Unmasking Deception: Identifying Fake Reviews using Machine Learning Algorithm

Sandeep Singh Papola
Software Developer
Descartes Systems Group
sandeeppapola98@gmail.com

Yashika Aggarwal
SDET
Info Edge India Ltd
yashikaaggarwal08446@gmail.com

Janvi Gautam
SDET
CVENT, India
janvigautam298@gmail.com

Poonam Daneva
Research Scholar
CRIIS, BVP, India
pdaneva2001@gmail.com

Abstract. With the profusion of e-commerce websites, feedback from users in the form of reviews is one of the important factors that influence users in buying/ consuming a product. The high influence of online reviews impact spammers to generate fake reviews for promoting and demoting purposes. Identification of such reviews is necessary to stop fraud and losses to online customers. Existing work explores various machine learning models but not all to detect fake reviews. In this paper, the need to identify fake reviews was studied and all supervised machine learning is analyzed and compared to accurately identify the fake reviews. Yelp dataset is used for experimentation and comparative analysis. Experimental results depict that Support Vector Machine outperforms all considered algorithms followed by SDGC with an accuracy of 89% and 87% respectively. Further, k-Nearest Neighbour shows the least accurate results in detecting fake profiles.

Keywords: Fake reviews, Machine learning algorithms, Detection techniques, Feature extraction, Classifiers.

I. INTRODUCTION

The reviews over online marketing have become more prominent and significant over the years, which plays a major role in determining how well certain products and services are for users [1]. Spammers, hackers, or competitors misuse the features of reviews by generating fake or misleading reviews. The motive behind writing fake reviews is to boost the reputation of their products or services or to harm that of their rivals. A study [2] reports about 8-15% of the reviews posted as fake which damage the reputation of the company, further, spammers use various techniques to generate fake reviews playing with existing methods of getting caught by the machine.

A basic approach to detect fake reviews is to analyze reviews manually which is a feasible technique for a finite number of reviews and is a time-consuming expensive process. However, witnessing various algorithms and the quantity of fake reviews generated, continuous research is required to detect and mitigate fake reviews to gain customer's trust and thus generate revenue. There exist many heuristic rules for identifying fake profiles. One such rule is to spot common patterns, if a review deviates from these patterns, it may be more likely to be fake. With such type of detection, one challenge is that heuristic rules might not always be accurate with their applications. Another issue is that spammers identify the detection algorithm and alters their behavior once they know, any invalidation or restriction. As Artificial Intelligence (AI)/ Machine Learning (ML) is ruling the world, to identify fake reviews in the dataset, several detection techniques have been suggested. In this work, supervised learning techniques have been used, including Support Vector Machine (SVM), k-nearest neighbors (kNN), Logistic Regression (LR), Multinomial Naïve Bayes (MNB), Decision Tree (DT), Random Forest (RF), and Stochastic Gradient Descent Classifier (SDGC) classification to study the patterns of reviews being generated by spammers and identifying them.

Based on the pattern studied, various algorithms are identified to detect fake reviews [3]. At first, the dataset is pre-processed to remove any noisy or unreliable data points. The pre-processed dataset is taken as input for tokenization where sentences are broken down into tokens. This step ensures no punctuation marks, white spaces, or any such characters that are irrelevant for the algorithm to be followed. Further, to facilitate the methodology, affixes are removed using lemmatization. For example, the stem of "cooking" is "cook. This preprocessed data is then converted into a set of features by applying certain parameters. Count-vectorizer method [4] is used to convert the words into a bag of words representation which has the words and their frequency count in the matrix form. Finally, this transformed data is used by various machine learning classifiers to evaluate their prediction and accuracy.

A comparative analysis among various models reveals SVM has categorized the fake reviews with an accuracy of 88.85%. This solution not only provides an efficient approach

to fraudulent reviews but also gives assurance of a reduction in the number of fake feedbacks and enhances the credibility and trustworthiness of online review platforms across diverse industries. The adaptability of ML corroborates that our system remains strong against misleading reviews. The ongoing research will enable the system to improve its accuracy and effectiveness in detecting fake reviews.

The major contributions of the work are as follows:

1. A comparative analysis of various algorithms for detecting fake reviews is done.
2. SVM detected fake reviews with an accuracy of 88.85%.

This research study is organized as follows: Section 2 provides the literature review of fake reviews on online platforms. Various methodologies utilized to identify the fake reviews were presented in section 3. Models and predictions were presented in section 4. Results and discussion were illustrated in section 5. The conclusions were drawn in section 6.

II. Literature Review

Over the past few years, online retailers have expanded rapidly, particularly after the worldwide pandemic. Using this opportunity fake review on the internet has also begun to increase. Nowadays, when fake reviews are used to defraud consumers, it is crucial to focus research efforts on the early detection of these reviews. Various machine learning approaches are imposed to accurately identify fake reviews on the online platform. SVM is one of the algorithms that every researcher utilizes when working on the identification of reviews. In the research paper [5] the author has classified the reviews into negative, positive, and neutral. They have verified the reviews based on user id and booking id. The effectiveness of sentiments and emotions was explored [6] to detect the misleading reviews that would hamper the purchases. Customers rely heavily upon the reviews; the sentiment analysis helps to uncover the hidden truth behind those fake reviews. The Sentiment Analysis (SA) has also been applied to detect fake reviews on various movie streaming sites [7]. The basis on which sentimental analysis is working is to determine the polarity of a given text in the document. Here in this context, Sand Classifier (SA) is applied to classify the document into real positive reviews real negative reviews, or fake positive and fake negative reviews. Researchers have utilized various ML algorithms [8] to detect fake reviews and mislead customers. Since the study of fake reviews via machine learning there have been many extensions of it [5, 8, 9]. Unfortunately, existing works have relied primarily on ad-hoc fake and non-fake labels for model development due to the unavailability of trustworthy or gold-standard fake review data. It was also observed in some online reviews that companies, as well as individuals, trick the reviewers into promoting their brand, to motivate their newly joined users, retain their old users, and demotivate their competitors. There have been many studies in the past [5, 10, 11] which has stated that such fake review competition exists and promotes unethical online shopping experience. Other articles [12] on machine learning in the detection of fake news have considered behavioral aspects in the reviews. These studies show that there are lakhs of reviews that are fake and were created just to impress the online audience. Such an alarming number and considerable threat to the trustworthiness of the consumers must be backtracked and traced so that they can stop spreading at the right time. The approaches to detect fake reviews lie in the heat of computer science i.e., machine learning algorithms.

Most of the known identification models developed were based on numeric values in the form of customer ratings. Also, the models considered only the overall rating of the product, which generally doesn't provide the accurate authenticity of the product. The customer may not be able to judge between real and fake. In this paper, authors tried to identify the fake reviews, textually written on the online platform. These are those reviews on which the customer's next purchase is entirely dependent. The paper also presents a comparative analysis of various algorithms which otherwise were employed independently in various papers.

III. Methodology

In this section, datasets were discussed along with the flow diagram of the paperwork. Data preprocessing, feature extraction, and various developed models have also been described briefly. Figure 1 illustrates the flow diagram being followed in this research work. The dataset is pre-processed first before any model development for fake review detection. The pre-processed data is then further cleaned by extracting important features and utilizing these extracted features for detecting fake reviews. The pre-processing is performed to increase the accuracy of the model and reduce the errors.

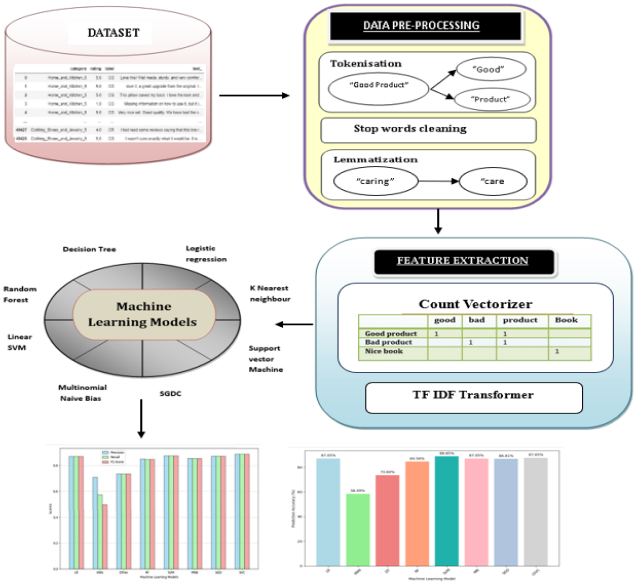

Fig. 1. Flow diagram of the fake review's detection.

IV. Data Utilised

The yelp dataset is used for investigation purposes. It hosts user-generated reviews for local businesses such as restaurants, shops, etc. The dataset consists of categories (home and kitchen, books, toys and games, movies and tv, etc.), rating (1.0

to 5.0), and text which is the review of the product. Table 1 describes the dataset whereas figure 2 shows the proportion of users giving various ratings.

TABLE I. DATASET DESCRIPTION

Category	Rating	Text
Home and Kitchen	5.0	I can't believe the low price for such quality! Love it!
Sports and outdoors	4.0	Used it for the first time yesterday and it worked great.
Movies and TV	5.0	Love it. One of the best. Great movie! Very good movie.
Books	4.0	good book if you are looking for a good read. I will admit that I read this one twice.
Electronics	1.0	Did not work for my Samsung Galaxy S3. Had to return.

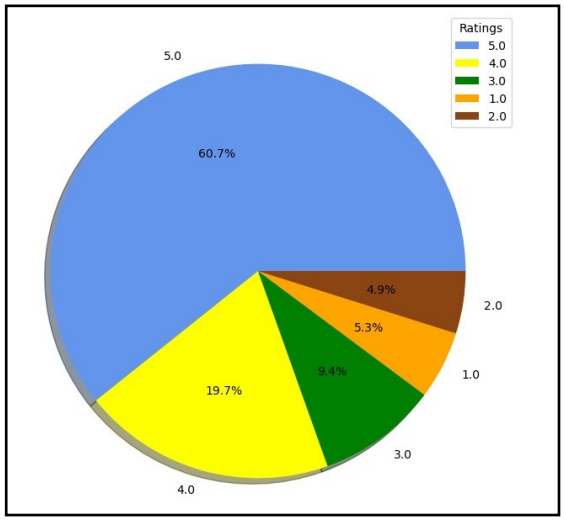

Fig. 2. Proportion of each rating

From Table 1 and Figure 2, it is clear that 60% of the reviews which are 5 stars, come from the category of Kitchen and Appliances used at home. Over 19% are from sports and outdoor activities which have got 4 stars while nearly 5% are those who got only one star and this belongs to the electronics category

A. *Data Preprocessing*

It is the essential step for ML as the data consists of irrelevant, redundant, and less frequently used words that are of no use in detecting fake reviews. This step is categorized further into-

Tokenization - It is a natural language processing technique that divides the whole text or review into individual words known as tokens. For example, if we have a review "I love the look and feel of this pillow" it divides it into tokens "I", "love", "the", "look", "and", "feel", "of", "this", "pillow".

Removal of Stop words - In this, we remove all the stop words (commonly used words) like "and", "this", "is", "the" etc. These words are removed because they do not contribute much to the overall meaning. For example, the above review is converted to "I love look feel pillow".

Lemmatization - This natural language processing technique is used to convert the words to their base or dictionary form. For example, lemmatization converts the word "caring" to "care".

Table 2 shows the pre-processed text after tokenization, removal of stop words, and lemmatization.

TABLE II. DESCRIPTION OF PRE-PROCESSED DATASET

Original text	After Pre-processing
I can't believe the low price for such quality! Love it!	I can't believe low-price quality love
Used it for the first time yesterday and it worked great.	use first time yesterday work great
Love it. One of the best. Great movie! Very good movie.	Love One best great movie very good movie
good book if you are looking for a good read. I will admit that I read this one twice.	good book looks good read I admit I read one twice
Did not work for my Samsung Galaxy S3. Had to return.	did work Samsung galaxy s3 had a return

B. *Feature Extraction*

In this step, two feature extraction tools are used and then various machine learning models are applied to predict whether the review is fake or not.

Count vectorizer - In this, the text is converted into a bag of words matrix representation. It represents text like a vocabulary where rows are review texts and columns are unique words. The value in the matrix is the frequency of the particular word in the text. So, each word becomes the feature and the value of the feature is the count of the frequency of that word in the text.

TF-IDF transformer - It converts the bag of words matrix into TF IDF values. These values combine to give the importance of the word in the text. The higher the TF IDF value, the importance of the word in the text is also considered to be higher. It is used to understand the significance of words in the review.

After data preprocessing, various machine learning models are used to generate predictions about reviews. These models were compared and discussed in the results section.

V. MODELS AND PREDICTION

Eight supervised machine learning models, namely, Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), Linear Support Vector Classifier (SVC), K-Nearest Neighbour (KNN), Random Forest (RF), Multinomial Naïve Bayes (MNB), and Stochastic Gradient Descent (SGDC) to classify reviews as fake or genuine. Each of the models has been described briefly below:

A. *Logistic Regression (LR)*

Logistic regression is a supervised ML algorithm that predicts the output of the categorical dependent variable that is labeled (CG/OR) by calculating the probabilistic values that lie

between 0 and 1 using a sigmoid mathematical function as shown in equation 1.

$$f(x) = \frac{1}{1 + e^{-\beta x}} \tag{1}$$

B. k-nearest Neighbors (k-NN)

k-NN is another supervised ML algorithm that works on similarity measures. It computes the similarity among data points to identify the category using equations (2) and (3).

$$dist(x, y) = \sqrt{\sum_{i=0}^{n} (x - y)^2} \tag{2}$$

$$dist(x, y) = \sum_{i=0}^{n} |x - y| \tag{3}$$

C. Decision Tree (DT)

DT is a classifier which appears like tree-structured, where internal nodes represent the features of a dataset i.e., bag of words, branches represent the decision rules and each leaf node represents the outcome that is whether the review is genuine or fake. We select the best attribute for the nodes of the tree by using two techniques, Information Gain (IG) and Gini Index.

$$Gini = 1 - \sum_{i=0}^{j} P(i)^2 \tag{4}$$

$$IG = Entropy(S) - [(Weighted\ Avg) \times Entropy(each\ feature)] \tag{5}$$

$$Entropy(S) = -\sum_{i=1}^{n} Pi \log_2 Pi \tag{6}$$

where,

P_i is the proportion of instances of class i in the set S

n is the number of unique classes

D. Random Forest (RF)

Random forest is a method which follows ensemble learning which includes combining various decision trees to reduce overfitting, improvise prediction, and mainly accuracy. It works on huge datasets well and can produce efficient results. According to our dataset, it assumes a random number of subsets and on those subsets of dataset it takes decisions and forms multiple decision trees. Then, finally it takes the average of all the results to make the final prediction.

E. Support Vector Machines (SVM)

SVM is used for regression as well as classification that finds an optimal hyperplane that separates the data into different classes. SVM can handle linear and non-linear data separability using different kernel functions. In this work, non-linear data separation is done by first training our model on our features extracted from bag-of-words and finding the support vectors (extreme cases). So, on the basis of extreme cases it will create decision boundary for the label (CG/OR).

F. Multinomial Naive Bayes (MNB)

It is a probabilistic learning method based on the Bayes theorem that is used for text data analysis. It calculates the probability of a review being genuine or fake based on the prior knowledge of extracted features. It uses a bag of words matrix to find the likelihood of each feature and the prior probability of each class. After training, it then calculates the final probability of review as fake or original using the equation (7).

$$P\left(\frac{C}{X1 \cap X2 \cap \ldots \cap Xn}\right) = \frac{P\left(\frac{X1}{C1}\right) \cdot P\left(\frac{X2}{C1}\right) \ldots P\left(\frac{Xn}{C1}\right)}{P(X1) \cdot P(X2) \ldots P(Xn)} \tag{7}$$

To perform the classification, each feature of bag-of-words representation are represented by a vector $\mathbf{X}$ (X1, X2,…, Xn) .

G. Stochastic Gradient Descent Classifier (SGDC)

SGDC minimizes an objective function to train a linear classifier. Instead of the complete dataset, SGDC uses small random datasets to compute the gradient and update the model's parameters in each iteration. It is computationally efficient and suitable for large-scale and sparse datasets like in our text classification of reviews and natural language processing.

VI. Results and Discussion

This section discusses the various evaluation matrices used along with the results analysis.

A. Evaluation Matrix

In this work, precision, recall, and F1-score are used to evaluate various classifiers.

Precision: It is the ratio of the correctly identified fake reviews (True Positive) to the total number of reviews that are predicted as fake.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{8}$$

Recall: It is the ratio of the correctly identified fake reviews to the total number of actual fake reviews.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{9}$$

F1-score: It is the harmonic mean of recall and precision .

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

B. *Result and Analysis*

The performance of various classifiers is evaluated by dividing the dataset into two sub parts i.e., 80% for training and 20% for testing. All classifiers are evaluated under ideal conditions. Table 3 shows the predicted values for all the classifiers where True Positive (TP) depicts a number of times a model correctly classifies a fake review as fake. True Negative (TN) tells a number of times the model correctly classifies genuine review as genuine. False Positive (FP) depicts a number of times a model classifies incorrectly a genuine review sample as fake. False Negative (FN) tells the proportion of incorrectly classified positive samples as negative.

TABLE III. COMBINED PREDICTED VALUES FOR ALL CLASSIFIERS

Classifiers	TP	FP	TN	FN
LR	3572	576	3468	471
kNN	790	104	3940	3253
DT	2943	1035	3009	1100
RF	3262	468	3576	781
SVM	3651	510	3534	392
MNB	3572	576	3468	471
Linear SVC	3527	483	3561	516
SDGC	3590	614	3430	453

Figure 3 along with Table 4 represents the comparative analysis of various classifiers based on precision, recall, F1-score, and accuracy. It can be observed that SVM outperforms other models with a precision of 0.89 followed by SDGC and LR with a precision of 0.87. DT and kNN classifiers have the found to be worst worst-performing models with a precision of 0.72 and therefore should not be considered to identify fake reviews.

TABLE IV. COMPARATIVE ANALYSIS OF VARIOUS ML MODELS

Classifiers	Precision	Recall	**F1-Score**	Accuracy
LR	0.87	0.87	0.87	0.87
kNN	0.72	0.58	0.51	0.57
DT	0.74	0.74	0.74	0.73
RF	0.85	0.85	0.85	0.84
SVM	**0.89**	**0.89**	**0.89**	**0.88**
NB	0.87	0.87	0.87	0.87
Linear SVC	0.88	0.88	0.88	0.87
SDGC	0.87	0.87	0.87	0.86

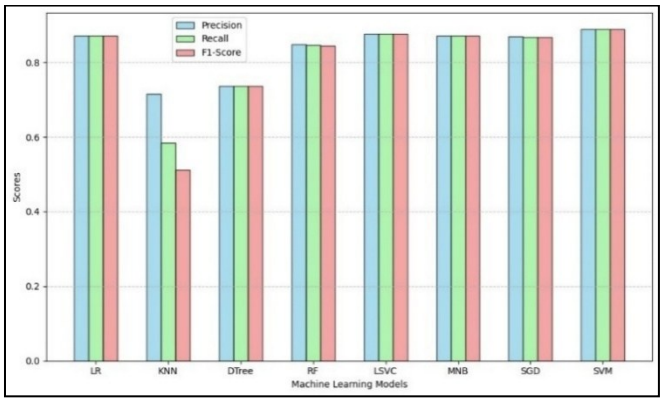

Fig. 3. Comparative result analysis of various classifiers

From Figure 3 and Figure 4, it can be observed that SVM classifies reviews correctly with an accuracy of 88.85% followed by LR and kNN with an accuracy of 87.05%. The least accurate results are shown by kNN depicting its insignificance in classifying reviews as fake or genuine.

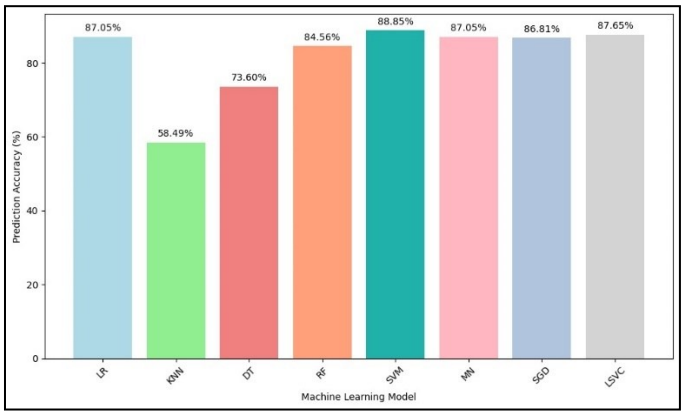

Fig. 4. Accuracy Measures of various considered algorithms

VII. CONCLUSION

Fake review detection is crucial for maintaining the integrity of online business systems and protecting consumers from misleading information. Through the utilization of advanced technologies and methodologies, such as natural language processing and machine learning, it is possible to develop automated systems that can effectively identify and flag fake reviews. It benefits consumers, businesses, and online platforms by enabling informed decisions, protecting reputations, and fostering trust. Based on the experimental evaluation both linear SVC and support vector machines with precision 0.88 and 0.89 respectively demonstrated strong performance in detecting fake reviews.

REFERENCES

[1] Ott, M., Cardie, C., Hancock, J.: Estimating the prevalence of deception in online review communities. In: Proceedings of the 21st international conference on World Wide Web. pp. 201–210. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2187836.2187864.

[2] Mohawesh, R., Xu, S., Tran, S.N., Ollington, R., Springer, M., Jararweh, Y., Maqsood, S.: Fake Reviews Detection: A Survey. IEEE Access. 9, 65771–65802 (2021). https://doi.org/10.1109/ACCESS.2021.3075573.

[3] Choi, Y., Farnadi, G., Babaki, B., Broeck, G.V. den: Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns. Proceedings of the AAAI Conference on Artificial Intelligence. 34, 10077–10084 (2020). https://doi.org/10.1609/aaai.v34i06.6565.

[4] Birim, Ş.Ö., Kazancoglu, I., Kumar Mangla, S., Kahraman, A., Kumar, S., Kazancoglu, Y.: Detecting fake reviews through topic modelling. Journal of Business Research. 149, 884–900 (2022). https://doi.org/10.1016/j.jbusres.2022.05.081.

[5] Poonguzhali, R., Sowmiya, S.F., Surendar, P., Vasikaran, M.: Fake Reviews Detection using Support Vector Machine. In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). pp. 1509–1512 (2022). https://doi.org/10.1109/ICSCDS53736.2022.9760747.

[6] Melleng, A., Jurek-Loughrey, A., P, D.: Sentiment and Emotion Based Representations for Fake Reviews Detection. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 750–757. INCOMA Ltd., Varna, Bulgaria (2019). https://doi.org/10.26615/978-954-452-056-4_087.

Mir, A.Q., Khan, F.Y., Chishti, M.A.: Online Fake Review Detection Using Supervised Machine Learning And BERT Model, http://arxiv.org/abs/2301.03225, 2023. https://doi.org/10.48550/arXiv.2301.03225.

[7] Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews.

[8] Rout, J.K., Dash, A.K., Ray, N.K.: A Framework for Fake Review Detection: Issues and Challenges. In: 2018 International Conference on Information Technology (ICIT). pp. 7–10 2018. https://doi.org/10.1109/ICIT.2018.00014.

[9] Elmogy, A.M.: Fake Reviews Detection using Supervised Machine Learning. International Journal of Advanced Computer Science and Applications. 12, 2021.

[10] Hajek, P., Hikkerova, L., Sahut, J.-M.: Fake review detection in e-Commerce platforms using aspect-based sentiment analysis. Journal of Business Research. 167, 114143, 2023. https://doi.org/10.1016/j.jbusres.2023.114143.

[11] Munzel, A.: Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus. Journal of Retailing and Consumer Services. 32, 96–108, 2016. https://doi.org/10.1016/j.jretconser.2016.06.002.