Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 61

Detecting the Fake Reviews Using Machine Learning

Mrs. P.Packiyalakshmi M.E.,
Department of Computer Science and Engineering
P.S.R.R College of Engineering
Sivakasi, Tamilnadu
p.packiya1982@gmail.com

GanaSree K
Department of Computer Science and Engineering
P.S.R.R College of Engineering
Sivakasi, Tamilnadu
Skumar140469@gmail.com

Dhanalakshmi V
Department of Computer Science and Engineering
P.S.R.R College of Engineering
rvds5566@gmail.com

Suvetha P
Department of Computer Science and Engineering
P.S.R.R College of Engineering
Sivakasi, Tamilnadu

ABSTRACT—Nowadays product (including service, e.g. hotel and restaurant.) reviews play an important role in consumers' online shopping activities. Buyer mostly focused only the positive reviews. The content in the form of reviews, ratings, and comments can be analyzed. Our data analysis and multi-agent simulations proved the feasibility of this framework. Product reviews are performed on the data extracted from Amazon reviews. Sentiment analysis of reviews extracts and aggregates fake and real opinions from product reviews. Millions of products and services are available in online marketing that generate huge amounts of information. we using various machine learning techniques for fake reviews. The fake review dataset is taken from the dataset repository. In our process, we have to analyze or classify the product review into fake reviews and real reviews by using the machine learning algorithms such as AdaBoost and logistic regression. Then, the experimental results show the accuracy, precision, recall, and f1 score.

KEYWORDS—Fake Reviews, Machine Learning-Adaboost, Logistic Regression, NLP.

I.INTRODUCTION

Consumers decide to buy a given product by looking at these ratings and reviews. Such content can be positive or negative reviews made by consumers who have previously used the product. The Machine Learning Algorithm can help us to a visual representation of the data and vectorize the data. Natural Language Processing, a sub-field of machine learning, is used to analyze text and identify positive or negative reviews given by consumers. This is also called Sensitive Analysis. The effect of traditional media can be estimated by established methods like audience rating. Meanwhile, the effect of the new strategies is hard to quantify.

Sentimental Analysis can be used to extract meaningful insights from customer reviews and ratings. Online reviews are used not only to post opinions about products, services and other issues, but also to analyze user opinions for making purchasing decisions and help companies in improving quality. Fake reviews have become an alarming issue as it misleads online users when making purchases and promotes or diminishes the reputation of competing brands.

Nowadays product (including service, e.g. hotel, and restaurant.) reviews play an important role in consumers' online shopping activities. Many people read these reviews before making a purchase decision. Buyers most focused only the positive reviews. This effect will bring a larger amount of business or lead to potential financial losses. Although lots of product reviews are posted by real consumers to express their views and share their shopping experiences with other people, more untruthful reviews appear on e-business websites because of financial reasons.

Fake consumer review detection has attracted much interest in recent years owing to the increasing number of Internet purchases. Current approaches to detecting fake consumer reviews use review content, product and reviewer information, and other features to detect fake reviewsNow the use of internet and online marketing has become very popular.Millions of products and services are available in online marketing that generate huge amounts of

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 61

information. Hence, it's difficult to find the best suitable services or products compatible with the requirement.

Customers directly take decisions based on reviews or opinions that are written by others based on their experiences. In this competitive world, any person can write anything, which raises the number of fake reviews. This process gives false input to new customers who want to buy products, so we need a system to detect fake reviews.

II.PROBLEM STATEMENT

Cyber technology's wide reach and fast spread contribute to its menace. The problem is to identify the authenticity of the reviews and online content. An equally important problem is to identify the bots involved in spreading false reviews. So, **to overcome the difficulties by using the different machine learning algorithms**.

III.METHODOLOGY

A. FETCH DATA:

The input data was collected from dataset repositories like the UCI repository. In our process, a fake review dataset is used. Data selection is the process of detecting or classifying product reviews into fake reviews and real reviews. The dataset contains information about the product such as reviews, IDs, ratings, product categories, comments, and so on. It provides useful information about the product such as its' title, description, image URL, sales rank, category, price, brand, and products that are most related to it.

B. DATA CLEANING:

Data cleaning(preprocessing) is the process of removal of unwanted data from the dataset. Pre-processing is the data transformation operations .It is used to dataset are transform into a structure suitable for machine learning. It is also includes cleaning the dataset , which makes it more efficient. Missing data removal. Encoding Categorical data.

Missing data removal: It is the process of the missing values and Non values are replaced by 0. Encoding Categorical data :a process of categorical data are converted into integer format ,so that the data with converted categorical values can be improve the predictions. Then, we have to remove the unwanted or unnecessary columns. Most machine learning algorithms wanted numerical input and output variables.

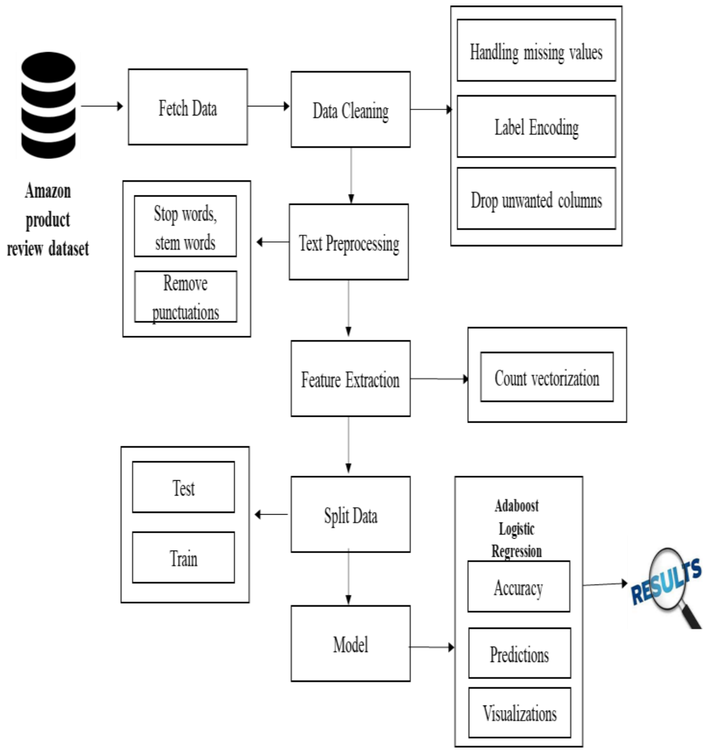

Figure: FLOW DIAGRAM

C. TEXT PREPROCESSING:

Natural language processing is uesd by computer to understand the human language . It is a component of artificial intelligence (AI).Preprocessing the data typically consists of some steps: First one is removing the punctuation: The punctuation removal process will help to treat each text equally. special characters does not add any values in the text , so we remove special characters in a text. Characters are removed depending on their uses. eg: How are you?->How are you. Tokenization: Tokenization is the start of the NLP process, converting sentences into understandable bits of data that a program . It gives structure to previously unstructured text. eg: Plata o Plomo->'Plata','o','Plomo'. Remove stopwords: There are too many stop words in any human language. By removing these words, we remove less information from our text to focus more on the important information. Stemming: Stemming is the process of reducing a word to its stem, which is attached to word roots called suffixes and prefixes or "lemmas".It removes suffices, like "ing", "ly", and "s", etc. by a simple rule-based approach.

D. FEATURE EXTRACTION:

In this step, we have to implement the vectorization method. Vectorizing is the process of encoding text as an integer . Here we have to use Count Vectorizer. Both are methods for converting text data into vectors as the model can process only numerical data. In count vectorizer, it converts the tokens into vectors.

E. SPLIT DATA:

Here we use machine learning process, In addition to the data required for training, test data are needed to evaluate the performance of the algorithm to see how well it works. In our process, we considered 80% of the spam dataset to be the training data and the remaining 30% to be the testing data. Dataset splitting into two portions for validation. One Part of the dataset is used to implement a predictive model and next part to evaluate the model's performance. total number of dataset is divided into two parts that is training set and a test set, train set contain large number of data most of the data and a testing dataset contain minimum number of data.

F. MODEL:

In our process, we have to implement the machine learning algorithms such as boost and logistic regression. This algorithm used with **AdaBoost** is decision trees with one level, which means Decision trees with only 1 split. These trees are also called Decision Stumps. An AdaBoost classifier is a meta-estimator ,that starts with fitting a classifier on the taken dataset and then fits additional copies of the classifier on the same dataset but where the incorrectly classified instances are adjusted such that focus more on struggle cases.

Logistic regression is a method used for analysis to predict a outcome, such as yes or no, based on taken of a data set. A logistic regression model find a data variable by analyzing the dataset relationship between one or other existing variables.

G. RESULT GENERATION:

The Final Result will be based on the overall performance and precision. Here we proposed the overall performance of this approach is to evaluated using,

Accuracy

The accuracy of the classifier is refers to the ability of the data. It predicted the class label on corrected value and the accuracy of the precision refers to how well a given precision can guess the correct value of the predicted attribute for new entry data.

AC= (TP+TN)/ (TP+TN+FP+FN)

Precision

Precision is used as the number of TRUE POSITIVES divided by the number of TRUE POSITIVES plus the number of FALSE POSITIVES.

Precision=TP/ (TP+FP)

Recall

Recall the number of corrected values divided by the number of correct value that should have been returned to next process. In binary classification, recall is called sensitivity data. It can be viewed as the orginal probability of that a relevant dataset is retrieved by the recall.

Recall=TP/ (TP+FN)

IV.EXPERIMENTAL RESULTS

A. *DATA SELECTION*

```
----------------------------------------
============ Data Selection ============
----------------------------------------
   Unnamed: 0                              Date  ...    AT Rev_Type
0           0     Reviewed in India on 10 August 2018  ...  13.0        1
1           1      Reviewed in India on 27 March 2018  ...  16.0        1
2           2   Reviewed in India on 19 December 2018  ...  22.0        1
3           3       Reviewed in India on 25 June 2020  ...  22.0        1
4           4    Reviewed in India on 27 October 2018  ...  28.0        0
5           5  Reviewed in India on 11 September 2020  ...  27.0        1
6           6    Reviewed in India on 9 December 2018  ...  23.0        1
7           7         Reviewed in India on 9 May 2019  ...  25.0        0
8           8  Reviewed in India on 22 September 2020  ...  24.0        0
9           9   Reviewed in India on 14 November 2018  ...  28.0        1

[10 rows x 26 columns]
```

Figure-2: AMAZON REVIEWS

B. PREPROCESSING

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 61

```
----------------------------------------
========= Checking missing values  ======
----------------------------------------
Unnamed: 0        0
Date              0
URL               0
Review_Title      0
Author            0
Rating            0
Review_text       0
Review_helpful    0
Sentiment         0
Subjectivity      0
Neg_Count         0
Word_Count        0
Unique_words      0
Noun_Count        0
Adj_Count         0
Verb_Count        0
Adv_Count         0
Pro_Count         0
Pre_Count         0
Con_Count         0
Art_Count         0
```

```
---------------------------------------------
============ Drop unwanted columns  ========
---------------------------------------------

1.Before drop unwanted columns : (9438, 26)


2.After drop unwanted columns  : (9438, 24)
```

Figure-3: MISSING VALUES

C. NLP TECHNIQUE

```
The original text

['A', 'really', 'awesome', 'keyboard', 'i', 'was', 'actually', 'going', 'for', 'the', 'cheap',
'gaming', 'keyboards', 'like', 'the', 'Blaze', 'but', 'This', 'it', 'may', 'be', 'a', 'bit',
'costly', 'for', 'most', 'of', 'you', 'all', 'but', 'trust', 'me', 'its', 'worth', 'it',
'this', 'is', 'the', 'best', 'RGB', 'membrane', 'keyboard', 'ever', 'It', 'may', 'not', 'be',
'mechanical', 'but', 'do', 'not', 'ever', 'go', 'for', 'the', 'cheap', 'gaming', 'keyboards',
'this', 'keyboard', 'gives', 'a', 'premium', 'stealthy', 'look', 'and', 'has', 'alot', 'of',
'game', 'lighting', 'profilesPros', 'durable', 'Nice', 'vibrant', 'colors', 'Nice', 'quality',
'membrane', 'keys', 'Almost', 'everythingCons', 'Not', 'many', 'lighting', 'effects', 'COULD',
'have', 'been', 'a', 'Mechanical', 'keyboard']
------------------------------------------------------------


After removing stopwords
['A', 'really', 'awesome', 'keyboard', 'actually', 'going', 'cheap', 'gaming', 'keyboards',
'like', 'Blaze', 'This', 'may', 'bit', 'costly', 'trust', 'worth', 'best', 'RGB', 'membrane',
'keyboard', 'ever', 'It', 'may', 'mechanical', 'ever', 'go', 'cheap', 'gaming', 'keyboards',
'keyboard', 'gives', 'premium', 'stealthy', 'look', 'alot', 'game', 'lighting', 'profilesPros',
'durable', 'Nice', 'vibrant', 'colors', 'Nice', 'quality', 'membrane', 'keys', 'Almost',
'everythingCons', 'Not', 'many', 'lighting', 'effects', 'COULD', 'Mechanical', 'keyboard']
```

Figure-4: REMOVE STOPWORDS

```
============ Before Applying NLP  ============
----------------------------------------------

0      A really awesome keyboard  i was actually go...
1      I know its costly but its worth your money  ...
2      I had been contemplating to buy this for a l...
3      SO   Very very small keys  For fast typers w...
4      Good to use keyboard while it is new  but it...
5      Have always used a Logitech as my keyboard f...
6      The short review  The Logitech G     membran...
7      Pros  cheaper than most other high end keybo...
8      This is a terrible product by Logitech  I am...
9      Pros    Dedicated media keys and Windows tog...
Name: Review_text, dtype: object
----------------------------------------------
============ After Applying NLP  =============
----------------------------------------------

0      a really awesome keyboard  i was actually go...
1      i know its costly but its worth your money  ...
2      i had been contemplating to buy this for a l...
3      so   very very small keys  for fast typers w...
4      good to use keyboard while it is new  but it...
5      have always used a logitech as my keyboard f...
6      the short review  the logitech g     membran...
7      pros  cheaper than most other high end keybo...
8      this is a terrible product by logitech  i am...
9      pros    dedicated media keys and windows tog...
Name: Summary_Clean, dtype: object
==============================================
```

Figure-5: UPPERCASE TO LOWERCASE

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7-Issue 1, January 2024
Paper: 61

```
==========================================
Vocabulary size: 10123


Example:


Sentence:
0        a really awesome keyboard  i was actually go...
1        i know its costly but its worth your money  ...
2        i had been contemplating to buy this for a l...
3        so   very very small keys  for fast typers w...
4        good to use keyboard while it is new  but it...
                              ...
9433              i have using all products of one plus
9434     everything is excellent but jio cinema not a...
9435                                           excellent
9436                                         i liked this
9437          best tv for good price  thanks for oneplus
Name: Summary_Clean, Length: 9438, dtype: object


After tokenizing :
[6, 214, 43, 1676, 16, 43, 119, 55, 73, 335, 93, 2195, 1770, 205, 175, 896, 67, 93, 1885, 64
266, 603, 43, 5, 321, 2, 251, 189, 20, 732, 2, 879, 23, 2, 113, 222, 114, 3, 39, 49, 1147, 1
19, 11, 1116, 298, 23, 2, 113, 2, 200, 540, 3634, 120, 19, 28, 1218, 68, 1, 41, 335, 3, 339,
20, 2648, 1117, 61, 23, 4819, 1472, 289, 1218, 4820, 2, 567, 474, 1, 733, 2381, 68, 1, 5, 11
307, 478, 16, 2, 577, 123, 39, 35, 355, 55, 567, 2196, 9, 90, 669, 2, 879, 23, 2, 200, 286,
3010, 3635, 494, 19, 2, 1116, 3, 308, 968, 69, 43, 4821, 68, 1, 3011, 2, 852, 251, 31, 5, 13
19, 860, 1607, 47, 404, 98, 1054, 6, 189, 4, 30, 119, 2, 276, 36, 99, 16, 1771, 4, 30, 23, 7
```

Figure-6: TOKENIZATION

D. DATA SPLITTING

```
=============================================
---------------- Data Splitting --------------
=============================================

Total No.of data's :  9438

Total No.of training data's :  7550

Total No.of testing data's :  1888
==========================================
```

Figure-7: SPLITTING DATA

E. ML ALGORITHM- ADABOOST

```
==========================================
----------------- AdaBoost ---------------
==========================================

=========== Performance Analysis ========

 Accuracy : 80.8792372881356 %

              precision    recall  f1-score   support

           0       0.89      0.72      0.79      1279
           1       0.58      0.80      0.67       609

    accuracy                           0.75      1888
   macro avg       0.73      0.76      0.73      1888
weighted avg       0.79      0.75      0.75      1888
```

Figure-8: ACCURACY

F. VISUALIZATION

```
========================

 The review is fake
========================

 The review is real
========================

 The review is fake
========================

 The review is real
========================

 The review is fake
========================

 The review is fake
========================

 The review is fake
========================

 The review is real
========================

 The review is fake
========================
```

Figure-9: FAKE OR REAL

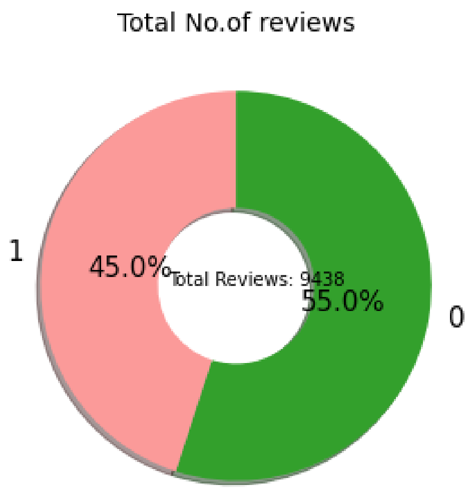

Figure-10: OVERALL PREDICTION

V. CONCLUSION

We conclude that, a machine-learning-based method for the detection of fake and real reviews on fake reviews dataset. The research in the paper adopted an approach based on

AdaBoost and logistic regression, which was chosen mostly for its simplicity and its known performance capabilities. The experimental results indicate that the proposed approach outperformed the machine learning algorithms and achieved the highest performance in terms of Accuracy, Precision, recall, and F1-score.

VI. REFERENCES

[1] Syed Mohammed Anas, Santoshi Kumari," Opinion Mining based Fake Product Review Monitoring and Removal System", Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT] 2021.

[2] Petr Hajek, AliaksandrBarushka, Michal Munk, "Fake consumer review detection using deep neural networks integrating word embedding and emotion mining", Accepted: 23 January 2020, Springer-Verlag London Ltd., part of 2020.

[3]S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef, and S. Nadali,‘‘Detecting deceptive reviews using lexical and syntactic features,’’ in Proc. 13th Int. Conf. Intellient Syst. Design Appl., Dec. 2013, pp. 53–58.

[4] T. Fornaciari and M. Poesio, ‘‘Identifying fake Amazon reviews as learn-ing from crowds,’’ in Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics, 2014, pp. 279–287.

[5] L. Nanni, ‘‘An ensemble of classifiers for the diagnosis of erythemato- squamous diseases,’’ Neurocomputing, vol. 69, nos. 7–9, pp. 842–845,Mar. 2006.

[6]J. Z. Lei and A. A. Ghorbani, ‘‘Improved competitive learning neural networks for network intrusion and fraud detection,’’ Neurocomputing, vol. 75, no. 1, pp. 135–145, Jan. 2012.

[7] U. Aslam, M. Jayabalan, H. Ilyas, and A. Suhail, ‘‘A survey on opinion spam detection methods,’’ Int. J. Sci. Technol. Res., vol. 8, no. 9, pp. 1–10, 2019.

[8]Y. Ren, L. Yin, and D. Ji, ‘‘Deceptive reviews detection based on language structure and sentiment polarity,’’ J. Frontiers Comput. Sci. Technol., vol. 8, no. 3, pp. 313–320, 2014.

[9]C. M. Yilmaz and A. O. Durahim, ‘‘SPR2EP: A semi-supervised spam review detection framework,’’ in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2018, pp. 306–313.

[10]G.Shan,L.Zhou,Examining review inconsistency for fakereview detection,’’ Decis. Sup-port Syst., vol. 144, May 2021.

[11] X. Wang, K. Liu, and J. Zhao, ‘‘Detecting deceptive review spam viaattention-based neural networks,’’ in Proc. Nat. CCF Conf. Natural Lang.Process. Chin. Comput.Dalian, China: Springer, 2017.

[12]F. Zhu and X. Zhang, ‘‘Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics,’’ J. Marketing, vol. 74, no. 2, pp. 133–148, Mar. 2010.

[13] J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015.

[14] R. Barbado, O. Araque, and C. A. Iglesias, ‘‘A framework for fake reviewdetection in online consumer electronics retailers,’’ Inf. Process. Manage.,vol. 56, no. 4, pp. 1234–1244, Jul. 2019.

[15] Z.-Y. Zeng, J.-J. Lin, M.-S. Chen, M.-H. Chen, Y.-Q. Lan, and J.-L. Liu,‘‘A review structure based ensemble model for deceptive review spam,’’Information, vol. 10, no. 7, p. 243, Jul. 2019.

[16] Wang, Guan, SihongXie, Bing Liu, and Philip S. Yu., "Identify Online Store Review Spammers via Social Review Graph", ACM Transactions on Intelligent Systems and Technology, 2012.

[17]R. Xu, Y. Xia, K.-F. Wong, and W. Li, ‘‘Opinion annotation in on-line Chinese product reviews,’’ in LREC, vol. 8, 2008, pp. 26–30.

[18]W. Zhang, Y. Du, T. Yoshida, and Q. Wang, ‘‘DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network,’’ Inf. Process. Manage., vol. 54, no. 4, pp. 576–592, 2018.

[19] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, ‘‘Spotting fake reviews via collective positive-unlabeled learning,’’ in Proc. IEEE Int. Conf. Data Mining, Dec. 2014, pp. 899–904.

[20] D. U. Vidanagama, T. P. Silva, and A. S. Karunananda, ‘‘Deceptiveconsumer review detection: A survey,’’ Artif. Intell. Rev., vol. 53, no. 2,pp. 1323–1352, Feb. 2020.