INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Novel Hybrid Feature Extraction Technique and Spam Review Detection using Ensemble Machine Learning Algorithm by Web Scrapping

Navin Kumar Goyal[1]*, Anil Pal[2], Bright Keswani[3], Dinesh Goyal[4], Mukesh Kr Gupta[5]

1 Research Scholar, Department of Computer Engineering and Information Technology, Suresh Gyan Vihar University, Jaipur, Rajasthan, India
2 Professor, Department of Computer Application, Suresh Gyan Vihar University, Jaipur, Rajasthan, India
3 Professor, Department of Computer Application, Poornima University, Jaipur, Rajasthan, India
4 Director, Poornima Institute of Engineering and Technology, Jaipur, 302022, Rajasthan, India
5 Professor, Department of Electrical Engineering, Suresh Gyan Vihar University, Jaipur, Rajasthan, India

OPEN ACCESS

Received: 19-06-2023
Accepted: 28-06-2023
Published: 05-08-2023

Citation: Goyal NK, Pal A, Keswani B, Goyal D, Gupta MK (2023) A Novel Hybrid Feature Extraction Technique and Spam Review Detection using Ensemble Machine Learning Algorithm by Web Scrapping. Indian Journal of Science and Technology 16(29): 2261-2268. https://doi.org/10.17485/IJST/v16i29.1500

* **Corresponding author**.

chikoo.1606@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop a novel hybrid method for feature generation and a novel dataset for experimenting and extracting the features for numerical representation. **Methods:** In the pursuit of the best spam review detection model, a four-stage process was undertaken. Initially, a dataset 'Fake reviews' was collected from Flipkart, containing 9926 samples from the home and kitchen products domain. Next, the data underwent pre-processing using the Natural Language Toolkit (NLTK) library. A novel Hybrid Feature Generator (HFG) was then developed, extracting informative features based on parameters like TF-IDF (Term Frequency - Inverse Document Frequency), sentiment analysis scores, and syntactic patterns. Finally, the model was trained on these generated features using Gaussian Naïve Bayes (GNB), Multinomial Naïve Bayes (MNB), and Bernoulli Naïve Bayes (BNB) algorithms. Performance evaluation was conducted using metrics such as accuracy, precision, recall, and F1-score, comparing the model's results to gold standard or known spam reviews. **Findings**: The feature generation technique was implemented on three different models, and the models were trained using 70% of the available data. The results of these experiments showed that GNB, NB, and NB achieved testing accuracies of 99.7%, 96.4%, and 99%, respectively. The performance of these models was compared with and without the inclusion of extracted product review features. The results demonstrated that the GNB algorithm outperformed the other methods in terms of accuracy and precision. **Novelty:** This study presents a novel HFG for feature extraction from review-text and a novel dataset that outperforms hitherto reported

approaches.

Keywords: Fake Reviews Detection; Ensemble Machine Learning; Feature Engineering; Naïve Bayes; Web Scrapping

1 Introduction

The growing popularity of e-commerce and the increased reliance on online shopping, especially during the Covid-19 pandemic, have highlighted the significance of product reviews in consumers' purchasing decisions. Positive reviews often attract more customers and drive sales, while negative opinions can lead to potential losses. However, the prevalence of fake reviews, both positive and negative, has become a concerning issue in recent years. Detecting and addressing these fraudulent reviews is crucial for ensuring the credibility and trustworthiness of online platforms [1–3].

Recent research has focused on developing frameworks for the detection of false reviews, particularly in the consumer electronics domain. These studies have achieved promising results, with an 82% F-Score on classification tests and identifying the Ada Boost classifier as the most effective [4]. Additionally, online platforms have implemented reporting systems for consumers to flag suspected fake reviews. However, distinguishing between authentic and fraudulent reviews remains challenging, as some fake reviews are skilfully crafted to resemble genuine ones [5,6].

Automatic detection techniques have emerged as a primary research focus to tackle this issue. The primary objective is to develop accurate methodologies that analyze reviews on e-commerce platforms, such as Amazon, by strengthening feature extraction techniques across various models. By leveraging advanced machine learning algorithms and feature extraction methods, it becomes possible to identify fake reviews more effectively [7–9].

One significant category of fake reviews includes undeservingly positive reviews aimed at promoting specific products or tarnishing the reputation of others. Another type is "non-reviews," which lack substantive judgments about the products. Recognizing and filtering out these spam reviews is crucial for establishing trust in online review systems [10,11]. The performance of several classifiers, such as KNN, Naive Bayes, SVM, Logistic Regression, and Random Forest, has been compared, considering different language models and behavioral features of reviewers. Notably, KNN has shown superior performance in terms of F-Score, achieving an impressive 82.40% accuracy, with a 3.80% increase when incorporating reviewers' behavioral features [12,13].

Moreover, the detection of fake reviews extends beyond the realm of consumer electronics. For instance, in the hotel industry, classifiers such as Naïve Bayes, Support Vector Machine, Random Forest, and Adaptive Boost have been utilized. Among these, Random Forest has shown superior performance, with 95% accuracy and F1-score. Researchers have also explored semi-supervised approaches, utilizing time series models and comprehensive feature sets, to detect fake reviews more efficiently [14].

Despite the advancements in fake review detection, there are still research gaps that need to be addressed. These include the need for more accurate methodologies, enhanced feature extraction techniques, and the incorporation of linguistic and contextual information from the reviews. Additionally, the increasing sophistication of fake reviews requires continuous improvements in detection techniques to ensure their effectiveness [15].

In light of these research gaps, our work aims to address these challenges by proposing an innovative approach that combines state-of-the-art feature extraction techniques, advanced machine learning algorithms, and linguistic analysis. By leveraging a comprehensive set of features and incorporating linguistic inquiry, we aim to improve the accuracy and reliability of fake review detection. Our methodology will also consider the evolving nature of fake reviews and strive to stay ahead of deceptive practices.

By developing a robust framework that accounts for both the content and context of reviews, our research endeavors to provide a more accurate and trustworthy system for detecting fake reviews in e-commerce platforms. The proposed approach will contribute to the integrity of online review systems, empower consumers to make informed decisions, and foster trust between consumers and platform owners.

1.1 Research gaps

- Existing research lacks a clear method for feature generation and selection in the context of fake review detection.
- While previous systems have achieved good results with standard datasets, there is a research gap in studying the problem of fake review detection specifically in the Flipkart home and kitchen products domain.
- The main objective of our research is to create a dataset by scraping Flipkart reviews in the home and kitchen domain.
- We aim to develop a novel hybrid feature extraction system that improves the precision of fake review detection techniques.
- Our research focuses on addressing the need for a more precise technique for detecting fake reviews on Flipkart or any other e-commerce platform in the home and kitchen domain

2 Methodology

In this proposed system, a novel hybrid method called HFG was developed for spam review detection as shown in Figure 1. A dataset was collected from Flipkart, specifically focusing on the ”cotton bedsheet” category in the home and kitchen products domain. A web scraper using Python's Beautiful Soup library was created to gather 9926 user reviews. Data preprocessing involved handling missing and noisy data, text tokenization, stop word removal, stemming, and lemmatization.

To understand the sentiment of the text reviews, vectorization was performed to convert the text into numeric values. Feature extraction techniques like Bag of Words (BOW) with uni-grams, bi-grams, and tri-grams were used to extract important text features. Domain-specific knowledge led to the inclusion of additional features such as positive votes, negative votes, ratings, capital letter count, punctuation count, emoji count, sentiment analysis, subjectivity, word count, unique word count, mean word length, repetitive words count, and length of text reviews.

These features were used in a classification model, and the performance of the model was evaluated using metrics such as accuracy, precision, recall, and F1-score. The findings suggested that the inclusion of these features improved the classifier's performance in detecting deceptive reviews.

Algorithm: Hybrid Review Text Feature Generator (HFG) (review r)

Input: Dataset

Output: Dataset with review features

Step-1 Calculating the sentiment score and subjectivity of the text review using TextBlob python library in Senti_Sent and Sub_Sent respectively.

Step-2 Finding negative and positive words in nNegWordCount, nPosWordCount of the review's text using TextBlob python library.

a. for each review r in the dataset (r1, r2,……rN)

b. for each word w in review r (w1, w2,……wM)

c. if (w.sent_score>0) then

d. nPosWordCount = nPosWordCount +1

e. else if(w.sent_score<0) then

f. nNegWordCount = nNegWordCount +1

g. else

h. nNetWordCount= nNetWordCount +1

Step-3 Finding and counting unique words in nUniqueWordCount for each review r in the dataset.

Step-4 Finding nCharCount, nWordCount, MeanWordLength, nSentenceCount, nPunCount, nHashTagCount for each review r.

Step-5 Finding the part of speech (POS) of review text using python's NLTK library.

nNounCount, nAdjCount, nVerbCount, nAdvCount, nProCount, nPreCount, nConCount, nArtCount, nNegaCount, nAuxCount for each review r in the dataset.

Step-6 Calculate the Authenticity of the text review using the following formula.

For each review r in the dataset (r1, r2 ,……rN

For each review r (w1, w2,……wM :

If (r posvote>r negvote AND r senimentscore>0 AND r rating>=3 then

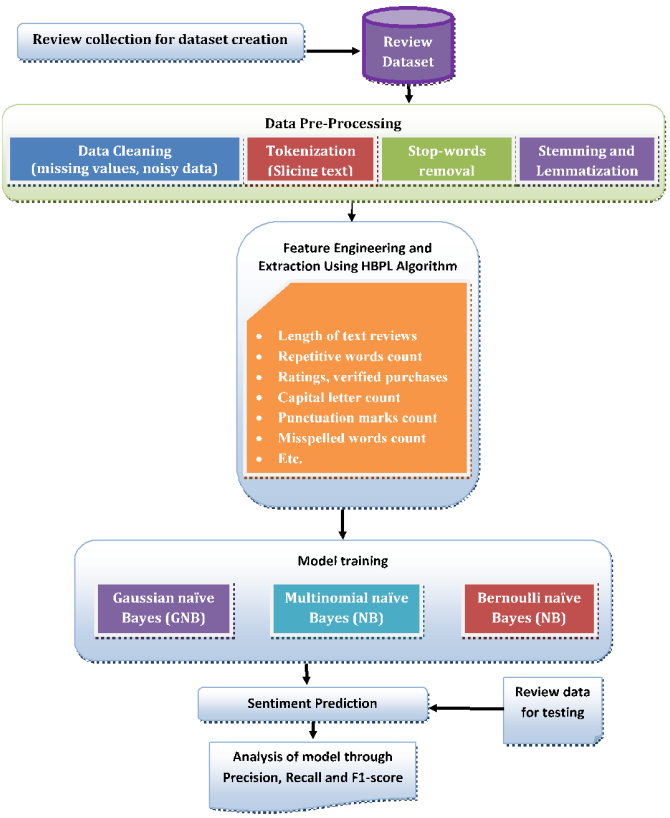

Fig 1. Proposed system

Authenticity = 1
Else:
Authenticity = 0
Step-7 Calculate the Analytical Thinking for each review r.
AT = 30 + (nArtCount + nPreCount - nProCount - nAuxCount - nConCount - nAdvCount - nNegaCount)
Step-8 combines all the features generated in steps 1-7.
Step-9 Sending the dataset with generated features for Spam Review Detection Model.

3 Results and Discussion

The dataset used in this study was generated through web scraping methods and focuses on reviews of Home and Kitchen products from Flipkart. The dataset comprises a total of 9926 reviews, with 7939 reviews identified as genuine and 1987 reviews classified as fake. Each review record includes various attributes such as product_id, product_category, reviewer_name, review_date, reviewer_address, rating, posvote, negvote, review, review_text, and review label. A summary of the dataset's statistical characteristics is given in Table 1. The longest review contains 106 words, while the shortest review consists of only 3 words. The average word count across all reviews is calculated to be 8.7 words. In total, the corpus encompasses 86470 words, and the vocabulary of the entire corpus contains 6449 distinct terms. For the purpose of analysis, the grid search technique was employed to determine the optimal parameters for all classifiers utilized in this study.

In conjunction with the statistical analysis of the dataset, we conducted an extraction of additional features to capture the behavioral characteristics exhibited in the reviews. The extracted features pertaining to user reviews are presented in Table 2. Moreover, to provide a comprehensive understanding of the data, we showcase a sample user text review in Table 3, along with the outcome resulting from the pre-processing steps applied to the review text.

In our dataset, reviewers express their opinions through written reviews. Notably, we have extracted several linguistic features to assess their impact on classifier effectiveness. These features encompass the nCapCount, which quantifies the total count of capital characters employed throughout the review content, the reviewer's usage of punctuation marks indicated by nPunCount,

Table 1. The Statistics of dataset

S. No.	Feature Name	Value(s)
1	Mean word length	5.4
2	minimum review length	3 words
3	Average review length of all reviews	8.7 words
4	Maximum review length	106 words
5	Maximum characters in a review	509
6	Minimum characters in a review	11
7	Average characters in all reviews	36.7
8	the total number of tokens	86470
9	the number of unique words	6449

Table 2. Featureengineering analysis of a review

S. No.	Feature Name	Value(s)
1	Rating	5
2	Posvote	182
3	Negvote	103
4	Review	worth every penny
5	review_text	worth every penny It looks very nice and comfy...
6	review_text_clean	worth every penni look nice comfi use bed vibr...
7	Sentiment	0.468333
8	Subjectivity	0.547917
9	Neg_Count	0
10	Pos_Count	9
11	Word_Count	42
12	Unique_words	32
13	mean_words_length	4.0
14	characters_count	219

Table 3. An illustration of review text after pre-processing

Original Text	Text after Pre-processing
a very good product within the minimum value its awesome but the TC/THREAD COUNT is low as per rough usage .. it's very good in design n fabric is too good but i think it would be better if THREAD COUNT was increased TOOO awesome in this price READ MORE.	good product within minimum value its awesome but count low per rough usage good design n fabric good but think would b better thread count increased TOOO awesome price

and the nEmojiCount, which measures the number of emojis utilized in each review. Our objective is to evaluate the influence of these user linguistic features on the effectiveness of classifiers. To provide a visual representation, a sample of the review text is presented in Figure 2.

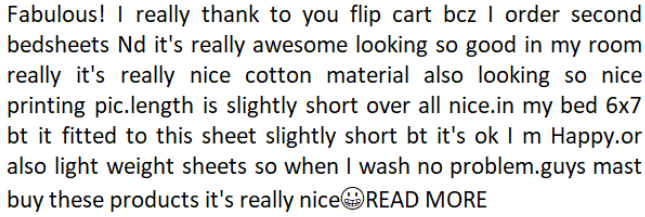

Fig 2. Snapshot of a review

The values of the linguistic features, namely nCapCount, nPunCount, and nEmojiCount, for the review text presented in Figure 2, have been computed a total 15 capital letters, 9 punctuation characters, and 1 emoji respectively. From this example, it can be observed that capital letters constitute 60% of the text, punctuation marks make up 36%, and emojis account for 4% of the text in terms of their respective proportions at the time of writing the review.

The relationship between Ham and Spam reviews characters count is shown in Figure 3. We can analyze the number of characters in genuine (Ham) reviews more than the fake (Spam) reviews.

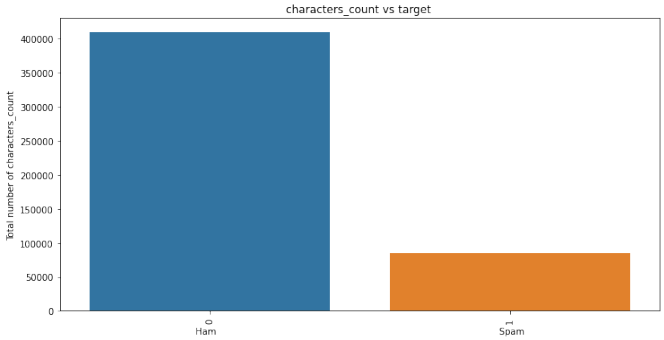

Fig 3. Character count in the Ham and Spam reviews

Using the feature generation algorithm proposed in this study, we applied it to the dataset, resulting in an extended dataset. We conducted two scenarios for evaluating the classifiers' performance. In the first scenario, 60% of the dataset was allocated for training purposes, while the remaining 40% was utilized for testing. Similarly, in the second scenario, we trained the classifiers with 70% of the dataset and evaluated their performance on the remaining 30% for testing.

Initially, the classifiers were tested without incorporating any of the additional features. Subsequently, in the second scenario, the classifier models were evaluated using linguistic features extracted from the reviews, obtained through the hybrid feature generation algorithm. The performance of the classifiers was compared in both scenarios to analyze any variations or improvements.

By conducting these evaluations, we aimed to assess the impact of the linguistic features and determine their contribution towards enhancing the classifiers' performance.

The recall, accuracy, and precision metrics achieved by the machine learning (ML) algorithms utilized in this study as presented in Table 4. Among the algorithms, the best accuracy was obtained using the BNB classifier, while the MNB classifier demonstrated the highest precision. On the other hand, the GNB classifier yielded the highest recall.

For feature extraction, BOW was employed, utilizing three different language models: 1-gram, 2-gram, and 3-gram. The classifiers were trained and evaluated using hyperparameters such as a maximum feature size of 1000, a training dataset size of 60%, and a testing dataset size of 40%. It's important to note that these evaluations were performed without considering the extracted user behavior parameters.

Table 4. Performance of different ML algorithms without using feature engineering

S. No.	Algorithm	Accuracy	Precision	Recall	F1-Score
1	GNB	66%	39%	**92%**	**85%**
2	MNB	88%	**76%**	81%	78%
3	BNB	**89%**	73%	79%	76%

The best accuracy, precision, and F1-score achieved by the GNB classifier, with slightly lower recall compared to the other two algorithms are depicted in Table 5. In this analysis, BOW was employed as the feature extraction method, utilizing uni-gram, bi-gram, and tri-gram representations. The classifiers were trained with a maximum feature size of 1000 and an additional set of 28 extracted features. The training dataset size was set to 60% of the total data.

Based on the results displayed in Table 6, the GNB classifier achieved the highest precision, accuracy, and F1-score. For the BOW feature extraction approach, employing tri-gram, bi-gram, and uni-gram representations, a maximum feature size of 1000 review features were utilized, along with an additional set of 28 extracted features. The dataset was divided into a training set of size 70% and a testing set of size 30%.

Table 5. Performance of different ML algorithms with feature engineering

S. No.	Algorithm	Accuracy	Precision	recall	F1-score
1	GNB	**99.5%**	**98%**	99%	**99%**
2	MNB	96%	86%	99%	92%
3	BNB	99%	96%	**100%**	98%

Table 6. Performance of different ML algorithms with extra features

S. No.	Algorithm	Accuracy	Precision	recall	F1-score
1	GNB	**99.7%**	**99%**	99%	**99%**
2	MNB	96.4%	86%	**100%**	92%
3	BNB	99%	97%	**100%**	98%

Upon comparing the results from both scenarios, it can be observed that all three-evaluation metrics, namely accuracy, precision, and recall, exhibit a slight increase. Notably, as the training and testing ratio increases, there is a corresponding increase in the accuracy of the Gaussian Naive Bayes (GNB) classifier.

3.1 Comparison with other Methods

The proposed model presented in this paper introduces a novel approach that has not been explored in previous research. To evaluate the effectiveness of our method, a comparison was conducted with the results reported by other studies, as depicted in Table 7. The comparison reveals that our method outperforms other approaches in terms of accuracy, specifically when utilizing the GNB method.

Table 7. Comparison between the proposed method and the methods suggested by previous workers

R ef.	Method	Accuracy
(14)	Random Forest	99.6%
(15)	GA+DNN	89%
Proposed Method	GNB	**99.7%**
Proposed Method	MNB	96.4%
Proposed Method	BNB	99%

4 Conclusion

Based on Flipkart's fake (spam) reviews, its three supervised machine learning methods, GNB, MNB, and BNB, were studied to identify fake reviews. By comparing the experimental result of classification models, we found the GNB classifier achieves better results in identifying deceptive reviews and surpassed other models by achieving a 99.7% accuracy and F1-score metric. Comparative analysis of fake review detection methods and datasets used, including feature extraction methods. NLTK library is used for cleaning up the review data. Classification uses a 70:30 train-test ratio.

Fake reviews mislead both buyers as well as sellers. Hence, the current study, focussing on identifying fake reviews, gains both academic and business interests.

Acknowledgement

The authors acknowledge with gratitude the assistance received from the administration of Suresh Gyan Vihar University Jaipur in conducting the study and preparing the report for communication.

References

1) Barbado R, Araque O, Iglesias CA. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*. 2019;56(4):1234–1244. Available from: https://doi.org/10.1016/j.ipm.2019.03.002.
2) Zhao H, Liu Z, Yao X, Yang Q. A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. *Information Processing & Management*. 2021;58(5):102656. Available from: https://doi.org/10.1016/j.ipm.2021.102656.

3) Guptta SD, Shahriar KT, Alqahtani H, Alsalman D, Sarker IH. Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques. *Annals of Data Science*. 2022. Available from: https://doi.org/10.1007/s40745-022-00379-8.
4) Aljabri M, Zagrouba R, Shaahid A, Alnasser F, Saleh A, Alomari DM. Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining*. 2023;13(1). Available from: https://doi.org/10.1007/s13278-022-01020-5.
5) Kaur G, Sharma A. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of Big Data*. 2023;10(1). Available from: https://doi.org/10.1186/s40537-022-00680-6.
6) Rayan A. Analysis of e-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique. *Computational Intelligence and Neuroscience*. 2022;2022:1–12. Available from: https://doi.org/10.1155/2022/2500772.
7) Elmogy AM, Tariq U, Mohammed A, Ibrahim A. Fake Reviews Detection using Supervised Machine Learning. *International Journal of Advanced Computer Science and Applications*. 2021;12(1). Available from: https://thesai.org/Downloads/Volume12No1/Paper_69-Fake_Reviews_Detection_using_Supervised_Machine.pdf.
8) Alsubari SN, Deshmukh SN, Alqarni AA, Aldhyani T, Alsaade FW, Khalaf OI. Data Analytics for the Identification of Fake Reviews Using Supervised Learning. *Computers, Materials & Continua*. 2022;70(2):3189–204. Available from: https://doi.org/10.32604/cmc.2022.019625.
9) Zhong M, Li Z, Liu S, Yang B, Tan R, Qu X. Fast Detection of Deceptive Reviews by Combining the Time Series and Machine Learning. *Complexity*. 2021;2021:1–11. Available from: https://doi.org/10.1155/2021/9923374.
10) Tang L, Mahmoud QH. A Survey of Machine Learning-Based Solutions for Phishing Website Detection. *Machine Learning and Knowledge Extraction*. 2021;3(3):672–694. Available from: https://doi.org/10.3390/make3030034.
11) Joni S, Chandrashekhar K, Ahmed MK, Jung SG, Bernard JJ. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*. 2022;64. Available from: https://doi.org/10.1016/j.jretconser.2021.102771.
12) Rosario C, Luca B, Nicola M, Vladimiro S, Massimo M, Hamido F, et al. A New Italian Cultural Heritage Data Set: Detecting Fake Reviews with BERT and ELECTRA Leveraging the Sentiment. *IEEE Access*. 2023;1. Available from: https://doi.org/10.1109/ACCESS.2023.3277490.
13) Dutta AK. Detecting phishing websites using machine learning technique. *PLOS ONE*. 2021;16(10):e0258361. Available from: https://doi.org/10.1371/journal.pone.0258361.
14) Gupta BB, Yadav K, Razzak I, Psannis K, Castiglione A, Chang X. A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications*. 2021;175:47–57. Available from: https://doi.org/10.1016/j.comcom.2021.04.023.
15) Deepa ST. Phishing Website Detection Using Novel Features and Machine Learning Approach. Turk. *Turkish Journal of Computer and Mathematics Education*. 2021;12:2648–2653. Available from: https://doi.org/10.17762/turcomat.v12i7.3638.