Online Fraud Review Detection Using Data Mining

[1]Pawar Amrita D., [2]Pawar Pooja S., [3]Dabhade Poonam V., [4]Kokane Tejaswini I., [5]Prof. B. A. Abhale

[1]B. E. Student, [2]Asst. Professor, Dept. of Information Technology. S.N.D. COE, Yeola. Maharashtra, India.

[1]amritapawar2018@gmail.com, [2]pawarpooja89@gmail.com, [3]poonamdabhade2017@gmail.com,
[4]tejaswinikokane@gmail.com

Abstract - Online reviews have great impact on today's business and commerce. Decision making for purchase of online products mostly depends on reviews given by the users. Nowadays, there are a number of people using social media opinions to create their call on shopping for product or service. Opinion Spam detection is an exhausting and hard problem as there are many faux or fake reviews that have been created by organizations or by the people for various purposes. They write fake reviews to mislead readers or automated detection system by promoting or demoting target products to promote them or to degrade their reputations, opportunistic individuals or groups try to manipulate product reviews for their own interests. This paper introduces some semi-supervised and supervised text mining models to detect fake online reviews as well as compares the efficiency of both techniques on dataset containing hotel reviews.

Keywords: Fake reviews, fake reviews detection, opinion mining, sentiment analysis, text mining, ICF.

I. INTRODUCTION

The identified challenges motivate to bring up a solution to all the problems stated in the above problem statement section. Following are the objectives of the proposed approach and this thesis work: To implement different algorithm to get better Spam Detection i.e.; IP Address, Account used, Negative Word Dictionary using Senti-strength, Ontology. Graphical representation of work. To deals with 6 different types of Spam Reviews. To presents Opinion Mining on Spam Filtered Data. To implement Ontology in Spam Detection to present an algorithm that does Opinion Mining with Spam Detection. For detection of fake online reviews, we start with raw text data. We have used a dataset which was already labelled by the previous researchers. We remove unnecessary texts like article and prepositions in the data. Then these text data are converted into numeric data for making them suitable for the classier. Important and necessary features are extracted and then classication process took place.

One of the very rapid growth area is ecommerce. Generally e-commerce provide facility for customers to write reviews related with its service. The existence of these reviews can be used as a source of information. For examples, companies can use it to make design decisions of their products or services, while potential customers can use it to decide either to buy or to use a product. Unfortunately, the importance of the review is misused by certain parties who tried to create fake reviews, both aimed at raising the popularity or to discredit the product. This research aims to detect fake reviews for a product by using the text and

rating property from a review. In short, the proposed system (ICF++) will measure the honesty value of a review.

Due to the absence of reliable data that has been labeled fake or genuine for training, Jindal and Liu (2008) utilized duplication in the review. In their study of 5.8 million reviews and 2.14 million reviewers of amazon.com, they found a large number of duplicate and near duplicate review, which showed that the presence of fake reviews is widespread. This is likely due to the difficulty of writing some new review when the spammers never buy the product or use the service.

Automatic detection of spammers is a very important work but still lacks in research. Unlike other types of spam, such as web spam or email spam, spam on a review is far more difficult to detect. The main reason is that spammers can easily disguise themselves. Thus it is difficult, for users to recognize, while web spam or email spam, one can determine spam or not without much difficulty.

Due to the absence of reliable data that has been labelled fake or genuine for training, Jindal and Liu (2008) utilized duplication in the review.

They tend to use the same review or slightly revised for different products. This duplication can be divided into four categories, that is
(1) duplicates from the same customer id on the same product,
(2) Duplicates from different customer id on the same product,
(3) Duplicates from the same customer id on different products,

(4) Duplicate from different customers id on different products.

In Our implementation we are using classification algorithm and new tools of opinion mining.

In our implementation we are taking example of books ecommerce here publisher can sell their books, view Books here publisher can check user fake reviews.

II. LITRATURE SURVEY

[1] Many approaches and techniques have been proposed in the eld of fake review detection. The following methods have been able to detect fake online review with higher accuracy. Sun et al. [1] divided these approaches into two categories. a) Content Based Method: Content based methods focus on what is the content of the review. That is the text of the review or what is told in it. Heydari et al.

[2] Have attempted to detect spam review by analyzing the linguistic features of the review. Ott et al.

[3] Used three techniques to perform classication. These three techniques are-genre identication, detection of psycholinguistic deception and text categorization [1]–[3]. 1) GenreIdentication: The parts-of-speech (POS) distribution of the review are explored by Ott et al. they used frequency count of POS tags as the features representing the review for classication. Detection of Psycholinguistic Deception: The psycholinguistic method approaches to assign psycholinguistic meanings to the important features of a review. Linguistic Inquiry and Word Count (LIWC) software was used by Penne baker ET al.to build their features for the reviews.

3) Text Categorization: Ott et al. experimented n-gram that is now popularly used as an important feature in fake review detection. Other linguistic features are also explored. Such as, Feng et al.

[5] Lexicalized and lexicalized syntactic features by constructing sentence parse trees for fake review detection. They show experimentally that the deep syntactic features improve the accuracy of prediction. Li et al.

[6] Explored a variety of generic deceptive signals which contribute to the fake review detection. They also concluded that combined general features such as LIWC or POS with bag of words will be more robust than bag of words alone. Metadata about reviews such as reviews length, date, time and rating are also used as features by some researchers.

b) Behavior Feature Based Methods: Behavior feature based study focuses on the reviewer that includes characteristics of the person who is giving the review. Lim et al.

[7] Addressed the problem of review spammer detection, or nding users who are the source of Spam reviews. Peoplewhopostintentionalfakereviewshavesignicantlydiffer entbehavior than the normal user. They have identied the following deceptive rating and review behaviours.

Giving unfair rating too often: Professional spammers generally posts more fake reviews than the real ones. Suppose a product has average rating of 9.0 out of 10. But a reviewerhasgiven4.0rating. Analysing the other reviews of their viewer if wend out tha the often gives this type of unfair ratings than we can detect him as a spammer.

Giving good rating to own country's product: Sometimes people post fake reviews to promote products of own region. This type of spamming is mostly seen in case of movie reviews. Suppose, in an international movie website an Indian movie have the rating of 9.0 out of 10.0, where most of the reviewers are Indian. This kinds of spamming can be detected using address of their viewers. Giving review on a vast variety of product: Each person has specie interests of his own. A person generally is not interested in all types of products. Suppose a person who loves gaming may not be interested in classic literature. But if we some people giving reviews in various types of products which exceeds the general behaviour then we can intuit that their reviews are intentional fake reviews. Deceptive online review detection is generally considered as a classication problem and one popular approach is to use supervised text classication techniques [5]. These techniques are robust if the training is performed using large datasets of labeled instances from both classes, deceptive opinions (positive instances) and truthful opinions (negative examples)

[8] Some researchers also used semi-supervised classication techniques. For supervised classication process ground truth is determined by–helpfulness vote, rating based behaviours, using seed words, human observation etc. Sun et al. [1] proposed a method that offers classication results through a bagging model which bags three classier including product word composition classier (PWCC), TRIGRAMSSV M classier, and BIGRAMSSV M classier. They introduced a product word composition classier to predict the polarity of the review. The model was used to map the words of a re view into the continuous representation while concurrently integrating the product-review relations. To build the document model, they took the product word composition vectors as input and used Convolutional Neural Network CNN to build the representation model. After bagging the result with TRIGRAMSSV M classication, and BIGRAMSSV M classication they got F-Score value 0.77. However supervised method has some challenges to overcome. The following problems occur in case of supervised techniques.

1. Assuring of the quality of the reviews is difficult.

2. Labeled data points to train the classier is difcult to obtain.

3. Human are poor in labeling reviews as fake or genuine. Hence Jitendra et al.

[2] S. R. Nandurkar ; V. R. Thool ; R C. Thool ,Crop farming in India is labour intensive and obsolete. Farming is still dependent on techniques which were evolved

hundreds of years ago and doesn't take care of conservation of resources. The newer scenario of decreasing water tables, drying up of rivers and tanks, unpredictable environment present an urgent need of proper utilization of water. We have the technology to bridge the gap between water usage and water wastage. Technology used in some developed countries is too expensive and complicated for a common farmer to understand. Our project is to give cheap, reliable, cost efficient and easy to use technology which would help in conservation of resources such as water and also in automatizing farms.

Unlabelled data both are trained together. They proposed to use semi-supervised method in the following situations. 1) When reliable data is not available. 2) Dynamic nature of online review. 3) Designing heuristic rules are difficult. They proposed several semi-supervised learning techniques which includes Co-training, Expectation maximization, Label Propagation and Spreading and Positive Unlabelled Learning

[8]. They used several classier which includes k-Nearest neighbour, Random Forest, Logistic Regression and Stochastic Gradient Descent. Using semi-supervised techniques they achieved highest assurance of 84percentage. Conclusion: It will help us to find out fake reviews by using NaıveBayesas algorithm. To find out fake review in the website this "Fake Product Review Monitoring System" system is introduced. This system will findout fake reviews made by the customers and it will block the users.

III. PROPOSED SYSTEM

The main goal of the project is to provide the original and real reviews to the customer who want the real opinion for the product which they are want to purchase.

1. Separates the reviews of customer in positive and negative. 2. Train the keyword for the separation purpose.

3. Uploads the Reviews.

4. Provides the real reviews.

Architecture
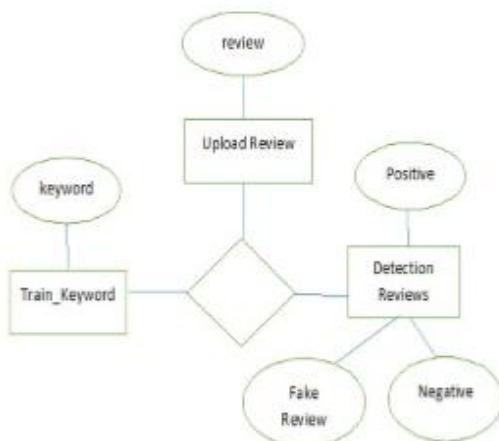
Project scope

In our research work we have worked on just user reviews. In future, user behaviours can be combined with texts to construct a better model for classication. Advanced prepossessing tools for tokenization can be used to make the dataset more precise. Evaluation of the effectiveness of the proposed methodology can be done for a larger data set. This research work is being done only for English reviews. It can be done the several other languages

Module:

Admin Login:

Admin can as publisher he can publish his book on our application so he can sell their books online and generate finance also admin can view users who are interested in his books.

View Users:

In this module admin can view register users in the system.

View Books:

Admin can check added books and alter data of books.

View Purchase Books:

Admin can view purchased Books.

View User All Review:

Admin can check all users review.

View Positive/Negative Review:

Admin can also check positive and negative review.

User Register:

User Have to register first in system then user will be authorized user of system.

User Login:

After register successfully user can be login in system.

View All Books:

User can all category books on site.

Search Book:

User can search book by keyword and user will get list of books relevant to keyword.

Add Review:

User can add review on books but if user simultaneously added negative review or fake review user will blocked automatically.

Buy Book:

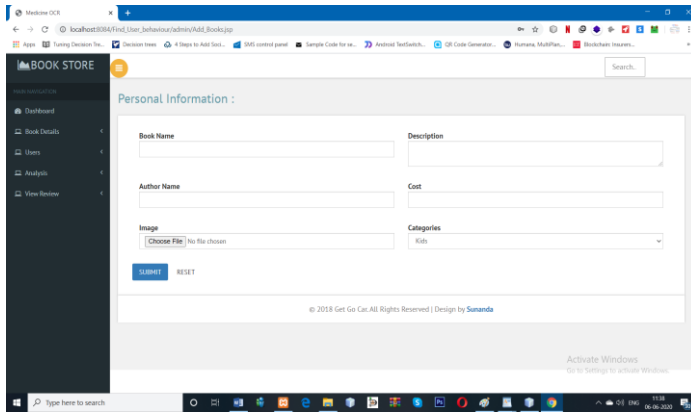User can purchase books from site**.**

IV.   RESULT

Fig.1 Add Books

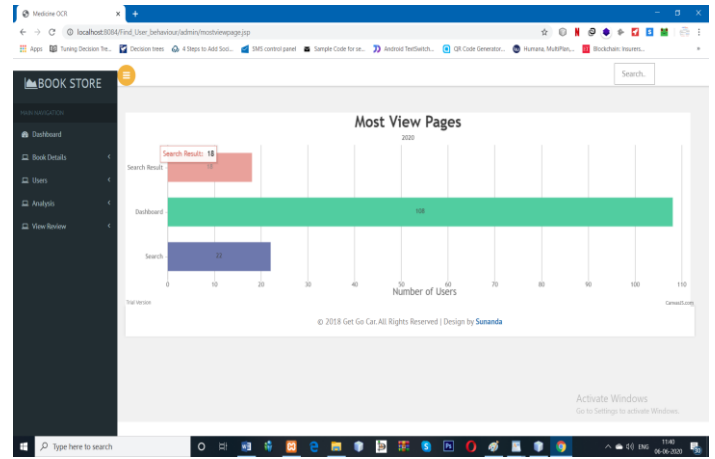In above figure Admin can add book into the system with parameters book name, description, author category and image.
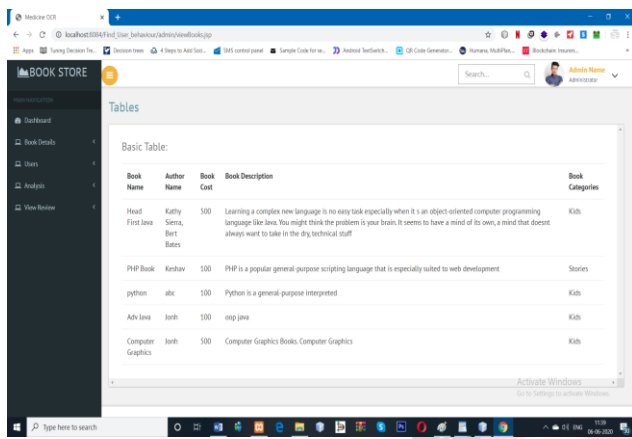

Fig.2 View Books

In above figure admin can view book details name, author, description and category.

Fig.3 Most Purchase Category

In above you can check analysis of most purchase category.

Fig.4 Most View Pages

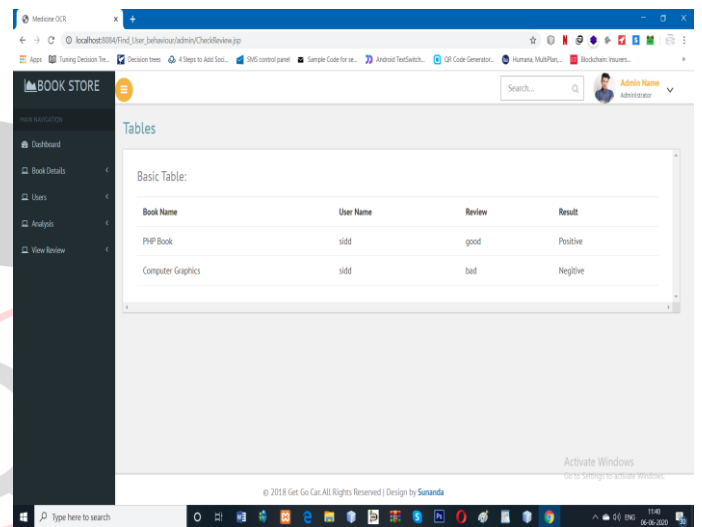In above you can check analysis of most View pages.

Fig.5 Positive/Negative Review

In above you can check analysis of positive and negative review.

Algorithm:

INPUT:

Labelled instance set L and unlabeled instance set U

OUTPUT:

Deploy-able classifier, C.

1 C ¡- train (L);

2 PU = NULL;

3 while True do

4 PU = predict(C, U)

5 if, PU is same as in previous iteration,

Then

6 return C;

7 end if

8 C ¡- train (L U PU)

9 end while
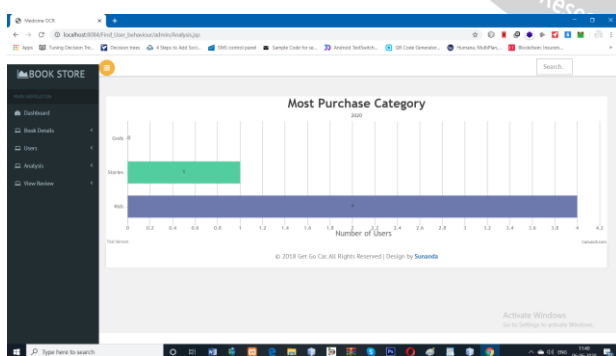
V. CONCLUSION

Other classifier that were not used on the previous work. Thus, we have been able to increase the accuracy of previous semi supervised techniques done. We have also found out that supervised Naive Bayes classier gives the highest accuracy. This ensures that our dataset is labelled well as we know semi-supervised model works well when reliable labelling is not available. In our research work we have worked on just user reviews. In future, user behaviours can be combined with texts to construct a better model for classication. Advanced pre-processing tools for tokenization can be used to make the dataset more precise. Evaluation of the effectiveness of the proposed methodology can be done for a larger data set. This research work is being done only for English reviews.

REFERENCES

[1] Coenen, F., Goulbourne, G. & Leng, P., 2003. Tree Structures for Mining association Rules. Journal of Data Mining and Knowledge Discovery, Vol 8, No 1, pp.25-51

[2] Forman, E.H., 1993. Facts and fictions about the analytic hierarchy process. Mathematical and Computer Modelling, Volume 17, Issues 4–5, pp.19-26.

[3] Fournier-Viger, P., Gomariz, Gueniche, T., A., Soltani, A., Wu., C., Tseng, V. S. 2014. SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research (JMLR) 15, pp. 3389-3393

[4] Han, J., Pei, J., Yin, Y. & Mao, R., 2004. Mining Frequent Patterns without Candidate Generation: A FrequentPattern Tree Approach. Data Mining and Knowledge Discovery 8, pp.53–87.

[5] Heydari, A., Tavakoli, M.A., Salim, N., & Heydari, Z., 2015. Detection of review spam: A survey. Expert Systems with Applications, 42(7), pp.3634–42.

[6] Jindal, N., & Liu, B. 2008. Opinion Spam and Analysis. Proceedings WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining, 219-230.

[7] Khan, K., Baharudin, B., Khan, A., & Ullah, A. , 2014. Mining opinion components from unstructured reviews: A review. Journal of King Saud University - Computer and Information Sciences, 26(3), pp.258–75

[8] Liu, B., 2009. Opinion Mining. Encyclopedia of Database System, pp.1986-90.