International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

The Consequence of Fake Analyses on E-Commerce During and After Covid-19 Epidemic: SKL-Based Fake Analyses and Recognition

RAJALAKSHMI.C[1],
ASSISTANT PROFESSOR,
R. B. GOTHI JAIN COLLEGE FOR WOMEN

USHA J[2]
ASSISTANT PROFESSOR,
R. B. GOTHI JAIN COLLEGE FOR WOMEN

Abstract:
The epidemic of Covid-19 and the prosecution of lockdown, public isolation, and other defensive procedures lead to a global intensification in online shopping. The growing consequence of online shopping and widespread use of e-commerce has improved race among companies for online marketing. Highpoints that online evaluations show a important role in enhancing a business or smearing it. Product review is a vital feature in customers' decision-making, foremost to an powerful subject known as duplicitous or fake reviews recognition. Given these analyses' influence over a business, the traitorous acts of giving false analyses for individual improvements have enlarged with time. In our research, we proposed a fake review recognition model by using Text Sorting and techniques related to Machine Learning. We used classifiers such as Support K-Nearest Neighbor, logistic regression (SKL), and Vector Machine using a bigram model that detects duplicitous reviews based on the number of pronouns, sentiments, and verbs. Our future methodology for detecting fake online reviews outperforms on the squeal dataset and the TripAdvisor dataset compared to other state-of-the-art techniques with 95% and 89.03% accuracy.

Introduction

The global epidemic of Covid-19 at the start of the year 2020 leaves a substantial effect on everything and everyone. This eruption nerves the world and shifts the dynamics of e-commerce and online shopping. The prosecution of lockdown and public isolation lead the world to buy products online. One of the most persistent issues faced today is scam concerning customers' opinions on online products or services related to a brand or an organization [1], [2]. The matter has become more sophisticated and organized due to the profit achieved by such pursuit. This phenomenon is called "Estimation Spamming" [3], [4]. Dissimilar to other spam, opinion spam is a mite hard to detect as understanding the context is important to detect the deceptiveness of a review. These reviews are posted by people who are inexpert with the subject, which is why they are considered spam. Given the dynamic nature of the reviews, supervised learning techniques suffer from a few restrictions. [2], [5], [6]. Not until the "quality" of the review is known, a garbage-in-garbage-out [7] situation can transpire. In a study, [7] it was accentuated by the researchers that fake or genuine reviews are hard to label by humans. This confounds the exploration for the ground truth for given examples exactly. Due to the adaptable nature of these reviews and the lack of reliable data, according to the study [8], [9] methods were utilized to detect deceptive spam. Semi-Supervised techniques were used to improve classification. Millions of people are distributing their ideas on social media on several products, amenities, and events. Along with that, social media also consists of billions of short informal texts that may include SMS, tweets, messages, emails reviews, etc. This scenario has brought light upon the topic for researchers to look deep into Sentiment Analysis, Opinion Mining, and Review Analysis because these reviews are potent on any business's survival and downfall. For this reason, it is vital to detect their genuineness. As the popularity of the public web increases, several users will keep on spreading various kinds of content almost which lacks any trustworthy external source implying that there is no way of validating the content being posted. In the business section, this phenomenon affects an individual consumer and corrupts the confidence of a purchaser in online shopping. Identifying indicators of these fraudulent reviews based on the impostor's behavior is also an essential task. Due to this, a few scholars have utilized the techniques of Data Mining and Natural Language Processing (NLP) [8], [17] and other techniques such as data cleansing and database query processing to deal with raw data. However, these techniques did not efficiently solve the spam reviews problem. Lately, the reviewers have given plenty of new reviews every day. In this manner, information cleaning and repair will prompt flood in high business activity costs. As the legitimacy cannot be identified, it will not be in our interest to approve the database query process that filters those spam. Given the extensive use of social media, intense competition arises in which there is a vital role of consumer reviews which has a great impact on the online marketplace [8], [11], [18]. For improved decision-making, people and organizations need to improve decision-making before purchasing any product [9], [19], [20]. Writing duplicitous comments is mostly done by professionals the formations hire. These professionals are paid for which they post negative and positive comments on products or brands that are a major help in uplifting or defaming a targeted business [3]. However, these actions of a user could also end up being only a coincidence. One of the primary problems we are confronting today is detecting false reviews and the extraction of genuine emotion in an opinion. According to American research, 80% of buying performance be contingent on product comment. The problem is to determine if the feedback given is genuine or fraudulent. A supervised learning technique is proposed by initially studying the nature of the dataset. We did a thorough analysis of different types of approaches that are working in the same domain. Furthermore, we proposed a technique that shows more remarkable results than state-of-the-art methodologies. Fake reviews are the most pressing issue in the present era. It is one of the most intense topics because it impacts the business world considerably. The gain and loss of businesses partially depend on the feedback, especially in the e-commerce domain. Therefore, it is vital to determine their authenticity by using Machine Learning techniques such as K-Nearest Neighbor, Support Vector Machine, and Logistic regression (SKL). In presented a survey of existing models for fake reviews detection. According to this survey, SKL algorithms outperform the accuracy for the proposed problem. The Naive Bayes algorithm is one of the best classification algorithms of machine learning. However, the accuracy of the Naive Bayes algorithm for the detection of fake reviews is slightly less than SKL algorithms [21]. The proposed system includes the following modules;

1. Bi-gram philological model
2. Parts-of-Speech cataloging
3. Sentimentality Analysis
4. Review length and word count
5. Relationship word count
6. text classification based on Machine Learning.

False reviews detection techniques are widely used in the e-commerce domain, which plays an essential role in our economy as they can easily uplift or defame a product company or service. Since the purchase decision is firmly motivated by the reviews or ratings, the study shows that work has been concluded in detecting these fraudulent reviews, but spammers' demeanor is constantly developing. Spammers have been discreetly designing these false reviews to camouflage their malevolent intentions. Many businesses appoint professionals who write inappropriate positive and negative reviews for financial gains. These are fictitious comments that these professionals purposely write for the sake of seeming authentic. False reviews have a powerful impact as they directly influence customers' decision-making power. Relying on the feedback, customers either reject the product or decide to buy the product. False reviews are fictitious comments that are either machine-generated or user-generated. Both spams are challenging to identify. In the ongoing years, the use of e-commerce has increased drastically. There have been chances of fraudulent comments that play an essential role in deprecating or uplifting a business. Due to the strong opposition between organizations, it has become more erudite, and thus, many of them use the incorrect method to obtain potential revenue. Reviews on a product play a part in consumer decisions, and they build confidence in that particular product. However, they cannot be sure about the fallacy of these reviews. False reviews can either be deceptive or destructive. Destructive spams are easier to identify by a typical customer since they are non-review and contain ads and messages unrelated to the product. The latter, however, may contain sentimental reviews that may be positive or negative and, thus, problematic. Deceptive review, However, considering the deceptiveness of these reviews, these false reviews are being used to advance a business or tattle and harm the repute of businesses that are in great competition.

The existence of such reviews is crucial for the customer and the business. This concept is known as "Opinion Mining." To bring out the public's mood, Natural Language Processing is used regarding a specific product, service, or company. Untruthful Opinions are negative opinions given to damage the company's repute specifically or promote a business undeservingly. Likewise, positive opinions are given for an organization to gain fame inappropriately. Brand-specific reviews primarily target brands, get negative or positive reviews. Advertisements and inappropriate reviews have no sense and compromise of no opinion at all. Since the purchase decision is firmly motivated by the reviews or ratings, the study shows that work has been concluded in detecting these fraudulent reviews, but spammers' demeanor is constantly developing. Spammers have been discreetly designing these false reviews to camouflage their malevolent intentions. Many businesses appoint professionals who write inappropriate positive and negative reviews for financial gains. These are fictitious comments that these professionals intentionally write for the sake of seeming authentic. False reviews have a powerful impact as it directly influences the customer's decision-making power. Relying on the feedback, customers either reject the product or decide to buy the product. The product's price is a significant factor for a consumer, but feedback or reviews on those products are also considered seriously when purchasing something online. Building a trust factor is essential because most people rely on feedback to make a purchase online.

The subsequent aspects sum up the originality of this research: First, feature choice is based on a multi-level feature abstraction system. Besides the usual Natural Language Processing (NLP) on the corpus to abstract and feed features to the classifiers, this research projected numerous feature engineering methods to extract several behaviors of the reviewer himself and reviews. Additional, behavior features were also extracted for feature engineering. Behavioral features represent the statistical consequence of a user's analysis. They may not directly subsidize to the classification correctness; instead, they have connected to the reviewer himself. For example, writing style, review time, use of punctuation, verb/noun count in a review, and relationship words. All these features contributed to the results' overall classification accuracy and authenticity. Secondly, we tried to get the best fit training and testing dataset examples to get the best classifiers results.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Projected Methodology and Research Setup

We projected a support vector machine, K-Nearest Neighbor and Linear Regression (SKL) based algorithm for false reviews detection in the e-commerce industry. To fulfill our objective, we observed a dataset on hotel reviews and applied machine learning methods and text classification procedures to detect reviews that are not genuinely made. Figure 1 shows the phases of the projected methodology.

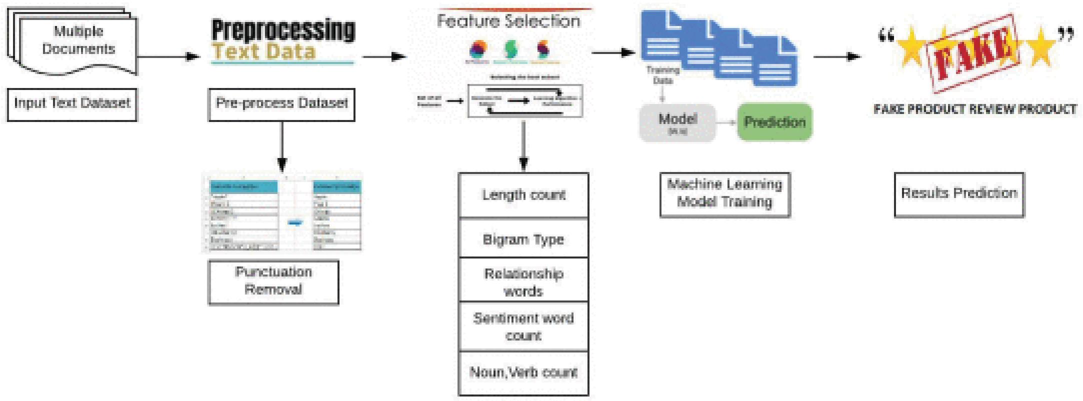

FIGURE 1.

Block Diagram of projected SKL base false reviews detection methodology.

Show All

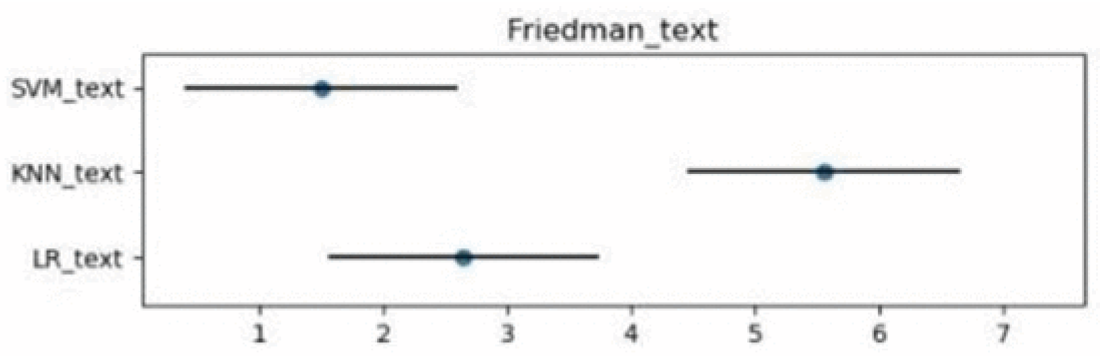

FIGURE 2.

Statistical Analysis based on Friedman test chart. The average order value is on the horizontal axis, while each algorithm is on the vertical axis. Each algorithm's average order value is revealed by a dot.

A. Preprocessing

Preprocessing data phase includes the filtering procedure. It represents the portion where we get rid of the text's less valuable parts, such as punctuation symbols. Punctuation marks such as,".! ?;.. etc. are eradicated because it lowers the overall correctness of the classification process. Their removal results in better output by the algorithm used. In order to complete this process, Natural Language Toolkit (NLTK) package is used. After successfully removing the punctuation words, word count is calculated. Selecting variables or identifying attributes to construct an efficient model is called feature selection. The objective of this process is to achieve a higher level of correctness. In our projected method, feature choice is based on the following parameters

1. Length count
2. Bigram Type
3. Relationship words
4. Sentiment word count
5. Noun, Verb count

Firstly, the total length of review is calculated, and then by using the bigram probability model, the probability of the next coming word is calculated. This is also called the Markov model, where you can define the probability of the next coming word without looking at it in the complete document. Some words describe the relationship just like husband, wife, sister, niece, etc. In-text classification analysis, we called these words relationship words. SKL selects features by considering relationship words. A list of relationship words is formed and used for SKL based projected solution. Feature selection also depends on the sentimentality of word count, whether it is a positive or negative word in a review. Corpora or bag of the word is created with positive and negative words. In order to figure out the sentimentality of a review word, those words match the pre-calculated corpora (positive or negative) of the given dataset. For this purpose, we have created positive and negative corpora, which contain approximately 2006 positive words and 4783 negative words. Part of speech (POS) tagging marks the corresponding word or part of speech in the given sentence. In our projected model, NLTK is used to tokenize the sentence, and then by using POS, they are tagged as noun-verb or adverb, etc. In the proposed model, Noun and verb counts are also calculated as part of the feature selection process. We split the dataset in an 80–20 ratio for training and testing samples, respectively. For more refined results, 10-fold cross-validation is done, leading to 95% and 89.03% overall classification correctness on the Yap and the TripAdvisor dataset.

B. Classification Algorithms

Classifying data into two or more than 2 classes/Labels is called classification. Machine Learning comes with many classification algorithms. In a recent study [21], they came up with a study report and determined that machine learning algorithms perform well on medium-sized datasets as related to Artificial Neural Network (ANN) and deep learning models, which are a good fit for large size datasets. Feature engineering is another reason ANN and deep learning models do not show better results for false review detection. Feature selection is an integral part of machine learning models training and plays a vital role in getting better classification results. However, in ANN, there is not a straightforward process for feature selection. ANN converts the word vector of analyses into a matrix. This matrix is fed to the convolutional layer by following different filters and forward outcomes to the pooling layer. ANN is a whole “black box” without any information about feature selection. Therefore, the classification algorithms we utilized are Support Vector Machine (SVM) with linear SVC (Support Vector Classifier) kernel to predict that either given product review is false or genuine. SVM is a pattern detection model in supervised learning, also associated with learning algorithms used for classification and prediction. It draws a decision boundary, also known as a hyperplane, near the extreme points of the dataset after identifying those extreme points inside the dataset. The K-Nearest Neighbors (K = 5) is also used for pattern recognition and classification. It is a straightforward algorithm. Its performance depends on many factors, such as the k parameter, an acceptable measure distance, and a majority voting scheme. In statistics, logistic regression (LR) is also a part of machine learning. It is a method that is used in binary classification. It uses the logistic function for prediction purposes. We further used logistic regression to predict the nature of a review. All three classification algorithms are part of false reviews detection.

Algorithm 1 Feature Selection for SKL Based Model.

Input:

True labeled training information t-train.txt as T

False labeled training information f-train.txt as F

Output:

Selected feature as SF.txt

Read T

Remove Punctuation

Produce genuine Review list as G

Read F

Eliminate Punctuation

Produce falseReview list as G′

Combine G and G′ as U

for *each review as x in U*, $\forall\ x \in U$ **do**

List.append(GetLengthOfReview)

List.append(GetNumberofBigramTypes)

List.append(GetNumberofRelationshipWords)

List.append(GetSentimentWordCount)

List.append(GetNumberofPronounsAndVerbs)

end for

trainfeatures.append(List)

if *rev*[‘‘*sentiment*″]==‘‘*True*″ **then**

trainfeatures.append(1)

else

trainfeatures.append(0)

end if

return SF.*txt*

C. Detection Procedure

After completing the training phase, the dataset will exam the model to envisage the output. Model is being skilled by SVM, KNN, and Logistic regression. The comparison table shows which algorithm outperforms for the certain process.

D. Experimentation Design/Details

We based our experiment on determining false reviews that play an essential role in the progress of online businesses. Dataset used in our proposed research is self-extracted using filtering method from oyo.com [27]. False reviews extracted from the oyo website are more realistic than deceptive datasets representing semi-real data. Moreover, false review recognition is more challenging with realistic datasets with overlapping between genuine and false review data. [21].

[24] developed a dataset for false reviews detection, they bypassed reviews having a length less than 150 characters. We assumed these facts from literature and created a considerably more extensive dataset, including mixed length reviews that vary from 452 words (1808 characters) to 15 words (60 characters). oyo dataset [27] is an imbalanced dataset. Moreover, this dataset is biased to positive reviews at the expense of detecting negative false reviews. We developed our self-extracted dataset from oyo’s data file to this effect. We used the filtering method to extract a subset of Yelp’s dataset and validated it with human judgments.

Two factors were assessed in the data abstraction process—first, mixed reviews. Secondly, biasness of any class (positive/negative). We confirm to take an equal number of positive and negative reviews to evade imbalance dataset problems.

1) Dataset Collection

We collected a dataset of 20 hotel reviews from oyo.com [27]. It includes 1800 reviews in which 800 are harmful, and the rest of the 1000 are positive reviews, which makes a balanced dataset. For building a model with good generalization performance, the best data splitting strategy is essential for every cataloging model, which is critical for model justification. For the performance valuation of a model, it is a practice to divide the available dataset into two subsets in the ratio of 4:1 for the training dataset and testing dataset. However, we may need to readjust this ratio most of the time to get improved performance. Moreover, the ratio may differ from one classification procedure to the other [30]. We split the dataset in an 80–20 ratio, of which 80% is the training set, and 20% is the test set. To assess the performance of the best suitable model for the proposed problem, we then split our data into 75–25 and 85–15 ratio split.

One of the challenging issues in false reviews detection is the availability of labeled datasets. It is observed that most of the available datasets are constructed based on a crowdsourcing framework. developed an opinion spam dataset with gold standard deceptive opinion. This self-generated dataset consists of 800 reviews from TripAdvisor. They collect positive reviews from the 20 most popular hotels from the Chicago area, including 5-star truthful reviews. while deceiving ideas are collected from the same 20 hotels using Amazon Mechanical Turk (AMT).

Conclusion and Future Work

Throughout this research, it has been observed that false reviews are certainly hard to block. Many studies have been working on this topic, but no study has given a one hundred percent outcome. Even in the current era, many ambiguities are not being addressed. We projected a methodology using text classification based on machine learning that helped determine whether the given remarks on a particular product/service are real or false. Our SKL technique proved to be more strong than already existing methodologies in the same field and proved to be more accurate. The results prove that SKL based false review provides 95% on Yelp dataset and 89.03% accurateness on TripAdvisor dataset compared to other state-of-the-art techniques. Since most of the researchers mainly intensive on a supervised learning process. Therefore, in the future, we would like to study Positive-unlabeled (PU) learning methods in depth, which is a semi-supervised learning method.