E-ISSN: 2707-6644
P-ISSN: 2707-6636
IJCPDM 2023; 1(1): 85-89
Received: 25-01-2023
Accepted: 02-03-2023

Krina Patel
Department of Computer Science, California State University East Bay, USA

Moayed D Daneshyari
Department of Computer Science, California State University East Bay, USA

Corresponding Author:
Krina Patel
Department of Computer Science, California State University East Bay, USA

Detection of fake customer reviews on e-commerce with supervised machine learning

Krina Patel and Moayed D Daneshyari

DOI: https://doi.org/10.33545/27076636.2023.v4.i1a.83

Abstract
According to consumer statistics, 74% of online consumers read reviews before purchasing a product. Consumers believe that the reviews are written by people who have purchased the same product. However, that is not the case sometimes. Many reviews are fake and generated by bots to enhance the brand value or demolish some product's image by posting a negative review for it. Limited research has been done in this area and automatic detection systems show partial success in detecting fake reviews. In this project, we discuss the issue of fake reviews and methods to detect them. The project experiments with three models – Naïve Bayes, Support Vector Machine, and Random Forest for classification. By observing the results of these models, we can surely say that human eyes cannot detect computer-generated reviews as accurately as machine learning techniques can. This web application can be implemented in any e-commerce platform to train, and test based on their data, and it can provide consumer protection and increase the credibility of reviews.

Keywords: Fake reviews, detection, naïve Bayes, support vector machine, random forest, e-commerce, machine learning, consumer protection

1. Introduction
Online reviews are also called as electronic Word-of-Mouth (eWOM). They are emerging as strong influencing factors for manipulating consumer minds [2]. There are 250 million reviews on Amazon and 224 million reviews on Yelp. This shows that there is a large set of customers who shops online and show their opinions after using that product to help other future shoppers. Reviews are important for both buyers and sellers as for future buyers it becomes guidance in decision making and for sellers, their product gets promoted. People prefer websites that have rich customer review data. As positive reviews can attract more customers and increase sales, negative reviews can lower the demand. Primarily, there are three types of non- original reviews: 1) [3] Fake review: These reviews are written in exchange for some benefit like a free product or cashback by a human who does not have used the product. This can be done both in a positive way as well as a negative way. 2) Review about brands only: The review is not about the product but about the brand or manufacturer. It can also be both positive and negative. For example, on LG laptop products, the review is about how good LG refrigerators are. 3) Non-reviews: These reviews are advertisements and irrelevant content without any proper opinion. Fake reviews can be defined as "deceptive reviews provided with an intention to mislead consumers in their purchase decision making, often by reviewers with little or no actual experience with the products or services being reviewed"
[4] Commonly, two types of features are used to identify fake reviews: Contextual and Behavioral features. For contextual features, natural language processing NLP is used and for behavioral features, sentiment analysis is implemented. The successful detection of fake reviews lies in the construction of feature extraction and the performance of the classifier. Other type of fake reviews does exist, and they are generated by machines also called as bots. There are several machine learning algorithms which generates text data as if they were written by human [12]. Many fraudulent people do this malpractice of creating fake reviews with the help of bots. These bots written reviews looks like original reviews and it is difficult to distinguish between human written reviews and bot generated reviews.

2. Literature review

Despite the gravity of this issue, limited e-commerce websites have implemented fake review detection algorithms. The following literature work is some of the papers using different machine learning approaches for detection.

Literature 1: Creating and detecting fake reviews of online products [8]

Joni Salminen along with his four fellow researchers conducted a research to spot fake reviews. They created the fake review dataset using GPT-2 and ULMFiT models which created bot generated reviews and they merged them with original human written reviews from amazon. They experimented with several models like OpenAI, RoBERTa (Robustly Optimized BERT Pre-training Approach) model and SVM model. They split the dataset in 80/20 train/test split and the accuracy of RoBERTa model was the highest.

Literature 2: Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques [5]

E.I Elmurngi and A. Gherbi proposed a method with the use of 'Weka tool' to classify fake and original reviews from amazon reviews. They categorized reviews in three categories called positive, negative, and neutral sentiments. The major challenge they faced in research was differentiating unfair positive/negative reviews from opinion reviews. Researchers compared the performance of different classifiers like Naïve Bayes (NB), Decision Tree (DT-J48), Logistic Regression (LR) and Support Vector Machine (SVM) for sentiment classification and used performance measures like accuracy, precision, and recall.

Literature 3: Review spam detection [6]

In early research in 2007, Jindal, *et al*. studied review spam and spam detection. They collected 2.14 million reviews from Amazon for their research work. In their study, they faced large number of duplicate and near-duplicate reviews written by the same reviewers on different products or by different reviewers on the same products or different products. They performed spam detection based on duplicate findings and classification. They used logistic regression to train a predictive model and they got an average area under the ROC curve (AUC) value of 78%.

3. Techniques

Artificial Intelligence contains Natural Language Processing Techniques which is used by machines to understand human language. The goal of NLP is to learn, understand and use human languages to contact other humans. To generate product reviews which looks like it is created by another human, is a difficult task for a machine or bot. The natural language processing use NLTK - Natural Language Toolkit to read and understand any language by applying techniques like punctuation removal, stop words removal, tokenization, converting words to its lemma, stemming, etc.

Feature Engineering is an important task in Natural Language Processing because the machines or models cannot take raw text as input. The machine learning algorithms needs numbers or vectors as input to classify the data. The model like naive bays, support vector machine, they search for a partition between two classes by finding a hyper plane or a hidden pattern that separates two classes from each other. This can be made possible with the help of feature extraction.

A. Data Preprocessing

Step 1: Removal of Duplicate values

Some reviews are found as duplicates of already existing reviews. For examples, some reviewers may submit a review twice. The first task in the preprocessing is to find the duplicate reviews are there are total 12 duplicates found. They are removed from the Fake Review Dataset.

Step 2: Removal of punctuation marks

Most of the time the punctuation marks are used to show the surprise or disgust emotions in reviews. In some cases the punctuation marks or emojis can change the meaning of text, for most of the cases they are just used to make the statement complete. Therefore, white spaces and punctuations marks used in reviews are removed from the text.

Step 3: Tokenization

Reviews contains lengthy sentences which use slang words as well as spelling or grammar mistakes in them. The model just needs to use the core meaning of words from the review. For example, the lemma for bought or buying is 'buy'. The toolkit knows that the 'ing' suffix doesn't change the word. Therefore, the review words are converted into its pure lemma form to be better understood by ML algorithms using Porter Stemmer toolkit. The tokenized words are then stored into a list by the Porter Stemmer so they can be retrieved whenever needed.

Step 4: Removal of Stop words

Stop words like this, that, the, Where, When doesn't make an impact in detection algorithms. It is better to remove stop words and then feed to data to model to get more precise results. The stop words are also removed using NLTK toolkit. The techniques used to preprocess the data are: removal of duplicates, null values, removal of stop words and punctuations, converting review words to their original lemma to better understand meaning of it and converting text to lowercase so it would become easier to convert it into vectors.

B. Feature Extraction

After preprocessing techniques, feature extraction techniques are used to convert text data into vectors so it becomes more feasible to feed data to models. Two techniques called Vectorization (BOW model) and TF-IDF transformer are suited for feature extraction.

The process of turning text documents into numerical feature vectors is called vectorization [10]. The classifier models prefer features rather than plain text data. The occurrence of each word is counted, and its frequency is given as input to the classifier, this process is done by bag of words (bow) model which is mostly used in classification tasks [6]. Another technique used to find frequency as well as weight of a word in a document is called TF-IDF. According to the formula of TF- IDF transformer, the importance of a word increases in proportion to the number of times it displays in a document but decreases in inverse proportion to the number of times it appears in the whole corpus [11]. It means if a word appears 3 times in a single review, then it is an important word and heavy weight is

given to it, however if the same word appears 100 times in the whole dataset, the weight is decreased. Hence, we can say that the TF-IDF transformer perfectly calculates the importance of each word in review and feeds this numerical vector to the model [11]. We can see that bag of words model and TF-IDF transformer does a similar job of converting text into feature vectors but according to research, it is best suitable to first convert text into numbers with the help of BOW model and then calculate their importance with TF-IDF to search in the best feature space and it can lead to model performing with better precision.

C. Machine Learning Algorithms

After feature extraction, the next and main step is to classify the review in fake or real. Classification technique separates the items in target classes. Each review in the dataset is given as input to a classification model and a probability score is calculated to divide items in classes. Three most suitable models for natural language processing are chosen for this project which are:

Naive Bayes Classifier, Support Vector Machine and Random Forest Classifier

A. Naive Bayes Classifier

Naive-Bayes classifier is widely used in Natural language processing and proved to give better results for spam filtering because it uses conditional probability [6]. Naive Bayes technique uses Bayes theorem to determine probabilities.

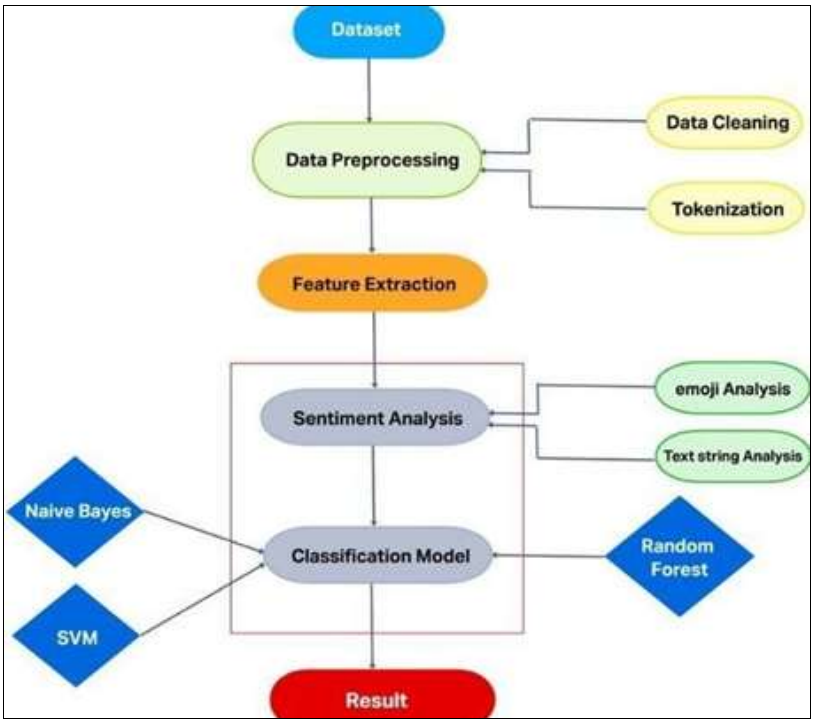
B. Support Vector Machine Classifier

SVM performs better classification by finding the hyper-plane that differentiates the classes plotted in n-dimensional space. Given labeled training data in supervised learning, SVM produces an optimal hyperplane that categorizes each new incoming review into two classes fake or original [9].

C. Random Forest Classifier

A random forest produces efficient predictions for big datasets that can be understood easily. The main concern of this algorithm is to create a forest with several trees. It creates an inner unbiased estimate of the simplification error as the forest building grows. It has an effective method for estimating lost data and maintains accuracy when a large proportion of the data are missing. Also, Random forests do not over fit [9].

4. Methodology

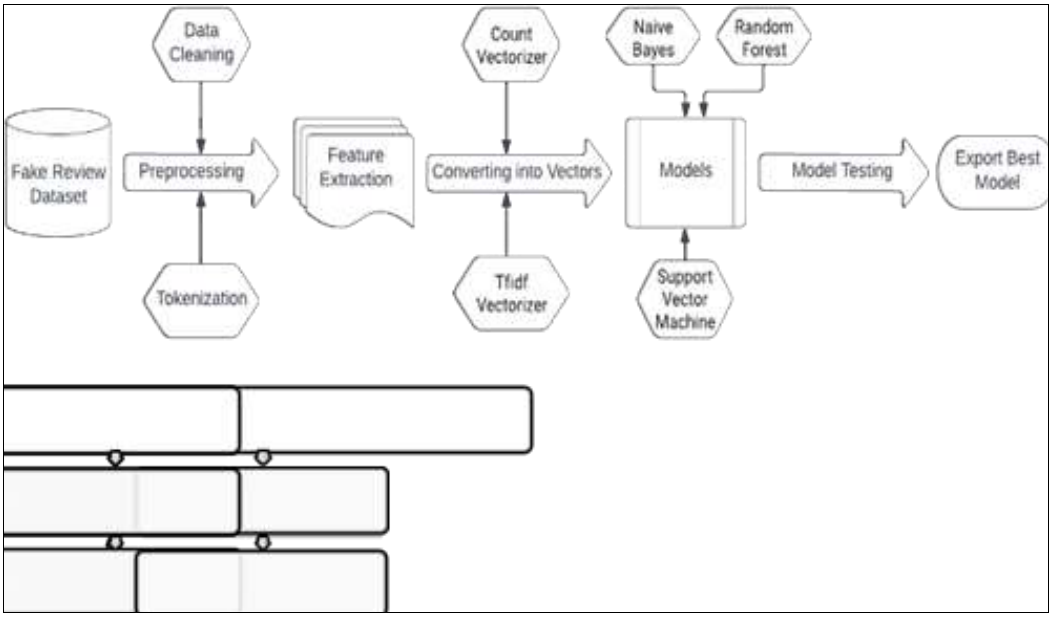


A jupyter notebook is used to record the methodology used in this research. The procedure of selection of best model follows six main steps:

- Use Fake Review Dataset and preprocess it to clean the data followed by saving it to a new csv file.
- Apply feature extraction techniques – Count Vectorization and TF-IDF transformation to convert text reviews into feature vectors.
- Split the fake review dataset into 60% training data and 40% testing data.
- Create a pipeline which includes feature extraction techniques and three models, which are Multinomial Naïve Bayes Classifier, Support Vector Classifier and Random Forest Classifier.
- Apply classifiers to training data so that using supervised machine learning techniques, classifiers can learn from the training data.
- Apply classifiers to test data and measure Accuracy, Prediction, recall score, F-1 score and confusion matrix to analyze the performance of three classifiers.
- Select the best performing model and export it using the pickle library.
- Using Flask framework, create a web application using HTML/CSS, Python and deploy it on server.
- Import the best model and load it into the web application where the user manually enters a review
- Let the model predict the review being fake or real.

5. System architecture

The system architecture diagram shows the backend procedure happening in jupyter notebook which includes data preprocessing like cleaning the data and tokenization techniques, converting text into vectors with feature extraction techniques like count vectorizer and tfidf transformer, training models like naïve bayes, support vector machine and random forest classifier, testing the models and after reviewing the results, exporting the best performing model.



The system architecture diagram also shows the frontend procedure including web application, the html form, input field, predict button, the function which takes text as input data and does preprocessing on it so it is ready to feed the imported best model.

6. Result

The classifiers run on test data and gives the accuracy score, precision, recall and f1-score. The confusion matrix for each classifier is also printed for better understanding of its performance.

The naïve bayes classifier is well known for spam detection and it is showing accuracy of 84.5%, recall of 84.5% and f-1 score of 84.5% too. Overall Naïve Bayes performed better on training as well as testing data.

The support vector machine classifier is well known for classifying data accurately by drawing a hyperplane between two classes. Here SVC shows accuracy of 89.16%.

The last classifier, Random Forest classifier is well known for large datasets as it uses large number of small decision trees to classify data. Here, the random forest classifier shows the accuracy of 86.10% which is lower than the SVM classifier.

Table 1: Support vector model, accuracy precision recall and f1 score

Model	Accuracy	Precision	Recall	F1-score
Naïve Bayes	85.13%	0.8210	0.8991	0.8583
Support Vector Machine	89.05%	0.9097	0.8673	0.8880
Random Forest	86.07%	0.8367	0.8969	0.8657

As we can observe from the results table, Support Vector Machine is the best performing classifier for the fake review detection dataset as it shows best accuracy, precision, recall and f1-score.

The SVM model is saved along with the pipeline of count vectorizer and tfid transformer in it. This model is now exported as best model using pickle library. Then it is imported into the web application.

Conclusion

- Detection of fake reviews is most relevant problem for today's online platforms.
- Fake review dataset showcased that reviews generated by bots appear so realistic that it is challenging for a human to detect them. The data visualization techniques helped in better understanding the dataset.
- However, machine learning classifiers like Naïve Bayes Classifier, Support Vector Machine and Random Forest Classifier provided high accuracy in detecting reviews generated by other machines. Support Vector Classifier detected fake reviews with highest accuracy of 89.16%.
- This research concludes that practically we can use AI to fight the issue of fake reviews.

References

1. Biswas B, Sengupta P, Kumar A, Delen D, Gupta S. A critical assessment of consumer reviews: A hybrid NLP-based methodology. Decision Support Systems. 2022 Aug 1;159:113799.
2. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. Journal of Big Data. 2015 Dec;2(1):1-24.
3. Liu B. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press; c2015.
4. Junylou Daniels AC. 2021. using-machine-learning-to-detect-bot-generated-fake-reviews. Objection.co.
5. Li Y, Feng X, Zhang S. Detecting fake reviews utilizing semantic and emotion model. In 2016 3rd international conference on information science and control engineering (ICISCE). IEEE. c2016. p. 317-320.

6. Rusland NF, Wahid N, Kasim S, Hafit H. Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. InIOP conference series: materials science and engineering. IOP Publishing. c2017;226(1):012091.
7. Zhou H. Research of Text Classification Based on TF-IDF and CNN-LSTM. Journal of Physics: Conference Series, International Conference on Computer, Big Data and Artificial Intelligence; c2021, 2171. (ICCBDAI 2021) 12/11/2021 - 14/11/2021 Beihai; c2021, 2171.
8. Joni Salminen CKG. Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services; c2021.
9. Kashti1 RP. Analysis of classifiers for fake review detection, International Journal For Technological Research In Engineering. 2019, 6(9).
10. Zhou V. Explanation of Bag-of-Words Model. Towards Data Science; c2019.
11. Baad D. sentiment-classification-for- restaurant-reviews-using-tf-idf. The Startup; c2020.
12. McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification, in Proceedings of AAAI-98 Workshop on Learning for Text Categorization, Pittsburgh, PA, USA. 1998;752(1):41-48.
13. Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F. Supervised Learning for Fake News Detection, IEEE Intelligent Systems. 2019;34(2):76-81.
14. Mukherjee A, Venkataraman V, Liu B, Glance NS. What yelp fake review filter might be doing?, in Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM; c2013. p. 409-418.
15. Wang G, Wang T, Zheng H, Zhao BY. Man VS. machine: Practical adversarial detection of malicious crowdsourcing workers", in Proceedings of the 23rd USENIX Security Symposium, San Diego CA; c2014.