support Vector Machine.

I. INTRODUCTION

With the rise of e-commerce platforms and their wide variety of product offerings, online reviews have become a major source of information about the quality of said products for consumers. They heavily rely on these reviews before making a purchasing decision [1] which makes their credibility important for consumer trust and platform's reputation. However, due to online reviews playing a major role in the sales of these products, fake reviews are made to artificially boost or demote their impression on customers [2]. These fake reviews are written to resemble like a genuine review making them difficult to detect manually [3].

Recent machine learning approaches used to detect fake reviews use textual features which achieve high accuracy on specific datasets but often fail to generalize to new data sets due to vocabulary and writing style differences [4]. In this paper, a hybrid fake review detection framework is proposed. The method combines text representations in terms of TF-IDF with sentiment and behavioral meta-feature to enhance strength without being too heavy or incomprehensible [5].

The aim of this paper is to categorize an online product review into fake or genuine in terms of textual content and rating information attached to it. The fake reviews can either have exaggerated sentiment, unnatural pattern of writing, or lack coherence of rating and textual sentiment. The main issue is to differentiate between fake and genuine reviews which are linguistically similar and make sure that the detection model would be generalized to various datasets and fields. It is stated as a binary classification problem where the reviews should be classified as fake or genuine. The solution must not only be effective even as little auxiliary information is available, but it must also be highly generalizable to data sets having varying characteristics.

II. RELATED WORK

The methods of detecting fake review have been suggested in a number of approaches in the recent years. Initial research was mainly designed to analyses the text and take advantage of linguistic indicators, including the use of words, the degree of sentiment of the review, and syntactic features to differentiate between fake and authentic reviews. These techniques proved that counterfeit reviews might vary in the tone and format to the real reviews [3], [6]. The next work presented behavioral and metadata-related elements, such as the patterns of activity of reviewers, time bursts, and distribution of ratings. These capabilities offered more contextual information than textual content and contributed to detecting suspicious reviewing patterns that the text-only models were unable to detect [7]. Hybrid schemes that utilize textual data along with sentiment and behavioral indicators have been discussed in recent researches. These approaches are meant to enhance the accuracy along with the robustness of detection through the use of complementary information sources. Nevertheless, most of the currently available studies consider model evaluation on one dataset, and the problem of cross-dataset generalization remains uncovered. This paper expands on the hybrid feature-based methods and explicitly looks at model robustness when tested on another dataset [4], [8].

III. METHODOLOGY

A. System Workflow

The suggested system is multi-staged. The input is first raw review text and rated information. Text features are represented as TF-IDF and sentiment-based and behavioral Meta features are calculated simultaneously. These features are combined into a hybrid representation for classification

ISBN: 978-81-689089-0-1

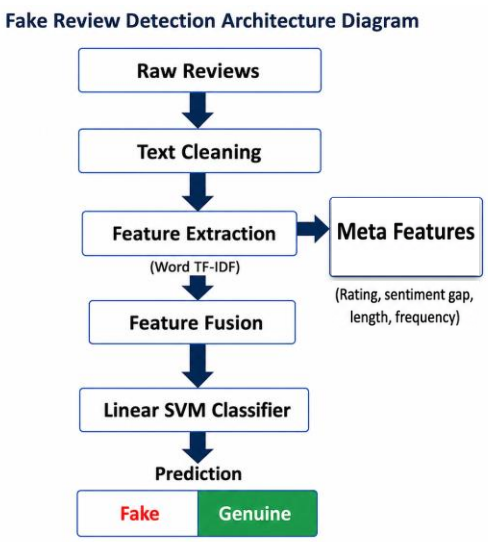

Fig. 1. System Architecture

The proposed framework works as the Fig. 1 shows that raw review text and rating information are processed by parallel feature extraction modules, and further processed by feature fusion along with final classification.

B. Text Cleaning

There is a minimum amount of preprocessing that maintains the stylistic and structural indicators which could be of value in identifying fake reviews. The reviews are translated to the text format, absent values are treated, and a simplistic normalization of the same is done. Intensive text cleaning is not done when the deceptive reviews could have unique writing patterns that can be used to classify them.

C. Textual Feature Representation

Term Frequency -Inverse Document Frequency (TF-IDF) representation is used to extract textual features. TF-IDF reflects the significance of the terms in a review as compared to the whole corpus [9]. Word-level and character-level TF-IDF are investigated. The word-level TF-IDF is sensitive to semantic content and the character-level TF-IDF is more efficient in picking up the stylistic patterns and can also withstand a vocabulary shift in different datasets [4], [10].

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right) \quad (1)$$

D. Meta Feature Engineering

In order to supplement textual features, a number of Meta features are designed. These are sentiment score which has been extracted in the review text, normalized rating, review length, number of sentences, and average word length. Moreover, the rating sentiment gap feature is also calculated so as to measure inconsistency between numerical rating and textual sentiment. These inconsistencies are typical to the false reviews, where inflated ratings cannot reflect the sentiment expressed [11].

$$\text{Gap} = |\,\text{Rating}_{\text{normalized}} - \text{Sentiment}_{\text{compound}}\,| \quad (2)$$

E. Feature Fusion Strategy

The textual TF-IDF vectors and the Meta features vectors are joined together to create a hybrid feature representation in a form of one. The combination of both linguistic patterns and behavioral cues in this fusion makes this model learn to detect better and be more robust.

The hybrid feature fusion enhances the performance of classification in that it incorporates semantic information in textual features with behavioral inconsistencies in Meta features. Whereas textual features characterize what is written, m E. Feature Fusion Strategy eta features characterize how it is written, and whether it correlates with rating behavioral. This combination improves robustness, particularly across different datasets.

F. Classification Model Selection

A number of classical machine learning classifiers have been tested, which include Naive Bayes, Logistic Regression and Linear Support Vector Machine (SVM). The reason behind the final hybrid model being linear SVM is that it is effective in high dimensional sparse feature space and it can incorporate heterogeneous feature types without a stringent independence assumption.

G. Algorithm Workflow

The general process of the proposed hybrid fake review detection system can be stated as follows:

- Step 1: Input raw review text and rating information
- Step 2: Preprocess it with cleaning, normalization and tokenization.
- Step 3: Extract textual features using TF-IDF (word and character level).
- Step 4: Extract meta features that include length of review, number of sentences, and rating-sentiment gap.
- Step 5: Fuse textual and meta features with feature fusion.
- Step 6: Linear SVM classifier: Train the hybrid feature set.
- Step 7: Estimate the validity of a review as a fake or authentic one.

IV. EXPERIMENTAL SETUP

A. Dataset Description

This study uses two datasets to determine in-dataset and cross-dataset performance. Dataset-1 is a balanced dataset with 8,087 reviews, whereas Dataset-2 has a bigger dataset with 40,431 reviews. These datasets allow to assess the model resilience to various data distributions [12].Both datasets used in this study are obtained from a public fake review dataset [13]. Dataset-1 is a cleaned and balanced subset which we use for training and in-dataset evaluation, while Dataset-2 is the larger version we used for cross dataset testing.

ISBN: 978-81-689089-0-1

TABLE I.
DATASET STATISTICS

Dataset	Total Reviews	Fake	Genuine
Dataset-1	8087	4,044	4,043
Dataset-2	40,431	20,216	20215

Cross-dataset testing and controlled in-dataset evaluation is also possible due to the use of two datasets. Although Dataset-1 aids the testing of the baseline performance, Dataset-2 can be used to test the generalization capacity of the model that is essential to the actual fake review detection systems. Deep learning and transformer-based fake review models have also been studied recently. Although these methods are highly performing, they tend to consume a lot of training data, computer power, and are incomprehensible. By contrast, the classical machine learning models with well-designed hybrid features provide the practical trade-off between the performances, efficiency, and explain ability, which is why they are useful in lightweight applications and the research applications [14].

B. *Experimental Environment*

The entire experimentation is done in Python and common machine learning packages, such as scikit learn pandas, and NumPy. The experiments are performed on a system based on a CPU, which further guarantees that the proposed practice is computational lightweight and reproducible.

C. *Evaluation Metrics*

Accuracy, precision, recall and F1-score are used to evaluate model performance. Accuracy is all that is correct in the general forecasts, whereas precision and recall are indicators of how the model is able to detect the fake reviews. F1-score presents a rooted evaluation, in that it offers a combination of precision and recall.

D. *Training Details*

The Linear SVM classifier was trained using a linear kernel with default hyper parameters. TF-IDF vectorization used n-gram range (1,2) and feature limitation for dimensionality control. The dataset was split using an 80:20 train-test ratio with stratified sampling. All experiments were conducted on CPU-based systems to ensure reproducibility.

V. RESULTS AND DISCUSSIONS

A. *In-Dataset Performance*

TABLE II
IN-DATASET PERFORMANCE ON DATASET-1

Model	Feature Type	Accuracy %	Precision	Recall	F1 Score
Naïve Bayes	TF-IDF	99	0.99	0.99	0.99
Logistic Regression	TF-IDF	99	0.99	0.99	0.99
Linear SVM	TF-IDF	99	0.99	0.99	0.99
Hybrid SVM	TF-IDF+META	99	0.99	0.99	0.99

Table II represents the in-dataset performance of different machine learning models evaluated on Dataset-1. Naive Bayes, Logistic Regression and Linear SVM models all have very high accuracy, precision, recall and F1-score. This implies that Dataset-1 is quite clean, balanced and has clear patterns that are easily learnt by text-based classifiers. The findings also indicate that TF-IDF attributes alone can be trained and tested to obtain close to perfect classification in both training and testing processes assuming the training and testing are carried out using a single dataset. Nonetheless, high performance in this manner must be taken with caution, in that it is not always an indicator of effectiveness in the real world. Models can receive data specific vocabulary, styles of writing or review patterns, which causes misleading optimism results. Therefore, while Table II confirms that the implemented models are correctly trained and functional, it does not provide sufficient evidence of robustness or generalization capability.

B. *Cross-Dataset Performance*

TABLE III.
CROSS-DATASET PERFORMANCE

Model	Feature Type	Accuracy %	Precision	Recall	F1 score
Naïve Bayes	Word TF-IDF	52	0.73	0.53	0.39
Logistic Regression	Word TF-IDF	52	0.74	0.51	0.35
Linear SVM	Word TF-IDF	53	0.75	0.52	0.38
Fine-Tuned SVM	Char TF-IDF	92	0.92	0.92	0.92

ISBN: 978-81-689089-0-1

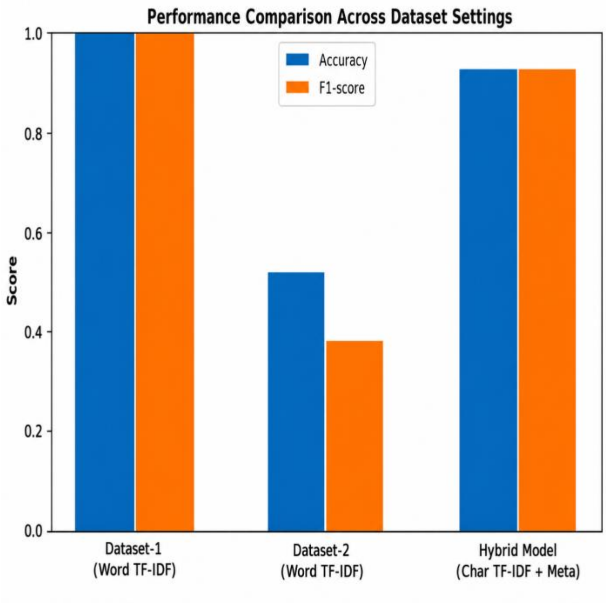

Fig. 2. Comparison of the in-dataset, cross-dataset and hybrid model performance based on Accuracy & F1-score.

As Fig. 2 shows, at the word-level TF-IDF models demonstrate almost the same level of accuracy in Dataset-1, their accuracy significantly decreases on Dataset-2. The hybrid model proposed can overcome this performance gap by a significant margin and has a better robustness.

Table III documents the cross-dataset performance with the models being trained using Dataset-1 and tested using Dataset-2. This assessment environment is more realistic and difficult because it enables the models to be exposed to unknown data using various vocabulary, writing styles, and review patterns.

The findings indicate that there is a major decline in performance of word-level TF-IDF models in all the classifiers. The value of accuracy that is nearly similar to random guessing would imply that text-only models are not effective in making generalizations outside the data they have been trained on. This presents an important weakness of using word-level textual features of fake reviews to detect them. Conversely, a character-level TF-IDF model that is trained with Linear SVM has significantly greater accuracy. Character-level representations are less sensitive to shifts in vocabulary and they are better in capturing stylistic patterns used in fake reviews. Consequently, the suggested method shows enhanced strength and extrapolation across sets of data.

The superior performance of the hybrid model can be attributed to the integration of textual and behavioral features. While textual features capture semantic patterns, Meta features such as rating–sentiment gap capture inconsistencies commonly found in fake reviews. This combination allows the model to detect deceptive patterns more effectively and improves generalization across datasets.

The results were consistent across multiple experimental runs, indicating stability and reliability of the proposed model.

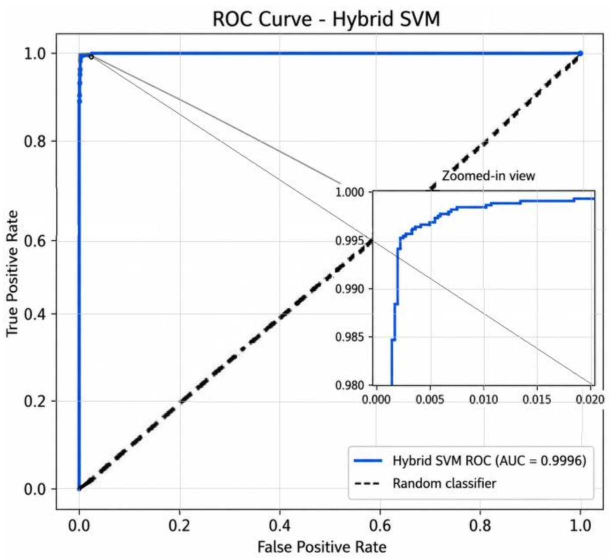

Fig.3. ROC Curve of Hybrid SVM Model showing strong classification performance with high AUC.

As shown in Figure 3, the Receiver Operating Characteristic (ROC) curve is used to illustrate the trade-off between the true positive rate and false positive rate. The high Area under Curve (AUC) value indicates strong classification performance and demonstrates the effectiveness of the proposed hybrid model.

Cross-dataset analysis shows the restriction of text-only models and how hybrid and character-level representations can be very important in real-life performance.

Practically, cross-dataset strength is a necessary condition of the real-world implementation, because e-commerce platforms are constantly being reviewed on the products and by different users. Incorrect moderation decisions and the lack of user trust can be caused by the usage of models that do not generalize. The provided hybrid solution will resolve this issue by incorporating textual and behavioral signals.

VI. CONCLUSION AND FUTURE SCOPE

This paper proposed a hybrid machine learning model to detect fake reviews that incorporates both textual TF-IDF based features and sentiment-based and behavioral Meta features. The proposed hybrid model achieved 99% accuracy on Dataset-1 and 92% accuracy on Dataset-2, demonstrating strong robustness and real-world applicability.

The further development of work can be the implementation of more comprehensive user-level behavioral capabilities, explainable artificial intelligence methods, and multilingual fake review detection to enhance robustness and applicability.

ISBN: 978-81-689089-0-1

REFERENCES

[1] Y. Liu, J. Huang, and A. An, “Mining opinion influence in online reviews,” IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 6, pp. 2802–2815, 2022.

[2] Tripadvisor, “Review Transparency Report,” 2024.

[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in Proc. ACL, 2011.

[4] S. Panda, A. Ray, and M. Islam, “Cross-domain deception detection using hybrid linguistic representations,” ACM Transactions on Intelligent Systems and Technology, vol. 13, no. 4, 2022.

[5] A. Alshehri and M. Alsubaie, “Hybrid machine learning approach for fake review detection using linguistic and behavioural features,” Expert Systems with Applications, vol. 211, 2023.

[6] N. Jindal and B. Liu, “Opinion spam and analysis,” in Proc. WSDM, 2008.

[7] G. Fei, Z. Mukherjee, B. Liu, M. Hsu, and R. Ghosh, “Exploiting burstiness in reviews for review spam detection,” in Proc. ICWSM, 2013.

[8] M. Ennaouri, Y. Sayah and A. K. Tiwari, Machine learning methods for fake review detection: A systematic literature review, Journal of Web Engineering, vol. 22, no. 3, pp. 1021-1056, 2023.

[9] A. Sharma and R. Kumar, Opinion spam detection by means of TF-IDF and classical machine learning classifiers, Intelligent Systems with Applications, vol. 18, 2023.

[10] M. Alsubaie, A. Alshehri and S. Alqarni, "Robust fake review detection using character-level representations" IEEE Access vol. 11, pp. 125489-125501, 2023.

[11] R. Singh, S. Verma, and P. K. Gupta, "Explainable fake review detection based on sentiment inconsistency features," IEEE Access, vol. 10, pp. 118344-118356, 2022.

[12] R. Gupta, A. Mehta, and P. Kumar, Cross-dataset fake review detection: Challenges and evaluation Knowledge-Based Systems, vol. 284, 2024.

[13] Fake Reviews Dataset, Kaggle [online]. https://www.kaggle.com/datasets/harshdecipher/amazon-enhanced-fake-reviews-datasets as of 30 January 2026.

[14] J. Hou, Y. Zhang and H. Liu, "Transformer-based deceptive opinion and spam detection," Information Fusion, vol. 95, pp. 34-47, 2024.

Cite this article as:

Divya, Aman and et. al. "A Hybrid Machine Learning Approach for Fake Review Detection in E-Commerce", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026 (Vol-II)

Displayed as online on 18th July 2026.

Link: http://actsoft.org/science/micro2026-pro/308-micro2026.pdf

ISBN: 978-81-689089-0-1