ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

Spam or Ham Prediction

Mrs.Savithri.S
Department of Mca
Dhanalakshmi College of Engineering and Technology

Ms.Sangeetha Priya.R
Department of Mca
Dhanalakshmi College of Engineering and Technology

Abstract: Nowadays, we use frequently e-mails, one of the communication channels, in electronic environment. It play an important role in our lives because of many reasons such as personal communications, business-focused activities, marketing, advertising etc. E-mails make life easier because of meeting many different types of communication needs. On the other hand they can make life difficult when they are used outside of their purposes. Spam emails can be not only annoying receivers, but also dangerous for receiver's information security. Detecting and preventing spam e-mails has been a separate issue. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. To propose a machine learning-based method to classify the email in the form of spam or ham by best accuracy from comparing supervised classification machine learning algorithms.

I. INTRODUCTION

The growing consumerism has led to the importance of online reviews on the Internet. Opinions voiced by these reviews are taken into consideration by many consumers for making financial decisions online. This has led to the development of opinion spamming for profitable motives or otherwise. This work has been done to tackle the challenge of identifying such spammers, but the scale of the real-world review systems demands this problem to be tackled as a big data challenge. A big data approach was applied for the detection of review spammers. A metadata-based rating model for detecting review spammers was implemented on large datasets to study the effect of scale on such models. The discrepancies were identified, and mitigations for the same in the form of exponential smoothing were proposed. A distributed computing platform was set up and heterogeneous methods were applied to compute the various spam indicators.

DRAWBACKS

- Machine Learning concept is not implemented.
- Accuracy and performance metrics are not calculated.

PROPOSED SYSTEM:

EXPLORATORY DATA ANALYSIS

Machine learning supervised classification algorithms will be used to give the given dataset and extract patterns, which would help in classifying the reviews, thereby helping the apps for making better decisions of their features in the future.

DATA WRANGLING

In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis.

DATA COLLECTION

The data set collected for classifying the given data is split into Training set and Test set. Generally, 70:30 percentage are applied to split the Training set and Test set. The Data Model which was created using the SMLT is applied on the Training set and based on the test result accuracy, Test set prediction is done.

BUILDING THE CLASSIFICATION MODEL

The prediction of the email spam or ham is good in machine learning algorithm prediction model and is effective because of the following reasons the following reasons: It provides better results in classification problem.

- It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.
- It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

ADVANTAGES

- Machine Learning method is implemented.
- Pre-Processing and analysing the data.
- Performance metrics of different algorithm are compared and the better prediction is done.

MODULE DESCRIPTION

All Rights Reserved © 2022 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

DATA PRE-PROCESSING

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different **data cleaning** tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, **missing values** and it able to **more quickly clean data**. It wants to **spend lesstime cleaning data**, and more time exploring and modeling.

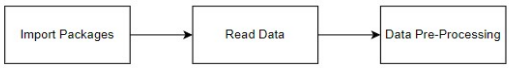
Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose and read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data
output : removing noisy data

DATA VALIDATION/ CLEANING/PREPARING PROCESS

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an

All Rights Reserved © 2022 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

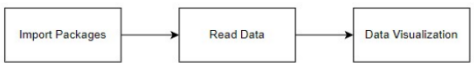
EXPLORATIONDATAANALYSISOF VISUALIZATION

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

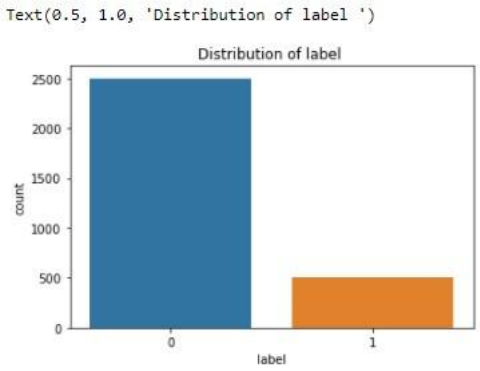
Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

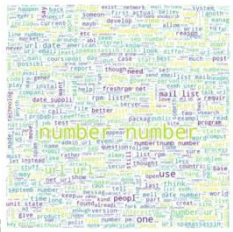
MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT



input : data
output : visualized data



Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

COMPARING ALGORITHM WITH PREDICTION IN THE FORM OF BEST ACCURACY RESULT

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below 4 different algorithms are compared:

- Logistic Regression
- Random Forest

The K-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

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

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

All Rights Reserved © 2022 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

3.4 ALGORITHM AND TECHNIQUES

ALGORITHM EXPLANATION

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Used Python Packages

sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

NATURAL LANGUAGE PROCESSING (NLP)

Natural language processing (NLP) allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of natural language processing include information retrieval, text mining, question answering and machine translation. Many current approaches use word co-occurrence frequencies to construct syntactic representations of text. “Keyword spotting” strategies for search are popular and scalable but dumb; a search query for “dog” might only match documents with the literal word “dog” and miss a document with the word “poodle”. “Lexical affinity” strategies use the occurrence of words such as “accident” to assess the sentiment of a document. Modern statistical NLP approaches can combine all these strategies as well as others, and often achieve acceptable accuracy at the page or paragraph level. Beyond semantic NLP, the ultimate goal of “narrative” NLP is to embody a full understanding of commonsense reasoning.

1. Remove stopwords: There are a few words which are very commonly used when humans interact, but these words don’t make any sense or add any extra value. Additionally, there might be few words which are not required for the business case given in hand. So, these words need to be deleted from the data.The NLTK package has a defined set of stopwords for different languages like English. Here, we will focus on ‘english’ stopwords. One can also consider additional stopwords if required.

2.Tokenization : This is one of the common practices while working on text data. This helps to split a phrase, sentence, or paragraph into small units like words or terms. Each unit is called a token. There are different types of tokenization. We have already used this in above examples for stemming, POS tagging, and NER. Below are different ways to tokenize the text.

Vectorization/Word Embedding

Once cleaning and tokenization is done, extracting features from the clean data is very important as the machine doesn’t understand the words but numbers. Vectorization helps to map the words to a vector of real numbers, which further helps into predictions. This helps to extract the important features.

LOGISTIC REGRESSION

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous

All Rights Reserved © 2022 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

RANDOM FOREST CLASSIFIER

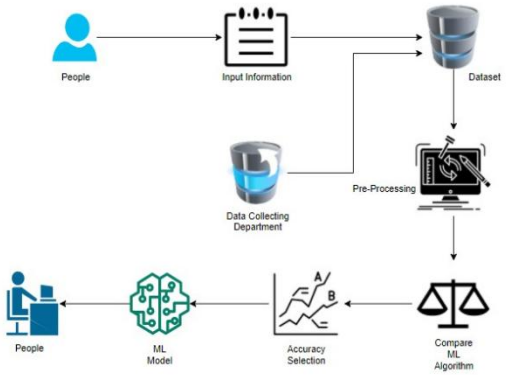
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

SYSTEM DESIGN AND DEVELOPMENT

5.1 SYSTEM ARCHITECTURE



System Architecture

SYSTEM TESTING

TESTING

System Analysis and Design process including Requirement Analysis, Business Solution Options, Feasibility Study, Architectural Design was discussed in previous chapter.

Generally Software bugs will almost always exist in any software module. But it is
not because of the carelessness or irresponsibility of programmer but because of the complexity. Humans have only limited ability to manage complexity. This chapter discusses about the testing of the solution and implementation methodologies.

UNIT TESTING

Software Testing is the process of executing a program or system with the intent of finding errors. The scope of software testing often includes examination of code as well as execution of that code in various environments and conditions. Testing

All Rights Reserved © 2022 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

stages of the project can be explained as below and system was tested for all these stages.

● **Component or unit testing**

- Individual components are tested independently;

- Components may be functions or objects or coherent groupings of these
entities.

SYSTEM TESTING

- Testing of the system as a whole. Testing of emergent properties is
particularly important.

ACCEPTANCE TESTING

- Testing with customer data to check that the system meets the customer's
needs.

TESTING METHODS AND COMPARISON

BLACK BOX TESTING

Black Box Testing is testing without the knowledge of the internal workings of the
item being tested. When black box testing is applied to software engineering, the
tester selects valid and invalid input and what the expected outputs should be, but not how the program actually arrives at those outputs.Black box testing methods include equivalence partitioning, boundary value analysis, all-pairs testing, fuzz testing, model-based testing, traceability matrix, exploratory testing and specification-based testing. This method of test design is applicable to all levels of software testing: unit, integration, functional testing, system and acceptance.

WHITE BOX TESTING

White box testing (glass box testing) strategy deals with the internal data structures and algorithms. The tests written based on the white box testing strategy incorporate coverage of the code written, branches, paths, statements and internal logic of the code etc. These testers require programming skills to identify all paths through the software.

Types of white box testing includes code coverage (creating tests to satisfy some criteria of code coverage.),mutation testing methods, fault injection methods, static testing.

CONCLUSION AND FUTURE ENHANCEMENT

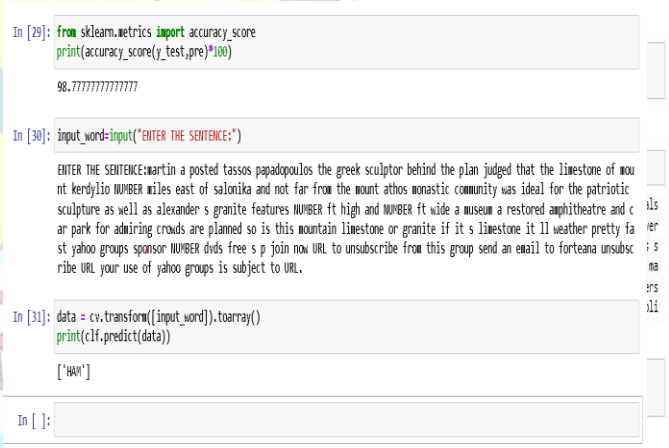
CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of email Spam or Ham.

FUTURE ENHANCEMENT

- Email spam or ham prediction to connect with cloud.
- To optimize the work to implement in Artificial Intelligence environment.

SAMPLE SCREENS



REFERENCES

[1] N. Hajli, Y. Wang, M. Tajvidi, and M. S. Hajli, "People, technologies, and organizations interactions in a social commerce era," IEEE Trans. Eng. Manage., vol. 64, no. 4, pp. 594–604, Nov. 2017.

[2] Y. Wang, W. Dai, and J. Ma, "BciNet: A biased contest-based crowdsourcing incentivemechanismthrough exploiting social networks," IEEE Trans. Syst., Man, Cybern.: Syst., vol. 50, no. 8, pp. 2926–2937, Aug. 2020.

[3] Google updates spam detection for reviews, warns SEOs, 2013. Accessed: Jul. 18, 2013. [Online]. Available: http://www.webpronews.com/googleupdates-spam-detection-for-reviews-warns-seos-2013-02/

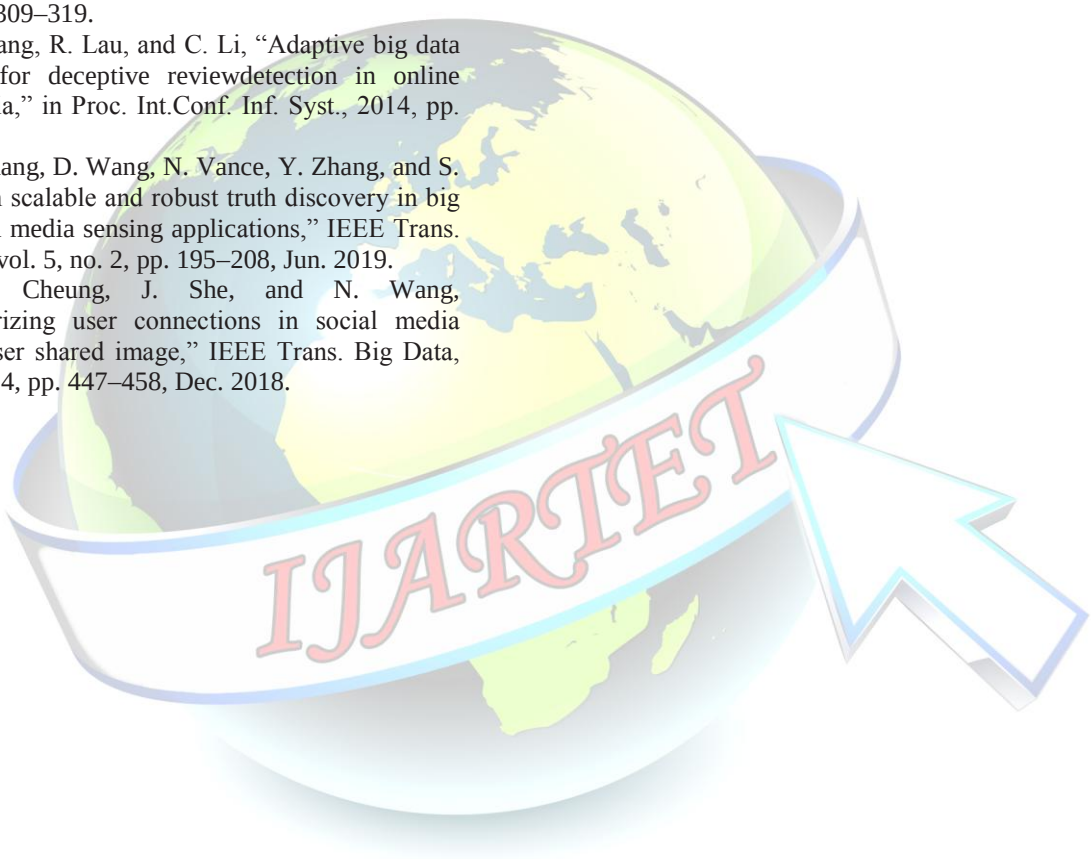
[4] N. Jindal and B. Liu, "Opinion spam and analysis," in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 219–230.

All Rights Reserved © 2022 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 9, Issue 8, August 2022

[5] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. A. Najada, "Survey of review spam detection using machine learning techniques," J. Big Data, vol. 2, pp. 1–24, 2015.
[6] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," Expert Syst. Appl., vol. 42, pp. 3634–3642, 2015.
[7] R. Mohawesh et al., "Fake reviews detection: A survey," IEEE Access, vol. 9, pp. 65771–65802, 2021.
[8] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Lang. Technol., 2011, pp. 309–319.
[9] W. Zhang, R. Lau, and C. Li, "Adaptive big data analytics for deceptive reviewdetection in online socialmedia," in Proc. Int.Conf. Inf. Syst., 2014, pp. 1–19.
[10] D. Zhang, D. Wang, N. Vance, Y. Zhang, and S. Mike, "On scalable and robust truth discovery in big data social media sensing applications," IEEE Trans. Big Data, vol. 5, no. 2, pp. 195–208, Jun. 2019.
[11] M. Cheung, J. She, and N. Wang, "Characterizing user connections in social media through user shared image," IEEE Trans. Big Data, vol. 4, no. 4, pp. 447–458, Dec. 2018.



All Rights Reserved © 2022 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

All Rights Reserved © 2022 IJARTET