iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45206

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 VII July 2022- Available at www.ijraset.com

Fake Review Detection on Using Machine Learning on Online Product Selling Platform

Jayalakshmi. L[1], Sneha. S[2], Subha Ilakiya. P[3], Kavi Bhaarathy. DA[4], Bhavani.N[5]
[1]*Head Of The Department Of Information Technology, Saranathan College of Engineering, Trichy, Tamilnadu, India*
[2, 3, 4, 5]*Department of Information Technology, Saranathan College of Engineering, Trichy, Tamilnadu, India*

Abstract: ***The increasing popularity of online review systems motivates malevolent intent in competing sellers and service providers to manipulate consumers by fabricating product/service reviews. Immoral actors use Sybil accounts, Bot farms, and purchase authentic accounts to promote products and vilify competitors. Facing the continuous advancement of review spamming techniques, more research work is been carried out to assess the approaches explored to date to combat fake reviews, and regroup to define new ones. Fake reviews detection attracts many researchers' attention due to the negative impacts on the society. Most existing fake reviews detection approaches mainly focus on semantic analysis of review's contents. This project is aimed at fake review detection in online platform, to prevent damage due to deceptive reviews.We propose a Novel Fake Reviews Detection based on Logistic Regression technique.***

Keywords: *Sybil Accounts , Authentic Accounts, Novel Fake Reviews, Logistic Regression*

I. INTRODUCTION

Machine Learning is defined as an application of artificial intelligence where available information is used through algorithms to process or assist the processing of statistical data. While Machine Learning involves concepts of automation, it requires human guidance. Machine Learning involves a high level of generalization in order to get a system that performs well on yet unseen data instances. Machine learning is a relatively new discipline within Computer Science that provides a collection of data analysis techniques. Some of these techniques are based on well-established statistical methods (e.g. logistic regression and principal component analysis) while many others are not. Machine learning might be able to provide a broader class of more flexible alternative analysis methods better suited to modern sources of data. It is imperative for statistical agencies to explore the possible use of machine learning techniques to determine whether their future needs might be better met with such techniques than with traditional ones.

II. OBJECTIVE

Online shopping is the technique where the customer buy the goods directly from the seller, over the internet. So, the customers make the decision of buyingthe product based on the reviews posted. The fake reviews can mislead the customers on buying the wrong product. The main objective here is to find the fake reviews. Most statistical techniques follow the paradigm of determining a particular probabilistic model that best describes observed data among a class of related models. Similarly, most machine learning techniques are designed to find models that best fit data (i.e. they solve certain optimization problems), except that these machine learning models are no longer restricted to probabilistic ones. Modern sources of data. It is imperative for statistical agencies to explore the possible use of machine learning techniques to determine whether their future needs might be better met with such techniques than with traditional ones.

III. LITERATURE SURVEY

Uttara M. Ananthakrishnan , Beibei Li , Michael D. Smith, A Tangled Web: Should Online Review Portals Display Fraudulent Reviews.The growing interest in online product reviews for legitimate promotion has been accompanied by an increase in fraudulent reviews. However, beyond algorithms for initial fraud detection, little is known about what review portals should do with fraudulent reviews after detecting them. In this paper, the question is addressed by studying how consumers respond to potentially fraudulent reviews and how review portals can leverage this knowledge to design better fraud management policies. Theoretical development from the trust literature is combined with randomized experiments and statistical analysis using large-scale data from Yelp. It is found that consumers tend to increase their trust in the information provided by review portals when the portal displays fraudulent reviews along with non-fraudulent reviews, as opposed to the common practice of censoring suspected fraudulent reviews.

©IJRASET: All Rights are Reserved

The impact of fraudulent reviews on consumers' decision-making process increases with the uncertainty in the initial evaluation of product quality. It is also found that consumers do not effectively process the content of fraudulent reviews (negative or positive). [7]

Daria Plotkina, Andreas Munzel, Jessie Pallud. Illusions of truth Experimental insights into human and algorithmic detections of fake online reviews. The issue of fake online reviews is increasingly relevant due to the growing importance of online reviews to consumers and the growing frequency of deceptive corporate practices. It is, therefore, necessary to be able to detect fake online reviews. An experiment with 1041 respondents two pools of reviews are created (fake and truthful) and compared them for psycholinguistic deception cues. The resulting automated tool accounted for review valence and incentive and detected deceptive reviews with 81% accuracy. A follow-up experiment with 407 consumers showed that humans have only a 57% accuracy of detection, even when a deception mindset is activated with information on cues of fake online reviews. Therefore, micro-linguistic automated detection can be used to filter the content of reviewing websites to protect online users. The independent analysis of reviewing websites confirms the presence of dubious content and, therefore, the need to introduce more sophisticated filtering approaches.

Yuanyuan Wu,Eric W.T.Ngai, Pengkun Wu, Chong Wu. Fake online reviews: Literature review, synthesis, and directions for future research . Fake online reviews in e-commerce significantly affect online consumers, merchants, and, as a result, market efficiency. Despite scholarly efforts to examine fake reviews, there still lacks a survey that can systematically analyze and summarize its antecedents and consequences. This study proposes an antecedent– consequence–intervention conceptual framework to develop an initial research agenda for investigating fake reviews. Based on a review of the extant literature onthis issue, we identify 20 future research questions and suggest 18 propositions. Notably, research on fake reviews is often limited by lack of high- quality datasets.

Manqing Dong, Lina Yao,Xianzhi Wang, Boualem Benatallah. Opinion fraud detection via neural auto encoder decision forest. Online reviews play an important role in influencing buyers' daily purchasd decisions. However, fake and meaningless reviews, which cannot reflect users' genuine purchase experience and opinions, widely exist on the Web and pose great challenges for users to make right choices. Therefore, it is desirable to build a fair model that evaluates the quality of products by distinguishing spamming reviews. An end-to-end trainable unified model to leverage the appealing properties from Auto encoder and random forest is presented. A stochastic decision tree model is implemented to guide the global parameter learning process. Extensive experiments were conducted on a large Amazon review dataset. The proposed model consistently outperforms a series of compared methods.[11]

Dushyanthi U. Vidanagama, Thushari P. Silva & Asoka S. Karunananda. Deceptive consumer review detection. Consumer reviews are considered to be of utmost significance in the field of e- commerce, for they have a stronghold in deciding the revenue of a business. When arriving at a purchasing decision, a majority of online consumers rely on reviews since they offer credible means of mining opinions of other consumers regarding a particular product. The trustworthiness of online reviews directly affects a company's reputation and profitability. Such generation of deceptive reviews which manipulate the purchasing decision of consumers is a persistent and harmful issue.

IV. EXISTING SYSTEM

The objective of the support vector machine algorithm is to find a hyper plane in an N-dimensional space (N-the number of features) that distinctly classifies the data points. This algorithm consist of a target/outcome variable or dependent variable which is to be predicted from a given set of predictors (independent variables).Using these set of variables, we generate a function that map inputs to desired outputs.The training process continues until the model achieves a desired level of accuracy on the training data.

V. PROPOSED METHODOLOGY

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logistic regression)is estimating the parameters of a logistic model (a form of binary regression). It isa classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logistic function. Hence, it is also known as logistic regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected). Mathematically, the log odds of the outcome are modeled as a linear combination of the predictor variables [2].

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

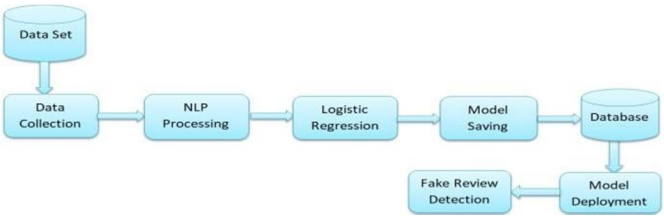
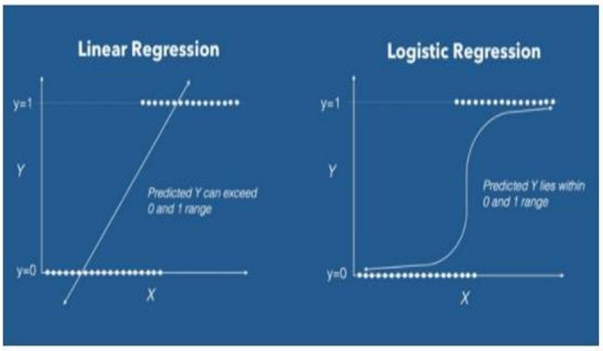


Fig 1.1 Architecture Diagram

The steps involved in system architecture are data extraction from an online product selling application(API) and a dataset having positive and negative tweets then using NLP preprocessing to tokenization separate the each word from sentences and removing the white spaces, null values and punctuations. We implement the logistic regression algorithm used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s).Build the model based on sentimental collection of positive negative tweets and stored in database then, accuracy and validation graph is analyzed in proposed system and validate the result to detect the fake reviews.

A. Logistic Regression

Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).



B. Preprocessing

The preprocessing steps are done based on the tokenization and stop words.

C. Tokenization

Processing and refining the data by removal of irrelevant and redundant information as well as noisy and unreliable data from the review dataset. Sentence tokenization The entire review is given as input and it is tokenized into sentences. Removal of punctuation marks and used at the starting and ending of the reviews are removed along with additional white spaces. Word Tokenization Each individual review is tokenized into words and stored in a list for easier retrieval.

D. Stop-words

Removing meaning less word from the sentence. One of the most important aspects of analyzing data is to ensure that our data is being understood by machines. Machines do not understand text, images, or videos , they can comprehend only 1's and 0's. To be able to provide an input consisting of 1's and 0's is a multistep process. Pre-processing the data is an absolute necessity and calls for a technique called data cleaning which involves transforming raw data into a machine-understandable format we are delete the rows and columns of data from data set. For example, the stem of "cooking" is "cook", and the stemming algorithm knows that the "ing" suffix can be removed.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. CLUSTERING

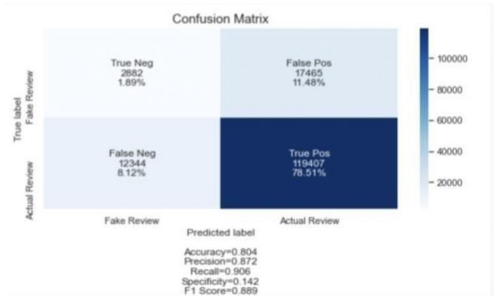
K means segmentation algorithm is implemented for clustering the positive reviews and negative reviews .Data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster. Because of the iterative nature of K-Means and the random initialization of centroids, K-Means may become stuck in a local optimum and fail to converge to the global optimum. As a result, it is advised to employ distinct centroids ' initializations .Based on group of collection we can cluster the group of words and sentence.

VII. FEATURE EXTRACTION

After collection of data set, stemming tokenization is applied based on sentimental words collection were build the model. A reviewer posting multiple reviews with the same Reviewer ID. Fake reviews in most scenarios have 5 out of 5 stars to entice the customer or have the lowest rating for the competitive products thus it plays an important role in fake detection.Purchase reviews that is fake have lesser chance of it being verified purchase than genuine reviews. Thus the combination of features are selected for identifying the fake reviews. This in turn improves the performance .We build the model based on sentimental collection of positive and negative tweets.

VIII. DIFFERENTIATING THE FAKE AND REAL CONTENT

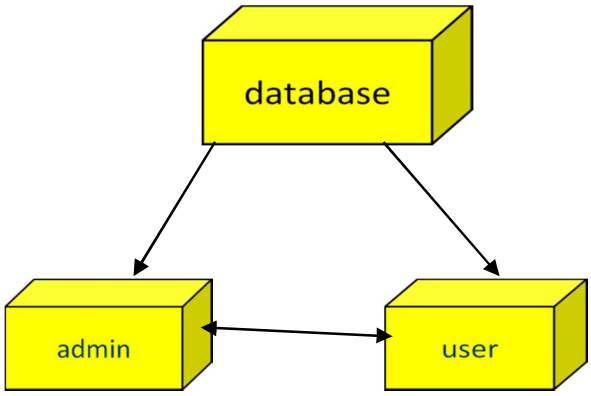
The output is the differentiation of fake and real review in the online productplatform as form of confusion matrix.



Once the user gives the review as the input, data undergoes the preprocessing stage, where the words get stemmed and tokenized. Next, the data is given to the Sentimental Analyzer from where the sentiment of the review is detected based on the positivity of the review.

IX. FAKE REVIEW DETECTION

Online reviews are less trustworthy than we think. The credibility of all reviews even real ones is questionable. A 2016 study published in The Journal of Consumer Research looked at whether online reviews reflected objective quality as rated by Consumer Reports. The researchers found very little correlation.

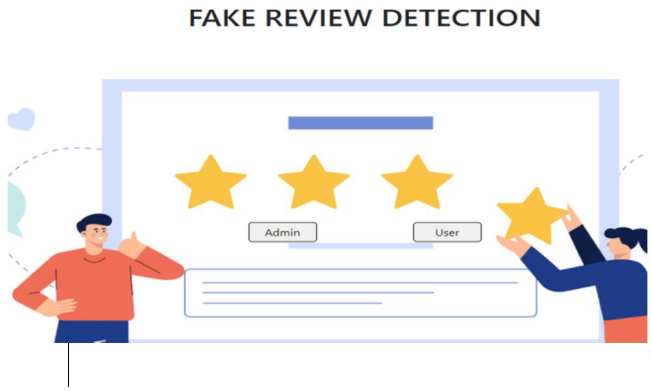


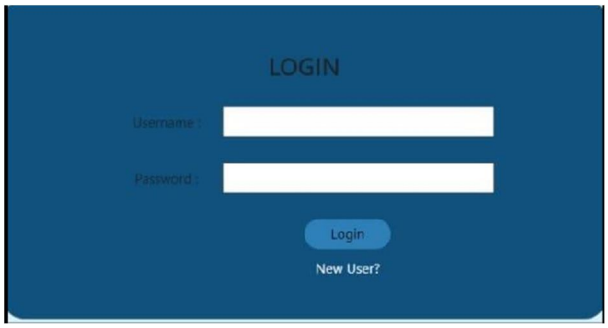
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

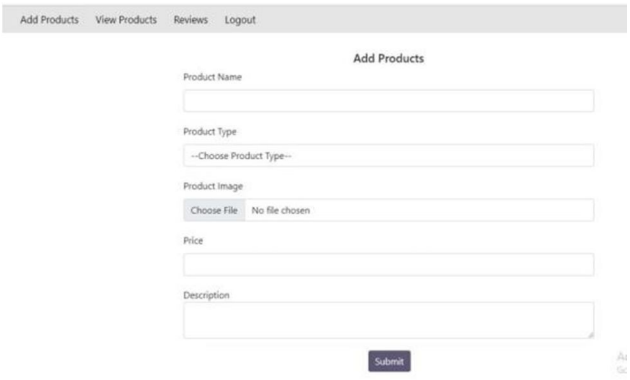
X. ADVANCED PREDICTIVE ANALYTICS

Predictive analytics tools are powered by several different models and algorithms that can be applied to wide range of use cases. Determining what predictive modeling techniques are best for the organization is a key to getting the most out of a predictive analytics solution and leveraging data to make insightful decisions in the statistical context.Typically, an organization's data scientists and IT experts are tasked with the development of choosing the right predictive models or building their own to meet the organization's needs. Today, however, predictive analytics and machine learning is no longer just the domain of mathematicians, statisticians and data scientists, but also that of business analysts and consultants. More and more of a business' employees are using it to develop insights and improve business operations – but problems arise when employees do not know what model to use, how to deploy it, or need information right away. At Statistical Analysis System (SAS), sophisticated software is developed to support organizations with their data governance and analytics. The data governance solution helps organizations to maintain high-quality data, as well as align operations across the business and pinpoint data problems within the same environment. Our predictive analytics solutions help organizations to turn their data into timely insights for better, faster decision making. These predictive analytics solution is designed to meet the needs of all types of users and enables them to deploy predictive models rapidly.

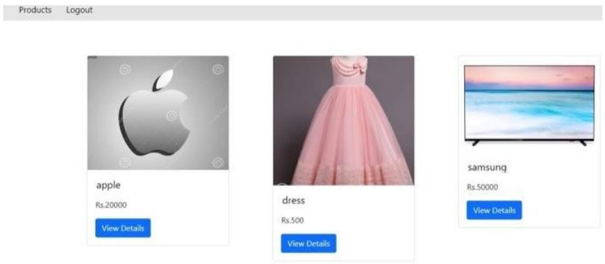
XI. RESULTS





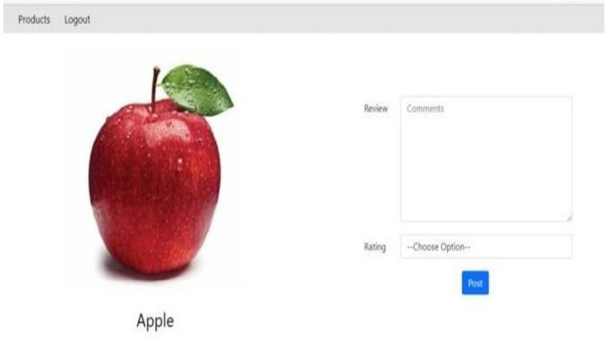


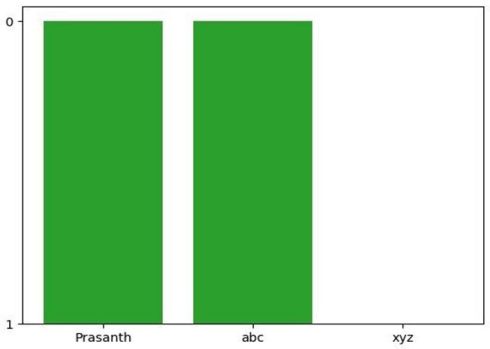
©IJRASET: All Rights are Reserved







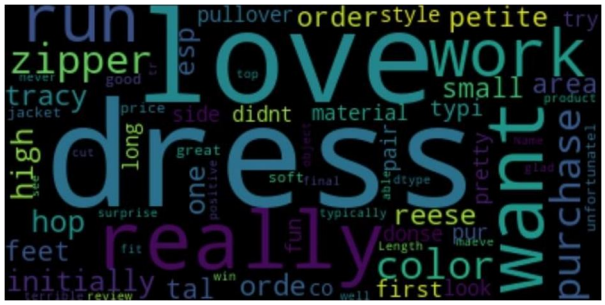




©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



Word Cloud of the Recommended Reviews



Word Cloud of the Non Recommended Reviews

XII. CONCLUSION

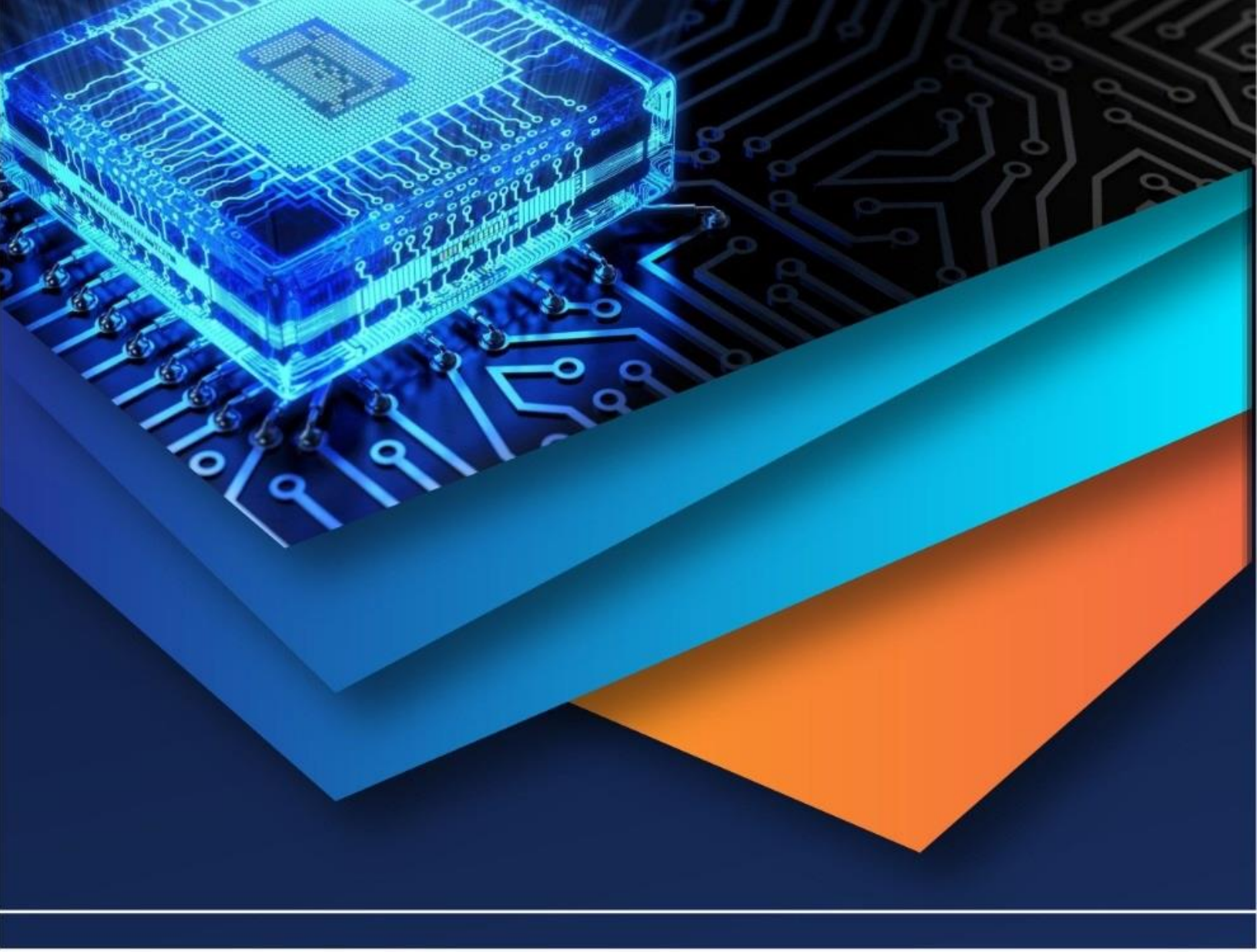
The Fake reviews detection attracts many researcher's attention due to the negative impacts on the society. LOGISTIC REGRESSION classification provided a better accuracy of classifying than the Svm classifier for testing dataset. Revealing that it can generalize better and predict the fake reviews efficiently .In Fake review detection ,the problem of classifying fake review using machine learning are implemented. Data processing is done by Stemming, Tokenization and vectorization is done by using TF-IDF vectorizer. Logistic regression technique is used to train the model.

XIII. FUTURE ENHANCEMENT

Fake review detector detects only the linguistic-based information as real or fake.In this , reviews are given as the input and the output is delivered as real or fake.The future enhancement of this project is to develop a similar process for unsupervised learning for unlabeled data to detect fake reviews and to get more accuracy on the fake review detection. Finally, additional research and work to identify and build additional fake review classification is ongoing and should yield a more refined classification scheme for fake reviews in a dynamic manner.

REFERENCES

[1] Aghakhani H, MacHiry A, Nilizadeh S, Kruegel C, Vigna G (2018) Detecting deceptive reviews using generative adversarial networks. In: Proceedings of 2018 IEEE symposium on security and privacy workshops, pp 89–95. 1805.10364v1.
[2] Akoglu L, Chandy R, Faloutsos C (2013) Opinion fraud detection in online reviews by network effects.In: Proceedings of 7th international conference on weblogs and social media, pp 2–11.
[3] Allahbakhsh M, Ignjatovic A, Benatallah B, Beheshti SMR, Bertino E, Foo N (2013) Collusion detection in online rating systems. In:Proceedings of Asia-Pacific web conference, Springer, vol 7808 LNCS, pp 196–207.
[4] Algur SP, Patil AP, Hiremath PS, Shivashankar S (2010) Conceptual level similarity measure based review spam detection. In: Proceedings of 2010 international conference on signal and image processing, pp 416–423.
[5] Dushyanthi U. Vidanagama, Thushari P. Silva & Asoka S.Karunananda (2020) Deceptive consumer review detection: a survey.
[6] Daria Plotkina, Andreas Munzel, Jessie Pallud(2020) Illusions of truth Experimental insights into human and algorithmic detections of fake online reviews.
[7] Fusilier DH, Montes-y Gómez M, Rosso P, Guzmán Cabrera R(2015a) Detecting positive and negative deceptive opinions using PU-learning. Inf Process Manage 51(4):433–443.
[8] Hai Z, Zhao P, Cheng P, Yang P, Li XL, Li G (2016) Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: Proceedings of 2016 conference on empirical methods in natural language processing, pp 1817–1826.
[9] Manqing Dong, Lina Yao,Xianzhi Wang, Boualem Benatallah,Chaoran Huang, Xiaodong Ning(2020) Opinion fraud detection via neural auto encoder decision forest.
[10] Uttara M Ananthakrishnan, Beibei Li, Michael D Smith(2020) A tangled web: should online review portals display fraudulent reviews.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)