IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Special Issue 2, April 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

Fake Product Review Analysis Using Machine Learning

Inayathullah S[1], Indiran B[2], Mukku Uday Sankar[3], Dr. P.S Smitha[4]

U.G. Student, Department of Computer Science and Engineering, Velammal Engineering College, Chennai, Tamil Nadu, India[1-3]

Associate Professor, Department of Computer Science and Engineering, Velammal Engineering College, Chennai, Tamil Nadu, India [4]

ABSTRACT:.In the modernization process, humans have always purchased desirable products and commodities. We frequently give advice on things to buy and to avoid to our friends, family, and acquaintances based on our own experiences. Similar to this, when we want to purchase something we have never done before, we speak with others who have some knowledge in that field. Manufacturers have also relied on this technique of getting client reviews to choose the products or product features that will best delight consumers.In today's world, reviews on online websites play a vital role in sales of the product because people try to get all the pros and cons of any product before they buy it.Product reviews are having a greater effect on the consumers than there used to be. Before buying a product almost everyone checks for star rating and reviews for the particular item. The consumer wants to know if the product is good or not. The main problem is that the fake reviews that are present enormous in the e-commerce websites. So, we use supervised machine learning to find out these fake reviews present in the e-commerce websites so that it will help consumers to avoid falling for such deceptions.

KEYWORDS:Fake review, Corpus, Feature Engineering, Random Forest Classifier.

I. INTRODUCTION

E-commerce sites receive large amounts of user generated data such as reviews. Reviews are statements which express suggestion, opinion or experience of someone about any market product. These opinions have both pros and cons while providing the right feedback to reach the right person which can help fix the issue .These reviews are an integral part of an e-commerce site as it decides an important part in the sales of a product. This is considered mainly because user decides to buy a product mostly by viewing user reviews. Some people and organizations take advantage of the flaws in this system by submitting fake reviews through the use of bots consequently affecting the sales of the product. These unethical actions are usually done by competitors or hackers for their own reasons. In order to prevent this, we have decided to create a solution for this problem using machine learning with python. The fake product review analysis is done by using Random Forest Classifier Algorithm.

Paper is organized as follows. Section II describes fake review detection usingmachine learning models like sentimental analysis, data mining and opinion mining. The flow diagram represents the step of the algorithm. After data processing, how corpus is formed and implemented in feature extraction that is given in Section III. Section IV presents experimental results showing results of Random Forest Classifier tested. Finally, Section V presents conclusion.

II. RELATED WORK

'Weka tool' is a method which was proposed by E.I Elmurngi and A. Gherbi [1] is an open source software to implement machine learning algorithms using sentiment analysis to classify biased and unbiased reviews from amazon reviews based on positive, negative and neutral words. The fake reviews are identified by only including the votes voted by the customers along with the rating deviation are considered which limits the overall performance of the system. Also, as per the researcher's observations and experimental results, the existing system uses Naive Bayes classifier for spam and nonspam classification where the accuracy is quite low which may not provide accurate results for the user. Initially N. O'Brien [2] and J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto [3] have

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

proposed solutions that depends only on the features used in the data set with the use of different machine learning algorithms in detecting fake news on social media. Though different machine learning algorithms the approach lacks in showing how accurate the results are. B. Wagh,J.V.Shinde, P.A.Kale [4] worked on twitter to analyze the tweets posted by users using sentiment analysis to classify twitter tweets into positive and negative. They made use of K-Nearest Neighbor as the algorithm to allot them sentiment labels by training and testing the set using feature vectors. But the applicability of their approach to other type of data has not been validated. The work in this paper is divided in three stages. 1) Data pre-processing 2) Corpus Formation 3) Feature Extraction. Data pre-processing removes the null values from the dataset to make it more clear and on-point. Corpus creation focuses on the important terms in the review to identify true and fake reviews.Feature extraction helps machine learning understand the data.

III. METHODOLOGY

The proposed methodology consists five important phases as follows:

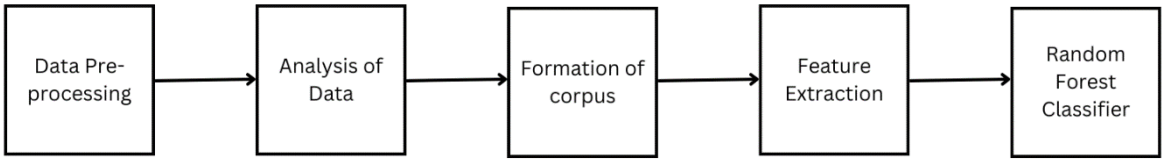

Figure 1

The data pre-processing stage removes all the unwanted values and null values from the review dataset.It usually refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data pre-processing is to improve the quality of the data and to make it more suitable for the specific task. Analysis of data gives a clear view of the data which we are handling through visualization features which are available in python. It provides a clear understanding of how data is available in the dataset and can know the nature and tone of the review written in the dataset. After data visualization,string operations like removal of punctuations, prefixes and suffixes are done in order to form a corpus which will make the work for machine learning model more easier. This is called as formation of corpus. Tokenization, stop-word removal and stemming removes the unwanted words from the review for further steps. Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data. Bag of Words Model is used in the feature extraction to change the tokenized words into Boolean form because the complexity and time taken for training the model will be lowered. Dummy variables are created for a numerical dataframe out of the dataset where both the dataframe is added to form the final dataset.Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. We use the train test split function to form a random forest classifier model. By using confusion matrix and classification report we check accuracy which defines the model.

IV. EXPERIMENTAL RESULTS

Figures 2a,2b,2cshows the results of data pre-processing of the dataset. Figure 2a shows the original or raw dataset. Figure 2bshows the removal of 'unnamed' column which is not necessary for the dataset. Figure 2c shows the usage of heatmap function to check for null values.

	Unnamed: 0	category	rating	label	text_
0	0	Home_and_Kitchen_5	5.0	CG	love well made sturdi comfort i love veri pretti
1	1	Home_and_Kitchen_5	5.0	CG	love great upgrad origin i 've mine coupl year
2	2	Home_and_Kitchen_5	5.0	CG	thi pillow save back i love look feel pillow
3	3	Home_and_Kitchen_5	1.0	CG	miss inform use great product price i
4	4	Home_and_Kitchen_5	5.0	CG	veri nice set good qualiti we set two month

Figure 2a

	category	rating	label	text_
0	Home_and_Kitchen_5	5.0	CG	love well made sturdi comfort i love veri pretti
1	Home_and_Kitchen_5	5.0	CG	love great upgrad origin i 've mine coupl year
2	Home_and_Kitchen_5	5.0	CG	thi pillow save back i love look feel pillow
3	Home_and_Kitchen_5	1.0	CG	miss inform use great product price i
4	Home_and_Kitchen_5	5.0	CG	veri nice set good qualiti we set two month

Figure 2b

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

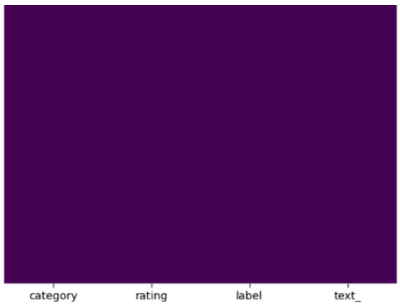

Figure 2c

Figures 3a,3b shows the analysis of the data for the reviews of the dataset. Figure 3a shows the bar chart representation of ratings with respect to the count of reviews. Figure 3b shows a chart representation of labels with respect to the count of the reviews.

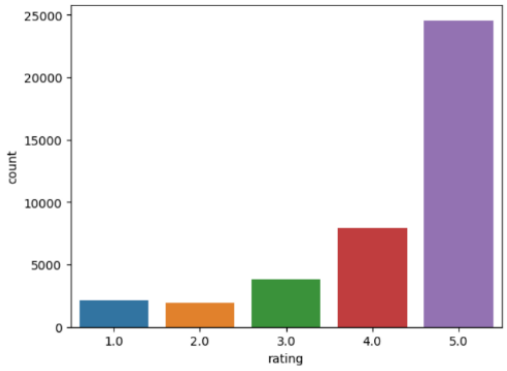

Figure 3a

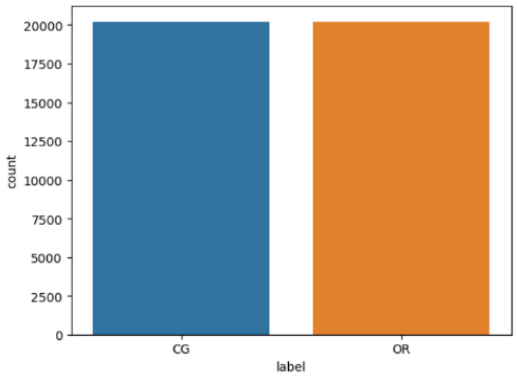

Figure 3b

Figure 4 shows the formation of corpus with respect to the review text. It is formed by removing punctuations,prefixes and suffixes, stopwords with the help of methods like stemming, tokenization etc.,

	index	rating	label	text_	CORPUS
0	0	5.0	CG	love well made sturdi comfort i love veri pretti	love well made sturdi comfort love veri pretti
1	1	5.0	CG	love great upgrad origin i 've mine coupl year	love great upgrad origin mine coupl year
2	2	5.0	CG	thi pillow save back i love look feel pillow	thi pillow save back love look feel pillow
3	3	1.0	CG	miss inform use great product price i	miss inform use great product price
4	4	5.0	CG	veri nice set good qualiti we set two month	veri nice set good qualiti set two month
...	...	...	...	...	...
4051	4051	5.0	OR	i bought vacuum cleaner month ago absolut love...	bought vacuum cleaner month ago absolut love l...
4052	4052	5.0	CG	i whole set contain look like one i use i repl...	whole set contain look like one use replac cou...
4053	4053	5.0	OR	i purchas friend recommend i power baker i mer...	purchas friend recommend power baker mere make...
4054	4054	5.0	CG	well i 've curiou set plea the dimens right th...	well curiou set plea dimens right blade nice t...
4055	4055	5.0	OR	i 've one year would still buy yeh 's littl he...	one year would still buy yeh littl heavi trade...

4056 rows × 5 columns

Figure 4

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

Figures 5a,5b shows the feature extraction of the corpus which was obtained. Figure 5a shows the count vectorize function to form a bag of words model as given below. Figure 5b shows the formation of numerical dataframe which is in Boolean values converted from string values.

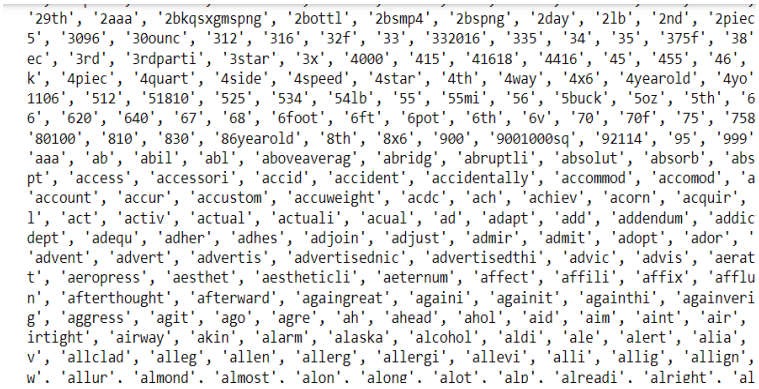

Figure 5a

	01	01oz	02	032018	045day	10	100	1000	10000	100g	...	ziploc	zipper	ziti	zojirushi	zoo	zoodl	zucchini	zuchinni	zyliss	CG
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	CG
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	CG
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	CG
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	CG
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	CG
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4051	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	OR
4052	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	CG
4053	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	OR
4054	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	CG
4055	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	OR

Figure 5b

Figure 6 shows the result of Random Forest Classifier Algorithm. We use the train test split function in the final dataset obtained to declare random forest model. The result provides the accuracy of the model.

```
[[564  56]
 [139 458]]
```

	precision	recall	f1-score	support
CG	0.80	0.91	0.85	620
OR	0.89	0.77	0.82	597
accuracy			0.84	1217
macro avg	0.85	0.84	0.84	1217
weighted avg	0.85	0.84	0.84	1217

Figure 6

V. CONCLUSION

.We have implemented a fake product review analysis and detection system using random forest classifier algorithm. Through this implementation, we detect and analyse the presence of fake reviews in e-commerce websites. We use Random Forest Classifier method due to it's accuracy, time and memory efficiency compared to other machine learning models. By achieving this, user can avoid falling for the trap of fake reviews and correct their knowledge about the review of the product.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

REFERENCES

[1] Shashank Kumar Chauhan, Anupam Goel, Prafull Goel, Avishkar Chauhan and Mahendra K Gurve "Research on Product Review Analysis and Spam Review Detection" 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)

[2] Piyush Jain, Karan Chheda, Mihir Jain, Prachiti Lade, "Fake Product Review Monitoring System", International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3, Issue-3, April 2019, pp. 105-107

[3] Chengai Sun, Qiaolin Du, and Gang Tian "Exploiting Product Related Review Features for Fake Review Detection" College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

[4] Rajashree S. Jadhav, Prof. Deipali V. Gore, "A New Approach for Identifying Manipulated Online Reviews using Decision Tree ". (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), pp 1447-1450, 2014

[5] Eka Dyar Wahyuni and Arif Djunaidy, "Fake Review Detection from aProduct Review using Modified Method of Iterative ComputationFramework" MATEC Web of Conferences BISSTECH 2015

[6] Wu, Ting-Fan, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling., The Journal of Machine Learning Research 5 (2004): 975-1005.

[7] Ranks.nl, "Stopwords". [Online]. Available: http://www.ranks.nl/stopwords

[8] J. Li, M. Ott, C. Cardie and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, vol. 1, no. 11, pp. 1566-1576,November 2014.v

INNO SPACE
SJIF Scientific Journal Impact Factor

SJIF: 8.423

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details