iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42581

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Fake Reviews Detection Using Multidimensional Representations with Fine-Grained Aspects Plan

Mr. Abhale B. A.[1], Miss. Bachhav M. K.[2], Miss. patil Y. P.[3]
[1, 2, 3]*SND College of Engineering and Research Center YEOLA, Savitribai Phule Pune University*

Abstract: ***As the trend to shop online is growing day by day and lot of people are interested in purchasing the products of their need from the online stores. This way of shopping does not take a lot of time of a customer. In this case reviews on online websites play a important role in sales of the product because people try to get all the pros and cons of any product before they buy it. Most of the people needs genuine information about the product while online shopping. Before spending their money on particular product can analyse the various comments in the website. In this scenario, they did not recognize whether it may be fake or genuine. Customer place the order for particular product only by considering the reviews of that product. Here, it might be possible that reviews are fake. Now here query is which are fake reviews? Fake reviews may be good or bad compliment on the products. To detect such type of reviews we have developed the system. In this research, the dataset of different fake reviews provided by Flipkart are considered where reviews sentiments are included and using the LOGISTIC REGRESSION CLASSIFIER the reviews are classified into two categories i.e. fake and genuine. So user can save his/her time only by reading genuine reviews and gives accuracy about the product.***

Keywords: Fake reviews, review sentiment, logistic regression classifier, detection, feature extraction, web scrapping.

I. INTRODUCTION

Now a days because of pandemic situation, it is observed that there is very fast increase in e-commerce. Society prefers e-banking, online shopping, etc. for their convenience. E-commerce allows customer to give feedback about the service. And the presence of these feedback can become source of information to another new customer. In case of online shopping user buys the product only by reading the reviews of the particular product. That means reviews are playing very important role in online shopping. But in this scenario, if the reviews about the product are fake then it will definitely give wrong conclusion about the product. We know that reviews are of two categories i.e. genuine and fake. Fake reviews can be good or bad. There are different types of fake reviews like if seller post any product for selling he himself ask to his social members to comment on that product or sometimes user himself/herself did not buy the product just comment on it . so these type of reviews are fake. To detect such type of reviews the system is designed. The System can detect the fake reviews of the product by using the text properties of the review. The reviews dataset from the legal website flipkart is collected for the implementation which includes multiple attributes and number of rows. The logistic regression classifier is used to develop this system. Different techniques like pre-processing, feature selection, tokenization, web scraping, etc. are used while developing this system. Using this system user can differentiate the reviews of product in two categories i.e. fake or genuine. And only by reading genuine reviews list user first saves his/her time then get the accurate judgment about the product. And finally we proved the effectiveness of the system.

A. Purpose of Planned System

1) Emerging a user-friendly scheme foot detection of spam reviews.
2) The aim is to provide the user with the capacity to accurate judgment about the product.

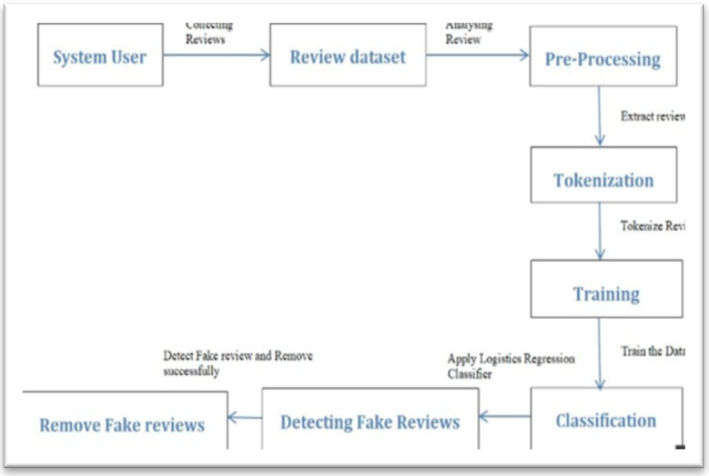
II. LITERATURE REVIEW

1) "A method for detecting of fake reviews based on temporal features of reviews and comments": In this paper, author studied the review records of online shopping sites and proposed a novel approach to detect fake reviews of products. This review detection method detects the type of products by analyzing the temporal trends of reviews and comments. Such perspective makes method more advantageous than some existing methods.
2) "A Framework for Fake Reviews Detection: Issues and Challenges", this paper puts Forward an abnormal scoring behavior analysis-based fake comment commodity recognition. On the basis of analyzing fake comment behavior, it adopts combination detection between static and dynamic feature to realize fake comment discovery. The experimental results show that this method can effectively detect fake comment target for online commodities.

3) "Fake product review monitoring system": Here the author studied the dataset provided by legal sites and after that performing different techniques like feature selection, data mining, data cleaning, web scrapping, etc. are used to develop the system which can differentiate fake and genuine reviews of the product.

III. PROPOSED SYSTEM

We projected this system which helps for detecting fake or spam reviews of the product. To put this into action, a variety of machine learning techniques should be used. A suitable dataset of reviews is used to build the model. The most accurate model i.e. the best model is utilized to categorize the reviews as genuine and fake. The model is trained and the different algorithms are used for classification. The algorithms includes Logistic Regression, Naïve Bayes. The features are extracted from pre-processed dataset. The best performing models after fitting all classifiers were chosen. Finally the chosen model was use in the identification of spam reviews with high accuracy and reliability. The following Architecture depicts the proposed system (Fig 1).



IV. METHODOLOGY

A. System User

System will register admin by default. Admin has to login the system and perform works he/she wants to do. Normal user should has to register first and then have to login the system to use it.

B. Dataset

User need to collect dataset of reviews from flipkart. The dataset contains near about 14 attributes and thousands of rows. This kind of dataset is used to train the model.

The attributes of dataset are:

1) URL: Scraped URL
2) Review bold: Title of the reviews
3) Ratings: Star ratings given by the review
4) Review: Review in paragraph format
5) Verified: reviewer is verified buyer or not.
6) Date: Date review was made
7) By: Name of user
8) Profile_id: Id number of profile
9) Most_rev: Max reviews made in day by profile
10) Byline: URLto profile
11) Helpful: Number of people who tagged helpful
12) Product: Product name
13) Product link: URL to product page

C. Pre-processing

Once dataset is collected, the pre-processing of the data is performed. Simply pre-processing is the process of converting raw data into suitable format which is understandable to machine and can be used for machine learning model.

1) *Feature Extraction:* Feature extraction is nothing but the process of selection relevant data and get rid from the noise. The features which are important and effective are taken under consideration only. Here, we have considered only two attributes i.e. reviews and reviews sentiment. The review attribute is consists of reviews about the product and sentiment attribute contains its sentiment in the form of float value. All other columns are removed.
2) *Data Cleaning:* Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Using the different NLTK libraries like stop words, punt, word net data is cleaned. The repeatable words like a, an, the, etc. Are removed from reviews, different punctuation marks are removed, lemmatization is done on reviews i.e. words with same meaning are considered only at once. So we will get proper organized data.

D. Tokenization

In Python tokenization basically refers to splitting up a larger body of text into smaller lines, words or even creating words for a non-English language. NLTK contains a module called tokenize () which further classifies into two sub-categories:

1) *Word Tokenize:* We use the word tokenize () method to split a sentence into tokens or words.
2) *Sentence Tokenize:* We use the sent tokenize () method to split a document or paragraph into sentences.

Here, the tokenization of features is performed i.e. reviews are splatted into small meaningful parts.

E. Training

Now the properly organized and cleaned data which is used to train the model is available. Training is the process of creating model (brain) based on the known information. The different algorithms like Logistic regression, Naïve Bayes Can be used for training the model.

F. Classification

Now the model is trained using the algorithms. The trained model means it has capacity to take the decisions. It is like human brain which considers the previous knowledge and experience and make decision. Now model is able to distinguish the reviews into two categories i.e. fake and genuine along with its truth probability.

G. Web Scrapping

Web Scrapping is the technique to obtain large amounts of data from websites. Most of this data is not in structured format in an HTML which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are various different ways to work on web scraping to obtain data from websites. Here, the web scraping is performed using beautiful soap in python.

H. Detecting Fake Reviews

Now the reviews from websites are fetched and are properly cleaned by removing punctuations, html parsers, etc. The system will detect the reviews provided by user to check whether genuine or fake. Using the different functions the prediction is made.

I. Removing the Fake Reviews

Here, the system has detected the fake reviews and further the system should remove them. That means the only genuine reviews are displayed in one list and remaining fake reviews are put into other side. So this is the removal of fake reviews from the list of genuine.

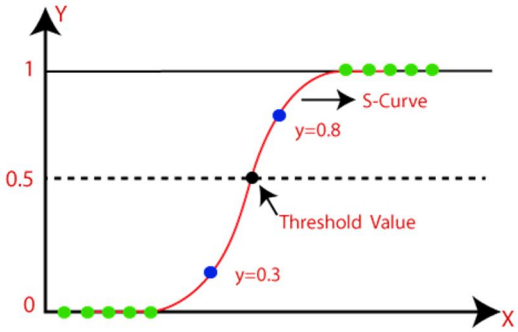
V. ALGORITHMS

A. Logistic Regression

Logistic Regression is the classification technique follows under the unsupervised learning. Logistic regression is the binary classification model which predicts the result into two values like true or false OR 1 or 0. It is used for predicting the categorical dependent variable using a given set of independent variables. When the system which results into two categories is developed the logistic regression is the best approach. It is mainly used for classification problems. The logistic regression contains two variables one is dependent variable and can call as x and another is dependent variable can call as y. x is input variable to algorithm and y is the output.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

In mathematical way,
y=f(x)



B. Assumptions for Logistic Regression

1) The dependent variable should be categorical in nature.
2) The independent variable must not have multi-collinearity.

C. Logistic Regression Equation

The Logistic regression equation can be derived from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- The equation of the straight line can be given as:

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

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

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

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

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

This is the final equation for Logistic Regression.

D. Steps in Logistic Regression

To implement the Logistic Regression using Python, we will use this steps:

1) Data Pre-processing step
2) Fitting Logistic Regression algorithm to the Training dataset
3) Predicting the test result
4) Checking the test accuracy of the result(Creation of Confusion matrix)
5) Visualizing the test set result

Naïve Bayes: Naïve Bayes algorithm is a type of supervised learning algorithm that is based on Bayes theorem. It is used for solving classification problems. This algorithm is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simplest and most effective Classification algorithms that helps for building the fast machine learning models which can make quick predictions. Based on the theory of Bayes, The classification technique is built and assumes feature is present in any class without affecting pf whether any other characteristic is present. It makes possible figure of final probability.

E. Bayes' Theorem

- Bayes' theorem is also called as **Bayes' Rule** or **Bayes' law**, that is used to find the probability of a hypothesis with previous knowledge. It depends on the conditional probability.
- The mathematical formula for Naïve Bayes' theorem is written as:

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

Where,

P(A|B): Probability of event A on the observed event B.
P(B|A): Probability of the evidence given that the probability of a hypothesis is true.
P(A): Probability of hypothesis before observing the evidence.
P(B): Probability of Evidence.

F. Steps to Implement

1) Data Pre-processing step
2) Fitting Naive Bayes algorithm to the Training dataset
3) Predicting the test result
4) Checking the test accuracy of the result(Creation of Confusion matrix)
5) Visualizing the test set result.

VI. RESULT ANALYSIS

The objective of this claim is to advance a system which knows about the type of reviews (genuine or fake) predict the reviews type with as a result with good accuracy. For that user need to give product reviews url as a input to the system. After that the system will perform all the processing on given input and using machine learning algorithm the predictions are made. The correctness of the system is given by testing accuracy which is 88% as shown in fig 3. The figure Shows the admin portal where the model is trained and accuracy is calculated

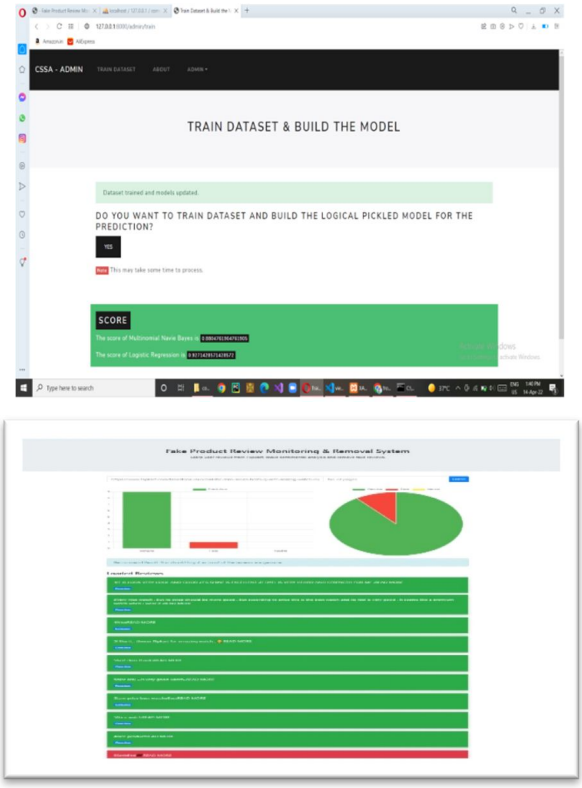


Fig.4

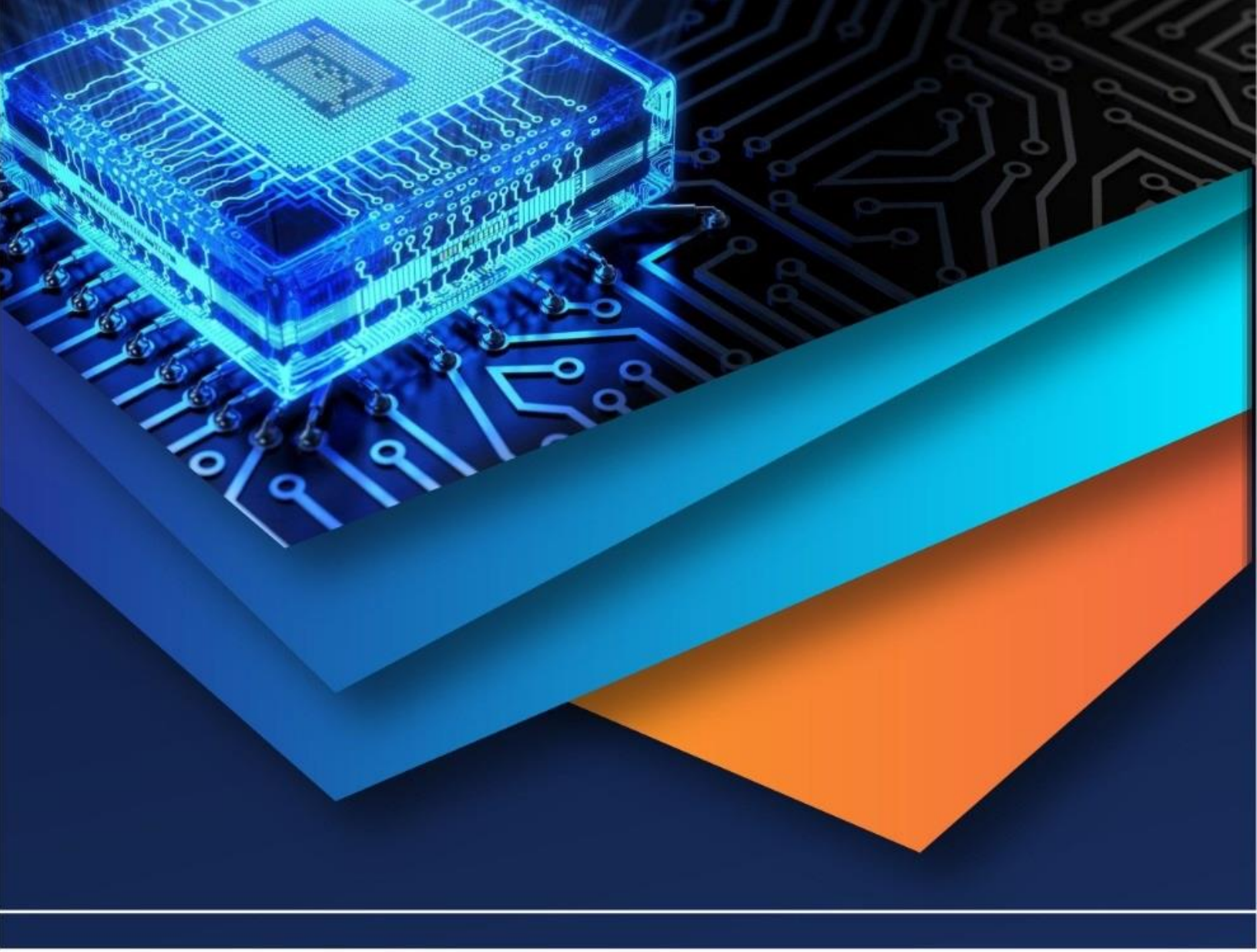
The fig.4 shows the actual result of the project in the form of bar graph and pie charts. The reviews highlighted by green colour are fake and the in red colour are fake. The green portion of bar graph and pie chart are the genuine and remaining are
Fake which are in red colour portion. Bar graph shows actual number of fake reviews and pie chart shows percentage format.

VII. CONCLUSION

The system has proved that the fake reviews detection model is works well. Fake reviews are not easy to +
Detect normally because they are made by someone purposefully. The technique for detecting such fake reviews is implemented by using matching learning. Machine learning simply computer data and model and makes prediction. Also here, the input is given to the system by user and the reviews are categorize in two categories fake or genuine. The suitable dataset is used to train the model. The goal of the project is to improve user satisfaction, as well as purchases to trustworthy. And also user saves money and time. System has proven its effectiveness by showing accuracy.

REFERENCES

[1] Wenqian Liu, Jingsha He, Song Han, “A method for detecting of fake reviews based on temporal features of reviews and comments” Faculty of Information TechnologyBeijing University of Technology, Beijing 100124, ChinaCorresponding author: znf@bjut.edu.cn
[2] Jitendra Kumar Rout “A Framework for Fake Reviews Detection : Issues and a Challenges,” KIIT Deemed to be University Bhubaneswar, INDIA. Email : jitu2rout@gmail.com
[3] Li Jing, “Online Fake Comments Detecting Model Based on Feature Analysis,” Guangxi University of Finance and Economics, Guangxi,Nanning,530003, 2018 IEEE
[4] Amit Sawan “Fake Product Review Monitoring And Removal”. , Department of Information Technology, Padmabhushan Vasantdada Patil Prathishtan’s College of Engineering, Mumbai, India
[5] Nazir M. Danish, “Fake Product Review Monitoring System.” 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE) Mexico City, Mexico. September 11-13, 2019
[6] Mupparam Sowjanya,K.Shnati latha,Ch.hyma,K.Naresh ,“Implementation of fake product review monitoring system and real review generation by using data mining mechanism.”



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY