IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 7, Issue 3, June - July, 2019
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

FEATURE SET EVALUATION MODEL FOR OPINION FRAUD DETECTION

Dr.K.Suvarchala[1], M.Pramodini[2], B.Rajkiran[3], Priyam Yadav[4]

[1]Institute of Aeronautical Engineering, Hyderabad, Telangana, India
[2]Institute of Aeronautical Engineering, Hyderabad, Telangana, India
[3]Institute of Aeronautical Engineering, Hyderabad, Telangana, India
[4]Institute of Aeronautical Engineering, Hyderabad, Telangana, India

Abstract
The volume of electronics has increased a lot over the years, mainly due to the popularity of e-commerce. Since this thunderstorm, we have seen a dramatic increase in the number of fraudulent cases, leading to a loss of billions of dollars each year worldwide. Therefore, it is important and necessary to develop and implement strategies that can help detect fraudulent Web sales. From a large amount of data generated from the transactions, getting the optimal set of features is a crucial task to identify fraud. Fraud detection is a specific anomaly detection application, characterized by significant imbalances between classes (e.g. fraud or dishonesty), which can be a destructive component of feature selection techniques. Further detection of fraud is a trivial and challenging problem
Therefore, manual labelling of reviews is difficult and ground truth information is often unavailable, which makes monitoring models less attractive to this problem. In this work we examine the effectiveness and impact of feature selection methods to detect deception in the Web Transaction environment, using supervised and semi-supervised techniques with the help of machine learning algorithms. We also aim to differentiate between the performances among the algorithms that are giving the best results for the datasets in detecting online reviews. We achieve the best performance in detecting fraud using the proposed method, reducing the number of symptoms.
Keywords: Fake reviews, online platform, Frauds, Machine learning algorithms, Products, Business, Supervised method, Semi-supervised method

Introduction
The effect of online critiques on corporations has grown considerably during ultimate years, being vital to decide business success in a big range of sectors, ranging from restaurants, resorts to e-commerce [10]. Unfortunately, a few users use unethical means to improve their online recognition with the aid of writing faux evaluations of their agencies or competitors. Advancements are evolving quickly. Old advancements are ceaselessly being supplanted by new and refined ones [3]. These new advancements are empowering individuals to have their work done efficiently. Such an advancement of innovation is online commercial centre.
We can shop and reserve spot utilizing on the web sites. Nearly, everybody of us looks at audits before buying a few items or administrations [5]. Henceforth, online surveys have become an incredible wellspring of notoriety for the organizations. Likewise, they have huge effect on notice and advancement of items and administrations [2]. With the spread of online commercial centre, counterfeit online surveys are getting extraordinary matter of concern. Individuals can make bogus audits for advancement of their own items that hurts the real clients. Additionally, serious organizations can attempt to harm every others notoriety by giving phony negative surveys. Previous research has addressed fake evaluation detection in a number of domains, inclusive of product or business critiques in eating places and hotels [2, 3].
Ongoing methodologies have proposed the extra utilization of highlights that consider the social structure of the system hidden the considered survey site. These methodologies, which are regularly founded on solo diagram based techniques, for the most part furnish more terrible presentation regarding managed arrangements.
In this paper, we make some classification tactics for detecting faux online reviews. Some of that are semi supervised and others are supervised. For semi-supervised learning, we use label Propagation and label spreading algorithms. K Nearest Neighbours, Random Forest and Multi layers algorithms are used as classifiers in our studies work to improve the overall performance of classification. We have especially focused on the content material of the overview based tactics.

Specialists have been reading about numerous methodologies for discovery of these phony online audits. A few methodologies are survey content put together and some are based with respect to conduct of the client who is posting audits. Content put together examination centres with respect to what is composed on the audit that is the content of the survey where client conduct put together technique centres with respect to nation, ip-address, and number of posts of the

analyst and so on [3]. The vast majority of the proposed approaches are administered classification models. Not many specialists, additionally have worked with semi-administered models. Semi-directed strategies are being presented for absence of solid naming of the surveys.

In the previous not many years, counterfeit survey identification has pulled in noteworthy consideration from both the modern associations and scholarly networks. In any case, the issue stays to be a difficult issue because of lacking of naming materials for administered learning and assessment [5]. The social Web and the expanding notoriety of online networking have prompted the spread of different sorts of substance created straightforwardly by clients, the purported client produced content. By methods for Web innovations, it is feasible for each person to diffuse substance via web-based networking media, nearly with no type of confided in outside control. This infers that there are no way to check, from the earlier, the unwavering quality of the sources and the acceptability of the substance created [8]. Right now setting, the issue of evaluating the believability of the data diffused by methods for internet based life stages is accepting expanding consideration from specialists. Specifically, this issue has been profoundly researched in audit destinations, where the spread of deception as supposition spam, and the negative outcomes that it brings, are especially unsafe for the two organizations and clients [9]. Right now, spam discovery targets distinguishing counterfeit surveys, counterfeit remarks, counterfeit web journals, counterfeit interpersonal organization postings, tricksters, and tricky messages and to make them promptly unmistakable.

Literature survey

Text mining model:

Here we have discussed the Detection of fake online reviews using semi-supervised and supervised learning and presented an efficient model. They introduced a text mining model. They also presented the comparison of the efficiency of technique on hotel review systems [10]. They presented the model based on the content of the review meaning the behavior of the user regularly. They approached this problem with a supervised classification model. Some are supervised and a few are semi-supervised. They used an expectation-maximization algorithm with a semi-supervised model and used naive-Bayes classifier [10]. They mainly focused on the content of the review with features like word frequency count, sentiment polarity, length of the review. The result was an improved accuracy than the previous approaches. Out of which naive Bayes classifier gives the very best accuracy [10].

Collective behaviour model:

Here we have discussed Collusive Fraud Detection in Online Reviews using an unsupervised model. It mainly focuses on the collective behaviors of the reviewers. One of the ways is to check for the commonly reviewed products. The drawback of this model is that it cannot learn from the existing model to make predictions [3]. They proposed a latent collusion model which is a statistical model it checks for the unique collective behavioral patterns. Not only can LCM perform collusion inference as unsupervised models, but it can also make collusion predictions as supervised [3]. There are a few problems associated with this approach, it uses unlabeled data as an unsupervised model, and it is difficult to handle uncertainty in the measures.

Regression technique model:

Here we have discussed Detecting Frauds in Restaurant Reviews this model mainly works for detecting legitimate and illegitimate customers they used auto regression moving average to predict the ratings. Then they classify the basic differences between customer class, predicted rating and actual rating by the customer [9]. This model has given and very high accuracy. they applied basic process such as checking whether the review is the only review of the user on the particular website if the reviews are too long in terms of content and if the rating is too low, if the reviews are extremely positive etc [9].it mainly focuses on data analysis such as probabilities, time serviced, user behavior then they classify and cluster to find patterns .this model is also turned out to be efficient.

Detection Using Processing Of Natural Language-NLP:

This fraud news detection model works to assist the user to know and analyze number of fraud news. this whole model is based the process working backwards where they first analyze the true and false news from the past [5] .we used natural language processing (NLP) for detection of deceive news. we are mostly focused on separating the serous and genuine news from the deceptive ones [5].this whole working model can be named as a predictive model. The whole information is drawed from the library and information science sections.

PROPOSED WORK

Current works made numerous endeavours to address this issue from the edges of commentator and survey. In any case, there has been little conversation about the item related audit highlights which is the principle focal point of our main Strategy.

In our project, we used both supervised and semi supervised methods in order to detect the online spammers. This paper proposes a novel convolutional neural system model to coordinate the item related audit includes through an item word creation model. To lessen over fitting and high difference, a sacking model is acquainted with pack the neural system model with two effective classifiers. Investigations on the genuine Amazon audit dataset exhibit the viability of the proposed approach. In our project, we used both supervised and semi supervised methods in order to

detect the online spammers. Word frequency, sentiment polarity and length of review are the features that are used in our project. In the proposed system, each review goes by each and every character. Then, unnecessary words are removed and candidate feature words are generated.

Each candidate feature words are checked against the dictionary and if its entry available in the dictionary then its frequency is counted and added to the column the feature vector that corresponds the numeric map of the word. Alongside with counting frequency, the length of the review is measured and added to the vector. Finally, sentiment score which is available in the data set is added in the feature vector.

The text mining models to detect fake online reviews as well as prediction the sentiment analysis for reviews using the machine learning on containing product reviews. In sentimental analysis, we will come to know the efficient algorithm.

	features	Algorithm type	Classifier used	accuracy
Hassan	Word frequency count, Sentiment score, Review size	Semi-supervised	Naive bayes	0.8521
			SVM	8.8134
		supervised	Naive bayes	0.8621
			SVM	0.8228
proposed	Term frequency, Average content, Negative ratio	Semi-supervised	Label propagation	0.8973
			Label spreading	0.9053
		Supervised	K-nn	0.7589
			Random forest	0.7586
			MLP	1.00

Comparative summary of semi-supervised and supervised learning techniques

Random forest algorithm

Random Forest is a machine learning algorithm that uses a packing procedure to create a decision tree with a random subset of data. A sample is trained several times in a random sample of data to achieve good estimation performance from the random forest method. This collective learning method combines the production of all decision trees in a random forest to form a final estimate. The final estimate of the random forest method can be obtained by voting the results of each decision tree or by predicting that the decision tree will have more time

K-Nearest Neighbour

The K-nearest neighbor (k-NN) algorithm is a technique utilized for arrangement and bunching. In the two cases, the info contains the nearest instances of preparing in the component field. Quit relies upon whether k-NN is utilized to disconnect or pack.

According to the k-NN class, the outcome is a class part. The thing is part by the votes of the vast majority of its neighbors, something given by the most well-known class among the nearest neighbors (k is something with an ordinary, modest number). On the off chance that k = 1, the thing is just relegated to the class of the closest neighbour**.**

Multilayer Perceptron

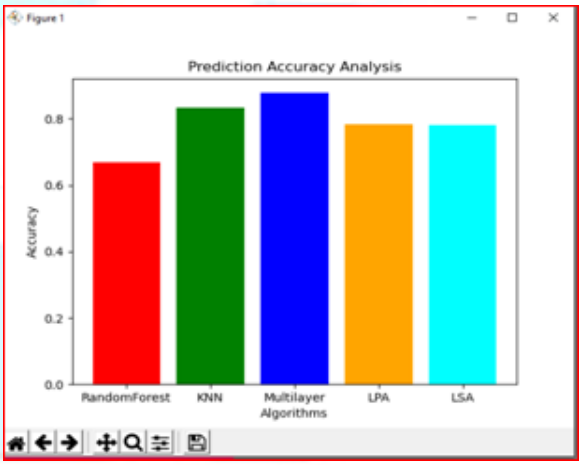
Multilayer perception algorithm is one of the neural network technique. The perceptron, that neural system whose name summons how the future looked from the point of view of the 1950s, is a basic calculation proposed to perform parallel grouping; for example it predicts whether info has a place with a specific class of intrigue or not: misrepresentation or not fraud, feline.

Label propagation algorithm

Label propagation algorithm is one of the semi supervised technique. The Label Propagation calculation (LPA) is a quick calculation for discovering networks in a diagram. It distinguishes these networks utilizing system structure alone as its guide, and doesn't require a pre-characterized target work or earlier data about the networks. It works by engendering names all through the system and shaping networks dependent on this procedure of mark proliferation.

Label spreading algorithm

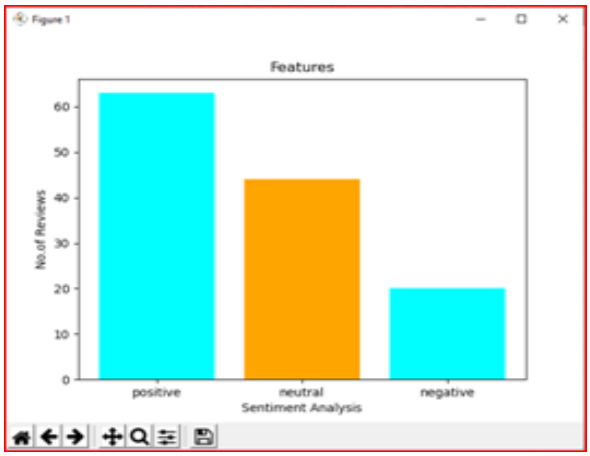
Label spreading algorithm depends up on the normalized graph laplacian. This
algorithm is used rarely in previous history. So, we introduced this algorithm in
our project to get better results



Comparison of accuracies

By using the proposed features, multilayer perceptron algorithm shows the highest accuracy when compared to remaining algorithms.

In the graph, we have taken algorithms on x-axis and their accuracies on y-axis.



Sentimental analysis graph

After uploading the review file which consists of positive, negative, highly positive and highly negative, it automatically undergoes sentimental analysis process and gives the sentimental analysis graph. Sentimental analysis graph will give the number of positive, negative and neutral reviews by analysing the entire dataset file.

Conclusion and future enhancement

We have worked with the classification approach for detecting fraud opinions. We have applied various combinations of classifiers and algorithms with supervised and unsupervised techniques. The advantage we have over the previous research works is that we have combined unusual combinations of classifiers with algorithms that have turned out to be efficient than the previous ones. the whole data have been trained on the basis of the previous history of the data .in the future advancement, we can also use the user's portfolio along with the review presented for making an optimal decision .we can use this model for large data sets as well .this whole model is designed based on the English language and we are looking forward to future advancements with other languages as well.

We have a high scope of future enhancements for the project we have presented. We are still trying to build enhancements with better working model. We are trying to include various attributes such as tracking IP addresses of the reviewers which will help us in building a model that can eradicate the spamming organisation and also make our model better and efficient.

References

[1] N. Jindal and B. Liu, "Opinion spam and analysis," in WSDM, 2008.

[2]. Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning, Naveen Kumar, Deepak Venugopal, Liangfei Qiu and Subodha Kumar, Journal of Management Information Systems, 2018, Volume 35, Number 1, Page 350.

[3]. Xu C, Zhang J. Towards collusive fraud detection in online reviews. In: 2015 IEEE international conference on data mining; 2015. p. 1051–6.

[4] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition,andyelpreviewfraud,"HarvardBusiness School NOM Unit Working Paper, no. 14-006, 2013.

[5] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in WWW, 2012.

[6] C. Xu, J. Zhang, K. Chang, and C. Long, "Uncovering collusive spammers in chinese review websites," in CIKM, 2013.

[7] M.Rahman,B.Carbunar, J.Ballesteros,G.Burri,and D.H.P. Chau, "Turning the tide: Curbing deceptive yelp behaviors," in SIAM SDM, 2014.

[8] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Identify online store review spammers via social review graph," ACM TIST, vol. 3, no. 4, p. 61, 2012.

[9] Weiwen Yang and Linchi Kwok," Detecting Frauds in Restaurant ReviewsConference: 2013 International Conference on Computer Sciences and Applications (CSA)".

[10] Rakibul Hassan and Md. Rabiul Islam," Detection of fake online reviews usingsemi-supervised and supervised learning" 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE).