Fake Product Reviews Detection Using Multidimensional Representations With AI

COMPUTER ENGINEERING

SND COLLEGE OF ENGINEERING AND RESEARCH CENTER YEOLA

SAVITRIBAI PHULE PUNE UNIVERSITY

Miss.Thombare Nayana. S[1]	**Mr. Daund Ramesh P.**[2]	**Dr.Pawar Umesh B.**[3]
ME Student	**PG Co-ordinator**	**PG Co-ordinator**
SNDCOE&RC	**SNDCOE&RC**	**SNDCOE&RC**

Abstract:

Due to the growing trend of online shopping, many user shows intrest in shoping the prpduct they need from these online retailers. Customers do not need to spend a lot of time shopping in this manner. Because consumers want to know all the advantages and disadvantages of a product before making a purchase, online reviews in this instance are crucial to the product's sales. When making an online purchase, the majority of people require accurate product information. Customers can examine the website's numerous comments before spending money on a specific product. They were unable to determine if this was a real or fake situation. The customer orders just that specific product. The efficacy of the phony review detection model has been proven by the system. Since phony reviews are written with intent, it is usually hard to spot them. Many studies have been done on this topic, but none have yielded a satisfactory solution. Even in the present era, there are still a lot of holes that need to be filled. We propose a system based on machine learning-based text categorization to determine the authenticity of comments made on a certain product or service. Compared to prior efforts in the same field, this strategy was found to be more dependable and accurate. One method for spotting these kinds of fake reviews is matching learning. The use of computer data for machine learning

Keywords:

Web scraping, sentiment analysis, detection, feature extraction, logistic regression classifier, and fake reviews.

Introduction:

Due to the current pandemic, there has been a noticeable and rapid rise in e-commerce. For convenience, the public favors online shopping, e-banking, and other services. Customers can provide feedback about the service through e-commerce. Additionally, the existence of these reviews may serve as a resource for information for a potential new client. When a user shops online, they only purchase the product after reading reviews of it., the product will undoubtedly be judged incorrectly. Reviews fall into two categories, as we all know: real and fraudulent. There are good and bad fake reviews. Fake reviews come in a variety of forms. For example, when a seller posts a product for sale, he may ask his followers on social media to leave comments about it, even though the user did not actually purchase the item. thus these reviews are fraudulent. The system is designed to identify reviews of this kind. By utilizing the textual characteristics of the reviews, the system is able to identify fraudulent product reviews. For the implementation, the flipkart legal website's reviews dataset—which has several attributes and a large number of rows—is gathered. This system is developed using a logistic regression classifier. Various methods such as feature selection, tokenization, web scraping, pre-processing, etc. are employed in the development of this system.. By using this system, users can distinguish between reviews for products that are phony and those that are real. And only by reading real reviews can a user save time and arrive at an accurate assessment of the product. Lastly, we demonstrated the system's efficacy.

The planned system aims to:

1. Identify spam reviews and develop a user-friendly scheme.
2. Giving the user the ability to make an informed decision about the product is the goal..

Literature Review:

1) " A technique to identify phony reviews that relies on the seasonal nature of comments and reviews ": The author of this paper examined the review histories of online retailers and put forth a novel method for identifying phony product reviews. Through an analysis of The present investigation uses a technique based on the temporal trends in reviews and comments. detection method is able to identify the types of products. This method is more advantageous than some other methods that are currently in use.
2) The paper " A Structure for Identifying False Reviews: Problems and Difficulties " presents a fake comment commodity recognition method based on abnormal scoring behavior analysis. In order to achieve phony statement discovery, it uses combination detection based on the examination of delusional speak the conduct between static and dynamic characteristics. . Based on the experimental results, it is possible to identify fake comment targets for online commodities with this method.
3) "Manipulative product evaluation tracking system": In this case, the author examined the dataset made available by lawful websites. Afterwards, various methods, including feature selection, data mining, data cleaning, and web scraping, were applied to design a framework that could distinguish between genuine and phony product reviews.

Proposed System:

We developed a system that assists in identifying fraudulent or spammy product reviews. A range of machine learning approaches should be applied in order to implement this. To construct the model, an appropriate reviews dataset is used. Reviews are classified as real or fake using the most accurate model, or the best model. Several algorithms are used for classification after the model has been trained. Among the algorithms are Naïve Bayes and Logistic Regression. The pre-processed dataset is used to extract the features. After fitting each classifier, the models with the best performance were selected. Ultimately, the selected model was highly accurate and reliable in identifying spam reviews. The following Architecture depicts the proposed system (Fig 1).

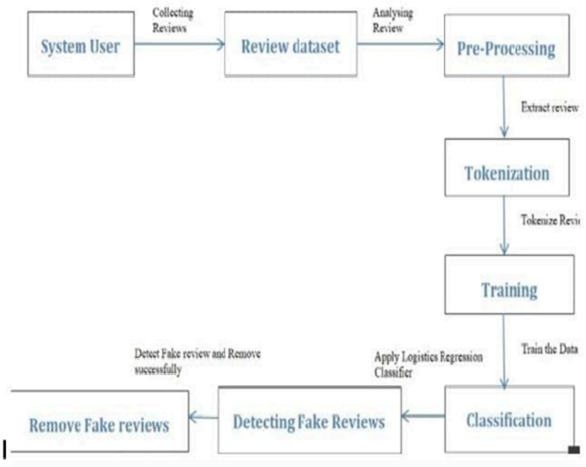


Fig.1 System Architecture

Fig1.System Architecture

Methodology:

1) **System User:**
Administrator is automatically registered by the system. The administrator must log in to the system and carry out any desired tasks. For normal users to access the system, they must first register and then log in
2) **Dataset:**

The user must gather the Flipkart review dataset. The dataset has thousands of rows and close to 14 attributes. The model is trained using datasets of this type. The attributes of dataset are:

1) URL: Taken from the Web

2) Review in bold: The reviews' titles

3) Stars: The review's assigned ratings.

4) Review: Summarize in a few paragraphs

5. Verified: Whether the reviewer is a confirmed purchase or not.

6) Date: The date of the review

7) By: User name

8) Profile_id: Id of profile

9) Most_rev: Maximum daily reviews per profile

10) Byline: Profile URL

11) Helpful: The quantity of individuals who annotated helpful

12) Goods: Brand name

13) URL to the product page for the product

3) Pre-processing:

Pre-processing of the dataset is done after it has been gathered. Pre-processing, put simply, is the process of transforming unprocessed data into a format that a machine can understand and use to build a machine learning model.

- **Feature Extraction:**

Simply put, feature extraction is the process of identifying pertinent data and eliminating noise. We prioritize the most important and practical features .We condense them into three concise lines. This ensures a focused and efficient approach . In this case, we have only taken into account two factors: reviews and reviews sentiment. The sentiment attribute holds the sentiment of the review in the form of a float value, and the review attribute is made up of product reviews. Every other column has been eliminated.

- **Data cleaning:**

Data cleaning refers to the procedure of rectifying or removing inaccurate, corrupted, improperly formatted, duplicate, or incomplete information from a dataset. Data is cleaned by using various NLTK libraries, such as stop words, punt, and word. Term repetitions such as "a," "an," and "the," and so forth are eliminated from reviews, along with various punctuation marks. Reviews also undergo lemmatization, which means that words with the same meaning are only taken into consideration once. As a result, we will receive well-organized data**.**

4) Tokenization:

Tokenization in Python essentially means dividing a longer text document into smaller lines, words, or even words for languages other than English. A module in NLTK named tokenize() further divides classification into two subcategories:

1) **Tokenize words:** To divide a sentence into tokens or words, we utilize the tokenize words() function.
2) **Sentence tokenization**: A document or paragraph can be separated using the message tokenize () function into sentences. This is where characteristics are tokenized, which means who reviews are broken up into digestible chunks.

5) Training:

The cleaned and arranged data needed In order to train the model, now to instruct tavailable. The procedure for building to instruct a model (brain) using previously acquired knowledge is called training. To train the model, various algorithms such as Naïve Bayes and logistic regression can be employed.

6) Classification:

The algorithms are now being used to train the model. Because the model has been trained, it can make decisions. It functions similarly to the human brain, which weighs past experiences and knowledge when making decisions. The model can now differentiate between reviews that are bogus and real, as well as their likelihood of being true.

7) Web Scrapping:

Web scraping is a method used to extract a large volume Extracting data from websites involves dealing with typically unstructured information presented in HTML format. This information is subsequently transformed into a structured format, such as a spreadsheet or another organized form.

This enables the data to be utilized in various applications. Web scraping has become an essential tool for both businesses and individuals, as it allows for the rapid and efficient collection of information from the internet. There are multiple approaches to web scraping, and in this case, the process is carried out using Beautiful Soup, a Python library known for its web scraping capabilities.

7) Detecting Fake Reviews:

Now that the website reviews have been retrieved, they have been thoroughly cleaned by eliminating punctuation, html parsers, etc. The system will identify user-provided reviews and determine whether they are authentic or fraudulent. It is predicted by utilizing the various functions.

8) Removing the fake reviews:

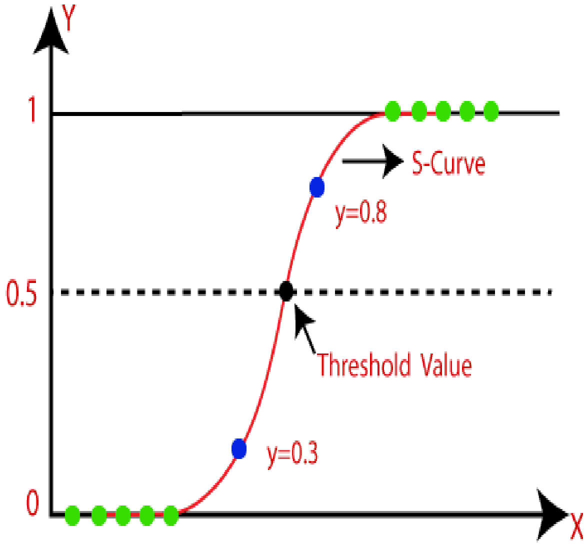
Now that the phony reviews have been identified by the system, it should be removed. This indicates that the remaining phony reviews are placed on the opposite side of the list containing the only real reviews. Thus, the phony reviews have been removed from the list of real.

Algorithms:

1) Logistic Regression:

The classification method used in unsupervised learning is called logistic regression. A binary classification model called logistic regression divides the outcome into two categories: true or false, or 1 or 0. It is used to forecast to use a set of various independent variables to train a dependent variable that is categorical. When developing a system with two categories, logistic regression is the best approach. It is mostly applied to categorization issues. One dependent variable, denoted as x in the logistic regression, and another dependent variable, denoted as y, are both present. The algorithm's input variable is x, and its output is y. In mathematical way,

y=f(x)



Assumptions for Logistic Regression:

- o The dependent variable ought to be classified.
- o Multicollinearity in the independent variable is not permitted.

Logistic Regression Equation:

Logistic Regression is a method employed in unsupervised learning for classification purposes. It serves as a binary classification model, predicting outcomes as either true or false, typically represented by 1 or 0. This technique is widely utilized for predicting categorical dependent variables by considering a specified set of independent variables. Logistic regression is especially useful in scenarios involving systems with two distinct categories. most of it employed in light addressing separating issues. There are two important variable in logistic regression independandent variable (x) the dependent variable (y). The input variable (x) is used as an input to the algorithm, while the output variable (y) represents the predicted outcome. Mathematically, this can be expressed as

y = mx + c

Where:

y is the dependent variable

x is the independent variable

m is the slope of the line

c is the y-intercept of the line

However, in logistic regression, we need to transform this equation to predict probabilities. To achieve this, we use the sigmoid function, which converts any real number to a value in the range of 0 and 1.efined as:

Sigmoid (z) = 1 / (1 + e^-z)

Where:

- z is the linear equation (mx + c)

By substituting the linear equation into the sigmoid function, we get the logistic regression equation:

P(y=1|x) = 1 / (1 + e^-(mx + c))

This equation shows the likelihood that, given the independent variable, the variable that is the dependent variable (y) will equal 1. (x). The logistic regression model then uses this probability to make binary classifications.

Also the linear equation can be expressed as follows:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- Since y can only be between 0 and 1 in logistic regression, let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

- To achieve a range from negative infinity to positive infinity, we can take the logarithm of the equation.

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots$$

This is the final equation for Logistic Regression.

Steps in Logistic Regression:

1) We will use the following procedures to implement the logistic regression using Python

2) Data pre-processing step

   - Fitting the training dataset to the Logistic Regression algorithm
   - Forecasting To verify the accuracy of the test results, we can create a confusion matrix.
   - Displaying the test set's outcome

3) Naïve Bayes:

The Naïve Bayes algorithm is a supervised learning method that uses the Bayes theorem for classification tasks. It is applied to the resolution of classification issues. The primary application of this algorithm is text classification with High-dimensional training data set This Naive Bayes Classifier, one of the simplest and most efficient classification algorithms, supports the rapid creation of machine learning models with fast predictive capabilities. Based on the Bayesian theory, the classification method is constructed with the assumption that a feature will always be present in any class, regardless of the presence or absence of other characteristics. It allows for the calculation of final probability.

Bayes' Theorem:

- The Bayes the theorem, commonly referred to as Bayes' Rule or Bayes' law, is a tool used to estimate the likelihood of a hypothesis being true. likelihood based on previous understanding. The probability with conditions determines this.

The Naïve Bayes theorem can be expressed mathematically as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

- **Probability of event A that is P(A|B)on observed event B is expressed as P(A|B).**
- **P(B|A): The likelihood that the proof supporting a hypothesis is true.**
- **P(A): Probability of hypothesis before observing evidence.**
- **Probability of Evidence, or P(B).**

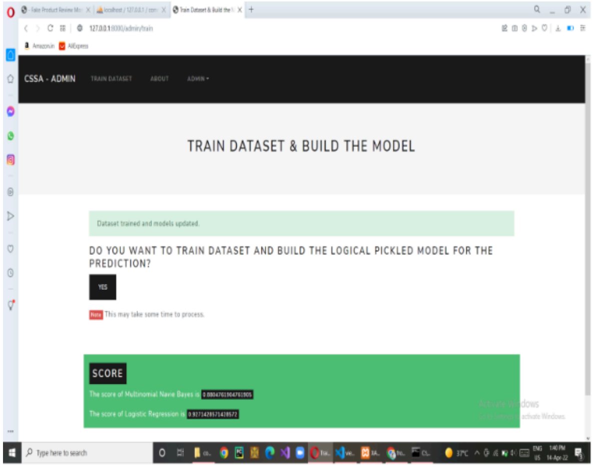
Put simply, we can determine the likelihood of when occurrence A will occur based on the fact that event B has already happened thanks to the Bayes theorem. It assists in revising our assumptions or probability in light of fresh data or observations. This is a commonly used theorem.

Steps to implement:

- Step of pre-processing data
- Adapting the method known as Naive Bayes to the Training dataset
- Determine the result which is tested.
- Verifying the test result precision (Confusion Matrix Creation)
- Presenting the test set result visually.

Result Analysis:

The goal of this claim is to develop a system that can accurately predict the kind of evaluates based on known review characteristics, such as originality or its untruthfulness The user must provide the system with the URL for the item being reviewed reviews. Following that, the system processes all of the input that has been provided and makes predictions using an algorithm that utilizes machine learning. The system's correctness is provided by testing accuracy, which, as Fig. 3 illustrates, is 88%. The admin portal serves as the training and accuracy determined is depicted in the figure3.



TRAIN DATASET & BUILD THE MODEL

Dataset trained and models updated.

DO YOU WANT TO TRAIN DATASET AND BUILD THE LOGICAL PICKLED MODEL FOR THE PREDICTION?

Note This may take some time to process.

SCORE

The score of Multinomial Navie Bayes is 0.8804761904761905

The score of Logistic Regression is 0.9271428571428572

Fig 2. Built model

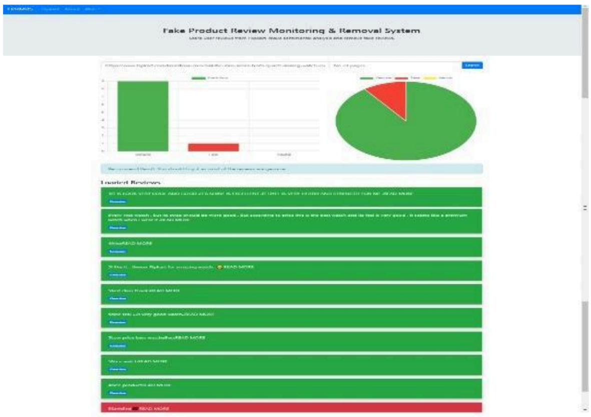


Fig.3 Review Detection

Figure 4 displays the project's actual outcome as a pie chart and bar graph. The reviews indicated in green are fraudulent, as are those highlighted in red. In a bar graph or pie chart, the green area represents the real data, while the red portion indicates fake data. A bar graph depicts the actual number of fake reviews, while a pie chart depicts the percentage format..

Conclusion:

The system has demonstrated the effectiveness of the fake review detection model. Since fake reviews are created on purpose, they are typically difficult to detectI.e. Numerous research have been conducted on this subject, but none have produced a perfect answer. There are still many gaps that are not being fixed, even in the modern day. To ascertain whether the comments left on a specific product or service are genuine or fraudulent, we put forth a methodology based on machine learning-based text classification. Compared to previously used approaches in the same field, this technique proved to be more reliable and accurate. Matching learning is used as part of the process to identify these types of phony reviews. Machine learning is the process of using computer data, models, and prediction. Also, the consumer provides participation to the system, and the reviews are classified into two types. categories **fake or genuine**. The model is trained using the appropriate dataset. The project's objective is to increase user satisfaction and trustworthy purchases. The individual who uses it also saves time and money. System precision has demonstrated the system's efficacy.

References:

[1] Wenqian Liu, Jingsha He, Song Han, Systematic Review of Deepfake Detection Literature BEDDHU MURALI2 SHOHEL RANA 1,2, (Member, IEEE), MOHAMMAD NUR NOBI3 , (Member, IEEE), , AND ANDREW H. SUNG2 , (Member, IEEE) date of publication February 24, 2022, Received January 25, 2022, accepted February 16, 2022, date of current version March 10, 2022

[2] Generating and identifying fraudulent reviews for online products. Chandrasekhar Kandpal, Joni Salminen a,b,* c , Ahmed Mohamed Kamel d , , Bernard J. Jansen a, Soon-gyo Jung a a Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar Turku School of Economics at the University of Turku, Turku, Finland Jaypee Institute of Information Technology, Noida, India Cairo University, Cairo, EgyptSource Journal: Journal of Retailing and Consumer Services

[3] Date of publication April 26, 2021Received April 1, 2021, accepted April 21, 2021, date of current version May 6, 2021. Digital Object Identifier 10.1109/ACCESS.2021.3075573 Fake Reviews Detection: A Survey ROBERT OLLINGTON1 RAMI MOHAWESH 1 , SHUXIANG XU 1, MATTHEW SPRINGER 1 , YASER JARARWEH 2 , AND SUMBAL MAQSOOD1, SON N. TRAN 1

[4] Date of current version March 10, 2022.Received January 17, 2022, accepted February 10, 2022, date of publication February 18, 2022, Digital Object Identifier 10.1109/ACCESS.2022.3152806 The Effect of Identifying Inauthentic Reviews in E-commerce Amidst and Post Covid-19: Detection Using SKL-Based Approaches for Fake Reviews M. USMAN ASHRAF , KHALID ALSUBHI , HINA TUFAIL , AND HANI MOAITEQ ALJAHDALI 4

[5] Hina Tufail, M. Usman Ashraf, Khalid Alsubhi, Hani Moaiteq Aljahdali. "The Effect of Fake Reviews on e-Commerce During and After Covid-19 Pandemic: SKL-Based Fake Reviews Detection" , IEEE Access, 2022 International Journal & Research Paper Publisher | IJRASET

[6] Meiling Liu, Yue Shang, Qi Yue, Jiyun Zhou. "Detecting Fake Reviews Using Multidimensional Representations With FineGrained Aspects Plan" , IEEE Access, 2021

[7] Survey on Various Tool for Analyzing and Detecting Fake Review by using AI Dr. Umesh B. Pawar1 , Prof. Daund Ramesh P.2 , Prof. Raindrop Pandit B.3 , Ms. Thom bare Nayna S4