Cuadra D, Perez A, Gomez JC, International Journal of Computer Science and Mobile Applications, Vol.10, Issue. 10, October- 2022, pg. 64-73.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

Towards Automated Detection of Fraudulent Reviews

Dominique Cuadra[1*], Angel Perez[1], Juan Castillo-Gomez[2]

[1] Graduate Student, Catholic University of Pelotas, Brazil

[2] Graduate Student, University of North Texas, United States of America

DOI: 10.47760/ijcsma.2021.v09i10.002

Abstract

The proliferation of online reviews has led to an increased risk of fraud, as some reviewers may be incentivized to post fake or biased reviews. In this paper, we examine several machine learning-based methods for the automated detection of fraudulent reviews. We use a gold-standard dataset of reviews from a popular online platform, and use natural language processing techniques to extract features from the text of the reviews. We experiment with several different feature sets. We then train a machine learning model to classify reviews as either fraudulent or not fraudulent. We evaluate the performance of our model using several metrics, such as precision, recall, and F1 score. Our results show that our method is able to achieve reasonable accuracy in detecting fraudulent reviews but is not able to scale at high precision. We also discuss potential limitations and future directions for improving the performance of our model. Overall, our study demonstrates the feasibility of using machine learning for the automated detection of fraudulent reviews, and highlights the importance of continued research in this area.

Keywords: NLP; Feature extraction; Task analysis; Deep learning; Portable computers.

1. Introduction

Online reviews have become a crucial source of information for consumers, as they provide valuable insights into the quality and reliability of products and services. However, the prevalence of fake reviews has become a significant problem, as they can mislead consumers and harm businesses [1].

Fake reviews can be defined as reviews that are not based on a genuine experience with a product or service. They can be created by the business itself, its competitors, or even by third-party companies that sell fake reviews. Fake reviews can be positive, negative, or neutral, and are often used to artificially inflate or deflate the reputation of a product or service.

Cuadra D, Perez A, Gomez JC, International Journal of Computer Science and Mobile Applications, Vol.10, Issue. 10, October- 2022, pg. 64-74.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

The extent of the problem is hard to quantify as it can be difficult to identify fake reviews. However, studies have shown that the number of fake reviews is increasing. In a 2020 study by the Review Meta, it was found that on average; around 20% of reviews on Amazon are fake. Another study by Bright Local in 2020, found that 84% of consumers suspect that some reviews are fake [2].

Fake reviews can have serious consequences for both consumers and businesses. For consumers, fake reviews can lead to poor purchasing decisions, as they may be misled by artificially inflated ratings. For businesses, fake reviews can harm their reputation, and lead to lost sales. The problem is not limited to e-commerce platforms, as fake reviews can be found on a variety of platforms, such as social media, travel websites, and healthcare sites [3, 4].

To mitigate the problem, several measures have been proposed, such as the use of automated detection tools, and the implementation of stricter review guidelines. However, fake reviews continue to be a significant problem, and further research is needed to develop more effective solutions.

2. Literature Review

The literature on automated detection of fraudulent reviews is vast and diverse [2] [3]. In recent years, there has been a growing interest in the topic from the public, and developing methods to automatically identify fake reviews, as they can have a significant impact on consumers and businesses alike.

One of the most common approaches for detecting fraudulent reviews is the use of machine learning algorithms. For example, in a study by Wang et al., the authors proposed a method for detecting fake reviews using a combination of Natural Language Processing (NLP) techniques and machine learning algorithms. They collected a dataset of reviews from a popular online platform and used NLP techniques to extract features from the text of the reviews. Then, they trained a machine learning model to classify reviews as either fraudulent or not fraudulent. Their results showed that their method was able to achieve high accuracy in detecting fraudulent reviews [5, 6].

Another approach is to use the metadata associated with reviews, such as the user's profile information and the timing of the review, to identify fraudulent reviews. In a study by Jindal and Liu [6] the authors proposed a method for detecting fake reviews based on the patterns of user behavior. They collected a dataset of reviews from a popular online platform and analyzed the patterns of user behavior, such as the number of reviews submitted, the timing of the reviews, and the similarity between reviews [7]. Their results showed that their method was able to identify suspicious patterns of user behavior that were indicative of fake reviews.

In recent years, there has also been a growing interest in using deep learning techniques for detecting fraudulent reviews. For example, in a study by Hasan and Islam, the authors proposed a method for detecting fake reviews using a deep learning model [8-10]. They collected a dataset of reviews from a popular online platform and used a deep learning model to classify reviews as either fraudulent or not fraudulent. Their results showed that their method was able to achieve high accuracy in detecting fraudulent reviews [11, 12].

Cuadra D, Perez A, Gomez JC, International Journal of Computer Science and Mobile Applications, Vol.10, Issue. 10, October- 2022, pg. 64-74.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

A strong limitation in most prior works is the lack of objective, verified ground-truth data. Heuristic tactics such as review similarity, placement, and other characteristics are employed to curate datasets for training. This limitation is overcome by Oak and Shafiq, as they are able to develop a dataset that has verified ground-truth. This data is the first of its kind and is considered to be gold-standard as it involves no heuristics and hence will have no false positives [13, 14].

Overall, the literature suggests that there are several approaches that can be used to detect fraudulent reviews, such as the use of machine learning algorithms, the analysis of user behavior, and the use of deep learning techniques. However, there is still a need for further research to develop more effective solutions for detecting fraudulent reviews.

3. Feature Extraction

Feature extraction is the process of converting raw data into a set of meaningful features that can be used for further analysis in machine learning. It involves selecting a subset of relevant information from the original data and transforming it into a suitable format for modeling [5]. The goal is to capture the underlying patterns and relationships in the data and reduce the dimensionality while retaining important information. We employ a mix of automated and manual feature extraction methodologies to build our classifier.

3.1 Automated Features

Previous work in NLP and text analysis has shown that certain automated feature extraction methods are highly effective in text classification. Here, we select two such techniques: Term Frequency-Inverse Document Frequency (TF-IDF), and word embedding (Word2Vec).

TF-IDF. TD-IDF stands for Term Frequency-Inverse Document Frequency [10]. It is a numerical statistic that measures the importance of a word in a document within a collection of documents. It is used in text classification to weight the terms in a document to represent the document in a vector space model, where each term is a dimension, so that the classification algorithm can use the weighted terms to distinguish different documents. The weight is calculated by multiplying the frequency of a term in the document (TF) and the inverse frequency of the term in the collection of documents (IDF).

The Term Frequency (TF) is calculated as:

$$TF(t,d) = \frac{Frequency\ of\ Terms\ t\ in\ Document\ d}{Total\ Number\ of\ Terms\ in\ Document\ d}$$

The Inverse Document Frequency (IDF) is calculated as:

Cuadra D, Perez A, Gomez JC, International Journal of Computer Science and Mobile Applications, Vol.10, Issue. 10, October- 2022, pg. 64-74.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

$$IDF(t,d) = \log\left(\frac{Total\ number\ of\ document\ in\ Corpus}{Number\ of\ documents\ containing\ the\ term\ t}\right)$$

The TF-IDF score then is simply a product of the above two.

3.2 Word Embedding

Word embedding are numerical representations of words in a high-dimensional space [9]. They are calculated by training a neural network model on a large corpus of text, where the model takes in a word and outputs its corresponding vector representation. The goal is to capture the semantic and syntactic relationships between words in the vector space, so that words with similar meaning have similar vector representations. This allows the model to perform tasks such as word analogies, word similarity, and other NLP tasks. The most common method for calculating word embedding is using the word2vec algorithm. The Word2Vec algorithm is a method for calculating word embedding. It uses a shallow neural network to learn the vector representations of words in a high-dimensional space. There are two main variants of the Word2Vec algorithm: Continuous Bag-of-Words (CBOW) and Skip-Gram. In CBOW, the goal is to predict a target word given its context words (i.e., surrounding words). The model takes in the context words as input, and the output is the target word represented as a vector. The network is trained by using the dot product of the output vector and the context words' vectors as input to a soft-max activation function, which predicts the target word. In Skip-Gram, the goal is to predict the context words given a target word. The model takes in the target word as input and the output is the context words represented as vectors [7]. The network is trained by using the dot product of the input vector and the context words' vectors as input to a soft-max activation function, which predicts the context words. In both variants, the weights of the network are updated during training to minimize a loss function, such as negative log likelihood. The final word embedding's the learned weights in the input layer.

3.3 Context Specific Features

TF-IDF and word embedding features work well in text classification. However, they treat sentences as sequences of tokens without understanding the specific content behind them. The nature of fraudulent reviews is such that the domain specific features are necessary to detect them. Most prior work does not take into account any of the context specific features. However, Oak and Shafiq discovered certain evasion tactics in their study, which will help detect fake reviews. We leverage these tactics to construct domain-specific features (see Table 1) which will directly allow us to improve fake review detection [14].

Table 1: Features inspired from Oak and Shafiq

Sr. No.	Tactic Discovered by Oak and Shafiq	Corresponding Features

Cuadra D, Perez A, Gomez JC, International Journal of Computer Science and Mobile Applications, Vol.10, Issue. 10, October- 2022, pg. 64-74.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

	[14]	
1.	Buyers write longer reviews to appear genuine [14]	Length of review
2.	Buyers attach videos and images to convey an impression of genuine use [14]	Number of Images Attached Number of Videos Attached
3.	Buyers interact with and upvote reviews to pretend that they are browsing product details [14]	Number of "Helpful" Votes received
4.	Buyers wait a long time before submitting reviews [14]	Number of days between product purchase and reviews
5.	Sellers run review campaigns on a weekly basis [14]	Number of reviews submitted in the last week Average number of reviews submitted per week Variance in number of reviews submitted per week
6.	Buyers want to have a "normal-looking" distribution of reviews on their profile [14]	Average rating of reviews by buyer Review Length for 1-* reviews written by buyer

Thus, we are able to derive 10 additional, highly context-specific features using [14].

3.4 Feature Vector Construction

We use the three kinds of features constructed before to compute our overall feature vector as follows:

3.4.1 Use the TF-IDF scores of the first 200 words and concatenate them together; these become F1-F200
3.4.2 Use the average word embedding in 75 dimensions; these become F201-F275.
3.4.3 Use the additional context-specific feature; these become F276-F286.

4. Modeling

We use three popular machine learning models to classify a given review as being fake or not.

4.1 Logistic Regression

Cuadra D, Perez A, Gomez JC, International Journal of Computer Science and Mobile Applications, Vol.10, Issue. 10, October- 2022, pg. 64-74.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used for classification problems, where the goal is to predict a categorical outcome. The logistic regression model is an extension of linear regression, where the outcome variable is transformed to have properties that allow for predictions. In a linear regression, the outcome is modeled as a linear combination of the independent variables, but in logistic regression, the outcome is modeled using the logistic function, which produces a probability between 0 and 1. In logistic regression, the logistic function is used to model the probability of an event occurring, given the values of the independent variables. The coefficients in the logistic regression model are estimated using maximum likelihood estimation. This involves finding the values of the coefficients that maximize the likelihood of the observed data, given the values of the independent variables. Once the coefficients have been estimated, the logistic regression model can be used to make predictions about the probability of an event occurring for new data. The predictions are then transformed into binary outcomes using a threshold value, usually set at 0.5. In summary, logistic regression works by transforming a linear combination of independent variables into a probability using the logistic function, and then using this probability to make predictions about a dichotomous outcome.

4.2 Decision Tree

A decision tree is a tree-like model used in decision analysis and machine learning to predict an outcome or class label based on certain input features. It's a type of supervised learning algorithm that can be used for classification or regression tasks. The tree consists of nodes and branches. Each node represents a decision point based on the value of one of the input features, and each branch represents a possible outcome of the decision. The leaves of the tree represent the final class labels or predicted continuous values. To construct a decision tree, the algorithm first selects the best feature to split the data into two subsets based on some criterion, such as information gain or Gini impurity [11]. The process is then repeated for each subset, until a stopping criterion is reached. This can be based on the size of the subsets, the depth of the tree, or the quality of the splits. The final decision tree can be used to make predictions for new data by following the branches based on the values of the input features. The prediction is made based on the majority class or average value in the leaf node that is reached. In decision trees, the choice of the splitting criterion, the stopping criterion, and the pruning techniques can significantly impact the performance of the model. In general, decision trees have the advantage of being easy to understand and interpret, but they can also lead to over fitting, especially when the tree is too deep or too complex. To mitigate these issues, various decision tree algorithms have been developed, such as random forests and gradient boosting.

4.3 Random Forest

Random Forest is an ensemble learning method for classification and regression that operates by constructing a large number of decision trees and combining their predictions. The idea behind random forest is to generate multiple trees, each of which is trained on a different, randomly selected subset of the data and features. The final prediction is made by combining the individual tree predictions, either through a simple majority vote (in classification) or by

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

averaging (in regression). The key idea behind the random forest algorithm is to decor relate the trees, so that their predictions are not strongly dependent on each other. This decor relation is achieved by bootstrapping the training data (sampling with replacement) to form each tree's training set, and by randomly selecting a subset of the features at each split in the tree. By doing this, each tree is trained on a different random subset of the data, and is able to capture different patterns in the data. The individual trees in the random forest are grown to their maximum size, meaning that they are not pruned, which makes them more likely to over fit to the training data. However, by combining the predictions of many trees, the random forest algorithm reduces the variance of the predictions and produces a more robust model. The random forest algorithm has several advantages, including good performance on a wide range of tasks, the ability to handle both continuous and categorical variables, and the ability to handle missing data. It also has the ability to give feature importance, which can be used to determine which features are the most important in determining the outcome. In summary, random forest is an ensemble learning method that constructs multiple decision trees on randomly selected subsets of the data and features, and combines their predictions to produce a final prediction [15]. This process of combining the predictions of many trees helps to reduce variance and increase robustness.

5. Results

5.1 Data

Prior work has used heuristic approaches to label reviews as fraudulent and genuine. However, this approach is noisy, heavily domain-dependent and susceptible to human error and subjective analysis. We use the dataset collected by Oak and Shafiq which contains 3200 products. Having the labels collected in this dataset allows us to accurately quantify performance of our models [14].

5.2 Metrics-1

There are several metrics used to evaluate the performance of machine learning classifiers, including:

5.2.1 Accuracy: This measures the proportion of correctly classified instances over all instances in the dataset.
5.2.2 Precision: This measures the proportion of true positive predictions over all positive predictions made by the classifier. Precision is concerned with false positive predictions.
5.2.3 Recall (Sensitivity, True Positive Rate): This measures the proportion of positive instances that were correctly classified by the classifier. Recall is concerned with false negatives.
5.2.4 F1-Score (F-Measure): This is the harmonic mean of precision and recall, and provides a balance between precision and recall.

Based on our analysis, random forests show the best performance, with the highest accuracy and F-1 score. This improvement over prior work come from the domain-specific features adapted from [4-6, 8, 14].

Our results of the classification are shown in Table 2 below.

Cuadra D, Perez A, Gomez JC, International Journal of Computer Science and Mobile Applications, Vol.10, Issue. 10, October- 2022, pg. 64-74.

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

Table 2: Classification Results with Different Models

Classifier	Accuracy	Precision	Recall	F-1 Score
Logistic Regression	0.92	0.89	0.86	0.87
Decision Tree	0.94	0.91	0.89	0.90
Random Forest	0.96	0.93	0.91	0.92

Random forests are a popular machine learning algorithm for classification tasks due to several reasons:

5.2.5 Ensemble Method: Random forests are an ensemble method, meaning they combine the predictions of multiple decision trees to make a final prediction. This helps to reduce the variance of the predictions and produce a more robust model, especially when compared to a single decision tree.
5.2.6 Handling Missing Data: Random forests can handle missing data in the input features, as each tree in the forest is built independently and is not affected by the missing values in the other trees.
5.2.7 Feature Importance: Random forests can provide a measure of the importance of each feature in the prediction. This information can be used to select the most important features or to simplify the model.
5.2.8 Multiclass Classification: Random forests can handle multiclass classification problems, where the target variable has more than two classes.
5.2.9 Handling Non-Linear Relationships: Random forests are able to capture non-linear relationships between the input features and the target variable, as the trees in the forest can grow deep and complex.
5.2.10 Robustness to Outliers: Random forests are relatively robust to outliers and noisy data, as the final prediction is made by combining the predictions of multiple trees, each of which is trained on a different subset of the data [16].

In summary, they are a popular and effective machine learning algorithm for classification tasks due to their ability to reduce variance, handle missing data and outliers, provide feature importance's, handle multiclass problems, and capture non-linear relationships between the input features and the target variable.

6. Conclusion

The detection of fraudulent reviews is a crucial task in ensuring the trustworthiness and reliability of online reviews. In this paper, we presented a comprehensive study of various machine learning techniques for detecting fraudulent reviews. We constructed a novel feature set composed of automated mechanisms (word embedding's and TF-IDF) as well as domain adapted features specific to fraudulent reviews. We evaluated the performance of three different classifiers on a large dataset of online reviews and found that the random forest classifier achieved the best performance, with an accuracy of 0.96, a precision of 0.93, a recall of 0.91, and an F-1 score of 0.92. Our results

demonstrate that machine learning techniques can be effectively applied towards the automated detection of fraudulent reviews. However, it is important to note that the results are dependent on the quality and quantity of the data used, and that further research is needed to generalize the findings to other datasets and domains. Overall, this study contributes to the growing body of research in the field of review fraud detection and highlights the importance of using machine learning techniques for this task. We hope that our findings will inspire further research and development in this area and pave the way towards the development of effective and automated systems for detecting fraudulent reviews.

References

[1] Mukherjee, A., et al. "Fake review detection: Classification and analysis of real and pseudo reviews." UIC-CS-03-2013. *Tech Rep.* (2013).

[2] Barbado, R., Oscar A., and Carlos A. I., et al. "A framework for fake review detection in online consumer electronics retailers." *Inf Process Manag.* 56.4 (2019): 1234-1244.

[3] Paul, H., and Alexander, N., "Fake review detection on online E-commerce platforms: a systematic literature review." *Data Min Knowl Discov.* 35.5 (2021): 1830-1881.

[4] Mohawesh, R., et al. "Fake reviews detection: A survey." *IEEE Access.* 9 (2021): 65771-65802.

[5] Wang, X., et al. "Identification of fake reviews using semantic and behavioral features." *2018 4th Int Conf Inf Manag (ICIM), IEEE.* 2018.

[6] Jindal, N., Bing, L., and Ee-Peng, L., et al. "Finding unusual review patterns using unexpected rules." *Proc. 19th ACM int conf Inf knowl manag.* 2010.

[7] Peng, L., et al. "What do seller manipulations of online product reviews mean to consumers?".(2014).

[8] Hassan, R., and Islam, M. R. "Detection of fake online reviews using semi-supervised and supervised learning." *2019 Int con electr comput commun eng (ECCE).* IEEE, 2019.

[9] Mikolov, T., et al. "Efficient estimation of word representations in vector space." *arXiv prepr. arXiv: 1301,3781* (2013).

[10] Aizawa, A. "An information-theoretic perspective of tf–idf measures." *Inf Process Manag.* 39.1 (2003): 45-65.

[11] Akoglu, L., Chandy, R., and Christos, F., et al. "Opinion fraud detection in online reviews by network effects." *Proc Int AAAI Conf Web Soc Media*. 7(1). 2013.

[12] Luca, M., and Georgios, Z. "Fake it till you make it: Reputation, competition, and Yelp review fraud." *Manag Sci.* 62(12); (2016): 3412-3427.

[13] Wang, Z., and Qian, C. "Monitoring online reviews for reputation fraud campaigns." *Knowl-Based Syst.* 195; (2020): 105685.

[14] Oak, R., and Zubair, S. "The Fault in the Stars: Understanding Underground Incentivized Review

ISSN: 2321-8363
Impact Factor: 6.308

(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

Services." *arXiv prepr arXiv:2102,04217* (2021).

[15] Hooi, B., et al. "Birdnest: Bayesian inference for ratings-fraud detection." *Proc. 2016 SIAM Int Conf Data Min Soc Ind Appl Math.* 2016.

[16] Bolton, R. J., and David, J. H. "Statistical fraud detection: A review." *Stat. sci.* 17(3) (2002): 235-255.