Using Novel Data Clustering Technique, a Formal Improvement in the Prediction of Genuine Online Reviews Through a comparison of KNN and Logistic Regression

G.Priyanka[1], K. Malathi [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 machine learning by comparing KNN over Logistic Regression. **Materials and Methods:**The study contains two groups i.e KNN 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 KNN 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 KNN has a high accuracy of (88.8%) in comparison with the Logistic Regression algorithm (80.7%). There is a statistically significant difference between the study groups with Significant value= 0.936 (p< 0.05) and having a G power of 80%. **Conclusion:** Prediction in classifying an end user's fake reviews in online products shows that the KNN appears to generate better accuracy than the Logistic Regression algorithm.

Keywords: Fake Review Detection, Logistic Regression, Machine Learning, Formal Enhancement, Novel Data Clustering, K-Nearest Neighbors, Logistic Regression.

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 observed by a dramatic shift from in-person to online shopping using novel data clustering. 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 formal enhancement of online shopping is more convenient for the customers with respect to various scenarios based on the clustering using novel data clustering (Sihombing and Fong 2019). Customers search for various good products, this is due to the large number of products in the world.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) based on K-Nearest Neighbors. Similarly, when a product of any brand has got many negative reviews and very poor ratings for formal enhancement, 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 data clustering 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. Clustering algorithm and cluster validity are two highly correlated algorithm parts in cluster analysis. In this work, a novel idea for cluster validity and a clustering based on the validity index are introduced (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 using Novel Data Clustering. 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 in formal enhancement 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.(Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 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 KNN 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.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 product using machine learning and comparing ML techniques KNN over Logistic Regression (Del Bimbo, n.d.).

MATERIALS AND METHODS

The research work was performed in the Image Processing Lab, Department of computer Science and Engineering , Saveetha School of Engineering , SIMATS. Basically it is considered that two groups of classifiers are used, namely KNN and Logistic Regression algorithms, which are used to detect the fake reviews in online products. Group 1 is the KNN 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 (Blackstone, Fuhr, and Pociask 2014). Sample size has been calculated and it is identified that 10 samples/ group in total 20 samples with a standard deviation for KNN = 88.8% and Logistic Regression = 80.7%.

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.

K-Nearest Neighbors (KNN)

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The K-Nearest Neighbors algorithm assumes the similarity between the new case and available cases and puts the new case into the category that is most similar to the available categories. K-Nearest Neighbors algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K-Nearest Neighbors algorithm.

Pseudo code for K-Nearest Neighbors

```
Import pandas as pd
Import Matplotlib.pyplot as plt
Import KNN Classifier
import K-Nearest Neighbor as KNN
compare from sklearn.tree import KNN Classifier
skip from sklearn.linear_model import Decision Tree
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()
cv=count_vectorizer.fit()
cv.shape
KNN= K-Nearest Neighbor ()
KNN.fit(X_train,Y_train)
prediction_KNN=KNN.predict(X_test)
print(accuracy_score(prediction_KNN
actiom_KNN,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 analyzing 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 (Jnoub, Brankovic, and Klas 2021).

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 KNN and Logistic Regression varies for fake review detection with a mean value= 88.8 % , Std.Deviation = 1.22579 . 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 KNN 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 $p<0.05$ with an independent sample T-Test. In Table 1Fake review data set with five attributes which selects the random samples from a given dataset. In Table 2 the dataset is the independent and dependent variable. Fig. 1 represents the comparison of KNN 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 KNN has higher accuracy mean (88.8%) compared to Logistic Regression.

DISCUSSION

This research data clustering technique has proven to be a highly effective and versatile approach for fake review detection (Waikhom and Goswami, n.d.). 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 which will not produce the good results compared to KNN. Finally it is observed to find the results that validate KNN 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 data clustering technique by comparing KNN classifier over logistic regression classifier. The current study focused on machine learning algorithms, KNN 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 KNN shows 88.8 % higher accuracy than logistic regression 80.7 %.

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. Blackstone, Erwin A., Joseph P. Fuhr Jr, and Steve Pociask. 2014. “The Health and Economic Effects of Counterfeit Drugs.” *American Health & Drug Benefits* 7 (4): 216–24.
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. Del Bimbo, Alberto. n.d. *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*. Springer Nature.
6. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. “Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine.” *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
7. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. “Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea.” *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.

8. Fitzpatrick, Eileen, Joan Bachenko, and Tommaso Fornaciari. 2015. *Automatic Detection of Verbal Deception*. Morgan & Claypool Publishers.
9. 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.
10. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
11. 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.
12. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
13. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
14. 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.
15. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
16. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
17. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
18. 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.

19. Shu, Kai, and Huan Liu. 2019. *Detecting Fake News on Social Media*. Morgan & Claypool Publishers.
20. 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.
21. S., Neha, and Anala A. 2018. “Fake Review Detection Using Classification.” *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2018917316.
22. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. “Design and Analysis of Serial Drilled Hole in Composite Material.” *Materials Today: Proceedings* 45 (January): 5759–63.
23. 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.
24. Uganya, G., Radhika, and N. Vijayaraj. 2021. “A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms.” *Journal of Circuits Systems and Computers* 30 (05): 2130006.
25. 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.
26. Waikhom, Lilapati, and Rajat Subhra Goswami. n.d. “Fake News Detection Using Machine Learning.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3462938.

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	Mon Jan 5 16:35:06 2013	Plastic bottle	Akhila	Truthful
2	5	Tue Feb 12 12:45:23 2019	Laptop cover	Chandu	Fake
3	5	Wed Dec 5 13:12:56 2016	College bag	Lavanya	Fake
4	3	Sat Nov 3 18:09:12 2016	Night dress	Mounika	Truthful

Table 2. Group Statistics of KNN with Logistic Regression by grouping the iterations with Sample size 10, Mean = 88.8 , Standard Derivation = 1.22579 , Standard Error Mean =0.38763. 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	KNN	10	88.8310	1.22579	.38763
	LOR	10	80.7280	1.26761	.40085
Loss	KNN	10	2.1750	.39334	.12439
	LOR	10	3.1220	.23156	.07322

Table 3. Independent Sample Test of Accuracy and Loss (Calculate P-value = 0.001 and Significant value= 0.936 , Mean Difference= 8.10300 and confidence interval = (0.55762 - 0.14434). KNN 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 Variances assumed	.007	.936	14.531	18	<.001	8.10300	.55762	6.93148	9.27452
	Equal Variances Not assumed			14.531	17.980	<.001	8.10300	.55762	6.93139	9.27461
Loss	Equal Variances assumed	3.699	.070	-6.561	18	<.001	-.94700	.14434	-1.25024	-.64376
	Equal Variances Not assumed			-6.561	14.569	<.001	-.94700	.14434	-1.25544	-.63856

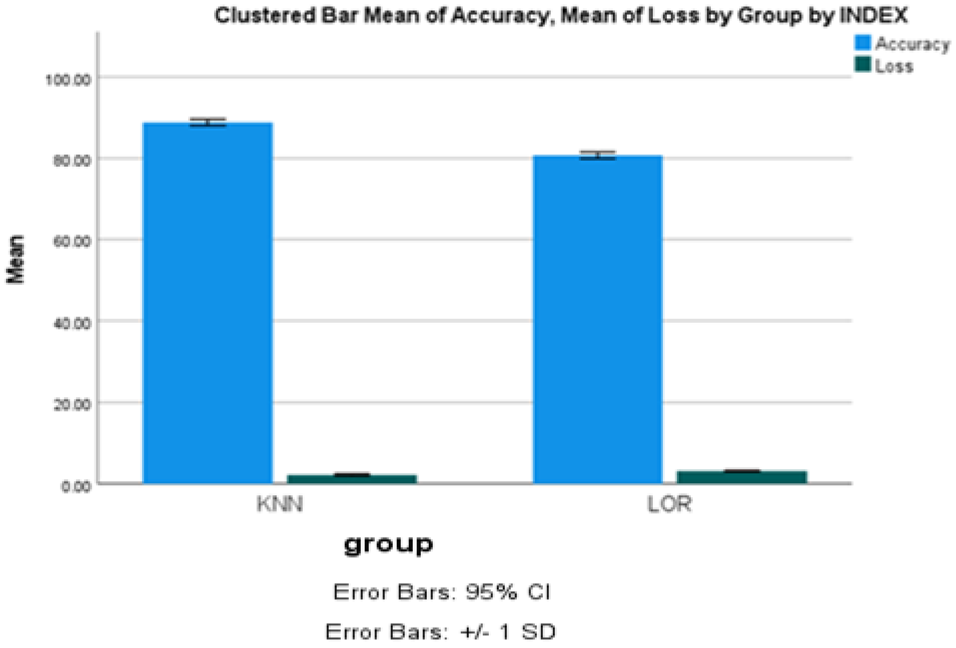

Fig. 1. Comparison of KNN over Logistic Regression in terms of mean accuracy. It explores that the mean accuracy is slightly better than Logistic Regression and the standard deviation is moderately improved compared to Logistic Regression. Graphical representation of the bar graph is plotted using groupid as X-axis KNN vs LOR, Y-Axis displaying the error bars with a mean accuracy of detection +/- 2 SD.