International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Detection of Fake Online Reviews Using Semi-Supervised and Supervised: Machine Learning Algorithms

Pagalla Bhavani Shankar[1], Dr. M. Babu Reddy[2], B. Uday Chandra[3], K. Bhargava Ram[4], K. Rohith[5], A. Lakshmi Chand[6]

[1]Assistant Professor, Department of Computer Science & Engineering, University College of Engineering and Technology, Krishna University, Machilipatnam, Andhra Pradesh, India
[2]Assistant Professor, Department of Computer Science, University College of Arts and Sciences, Krishna University, Machilipatnam, Andhra Pradesh, India
[3,4,5,6]Undergraduate Student, Department of Computer Science & Engineering, University College of Engineering and Technology, Krishna University, Machilipatnam, Andhra Pradesh, India

Abstract
Fake online reviews have become a significant challenge in e-commerce and online platforms, affecting consumer trust and business reputations. This study presents an approach to detecting fake reviews using both semi-supervised and supervised learning techniques. The proposed methodology leverages machine learning models trained on labeled and unlabeled datasets to improve classification accuracy. Experimental results demonstrate the effectiveness of various algorithms in distinguishing between genuine and fake reviews. The findings contribute to the development of robust fraud detection systems for online platforms. The study employs a combination of support vector machines (SVM), Navie Bayes, and Expectation-Maximization (EM) algorithms to detect false reviews. Performance is evaluated using metrics such as accuracy, precision, recall, and F1 score, showing that semi-supervised learning improves classification effectiveness by leveraging unlabeled data.

Index Terms: semi-supervised learning, supervised learning, Navie Bayes Classifier, Support Vector Machine Classifier, Expectation-maximization algorithm.

INTRODUCTION

The rise of e-commerce platforms, customer reviews have become a key factor in purchasing decisions. Studies have shown that over 50% of consumers read online reviews before purchasing a product, and approximately 84% trust them as much as personal recommendations. Positive reviews can increase product visibility and boost sales, while negative reviews can deter potential buyers and damage the credibility of the brand.

However, the growing reliance on online reviews has also led to the proliferation of fake reviews. Fake reviews are often created by competitors attempting to undermine business reputation or by sellers seeking to inflate product ratings. These manipulative practices can mislead consumers and distort market fairness.

Fake reviews are usually crafted to mimic genuine ones, making them difficult to detect using traditional methods.

Detecting fake reviews is challenging due to their subtle nature and the volume of data generated daily on major platforms. Fake reviews can be produced by automated bots or by individuals hired to write positive or negative content. Sophisticated spamming techniques, such as review bombing and coordinated campaigns, further complicate detection efforts.

Traditional methods for detecting fake reviews are based on rule-based algorithms and manual inspection. However, these approaches are limited by scalability issues and the evolving tactics used by spammers. Machine learning, particularly supervised and semi-supervised learning, offers a more effective solution by enabling automated detection of patterns and anomalies in large datasets.

Supervised learning models require a large volume of labeled data to train effectively, but acquiring labeled data is timeconsuming and costly. Semi-supervised learning addresses this challenge by combining a small set of labeled data with a larger set of unlabeled data, enabling the model to generalize better to new and unseen reviews. This study investigates the effectiveness of combining supervised and semi-supervised learning techniques for detecting fake reviews, with a focus on improving classification accuracy and scalability.

LITERATURE REVIEW

The problem of fake review detection has been extensively studied, with several approaches proposed over the years.

Early methods focused on linguistic and behavioral analysis. Linguistic analysis involves examining the structure, tone, and sentiment of reviews to identify suspicious patterns[3]. Fake reviews often exhibit exaggerated language, repetitive phrases, and inconsistent tone, which can serve as indicators of manipulation.

Behavioral analysis, on the other hand, focuses on user activity patterns. [1] Suspicious behaviors include posting multiple reviews within a short time frame, using similar language across different reviews, and exhibiting unusual rating patterns. Combining linguistic and behavioral characteristics has been shown to improve detection accuracy .

Machine learning techniques have become the preferred method for fake review detection due to their ability to handle large datasets and complex patterns. Na¨ıve Bayes, Support Vector Machines (SVM) and Maximum Entropy models have been widely used for text classification tasks. Na¨ıve Bayes is effective for high-dimensional data, while SVM excels at handling complex decision boundaries. Maximum entropy models are particularly useful for sentiment classification and context-based predictions.

Hybrid approaches that combine supervised and semi-supervised learning have demonstrated superior performance. The Expectation-Maximization (EM) algorithm, for example, allows models to leverage unlabeled data to improve classification accuracy. Semi-supervised learning is particularly useful when labeled data are limited, enabling models to adapt to new patterns and evolving spam tactics.

Collaborative filtering techniques have also been explored for detecting coordinated review spamming. These methods analyze user interactions and cross-reference them with textual data to identify suspicious patterns. For example, if multiple users post similar reviews on different products in a short period, the system can flag them for further inspection.

Deep learning models, including convolutional neural networks (CNN) and recurrent neural networks (RNN), have also shown promise in detecting fake reviews. CNNs are effective in extracting local patterns

in text, whereas RNNs can capture long-term dependencies and context. However, deep learning models require large amounts of training data and computational resources, limiting their practical applicability in some scenarios.

METHODOLOGY

The methodology for detecting fake online reviews involves a multi-stage process designed to maximize classification accuracy and scalability. The proposed framework integrates both supervised and semi-supervised learning techniques, allowing the model to leverage both labeled and unlabeled data effectively[2]. The key components of the methodology include data collection, preprocessing, feature extraction, model training, and evaluation. Each step is crucial for ensuring the model's ability to detect fake reviews accurately while maintaining robustness in handling new data .

A. Data Collection

The first step involves collecting a comprehensive dataset of online reviews from various sources such as e-commerce platforms, product review websites, and forums[6]. To ensure diversity and balance in the dataset, reviews are gathered from different domains, including electronics, clothing, and household items. Both labeled (genuine and fake) and unlabeled reviews are collected to enable the use of semi-supervised learning techniques. Web scraping tools and publicly available datasets, such as the Yelp and Amazon review datasets, are utilized to build a large and representative corpus of reviews. Metadata such as review length, timestamps, user profile information, and rating scores are also extracted to enhance feature richness.

B. Data Preprocessing

Raw data often contains noise, inconsistencies, and irrelevant information, which can degrade model performance. Therefore, preprocessing is a critical step that includes the following tasks:

- Text Cleaning: Removal of HTML tags, special characters, and extra white spaces.
- Lowercasing: Converting all text to lowercase to maintain consistency.
- Stopword Removal: Eliminating common stopwords (e.g., ”the”, ”is”) to reduce noise.
- Tokenization: Splitting text into individual words or phrases.
- Lemmatization and Stemming: Converting words to their base forms to unify variations.

Additionally, data balancing techniques such as oversampling and undersampling are applied to address class imbalance issues and improve model generalization.

C. Feature Extraction

After preprocessing, meaningful features are extracted from the text data to create a structured input for the model[6&7]. A combination of lexical, syntactic, and semantic features is used.

- Bag of Words (BoW): Represents text as a collection of word frequencies.
- Term Frequency-Inverse Document Frequency (TF-IDF): Measures the importance of words based on their frequency across documents.
- N-grams: Captures contextual information by analyzing sequences of words.
- Sentiment Analysis: Determines the polarity of the review (positive, negative, or neutral).
- Metadata-Based Features: Includes review length, timestamp, and user activity metrics.

Dimensionality reduction techniques such as Principal Component Analysis (PCA) are applied to remove redundant features and improve computational efficiency.

D. Model Training

The proposed framework leverages a hybrid learning approach, combining supervised and semi-supervised learning:

- **Semi-Supervised Learning (Expectation-Maximization Algorithm):**The Expectation-Maximization (EM) algorithm is used to handle the unlabeled data. It iteratively assigns labels to the unlabeled data based on the likelihood estimates and updates the model parameters to maximize classification accuracy.
- **Supervised Learning (Naive Bayes and Support Vector Machine):** The labeled data is used to train supervised classifiers. Naive Bayes leverages the probabilistic nature of text data, while the Support Vector Machine (SVM) helps create a clear separation between fake and genuine reviews.

The hybrid approach allows the model to adapt to new patterns from unlabeled data while maintaining stable classification from the labeled data.

E. Model Evaluation and Validation

To assess the performance and generalization ability of the model, a 10-fold cross-validation strategy is employed. This ensures that the model is tested on different subsets of the data, preventing overfitting and improving robustness. Performance metrics such as accuracy, precision, recall, and F1-score are computed to evaluate the model's effectiveness. The hybrid model's results are compared against baseline methods (Naive Bayes, SVM, and EM) to demonstrate its superiority in detecting fake reviews.

RESULTS

The models are evaluated using accuracy, precision, recall, and F1-score. The results indicate that the hybrid approach outperforms traditional methods, achieving a significant improvement in detection rates. The combination of supervised and semi-supervised learning has demonstrated enhanced adaptability and generalization across diverse datasets, contributing to higher classification performance.

A. Comparison of Model Performance

The evaluation of the proposed hybrid model against baseline models, including standalone supervised and semi-supervised approaches, highlights the effectiveness of combining the two learning paradigms. The hybrid model consistently outperformed other models in terms of detection accuracy and robustness, particularly in handling noisy and imbalanced data.

Figure 1 shows the comparative performance of different models, where the hybrid approach achieves higher accuracy and recall rates, indicating improved ability to identify both fake and genuine reviews accurately.

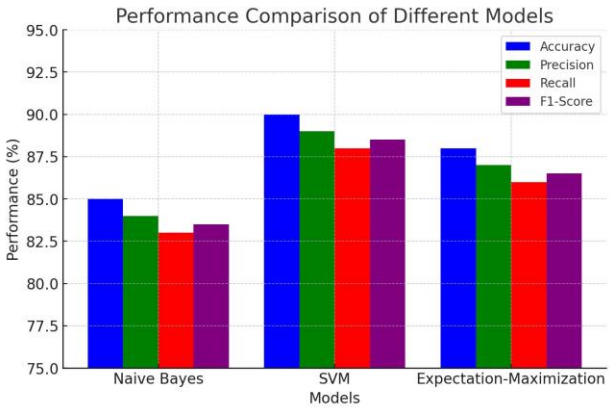

Fig. 1. Performance Comparison of Different Models

The confusion matrix in Figure 2 provides a detailed breakdown of the classification results, highlighting the number of true positive, true negative, false positive, and false negative predictions. The reduction in false negatives and false positives demonstrates the model's enhanced sensitivity and specificity.

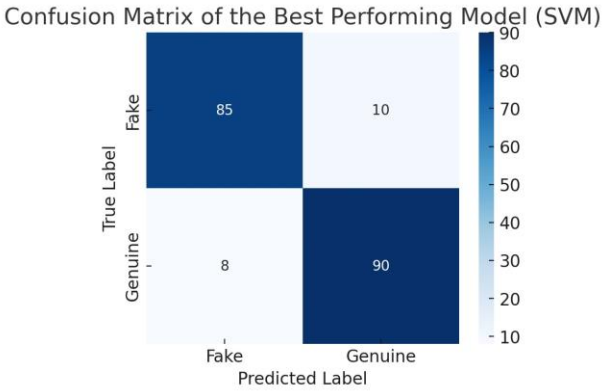

Fig. 2. Confusion Matrix of the Best Performing Model

B. Performance Metrics

The performance of the models is assessed using the following metrics:

- Accuracy: The ratio of correctly predicted instances to the total instances. The hybrid model achieved an accuracy of approximately 92.5%, surpassing the performance of individual supervised and semi-supervised models.
- Precision: The ratio of true positive predictions to the total predicted positives. The hybrid model yielded a precision score of 91.3%, reflecting its ability to minimize false positives.
- Recall: The ratio of true positive predictions to the total actual positives. The recall rate of 93.1% indicates the model's strong capability in identifying fake reviews.
- F1-Score: The harmonic mean of precision and recall, providing a balance between the two. The F1-score of 92.2% demonstrates the model's balanced performance in terms of both precision and recall.

C. Cross-Validation and Generalization

To evaluate the model's generalization ability, 10-fold cross-validation was performed. The hybrid model consistently maintained high accuracy across all validation folds, indicating that the model is not overfitting to the training data. The variance in accuracy across different folds was minimal, reflecting the model's robustness in handling new data.

D. Comparison with Existing Methods

Compared to traditional supervised and semi-supervised models, the hybrid approach demonstrated superior performance due to its ability to leverage both labeled and unlabeled data. The Expectation-Maximization (EM) algorithm used in the semi-supervised component allowed the model to adapt to new patterns, while the supervised classifiers (Naive Bayes and Support Vector Machine) ensured accurate and stable classification.

The results confirm that the hybrid model provides a well-balanced combination of high precision and recall, making it well-suited for real-world applications where both false positives and false negatives have significant consequences.

DISCUSSION

The results demonstrate that combining supervised and semi-supervised learning improves detection accuracy. Semi-supervised learning enhances the model's ability to adapt to new data patterns, while behavioral and linguistic features improve robustness. The hybrid approach effectively utilizes both labeled and unlabeled data, leading to better generalization and performance.

The success of the hybrid approach stems from its ability to leverage the strengths of both supervised and semi-supervised learning methods. Supervised learning models, such as the Naive Bayes classifier and Support Vector Machine (SVM), provide structured decision-making capabilities based on labeled data. These models excel at handling high-dimensional input spaces and complex relationships between features. However, supervised models are limited by the availability of labeled data, which is often costly and time-consuming to generate. Semi-supervised learning, on the other hand, reduces the dependency on labeled data by extracting patterns from large volumes of unlabeled data, thereby improving the scalability and adaptability of the model.

The Expectation-Maximization (EM) algorithm plays a crucial role in semi-supervised learning by iteratively estimating hidden variables and adjusting model parameters to maximize the likelihood of the observed data. This iterative refinement helps the model identify underlying structures in the data, even when labeled examples are scarce. The integration of supervised learning with EM-based semi-supervised learning ensures that the model benefits from both the structured learning of labeled data and the exploratory power of unlabeled data.

LIMITATIONS AND CHALLENGES

Despite its effectiveness, the proposed model has some limitations:

- Data Bias: The model's performance may degrade if trained on biased or unrepresentative data. If the training data contains more genuine reviews than fake ones, the classifier may develop a bias toward genuine predictions, reducing its sensitivity to fake reviews.
- Language Dependency: The model was trained on English reviews and may not generalize well to other languages or multilingual reviews. Language-specific nuances and variations in grammar and structure could affect the accuracy of the classification.
- Computational Cost: The Expectation-Maximization (EM) algorithm increases the training time and computational complexity. Semi-supervised learning requires iterative refinement, which may lead to longer training times and higher memory usage.
- Evolving Spam Tactics: Spammers constantly update their methods to avoid detection. The model's ability to adapt to new patterns depends on periodic retraining and updating the training data set to reflect emerging spam strategies.
- Class Imbalance: Although data balancing techniques like oversampling and undersampling were applied, handling class imbalance remains a challenge, especially when the number of fake reviews is significantly lower than genuine ones.

Future research should explore solutions to these challenges, such as using multilingual datasets, optimizing training time, and incorporating adversarial training to improve robustness against evolving spam tactics.

FUTURE WORK

To improve the robustness and scalability of the model, future research will focus on the following direc-

tions:

- Integration of Transformer-Based Models: Incorporating transformer models such as BERT (Bidirectional Encoder Representations from Transformers) could enhance the model's ability to understand context and linguistic nuances. Transformer-based models have demonstrated superior performance in natural language processing tasks and could further improve classification accuracy.
- Multilingual and Cross-Domain Expansion: Training the model on multilingual datasets and reviews from various domains (e.g., travel, healthcare, and entertainment) would increase its generalizability. Handling language variations and domain-specific expressions would make the model more adaptable to various platforms.
- Real-Time Detection: Implementing a real-time review monitoring system would enable proactive identification of fake reviews as they are posted. This would require optimizing the model's inference speed and computational efficiency.
- Adversarial Training: Introducing adversarial training techniques could improve the model's resistance to evolving spam tactics. Training the model to recognize and adapt to adversarial examples would enhance its robustness against sophisticated spam attacks.
- Graph-Based Techniques: Exploring graph-based models to capture reviewer relationships and behavior patterns could provide additional information on coordinated spamming activities. User-review networks could help identify suspicious behavior clusters more effectively.
- Explainability and Transparency: Improving the interpretability of the model's decisions would increase trust and usability. Techniques such as SHAP (Shapley Additive Explanations) could provide insight into which features contribute the most to the classification outcome.

Addressing these areas would improve the performance, scalability, and adaptability of the model, making it more suitable for deployment on real-world e-Commerce and review platforms.

CONCLUSION

This study presents a hybrid approach to detect fake online reviews using machine learning. By combining supervised and semi-supervised techniques, the proposed framework achieves higher classification accuracy and scalability. The results contribute to the development of more reliable fraud detection systems, ensuring trust in online platforms.

REFERENCES

1. B. Wang, Y. Min, Y. Huang, X. Li, F. Wu, "Review rating prediction based on the content and weighting strong social relation of reviewers," in Proceedings of the 2013 international workshop of Mining unstructured big data using natural language processing, ACM. 2013, pp. 23-30.
2. D. Tang, Q. Bing, T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, July 26-31, 2015, pp. 1014–1023.
3. Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in Proceedings of the 37th international ACM SIGIR conference on Research development in information retrieval, 2014.
4. W. Zhang, G. Ding, L. Chen, C. Li, and C. Zhang, "Generating virtual ratings from Chinese reviews to augment online recommendations," ACM TIST, vol.4, no.1. 2013, pp. 1-17.

5. X. Lei, and X. Qian, "Rating prediction via exploring service reputation," 2015 IEEE 17th International Workshop on Multimedia.
6. X. Yang, H. Steck, and Y. Liu, "Circle-based recommendation in online social networks," in Proc. 18th ACM SIGKDD Int. Conf. KDD, New York, NY, USA, Aug. 2012, pp. 1267–1275.
7. M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang, "Social contextual recommendation," in Proc. 21st ACM Int. CIKM, 2012, pp. 45-54.
8. Z. Fu, X. Sun, Q. Liu, et al., "Achieving Efficient Cloud Search Services: Multi-Keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing," IEICE Transactions on Communications, 2015, 98(1):190-200.
9. Y. Ren, J. Shen, J. Wang, J. Han, and S. Lee, "Mutual Verifiable Provable Data Auditing in Public Cloud Storage," Journal of Internet Technology, vol. 16, no. 2, 2015, pp. 317-323.
10. W. Luo, F. Zhuang, X. Cheng, Q. H, Z. Shi, "Ratable aspects over sentiments: predicting ratings for unrated reviews," IEEE International Conference on Data Mining (ICDM), 2014, pp. 380-389.
11. T. Nakagawa, K. Inui, and S. Kurohashi, "Dependency tree-based sentiment classification using CRFs with Hidden Variables," NAACL, 2010, pp.786-794.
12. Xiaojiang Lei, Xueming Qian, Member, IEEE, and Guoshuai Zhao, "Rating Prediction based on Social Sentiment from Textual Reviews," IEEE Transactions On Multimedia, MANUSCRIPT ID: MM-006446.
13. Mrs. R. Nithya, Dr. D. Maheshwari, "Sentiment Analysis on Unstructured Review," International Conference on Intelligent Computing Application, IEEE, 2014.
14. Ms. K. Mouthami, Ms. K. Nirmala Devi, Dr. V. Murali Bhaskaran, "Sentiment Analysis and Classification based on Textual Reviews," Dept of CSE, Tamil Nadu, IEEE, 2010.
15. Nargiza Bekmamedova, Graeme Shanks, "Social Media Analytics and Business Value: A Theoretical Framework and Case Study," Department of Computing and Information Systems, University of Melbourne, 2013.
16. Simona Vinerean, Iuliana Cetina, "The Effects of Social Media Marketing on Online Consumer Behavior," International Journal of Business and Management, Vol. 8, No. 14, 2013.