Online Genuine Review Prediction Using a Novel Ranking-Based Technique By comparing Logistic Regression with Random Forest.

G.Priyanka[1], Devi.T[2]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India, PinCode: 602105

[2]*Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu ,India, PinCode: 602105

ABSTRACT

Aim: The aim of the research work is to detect the fake reviews in online products using 1 machine learning by comparing Random Forest over Logistic Regression. **Materials and Methods:** The study contains two groups i.e Random Forest algorithm is developed in the first group and Logistic Regression is developed in the second group. The categorizing is performed by adopting a sample size of n = 10 in Random Forest and sample size n = 10 in Logistic Regression algorithms with a sample size = 10. **Results and Discussion:**The analysis of the results shows that the Random Forest has a high accuracy of (87.7%) in comparison with the Logistic Regression algorithm (81.2%). There is a statistically significant difference between the study groups with Significant value= 0.144 ($p<0.001$) and having G power of 80%. **Conclusion:** Prediction in classifying an end user's fake reviews in online products shows that the Random Forest appears to generate better accuracy than the Logistic Regression algorithm.

Keywords: Reviews, Fake review detection,Logistic Regression, Machine learning,Novel Ranking,Random Forest.

INTRODUCTION

The purpose of the research work is to detect the fake reviews in online products by implementing novel ranking using online characteristic technique(B and Mallikarjuna 2020). In this pandemic it was absorbed by a dramatic shift from in-person to online shopping. Online shopping can save time for both buyer and seller with a lot of reasons why customers today prefer shopping online (Sharma 2021). It is found to be important in today's world since online shopping is more convenient for the customers with respect to various scenarios (Sihombing and Fong 2019). Customers search for various good products, this is due to the large number of products in the world based on novel ranking .Customers observe the views of different people to make decisions. This helps the customer to find the good products by comparing the reviews which are posted below the product. The applications of online marketing are based on reviews and ratings which plays a major role. When a brand and its products are posted with many positive reviews and high ratings, the customers who wish to buy those products would be more likely to buy without any doubt (Muhammad and Ahmed 2019). Similarly

in online marketing , when a product of any brand has got many negative reviews and very poor ratings then any customer of the shopping site would try to avoid buying those products.

In distinguishing and forecasting fake review detection in online products using characteristic technique by comparing clustering 990 journals from IEEE Xplore digital library, 460 articles from ScienceDirect, 1360 articles from google scholar and 498 articles from Springer. In cluster analysis , the clustering algorithm and the cluster validity are highly correlated. In this work, the validity index is introduced as a ranking for cluster validity and a clustering based on it (V., Aishwarya, and Sultana 2020). Among all the articles and journals the most cited paper is nearly 1000 people cited these articles and this article is useful for future research. Every customer buys a product after viewing the reviews given by the previous customers. Many studies have shown that reviews play a major role in impacting the brand reputation in the market (Jnoub, Brankovic, and Klas 2021). To prevent this opinion spam, Amazon India has introduced a policy of limiting the number of reviews posted for a product in a day. There are many studies and many approaches for detecting fake reviews in the online review system. Previous studies prove that reviews and ratings prove to affect user sentiments to a phenomenal effect (Patel and Patel 2018) had done a pioneering effort in detecting extremist reviewers at the brand level and they attempted to identify groups involved in posting extreme reviews at the brand level which is used in various classification techniques and used various features for classification (S. and A. 2018). It includes average rating, average upvotes, average sentiment, review count, and many other features. The main objective of this research is to detect the fake reviews in online products.(Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020)

Previously our team has a rich experience in working on various research projects across multiple disciplines(Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).The methods which were used before have less accuracy and detection rate in detecting fake reviews in online products. The staging of the Logistic Regression method is lagging to find the fake reviews of a particular product or a brand. In order to sequence the methods and techniques in this research study Random Forest generally fares better than Logistic Regression (Shu and Liu 2019). It also takes more time to train a Logistic approach for analyzing fake review datasets which are based on the model to increase with the size of the datasets based on novel ranking. Its estimates and calculations are probably done using a characteristic technique.The aim of the research work is to detect the fake reviews in online products using machine learning and comparing ML techniques Random Forest over Logistic Regression.

MATERIALS AND METHODS

The research work was performed in the Image Processing Lab, Department

of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. Basically it is considered that two groups of classifiers are used, namely Random Forest and Logistic Regression algorithms, which are used to detect the fake reviews in online products. Group 1 is the Random Forest Algorithm with the sample size of 10 and the Logistic Regression is Group 2 with sample size of 10 and they are compared for more Accuracy score and Precision score values for choosing the best algorithm. The Pre-test analysis has been prepared by having a G power of 80% and threshold 0.05%, CI 95% mean and standard deviation (S.Akash 2021). Sample size has been calculated and it is identified that 10 samples/ group in total 20 samples with a standard deviation for Random Forest = 87.7% and Logistic Regression = 81.12%

The dataset is used for observing both the algorithms . The dataset is checked and verified for any empty values. A minimum of 4GB ram is required to use and install all the necessary applications, and a minimum of i3 processor to run all the applications and processes simultaneously. A minimum of 50GB hard disk space is required to install the required software and files, and an internet connection is required to download and install all the necessary software and development environment to run this Novel Effective framework. Python programming language is used for the application. The version of python used is 3.9, and the IDLE is used to run and execute the application.

Random Forest

Random forest is a Machine Learning algorithm which is used for classification problems, it is a predictive analysis algorithm and based on the concept of probability it will make an exact sequential identification using characteristic technique as shown below.

Pseudocode for Random Forest

Import random forest Classifier

Import random forest as rm

Compare from sklearn.ensemble import RandomForestClassifier

Skip from sklearn.linear_model import random forest

Data extraction from sklearn import svm

Initiate sklearn.metrics import accuracy score

Calculate sequence from sklearn.model_selection import train_test_split

Find results from sklearn.feature_extraction.text import CountVectorizer

count_vectorizer =CountVectorizer(stop_words='english')

cv = count_vectorizer.fit_transform(data['Clean_TweetText'])

cv.shape

Rm= Random forest()

Rm.fit(X_train,y_train)

prediction_Rm = Rm.predict(X_test)

print(accuracy_score(predi Rm action_rm,y_test))

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands

for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as Fake news detection,spam detection,Fake review detection. In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1

Pseudocode for Logistic Regression

Import LogisticRegression as LGR
Import pandas as pd
Import Matplotlib.pyplot as plt
Compare from sklearn.ensemble import Logisticregression
Data extraction from sklearn
Initiate sklearn.metrics import accuracy score
Calculate sequence sklearn.mode_selection import train_test_split
Find results from sklearn.feature_extraction.value import CountVectorizer
count_vectorizer=CountVectorizer()
cv.shape
lgr=LGRclassifier()
lgrc.fit(x_train,y_train)
prediction_lgr=lgrc.predict(x_test)
print(accuracy_score(prediction_lgr,y_test)

Statistical Analysis

The analysis was done using IBM SPSS version 21. It is a statistical software tool used for data analysis. For both proposed and existing algorithms 10 iterations were done with a maximum of 20 samples and for each iteration the predicted accuracy was noted for analysing accuracy. The value obtained from the iterations of the Independent sample T-test was performed. The independent data sets are groups, accuracy and loss. The Dependent values are product invoice and unit price. A detailed analysis has been done on these values for finding fake reviews in online e-commerce sites.

RESULTS

The dataset is provided which selects the random samples from a given dataset for fake review detection. The data is collected for a period of 15 days. As the sample sets are executed for a number of iterations the accuracy and precision values of Random Forest and Logistic Regression varies for fake review detection with a mean value= 87.7% , Std.Deviation = 0.59745. Thus the model is able to work efficiently to detect the fake reviews in online products.

The mean difference, standard deviation difference and significant difference of Random Forest based fake review detection and Logistic Regression based fake review detection is tabulated in Table 3,which shows there is a significant difference between the two groups since P<0.001 with an independent sample T-Test. Table 1Fake review data set with five attributes which selects the random samples from a given dataset. Table 2 the dataset is the independent and dependent variable. Figure 1 represents the comparison of Random Forest over Logistic Regression in terms of mean accuracy. The dependent variables in fake review detection are predicted with the help of the independent variables. The statistical analysis of two independent groups shows that the Random Forest has higher accuracy mean (87.7%) compared to Logistic Regression.

DISCUSSION

This research characteristic technique has proven to be a highly effective and versatile approach for fake review detection (V., Aishwarya, and Sultana 2020). Nowadays, fake reviews are deliberately written to build virtual reputation and attract potential customers. Thus, identifying fake reviews is a vivid and ongoing research area . Identifying fake reviews depends on the behaviors of the reviewers but not only on the key features of the reviews. This research work proposes a ML based semantic approach to identify deceptive or fake reviews in various e-commerce environments (Brown et al. 2020).

Similar findings for fake review detection are observed as the important factors with the evolution of E- commerce systems, online reviews are mainly a factor in building and maintaining a good reputation. Moreover, they have an effective role in the decision making process for the customers. Usually, a positive review for a target object attracts more customers and leads to a high increase in sales of the particular product (Shu and Liu 2019). However many people wrongly promote or demote a product of a particular brand by selling and buying fake reviews. Many websites have become a source of such opinion spam (Tsukerman 2019). In this pandemic , direct shopping has been reduced to a greater extent. This gives way for a greater number of people selecting online shopping as their best and safest way for buying their needs. This is used by opinion spammers to influence the customers to buy their products. It is intended to propose a supervised model that detects the extremist reviewer for a particular brand or product who intends to promote the sale of the brand or product (B and Mallikarjuna 2020). A Decision tree classifier is used to classify the extremist reviewers of a particular brand or product. Once detected we propose to disable the review option for that particular reviewer for that particular brand. When the reviewer continues to post extreme reviews for another brand, then it is focused to block the reviewer from logging into the shopping website. This prevents the reviewers from influencing the user sentiment in buying any product (Fitzpatrick, Bachenko, and Fornaciari 2015). In this the fake review detection is done using Logistic Regression technology with Novel Ranking technique which will not produce the good results compared to Random Forest. Finally it is observed to find the results that validate Random Forest as higher in comparison with the Logistic Regression.

Hence the study results produce clarity in performance with both experimental and statistical analysis, but it has some limitations to the proposed work such as threshold and precision. The accuracy level of detecting fake reviews in online products can still be improved by implementing artificial intelligence techniques to predict and analyze better results while comparing with existing ML techniques (Bansode* et al. 2021). In the future, the large dataset for customers can be considered to validate our proposed model with respect to recent scenarios.

CONCLUSION

The advanced fake review detection using characteristic technique by comparing Random forest classifier over logistic regression classifier. The current study focused on machine learning algorithms,

Random forest over logistic regression for higher classification in detecting fake reviews. It can be slightly improved based on the random data sets analysis in future. The outcome of the study Random forest shows 87.7% higher accuracy than logistic regression 81.2%.

DECLARATIONS

Conflict of Interests

No conflict of interest

Authors Contribution

Author GP was involved in data collection, data analysis, manuscript writing. Author DV was involved in the Action process, Data verification and validation, and Critical review of manuscript.

Acknowledgments

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this research work successfully.

Funding: We thank the following organization for providing financial support that enabled us to complete the study.

1. SNEW.AI Technologies, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

1. Bansode*, Manasi, Siddhi Pardeshi, Suyasha Ovhal, Pranali Shinde, and Anandkumar Birajdar. 2021. “Fake Review Prediction and Review Analysis.” *Regular Issue*. https://doi.org/10.35940/ijitee.g9042.0510721.
2. Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. “A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres.” *Materials Today: Proceedings* 33 (January): 798–805.
3. B, Mallikarjuna S., and S. B. Mallikarjuna. 2020. “Detection of Fake Online Reviews Using ML.” *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2020.30950.
4. Brown, Brandon, Alexicia Richardson, Marcellus Smith, Gerry Dozier, and Michael C. King. 2020. “The Adversarial UFP/UFN Attack: A New Threat to ML-Based Fake News Detection Systems?” *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. https://doi.org/10.1109/ssci47803.2020.9308298.
5. Fitzpatrick, Eileen, Joan Bachenko, and Tommaso Fornaciari. 2015. *Automatic Detection of Verbal Deception*. Morgan & Claypool Publishers.
6. Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. “Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition.” *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.
7. Jnoub, Nour, Admir Brankovic, and Wolfgang Klas. 2021. “Fact-Checking Reasoning System for Fake Review Detection Using Answer Set Programming.” *Algorithms*. https://doi.org/10.3390/a14070190.
8. Muhammad, Faisal, and Sifat Ahmed.

2019. “Fake Review Detection Using Principal Component Analysis and Active Learning.” *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2019919418.

9. Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. “Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia.” *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.
10. Patel, Nidhi A., and Rakesh Patel. 2018. “A Survey on Fake Review Detection Using Machine Learning Techniques.” *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. https://doi.org/10.1109/ccaa.2018.8777594.
11. Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. “Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City.” *Oral Health & Preventive Dentistry* 18 (1): 379–86.
12. Sathish, T., and S. Karthick. 2020. “Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite.” *Materials Today: Proceedings* 33 (January): 2555–58.
13. Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. “Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid.” In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.
14. Sharma, Udit. 2021. “Fake News Detection Using ML.” *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2021.37209.
15. Shu, Kai, and Huan Liu. 2019. *Detecting Fake News on Social Media*. Morgan & Claypool Publishers.
16. Sihombing, Andre, and A. C. M. Fong. 2019. “Fake Review Detection on Yelp Dataset Using Classification Techniques in Machine Learning.” *2019 International Conference on Contemporary Computing and Informatics (IC3I)*. https://doi.org/10.1109/ic3i46837.2019.9055644.
17. Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. “Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies.” *Materials Today Communications* 25 (December): 101603.
18. S., Neha, and Anala A. 2018. “Fake Review Detection Using Classification.” *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2018917316.
19. Tsukerman, Emmanuel. 2019. *Machine Learning for Cybersecurity Cookbook: Over 80 Recipes on How to Implement Machine Learning Algorithms for Building Security Systems Using Python*. Packt Publishing Ltd.
20. V., Abhinandan, C. A. Aishwarya, and Arshiya Sultana. 2020. “Fake Review Detection Using Machine Learning Techniques.” *International Journal of*

Fog Computing. https://doi.org/10.4018/ijfc.2020070104.

21. Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.

TABLES AND FIGURES

Table 1. Fake review data set with five attributes which selects the random samples from a given dataset. Implement Logistic Regression train step for each data point.

Target	Rating	Date	Product	User	Review text
1	4	Tue Nov 02 20:15:28 2010	Aroma Magic Face Wash	Srilekha	Truthful
2	5	Wed Jan 27 18:06:45 2014	Lakme Forever Matte Liquid Lipstick	Sapna	Fake
3	5	Sun Mar 14 10:24:56 2019	Good Vibes Toner	Archana	Fake
4	4	Fri Dec 04 22:27:27 2020	WOW skin science sunscreen	Deepika	Truthful

Table 2. Group Statistics of Random forest with Logistic Regression by grouping the iterations with Sample size 10, Mean = 87.7 , Standard Derivation = 0.59745, Standard Error Mean = 0.18893. Descriptive Independent Sample Test of Accuracy and Precision is applied for the dataset in SPSS. Here it specifies Equal variances with and without assuming a T-Test Score of two groups with each sample size of 10.

	Group	N	Mean	Std.Deviation	Std .Error mean
Accuracy	RF	10	87.7020	.59745	.18893
	LOR	10	81.0580	.45043	.14244
Loss	RF	10	2.1030	.10584	.03347
	LOR	10	2.9690	.28908	.09141

Table 3. Independent Sample Test of Accuracy and Loss (Calculate P-value = 0.001 and Significant value= 0.144, Mean Difference= 6.64400 and confidence interval = (0.2366 - 0.9735).Random Forest and Logistic Regression are significantly different from each other.

		F	**Sig**	**T**	**Df**	**Sig(2.tailed)**	**Mean difference**	**Std error difference**	**lower**	**upper**
Accuracy	Equal Variance Assume	2.339	.144	28.080	18	< .001	6.64400	0.23661	6.14690	7.14110
	Equal variance not assumed			28.080	16.733	< .001	6.64400	0.23661	6.14419	7.14381
Loss	Equal Variance Assumed	2.088	.166	-8.896	18	< .001	-.86600	.09735	-1.07052	-.66148
	Equal variance not assumed			-8.896	11.370	< .001	-.86600	.09735	-1.07941	-.65259

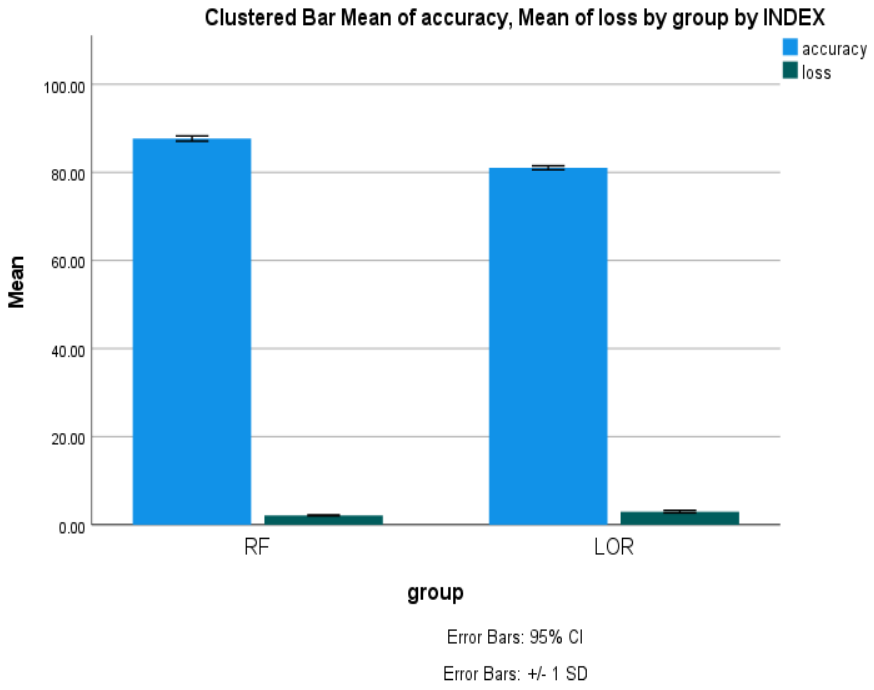

Fig. 1. Comparison of Random Forest over Logistic Regression in terms of mean accuracy. It explores that the mean accuracy is slightly better than Logistic Regression and the standard deviation is moderately improved compared to Logistic Regression . Graphical representation of the bar graph is plotted using groupid as X-axis RF vs LOR, Y-Axis displaying the error bars with a mean accuracy of detection +/- 1 SD.