Application of Machine Learning Techniques for Fake Customer Review Detection

Nupoor Shailendra Kangle
Department of Computer Engineering, PCCOE, Pune
Pune, India
nupoorkangle99@gmail.com

Dr. Rajeshwari Kannan
Department of Computer Engineering, PCCOE, Pune
Pune, India
kannan.rajeswari@pccoepune.org

Sushma Vispute
Department of Computer Engineering, PCCOE, Pune
Pune, India
sushma.vispute@pccoepune.org

***Abstract*—Now-a-days with the increasing demand of the web , online marketing is additionally becoming progressively popular. This is often because; tons of products and services are easily available online. Hence, reviews of these products and services are vital for customers also as sellers. But, to gain profit or promotion, scammers produce fake reviews. These fake reviews written by scammers prevent customers and sellers reaching actual opinion about the products. Hence, fake reviews or spam reviews must be detected and eliminated so as to prevent misleading potential customers. In our work, supervised and semi supervised learning techniques are applied to detect spam review.**

Keywords: Machine Learning, Logistic Regression Classifier, Random Forest Algorithm, SVM Algorithm.

I. INTRODUCTION

Machine Learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed, to act on their own without any human interruption[13]. It focuses on development of computer programs that can access data and use it to learn for themselves.

So, the primary aim is to allow the computers to learn automatically, without human intervention or assistance and adjust actions accordingly.

As internet availability and usability increases day by day, users mostly offer online marketing. In this case customers must check the review of the product or service before going to purchase but there is no guarantee that the reviews are true.

Customers may also feel inclined to review a product or service if they had an exceptionally good or bad experience with it. While online reviews can be helpful, blind trust of these reviews is dangerous for both the seller and buyer. Many refer to online reviews before placing any online order; however, the reviews may be contaminated or faked for profit or gain, thus any decision based on online reviews must be made cautiously.

The objective here is to compare the efficiency of different techniques useful for fake review identification and develop a system to accept reviews from authenticated users only. Once the reviews are gathered from trustworthy clients (Guanine users), the chosen machine learning algorithms are applied on the gathered data.

II. LITERATURE REVIEW

According to Dixit S, Agrawal AJ[13], there are three types of reviews:

1. Untruthful Reviews
2. Reviews on brands
3. Non-Reviews

1. Untruthful Reviews contain the reviews that are not believable.
2. Reviews on brands means reviews regarding brands directly, not for specific brands.
3. Non-Reviews contain unrelated reviews or not related materials like advertisements.

III. NECESSITY

Online reviews are a source of information as well as the experience of other customers which is helpful for drawing conclusions about the product or services. However the reviews may be faked in order to gain profit, thus any decision based on online reviews must be made cautiously[1]. Sometimes reviews contain unrelated words or advertisements of other products. This makes a misunderstanding on another customer and he or she may cancel buying it. Such situations are referred to as fake review or review spam.

Review spam can also negatively impact businesses due to loss in consumer trust. The issue is severe enough to attract the attention of mainstream media and governments and defame a particular product or service.

IV. METHODOLOGY

The aim of machine learning is to allow the computers to learn automatically, without human intervention or assistance and adjust actions accordingly.

Process[12]:

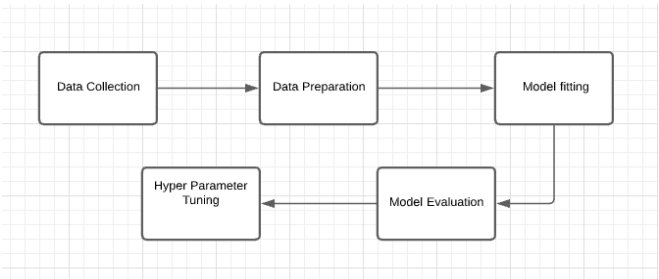

Fig. 1. ML Processing

1. **Data Collection**: Collect the data that the algorithm will learn from.

2. **Data Preparation**: Format and engineering the data into optimal format, extracting important features and performing dimensionality reduction.

3. **Training**: Also known as the fitting stage, this is where a machine learning algorithm actually learns by showing it the data that has been collected and prepared.

4. **Evaluation**: Test the model to see how it will perform.

5. **Tuning**: Fine tune the model to maximize its performance.

A. Dataset

It is an unprocessed collection of data which contains hotel name, review, remarks and pictures also.

Dataset contains 1500 rows which has 20 hotels data with tuples- Category, Hotel, Rating, Source and Text of review.

It contains both reviews that are positive and negative.

B. Preprocessing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.

In preprocessing procedures are performed including - tokenization & lowercase letters, removing stop words, removing punctuations, stemming etc.

Also worked on missing values.

According to the way of learning, machine learning mainly includes[4]:

A. Supervised Learning

The name itself indicates the presence of a supervisor as an educator . Basically supervised learning is learning during which we teach the machine using data which is well labeled, meaning some data is already tagged with the right answer .

The goal of supervised learning is to coach the model in order that it can predict output when it's given new data.

Commonly it includes decision tree, Bayesian classification ,least square regression, logistic regression, support vector machine, neural network algorithm.

Example: Train a machine to help predict how long it will take to drive to your destination from your current location. For this data includes whether conditions, time of the day, is it holiday.

Using this data machine can predict how much time is needed.

B. Unsupervised Learning

It is a machine learning technique, where you do not need any supervisor to supervise the model. Instead, need to allow the model to work on its own to discover the information.

Commonly it includes independent component analysis, k-means and apriori algorithm.

Example: In the case of a baby and her family dog. She knows and identifies the dog. Few weeks later a family friend brings another dog which has a different color and size but the baby can identify it as a dog.

C. Semi-supervised Learning

Semi-supervised learning is learning between unsupervised learning and supervised learning. Semi-supervised used for both classification and regression.

Algorithms Used[1]:

1) Support Vector Machine (SVM) [15]

Support vector means by considering the two points of opponent classes which are near to the decision boundary we draw the two parallel lines and now these two points are nothing but the support vector.

The goal of a support vector machine is not only to draw hyper-planes and divide data points, but to draw the hyperplane that separates data points with the largest margin, or with the most space between the dividing line and any given data point. Now the hyper plane is the decision boundary that segregates the data points.

Non-linear svm[15] is used for non-linearly separated data which means dataset cannot be classified by using a straight line and it is termed as non- linear data.

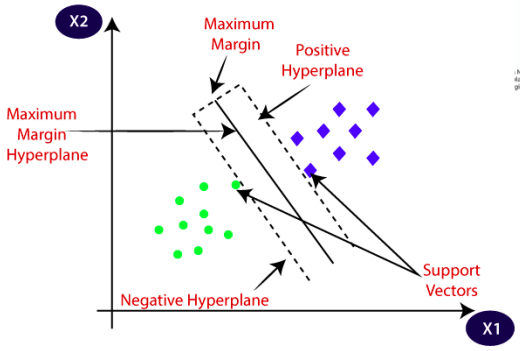

Fig. 2. Explaining SVM[16]

2) Random Forest

Random Forest algorithm is a supervised ML algorithm which can be used for regression as well as classification, but is mainly used for classification problems .The number of trees in the forest are directly proportional to the results it obtains: the more the number of trees, the higher is the accuracy.

When the obtained results for detecting fake reviews, this classifier yielded the highest accuracy along with precision and recall. Thus the final system for identifying fake reviews is built using the Random forest algorithm.

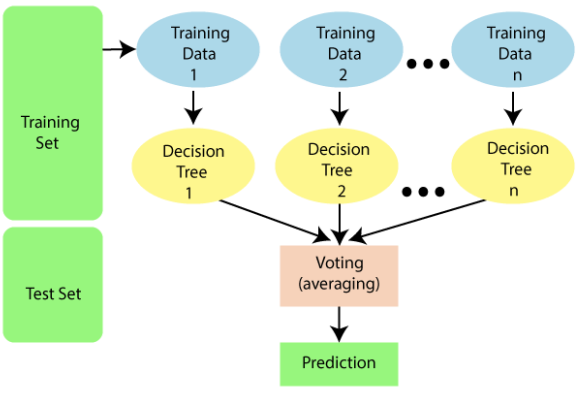

Fig. 3. Explaining Random Forest[17]

3) Logistic Regression Classifier

This supervised machine learning algorithm deals with probability to measure the relationship between dependent and independent variables.

It's goal is to find the best fitting model for independent and dependent variable relationships. Dependent variable is the response binary variable whose values are 0 and 1 and Independent variables are the predictor variables used to predict the response variable. Advantage of Logistic Regression is that it is easy to implement and very efficient to train[18]. Logistic regression is used for solving classification problems.

Logistic regression predicts the output of a dependent variable. Therefore the output must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

V. RESULT AND DISCUSSION

A. Outputs of Algorithms

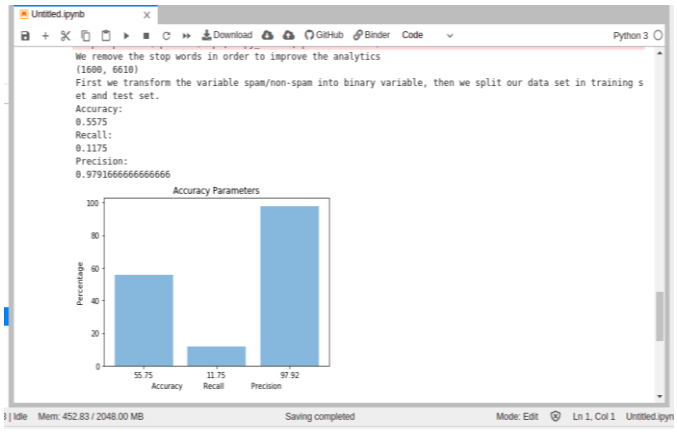

Fig. 4. Output of SVM

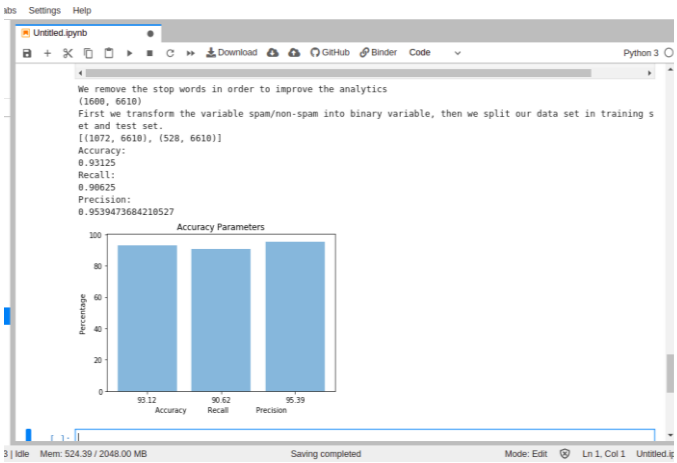

Fig. 5. Output of Random Forest

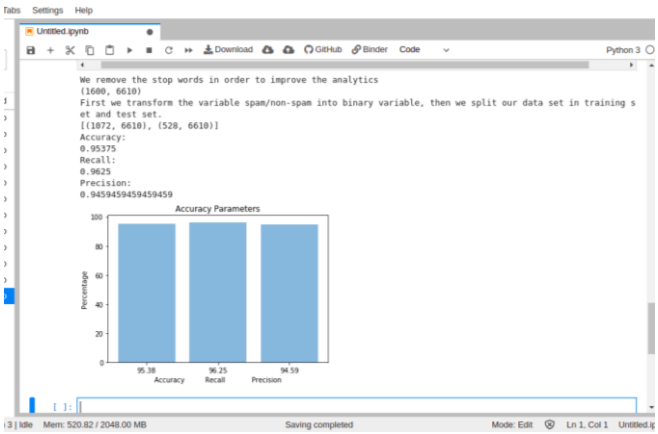

Fig. 6. Output of Logistic Regression

Comparison of Algorithms

Following table depicts the accuracy, precision and recall of the concerned algorithms when applied on raw dataset and on preprocessed dataset.

TABLE I. PERFORMANCE METRICS FOR DATABASE WITHOUT PREPROCESSING

CLASSIFIERS	ACCURACY	PRECISION	RECALL
SUPPORT VECTOR MACHINE	0.896015	0.912258	0.877171
LOGISTIC REGRESSION	0.952055	0.950556	0.954094
RANDOM FOREST	0.952055	0.953923	0.950372

TABLE II. :PERFORMANCE METRICS FOR DATABASE WITH PREPROCESSING

CLASSIFIERS	ACCURACY	PRECISION	RECALL
SUPPORT VECTOR MACHINE	0.923412	0.945241	0.899504
LOGISTIC REGRESSION	0.948318	0.942472	0.955335
RANDOM FOREST	0.945828	0.942189	0.950372

VI. CONCLUSION

This paper mainly focused on evaluating machine learning algorithms for reviewing spam detection.

The comparison is done on 3 different classifiers SVM, Logistic Regression and Random Forest. Of these, Logistic Regression gave best results with 95.20% accuracy, 95.40% recall and 95.05% precision.

Among SVM and Random Forest, Random Forest is the better one, because it has higher accuracy and recall than SVM.

REFERENCES

[1] DraskoRadovanovic and Bozo Krstajic , "Review spam detection using machine learning" ,IEEE 2018.

[2] Chirag Visani ,NavjyotsinhJadeja and Manali Modi, "Study on different machine learning techniques for spam review detection", IEEE 2017.

[3] Michael Crawford , Taghi M. Khoshgoftaar ,Joseph D.Prusa, Aaron N. Richter ,Hamzah Al Najada, "Survey of review spam detection using machine learning techniques", Journal of Big Data(Springer) 2015 .

[4] Naveed Hussain , Hamid Turab Mirza ,Ghulam Rasool , Ibrar Hussain , Mohammad Kaleem, "Spam review detection techniques :A systematic literature review", Applied Sciences ,2019.

[5] Ms. RajashriKashti ,Dr. Prakash Prasad, "Enhancing NLP Techniques for fake review detection", IRJET , volume 6 issue 02, Feb

[6] Mukherjee A, Venkataraman V, Liu B, Glance NS (2013) What yelp fake review filter might be doing? Boston, In ICWSM.

[7] N Jindal, B Liu, Proceedings of the 16th international conference on World Wide Web, 1189-1190

[8] N. Jindal and B. Liu. Analyzing and Detecting Review Spam.ICDM2007.

[9] A. Z. Broder. On the resemblance and containment documents. In Proceedings of Compression and Complexity of Sequences 1997, IEEE Computer Society, 1997

[10] Lai et al. (2010). Toward a language modeling approach for consumer review spam detection. In IEEE 7th international conference (pp. 8)

[11] Pennebaker et al. (2007). The development and psychometric properties of LIWC2007. Austin, TX, LIWC.Net.

[12] Mukherjee et al. (2013). What yelp fake review filter might be doing. In Seventh international AAAI conference on weblogs and social media.

[13] https://www.ibm.com/in-en/cloud/learn/machine-learning

[14] Dixit S, Agrawal AJ (2013) Survey on review spam detection. Int J Comput Commun Technol ISSN (PRINT) 4:0975–7449

[15] https://www.unite.ai/what-are-support-vector-machines/

[16] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[17] https://www.javatpoint.com/machine-learning-random-forest-algorithm

https://machinelearning-blog.com/2018/04/23/logistic-regression-101/