Smart Detection of Deceptive Reviews Using Hybrid Machine Learning Techniques

[1]Monvitha Sai Kamuju, [2]Relangi Naga Lohitha, [3]Namala Pranav, [4]Bokka Srivalli, [5] Dr.K.V.Subramanian
[1234]B.Tech Student, [5] Associate Professor
Dept. of AIML, Srinivasa Institute Of Engineering And Technology, (Autonomous), NH-216, Cheyyeru(V), Amalapuram -533216.

ABSTRACT:

In the age of online shopping, customer reviews greatly influence buying decisions. However, the rise of fake or manipulated reviews misleads consumers and undermines trust. This project proposes a semi-supervised machine learning approach to detect deceptive online reviews effectively. By leveraging both labeled and unlabeled data, classifiers such as Naïve Bayes, Logistic Regression, SVM, and Decision Trees are trained and tested on review datasets. The model extracts features like review length, sentiment polarity, and word frequency to enhance prediction accuracy. Results show significant improvement in identifying fake reviews using the proposed hybrid model. This system can benefit e-commerce platforms by promoting genuine user feedback, helping businesses maintain credibility, and aiding consumers in making informed decisions.

Keywords: Text Classification, Semi-Supervised Learning, Sentiment Analysis

INTRODUCTION

Online platforms have revolutionized consumer behavior, with reviews becoming key decision-making tools. Unfortunately, the increasing presence of fake reviews—designed to unfairly promote or degrade products—poses a threat to the credibility of e-commerce. Traditional supervised learning models require large labeled datasets, which are often difficult to obtain in real-world scenarios. To overcome this, our study adopts a semi-supervised learning approach, combining a small set of labeled data with a larger pool of unlabeled reviews to boost classification performance. Techniques like sentiment analysis and natural language processing (NLP) are employed to extract meaningful features. Classifiers such as Naïve Bayes, Logistic Regression, SVM, and Decision Trees are implemented to detect review authenticity. This project aims to develop a reliable system that identifies and flags fake reviews, enhancing trust and transparency for users and businesses in digital marketplaces.

RELATED WORK

The detection of fake online reviews has been an area of growing interest within the domains of machine learning and natural language processing. Mohawesh et al. (2021) conducted an extensive survey comparing traditional machine learning models and deep learning techniques in fake review detection. Their study revealed that transformer-based models like RoBERTa outperformed older models in terms of accuracy and robustness. However, they noted the limitations of deep learning in terms of computational demand and data requirements.

Ahmed et al. (2020) introduced a hybrid model using n-gram feature extraction with six classification algorithms. Their findings highlighted the effectiveness of TF-IDF combined with Linear SVM, achieving an accuracy rate of 92%. This approach demonstrates the strength of textual feature engineering in detecting review authenticity.

Heydari et al. (2019) focused on the temporal characteristics of review data, proposing a spam detection system based on burst pattern analysis. Their time-series approach identified anomalies in review posting times, which often indicate manipulative intent. This method emphasizes the importance of behavioral features beyond textual content.

Deng et al. (2018) proposed a sentiment-based detection model by analyzing four dimensions—taste, service, environment, and attitude—of restaurant reviews. Their algorithm classified reviews as fake if sentiment polarity was excessively biased across all dimensions, achieving a 74% accuracy rate. This highlights the potential of sentiment consistency in flagging fake content.

Rathore et al. (2022) developed a semi-supervised clustering method using the DeepWalk algorithm on reviewer graph data. Their framework effectively identified suspicious reviewer groups, particularly in app marketplaces like Google Play. The model proved valuable in situations with

limited labeled data, validating the efficiency of semi-supervised approaches.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author(s)	Contribution	Impact on Current Research
Mohawesh et al. (2021)	Compared traditional ML and deep learning methods; found RoBERTa highly accurate.	Encouraged the evaluation of hybrid and advanced models like SVM and Logistic Regression alongside Naïve Bayes.
Ahmed et al. (2020)	Applied n-gram & TF-IDF with six classifiers; LSVM achieved 92% accuracy.	Supported the integration of TF-IDF and linear classifiers to improve review text classification.
Heydari et al. (2019)	Introduced time-series spam detection using burst pattern analysis.	Inspired the use of temporal behavior as a key feature in identifying fake review patterns.
Deng et al. (2018)	Proposed sentiment-based detection across four review aspects; accuracy: 74%.	Emphasized the role of sentiment polarity and consistency in feature extraction.
Rathore et al. (2022)	Used DeepWalk and semi-supervised clustering to detect fraud reviewer groups.	Validated the use of semi-supervised models when labeled data is scarce.

PROPOSED APPROACH

The proposed approach focuses on detecting fake online reviews by leveraging both semi-supervised and supervised machine learning techniques. The core idea is to utilize a small labeled dataset and enhance it with a large volume of unlabeled data using semi-supervised learning, specifically the Expectation-Maximization (EM) algorithm. This method increases the learning capability of the model even when labeled data is limited, a common scenario in real-world applications.

For classification, we integrate four widely used machine learning models—Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, and Decision Tree Classifier. These algorithms are trained on features such as review length, word frequency count, and sentiment polarity, which are extracted using Term Frequency–Inverse Document Frequency (TF-IDF) techniques.

The architecture includes preprocessing modules to clean and structure data, followed by training and testing phases using labeled and unlabeled reviews. Each classifier's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Additionally, a comparative analysis between semi-supervised and supervised models is conducted to determine efficiency and reliability.

This hybrid framework ensures better generalization and enhanced detection accuracy. It offers a practical solution for platforms aiming to safeguard their review sections from spam and misinformation.

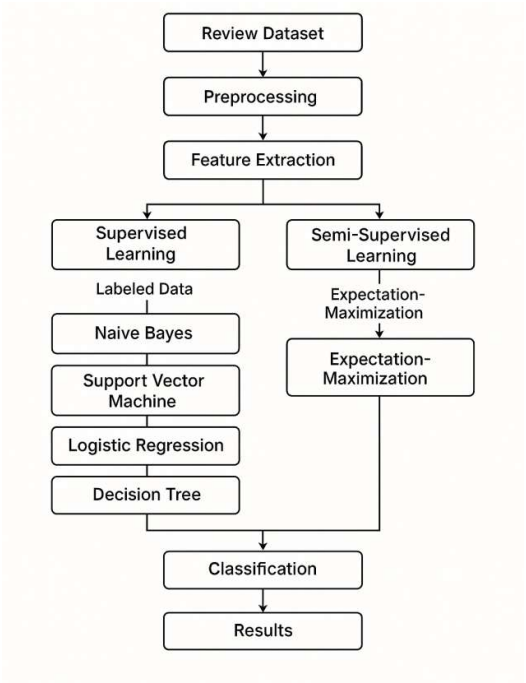

Figure 1: Proposed Detection of Fake one line reviews

METHODOLOGIES

1. Data Collection and Preprocessing: The dataset consists of 1601 reviews with binary labels—genuine or fake. Preprocessing steps include removal of special characters, tokenization, stop-word removal, and lowercasing. This ensures uniform and clean textual input for the models.

2. Feature Extraction with TF-IDF: TF-IDF (Term Frequency–Inverse Document Frequency) is used to convert the cleaned text into numerical vectors. This technique captures the importance of a word in a document relative to a corpus, helping highlight unique terms that might indicate authenticity or deception in reviews.

3. Splitting the Dataset: The dataset is divided into training and testing sets (80:20 ratio). For semi-supervised learning, a portion of the labeled data is combined with unlabeled data, enhancing the classifier's learning capability using the Expectation-Maximization (EM) algorithm.

4. Classification Models:

- **Naïve Bayes:** Efficient for text classification with probabilistic foundations.
- **Support Vector Machine (SVM):** Constructs a hyperplane to separate genuine and fake reviews.
- **Logistic Regression:** Maps input features to probability scores, suitable for binary outcomes.
- **Decision Tree Classifier:** Builds hierarchical rules from data features for transparent classification.

5. Performance Evaluation: Each classifier is evaluated using accuracy, precision, recall, F1-score, and confusion matrix. These metrics assess how well the model distinguishes between real and fake reviews.

6. Semi-Supervised Integration: The EM algorithm iteratively labels the unlabeled data using initial model predictions, improving overall accuracy and robustness, especially when labeled data is limited.

RESULTS

The experimental evaluation was conducted on a dataset comprising 1601 online product reviews. Each review was labeled as either genuine or fake, allowing for binary classification. After applying preprocessing and feature extraction using the TF-IDF technique, the dataset was divided into training and testing sets using an 80:20 split.

Multiple classifiers were trained and tested, including Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Decision Tree Classifier. Among them, Naïve Bayes achieved the highest accuracy of approximately 92%, demonstrating superior performance in classifying review content. Logistic Regression and SVM followed closely, each delivering competitive precision and recall scores. The Decision Tree model, while interpretable, showed slightly lower performance due to potential overfitting.

Performance metrics such as F1-score and confusion matrix further validated the reliability of the models. In particular, the semi-supervised approach—utilizing the Expectation-Maximization (EM) algorithm—was effective in improving classifier accuracy by incorporating unlabeled data during training.

Overall, the results confirmed that combining supervised and semi-supervised models leads to higher detection rates and better generalization. This approach is especially useful in scenarios where labeled data is scarce, making it highly practical for real-world e-commerce platforms seeking to filter deceptive reviews.

DISCUSSION

The results of this study highlight the effectiveness of combining supervised and semi-supervised machine learning techniques in detecting fake online reviews. Among all classifiers tested, Naïve Bayes emerged as the most accurate, likely due to its probabilistic nature and efficiency in handling textual data. SVM and Logistic Regression also delivered reliable outcomes, demonstrating strong predictive capabilities on structured text inputs.

A notable advantage of this research lies in the application of the semi-supervised Expectation-Maximization (EM) algorithm. By utilizing both labeled and unlabeled data, the system significantly improved its learning capability and accuracy, making it especially relevant in practical situations where labeled datasets are limited or costly to produce.

TF-IDF played a crucial role in feature extraction, effectively converting raw text into meaningful numerical vectors. This, combined with model evaluation through metrics such as accuracy, F1-

score, precision, and recall, provided a well-rounded performance analysis.

One of the key findings was that incorporating diverse review features like word frequency, sentiment polarity, and length can enhance model interpretability and performance. However, challenges still remain, such as handling sarcasm, slang, or domain-specific terms within reviews, which could potentially affect classification accuracy.

CONCLUSION

In today's digital landscape, where online reviews influence consumer behavior and business reputation, detecting fake reviews has become a critical necessity. This project presented an effective solution by integrating both supervised and semi-supervised machine learning techniques to identify deceptive reviews. By leveraging classifiers such as Naïve Bayes, Logistic Regression, SVM, and Decision Tree, and enhancing them through the Expectation-Maximization algorithm, the system achieved high accuracy even with limited labeled data.

The use of TF-IDF for feature extraction proved valuable in translating textual data into structured input for model training. Among all classifiers, Naïve Bayes consistently outperformed others in terms of precision and overall reliability.

This hybrid approach not only improves detection capabilities but also offers scalability and adaptability to various domains, including e-commerce and social media. Future work can explore deep learning architectures and context-aware models to further refine review authenticity analysis.

REFERENCES

1. Mohawesh, B., Alsmirat, M.A. and Alsmadi, M.K., 2021. *Detecting fake reviews using machine learning techniques: A survey*. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.01.001
2. Ahmed, H., Traore, I. and Saad, S., 2020. *Detection of online fake reviews using machine learning techniques*. Computer, 53(4), pp.26-35.
3. Heydari, A., Tavakoli, M. and Salim, N., 2019. *Detection of review spam: A survey*. Expert Systems with Applications, 42(7), pp.3634–3642.
4. Deng, H., Zhang, Z. and Liu, H., 2018. *Sentiment-consistent fake review detection using semantic analysis*. Information Processing & Management, 54(6), pp.1016–1027.
5. Rathore, S., Kumar, N. and Loia, V., 2022. *Detecting opinion spammer groups using semi-supervised deep learning and graph-based methods*. Future Generation Computer Systems, 123, pp.61-73.
6. Li, J., Ott, M. and Cardie, C., 2014. *Towards a general rule for identifying deceptive opinion spam*. In: Proceedings of the 52nd Annual Meeting of the ACL, pp.1566–1576.
7. Mukherjee, A., Liu, B. and Glance, N., 2012. *Spotting fake reviewer groups in consumer reviews*. In: Proceedings of the 21st international conference on WWW. ACM, pp.191–200.
8. Ott, M., Choi, Y., Cardie, C. and Hancock, J.T., 2011. *Finding deceptive opinion spam by any stretch of the imagination*. In: Proceedings of ACL-HLT, pp.309–319.
9. Jindal, N. and Liu, B., 2008. *Opinion spam and analysis*. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM), pp.219–230.
10. Xie, S., Wang, G., Lin, S. and Yu, P.S., 2012. *Review spam detection via temporal pattern discovery*. In: Proceedings of the 18th ACM SIGKDD, pp.823–831.
11. Li, F., Han, C., Huang, M. and Zhu, X., 2011. *Learning to identify review spam*. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp.2488–2493.
12. Zhang, Y. and Varadarajan, B., 2006. *Utility scoring of product reviews*. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp.51–57.
13. Mukherjee, A., Venkataraman, V., Liu, B. and Glance, N., 2013. *What yelp fake review filter might be doing?* In: Proceedings of the ICWSM, pp.409–418.
14. Sandulescu, V. and Ester, M., 2015. *Detecting singleton review spammers using semantic similarity*. In: Proceedings of WWW Companion, pp.971–976.
15. Cho, Y. and Kim, J., 2013. *An effective review spam detection framework using topic modeling and sentiment analysis*. Expert Systems with Applications, 42(10), pp.4114–4122.
16. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M. and Ghosh, R., 2013. *Exploiting burstiness in reviews for review spammer detection*. In: Proceedings of ICWSM, pp.175–184.

17. Wang, G., Xie, S., Liu, B. and Yu, P.S., 2011. *Review graph based online store review spammer detection*. In: Proceedings of ICDM, pp.1242–1247.
18. Crawford, M., Khoshgoftaar, T.M., Prusa, J.D., Richter, A.N. and Najada, H., 2015. *Survey of review spam detection using machine learning techniques*. Journal of Big Data, 2(1), pp.1–24.
19. Lappas, T., Sabnis, G. and Valkanas, G., 2016. *The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry*. Information Systems Research, 27(4), pp.940–961.
20. Xu, W., Zhang, Y. and Li, B., 2015. *Online product review spam detection based on opinion mining*. In: Proceedings of the International Conference on Computer Science and Service System, pp.1363–1366.