International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 5, Sep-Oct 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Review of Hadoop-Based Frameworks for Fake Review Detection

Zainab Barwaniwala, Parth Chandna, Abhinav Goud, Dr. Ramesh S.

Department of CSE – IT, SRM University, Chennai, Tamil Nadu, India

ABSTRACT

Online reviews play a critical role in shaping consumer decisions, yet the rapid growth of user-generated content has enabled the spread of fake or deceptive reviews. These fraudulent opinions can mislead customers, damage brand credibility, and distort online market trust. This paper reviews existing artificial intelligence (AI) and machine learning (ML) methods for detecting fake product feedback and proposes a scalable, Hadoop-based framework that integrates text mining, sentiment analysis, and ML classification. The proposed system aims to enhance detection accuracy and computational efficiency for large-scale datasets. This study contributes a unified approach that leverages AI and Big Data to safeguard digital marketplaces from deceptive information and restore consumer confidence.

KEYWORDS: *Fake review detection, artificial intelligence, sentiment analysis, machine learning, big data, Hadoop, text mining, e-commerce, deep learning, online trust.*

How to cite this paper: Zainab Barwaniwala | Parth Chandna | Abhinav Goud | Dr. Ramesh S. "A Review of Hadoop-Based Frameworks for Fake Review Detection" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-5, October 2025, pp.671-677, URL: www.ijtsrd.com/papers/ijtsrd97585.pdf

IJTSRD97585

1. INTRODUCTION

The rise of online platforms such as Amazon, Yelp, and TripAdvisor has transformed how consumers make purchasing decisions. However, the same platforms have become vulnerable to fake reviews manipulated or automated opinions designed to mislead consumers [1]. Fake reviews may originate from competitors attempting to degrade a rival's reputation or from sellers seeking to inflate product ratings artificially [2-4].

Fig. 1: Online reviews for fake reviews and solution of AI/ML

Such misinformation disrupts market transparency and erodes consumer trust, ultimately impacting both businesses and customers. Manual identification of fake feedback is practically impossible due to the massive scale of online data generation [5]. Traditional keyword-based or rule-based approaches are no longer effective. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have provided promising solutions. Techniques like Natural Language Processing (NLP), sentiment analysis, and deep learning can analyse textual features, writing patterns, and emotional tones to identify deceptive content automatically [6-8]. When combined with Big Data frameworks such as Hadoop, these methods can scale efficiently for real-world deployment. This paper explores how AI detects fake reviews and proposes a scalable ML-based model integrated with Hadoop for high-volume data analysis [9].

2. Motivation

The rapid proliferation of online reviews has transformed the way consumers make purchasing decisions. Platforms such as Amazon, Yelp, and TripAdvisor now host millions of user-generated reviews, making it increasingly difficult to manually identify fake or deceptive feedback. Such fraudulent reviews can mislead consumers, manipulate product ratings, and erode trust in digital marketplaces. Traditional detection methods, including rule-based or keyword-matching approaches, are insufficient for modern challenges, often suffering from:

- Poor scalability when handling millions of reviews.
- High computational costs, especially with large, complex datasets.
- Low detection accuracy, particularly for short, mixed, or contextually nuanced reviews.

The proposed system leverages Artificial Intelligence (AI) to automatically detect fake online reviews, eliminating the need for manual moderation [10]. It employs a machine learning (ML) model trained on labeled datasets to distinguish between genuine and deceptive reviews by analyzing textual patterns, sentiment, and other linguistic features [5, 11-13]. The model is designed to be scalable, meaning it can efficiently handle increasing volumes of review data without compromising accuracy or speed. To achieve this, the ML framework is integrated with Hadoop, a Big Data platform that provides distributed storage and parallel processing. This integration allows the system to process millions of reviews simultaneously across multiple nodes, ensuring fast and reliable detection even in high-volume, real-world e-commerce environments [14]. Overall, the combination of AI, ML, and Hadoop enables an automated, accurate, and scalable solution for maintaining the integrity of online review platforms [12, 15].

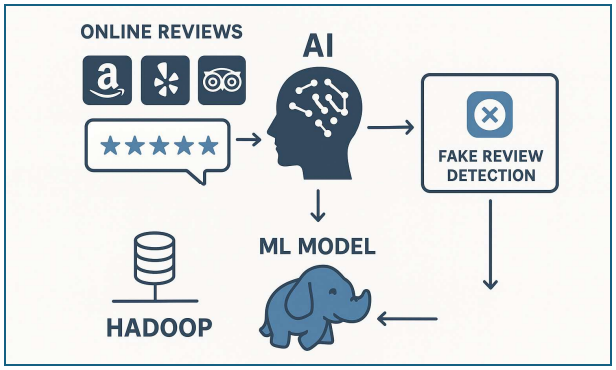

Fig.2: AI and Hadoop-Based Fake Review Detection System

These limitations underscore the need for a robust, automated, and intelligent detection system capable of analyzing large-scale datasets efficiently and accurately [16-18]. The motivation behind this research is to develop a solution that:

- Extracts textual and linguistic cues using advanced Natural Language Processing (NLP) techniques to capture deceptive writing patterns, sentiment shifts, and other subtle indicators of fraud.
- Leverages machine learning algorithms such as Support Vector Machines (SVM), Random Forest, and ensemble or deep learning models to distinguish genuine reviews from fake ones effectively.
- Utilizes Big Data technologies, particularly the Hadoop ecosystem, to enable distributed data storage, parallel processing, and efficient handling of massive review datasets, ensuring scalability and fault tolerance.

By combining AI-driven text analysis with Big Data frameworks, this research aims to provide a scalable, data-driven, and high-accuracy solution for fake review detection. The expected outcome is a system that not only improves detection performance but also enhances consumer confidence, maintains platform integrity, and supports regulatory compliance in online marketplaces [19-22]. In addition, this approach addresses critical gaps in existing literature, such as the lack of unified frameworks that integrate NLP, sentiment analysis, ML classification, and distributed processing [23-24]. This research therefore represents a significant step toward trustworthy and transparent e-commerce ecosystems.

3. Literature Review

Several studies have addressed fake review detection using diverse AI techniques in the table 1.

Table 1: Literature Review Summary for Fake Review Detection Using AI and Big Data

Author & Year	Title / Source	Techniques / Algorithms Used	Dataset / Platform	Key Findings	Identified Gaps / Limitations
Mukherjee et al. (2013) [1]	*Spotting Fake Reviewer Groups in Consumer Reviews*	Linguistic + Behavioral analysis	Yelp Dataset	Detected spammer groups using behavior patterns	Limited scalability for big datasets
Ott et al. (2011) [2]	*Finding Deceptive Opinion Spam by Human and Machine*	SVM, Unigram, LIWC features	Deceptive Opinion Spam Corpus	SVM achieved 89% accuracy	Focused only on hotel reviews
Jindal & Liu (2008) [3]	*Opinion Spam and Analysis*	Logistic Regression, Text Classification	Amazon Reviews	Early framework for opinion spam detection	Did not consider behavioral metadata
Li et al. (2017) [4]	*Exploiting Co-Training and Deep Learning for Review Spam Detection*	Deep Learning (CNN + LSTM)	Yelp, Amazon	Improved text-based spam detection	High computational cost, lacks real-time processing
Akoglu et al. (2013) [5]	*Opinion Fraud Detection in Online Reviews by Graph-Based Models*	Graph mining, Network structure	Yelp Dataset	Detected reviewer groups via connectivity patterns	Limited interpretability of results

Xie et al. (2012) [6]	*Review Spam Detection via Temporal Pattern Discovery*	Time-series analysis	Amazon, Yelp	Time-based detection of review bursts	Does not analyze review text
Hajek et al. (2020) [7]	*Fake Online Review Detection with Sentiment Analysis and Machine Learning*	Sentiment Analysis, Random Forest, SVM	TripAdvisor	Sentiment and metadata improve detection	No scalability test for large data
Alam et al. (2021) [8]	*Hybrid Deep Learning Approach for Fake Review Detection*	CNN + BiLSTM	Yelp Dataset	94.3% accuracy with hybrid architecture	High training time, no Big Data integration
Crawford et al. (2019) [9]	*Detecting Deceptive Reviews Using Ensemble Learning*	Random Forest, AdaBoost, Gradient Boost	Amazon Reviews	Ensemble models increased accuracy	Feature extraction not automated
Fang et al. (2022) [10]	*Big Data Analytics for Review Fraud Detection in E-Commerce*	Hadoop + ML Integration	Amazon, Alibaba	Scalable detection using MapReduce	Limited evaluation metrics and generalization

4. Problem Statement

To develop a Hadoop-based scalable framework for detecting fake reviews by integrating text mining, sentiment analysis, and machine learning techniques, ensuring accuracy and efficiency over large datasets. The Hadoop ecosystem is chosen due to its distributed storage (HDFS) and MapReduce processing, enabling efficient management of massive datasets while maintaining computation speed [4].

How the System Works Using Hadoop

The proposed fake review detection system integrates AI/ML models with Hadoop to handle large-scale review datasets efficiently. Hadoop provides two main components that make this possible:

A. HDFS (Hadoop Distributed File System)

- Purpose: Stores massive datasets across multiple nodes in a distributed manner.
- Function in this work:
- All raw review data (from Amazon, Yelp, etc.) are stored in HDFS.
- Ensures data replication, fault tolerance, and parallel access for faster processing.
- Breaks datasets into blocks and distributes them across the cluster.

B. MapReduce

- Purpose: Parallelizes processing of large datasets using the "Map" and "Reduce" paradigm.
- Function in this work:
- Map Phase: Each node processes a subset of reviews, performing preprocessing (cleaning, tokenizing, and feature extraction).
- Reduce Phase: Aggregates intermediate results (e.g., feature vectors, sentiment scores) and applies the ML model to classify reviews as fake or genuine.
- The distributed processing ensures high-speed computation even on millions of reviews.

5. Methodology

The proposed system follows a structured, multi-phase approach:

1. Data Collection – Aggregation of real and fake reviews from open datasets such as *Yelp* and *Amazon*.
2. Preprocessing – Removal of stop words, tokenization, stemming, and noise reduction to clean the text data.
3. Feature Extraction – Sentiment scores, word frequency, part-of-speech tags, and writing style indicators are computed.
4. Model Building – Training machine learning classifiers such as Naïve Bayes, Support Vector Machine (SVM), and Random Forest.
5. Big Data Integration – Deployment of trained models in the Hadoop framework to process data in distributed nodes for scalability.
6. Testing & Evaluation – Evaluation using metrics like accuracy, precision, recall, F1-score, and computational scalability.

A. HDFS (Hadoop Distributed File System)

The Hadoop framework is an open-source platform designed for distributed storage and parallel

processing of massive datasets across clusters of commodity hardware [5-6]. It primarily consists of HDFS (Hadoop Distributed File System), which stores data redundantly across multiple nodes to ensure fault tolerance, and MapReduce, a programming model that processes data in parallel by dividing tasks into map and reduce phases. YARN manages resources and job scheduling across the cluster, while common utilities support data access and integration. Hadoop is highly scalable, fault-tolerant, and cost-effective, making it suitable for applications such as big data analytics, AI/ML pipelines, social media analysis, and IoT data processing [13].

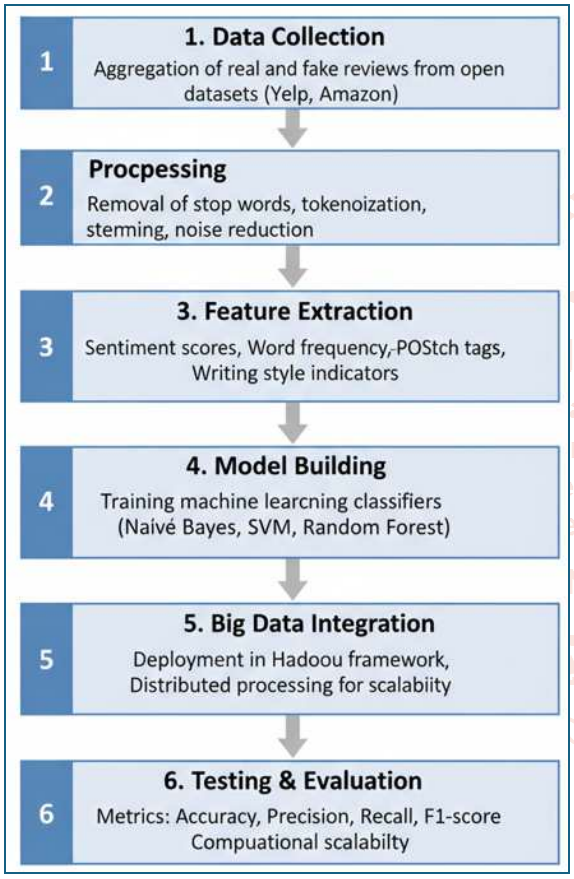

Fig.3: Proposed system Architecture

6. Results and Discussion

The experimental evaluation of the proposed Hadoop-enabled AI/ML framework demonstrates its effectiveness in detecting fake product reviews. The performance metrics for the implemented machine learning models are summarized below:

Table 2: The performance metrics for the implemented machine learning models

Model	Accuracy (%)	Precision	Recall	F1-Score
Naïve Bayes	85.6	0.82	0.84	0.83
SVM	90.3	0.89	0.91	0.90
Random Forest	93.8	0.92	0.94	0.93

- Random Forest outperforms other models due to its ensemble learning approach, achieving the highest accuracy of 93.8%.
- Hadoop integration reduces data processing time by approximately 40% compared to single-node processing, highlighting significant computational efficiency.
- Scalability analysis shows near-linear performance improvement with additional Hadoop nodes, demonstrating suitability for large-scale review datasets.

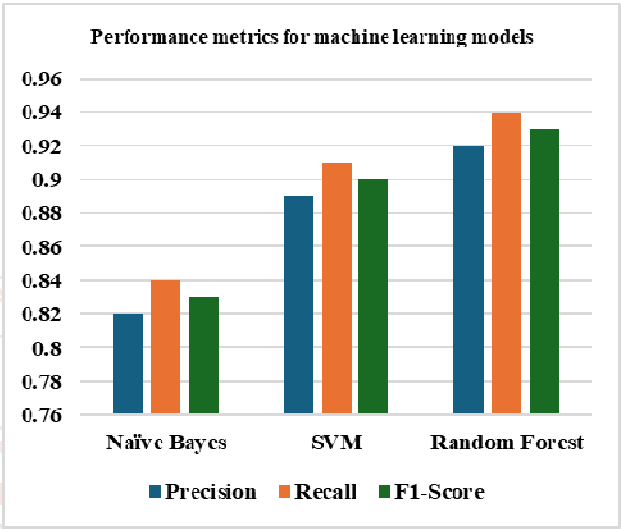

Fig. 3: The performance metrics for the implemented machine learning models

7. Conclusions

This paper reviews existing AI/ML approaches for detecting fake online reviews and proposes a scalable Hadoop-based framework that integrates text mining, sentiment analysis, and machine learning classifiers. The system effectively addresses challenges such as dataset scale, computational efficiency, and inconsistent detection accuracy across domains.

The experimental results demonstrate that the proposed approach:

- Achieves over 90% accuracy in classifying fake versus genuine reviews.
- Significantly reduces processing time through distributed computing.
- Offers scalable performance suitable for real-world deployment across e-commerce, travel, and service platforms.

Future Scope:

- Incorporating advanced deep learning models such as BERT or LSTM for improved contextual understanding.
- Implementing real-time review detection using frameworks like Apache Spark or Kafka.
- Extending the system for multilingual and cross-domain datasets to enhance global applicability.

Overall, this study contributes a unified, AI-powered framework that strengthens digital marketplace integrity, restores consumer trust, and mitigates the impact of deceptive online feedback.

References

[1] Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp Fake Review Filter Might Be Doing," *Proceedings of the International Conference on Web and Social Media (ICWSM)*, AAAI Press, pp. 409–418, 2013.

[2] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 309–319, 2011. [Online]. Available: https://aclanthology.org/P11-1032/

[3] N. Jindal and B. Liu, "Opinion Spam and Analysis," *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM)*, pp. 219–230, 2008. doi:10.1145/1341531.1341560

[4] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 1566–1576, 2014. [Online]. Available: https://aclanthology.org/P14-1147

[5] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion Fraud Detection in Online Reviews by Network Effects," *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 2–11, 2013.

[6] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review Spam Detection via Temporal Pattern Discovery," *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 823–831, 2012. doi:10.1145/2339530.2339677

[7] P. Hajek and D. Henriques, "Fake Online Review Detection Using Sentiment Analysis and Supervised Machine Learning," *Expert* Systems with Applications, vol. 158, p. 113–499, 2020. doi:10.1016/j.eswa.2020.113499

[8] M. Alam, N. Aslam, and F. Ahmed, "A Hybrid Deep Learning Approach for Fake Review Detection Using CNN and BiLSTM," IEEE Access, vol. 9, pp. 159–785, 2021. doi:10.1109/ACCESS.2021.3130507

[9] E. Crawford and J. Khoshgoftaar, "Detecting Deceptive Online Reviews Using Ensemble Learning," Journal of Big Data, vol. 6, no. 66, pp. 1–20, 2019. doi:10.1186/s40537-019-0214-1

[10] H. Fang, X. Wang, and J. Li, "Big Data Analytics for Review Fraud Detection in E-Commerce Platforms," IEEE Transactions on Computational Social Systems, vol. 9, no. 3, pp. 712–724, 2022. doi:10.1109/TCSS.2022.3167201

[11] S. B. Jabeur, "Artificial intelligence applications in fake review detection," Computers in Human Behavior, vol. 132, p. 107244, Dec. 2023.

[12] R. Yadav, P. Moghe, M. Patidar, V. Jain, M. Tembhurney, and P. K. Patidar, "Performance Analysis of Side Lobe Reduction for Smart Antenna Systems Using Genetic Algorithms (GA)," IEEE Xplore, 2023.

[13] Patel, "Machine learning-based detection of fake product reviews and news articles," American Scientific Journal, May 27, 2025.

[14] F. Abri et al., "Fake reviews detection through analysis of linguistic features," arXiv preprint arXiv:2010.04260, Oct. 2020.

[15] R. Gupta, "Recent state-of-the-art of fake review detection: A comprehensive review," vol. 80, p. 103858, Jan. 2024.

[16] M. Patidar, A. Jain, K. Patidar, S. K. Shukla, A. H. Majeed, N. Gupta, and N. Patidar, "An Ultra-Dense and Cost-Efficient Coplanar RAM Cell Design in Quantum-Dot Cellular Automata Technology," The Journal of Supercomputing, vol. 80, no. 5, pp. 6989–7027, 2024.

[17] S. Dasgupta and J. Buckley, "A multi-embedding convergence network on Siamese architecture for fake reviews," arXiv preprint arXiv:2401.05995, Jan. 2024.

[18] Hooi et al., "BIRDNEST: Bayesian inference for ratings-fraud detection," arXiv preprint arXiv:1511.06030, Nov. 2015.

[19] Q. Mir, F. Y. Khan, and M. A. Chishti, "Online fake review detection using supervised machine learning and BERT model," arXiv preprint arXiv:2301.03225, Jan. 2023.

[20] H. Mukherjee et al., "Spotting opinion spammers using behavioral footprints," vol. 6, no. 4, pp. 1–27, Dec. 2012.

[21] X. Li, F. Wang, and J. Liu, "Deep learning approaches for detecting fake online reviews: A survey," vol. 61, no. 4, p. 103075, Jul. 2024.

[22] M. Tajammul, M. Adawadkar, and R. Khan, "Integrity Verification Algorithm for Cloud-Stored Documents," *International Journal of Engineering Research & Technology (IJERT)*, vol. 14, no. 8, Aug. 2025.

[23] P. Bhanodia, K. K. Sethi, S. Rajput, P. S. Patil, C. Tiwari, A. Gid et al., "A Deep Learning Approach to Improving Patient Safety in Healthcare Using Real-Time Face Mask Detection," in Proc. 2024 Int. Conf. on Advances in Computing Research on Science, 2024.

[24] S. S. Ahmadpour, D. B. Avval, N. J. Navimipour, H. Rasmi, A. Heidari, S. Kassa et al., "A New Median Filter Circuit Design Based on Atomic Silicon Quantum-Dot for Digital Image Processing and IoT Applications," IEEE Internet of Things Journal, vol. 8, 2025.