Reputation Reporting System using Text Based Classification

Divyanshu Jalther, Priya G

Abstract: ***Reputation System is a system which allow users to rate and review an organization or a product so that other users or customers can judge an organization or a product be seeing the reviews and ratings of an organization or a product. But the predictive value of reputation reporting system can be manipulated either by buyers or competitors to promote or demote a product or an organization. Fake review detection has attracted significant research attention in recent years. Some research has been done using dataset produced by fake review generator which was found inefficient. Some research has been done using behavioral pattern of spammers and pattern of fake reviews written which has produced some better results. In this paper, we implemented an approach to detect biased feedback using supervised machine learning algorithm. We used data from yelp.com which contains labelled dataset of restaurants in New York to train and test different classifier. In the end, we compared the accuracy of different classifiers to conclude which classifier has worked best on textual data. This model can be used by any service or product provider companies to detect and deleted biased feedback from their website***

Index Terms: Machine Learning, Reputation System, Text based Classification, Online Fraud

I. INTRODUCTION

Product or organization reviews are becoming increasingly important in today's time. Good reviews of a product or an organization can lead to increased fame of an organization or increased sales of a particular product. On the other hand, this reputation system can be manipulated by imposter or competitor to defame an organization or a particular product. Manipulating a review by a person is called opinion spamming and the person who does this is called opinion spammer. As seen in the recent trends, this problem of giving fake reviews about product or organization is increasing. Competitor company or people are hiring people to write fake online reviews about a product or an organization to defame. Companies are also hiring people to write good reviews about their own organization or product to get an edge over the other competitors.

Establishment of trust is important between customer (or client) and service provider. Trust can only be established by seeing the past rating and behavior of the service/product provider. The main motive behind reputation system is to provide the genuine review of the provider using its past transaction history, reviews and ratings. Therefore, it is very important to protect the reputation system from imposters and competitors which gives a biased image of service/product provider.

II. AIM OF THE PROPOSED WORK

The aim of the proposed work is to provide the customer or client with a clear and a transparent image of the provider in terms of ratings and review on an online platform using machine learning algorithm. Biased review can be detected using machine learning algorithm and that particular review can be deleted so that it does not affect the overall rating of an organization or a product or a service. This project is applicable wherever client and service provider comes into play. The application of this work can be extended to multiple platforms.

III. LITERATURE SURVEY

Arjun Mukherjee, Vivek Venkataraman, Bing Liu and Natalie Glance did a study of real and Pseudo comments [1]. They did classification and analysis. Pseudo fake comments were produced from Amazon Mechanical Turk (AMT) system. They found out that classifying pseudo fake review is easier than real fake reviews. Pseudo fake reviews were giving an accuracy of 89.6%. While analysis of real fake comments gave an accuracy of 67.8%. They concluded that AMT dataset trained model are very weakly trained and are not efficient enough to detect real fake comments written by imposter or competitor to defame or promote a product or on an organization. They proposed a set of behavioral features which greatly increased the accuracy of the classifier.

Simran Bajaj, Niharika Garg and Sandeep Kumar Singh took into account the user characteristics instead of only reviews written by reviewers [2]. They considered the GPS location and internet protocol address of the reviewers along with what they have written in review. They set threshold on number of reviews that can be posted by a reviewer. Only 1 review can be posted from a unique email id. 2 reviews can be given from a specific device. 3 reviews can be given from a particular GPS location. They tested their model on size of 100 dataset which showed 65% accuracy.

Huayi Li, Zhiyuan Chen, Bing Liu, Xiaokai Wei and Jidong Shao did their research using dataset on dianping.com [3]. They worked on a dataset which had fake reviews and unlabeled reviews. Unlabeled reviews mean that the dataset can comprise of genuine as well as fake reviews. They proposed a model though which classification can be made precisely from positive and unlabeled data set. They concluded that accuracy of their proposed model is much better than baseline algorithm. Their proposed

Revised Manuscript Received on June 04, 2019.

Divyanshu Jalther, School of Computer Science and Engineering (SCOPE), VIT University, Vellore, Tamil Nadu, India.

Priya G, School of Computer Science and Engineering (SCOPE), VIT University, Vellore, Tamil Nadu, India.

algorithm was language independent and hence can be extended to reviews existing in any language.

In the paper written by Chrysanthos Dellarocas, He identified many scenarios where buyers and sellers try to demote or promote a product/service or an organization [4]. Some of those ways to defame was ballot stuffing, bad mothing and positive seller discrimination. These methods resulted in biased review about some product/service or on organization. To deal with the above-mentioned scenarios, he proposed two models- cluster filtering and controlled anonymity. Also, he tested the effectiveness of cluster filtering by changing some parameters. So basically, he tested using different flavors of cluster filtering and found out that combination of cluster filtering and controlled anonymity is very effective in protecting the online reputation of seller/product/service or an organization.

The work of Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu and Jidong Shao addressed the problem of opinion spamming [5]. They worked on a very large dataset of dianping.com. It is a website which contains reviews with their classification. They considered many parameters before classifying a review as fake or genuine. Some of those parameters were - if the user is registered on the website or not, if the user has registered himself between Tuesday and Thursday, number of Unique internet protocol address used by a user, number of cookies which are unique in nature used by a user. On the bases on their analysis, they proposed novel temporal and spatial features for supervised machine learning algorithm to detect fake reviews.

Li, Huayi, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao studied the behavior of spammers on Dianping's dataset [6]. They studied the behavior of the spammers on the basis of number of times reviews written and number of reviews written by the spammer. Posting rates were found to be bimodal. Spammers were found to write fake reviews on a single product/organization in a short period of time which they called as co-bursting. They proposed a two-mode Labeled Hidden Markov Model to model spamming using only individual reviewers' review posting times. Then they extended it to the Coupled Hidden Markov Model to capture both reviewer posting behaviors and co-bursting signals. Their results showed better results in comparison to existing models.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., & Ghosh, R. proposed an unsupervised machine learning method based on bayes settings [7]. They claimed that there is visible behavioral distribution pattern between spammer and non-spammers. They studied this behavior and incorporated their study in their proposed model on Amazon review dataset. Their proposed Author Spamicity Model (ASM) showed effectiveness and it performed better than existing model to detect opinion spam.

Lin, Y., Zhu, T., Wu, H., Zhang, J., Wang, X., & Zhou A worked on fake review detection [8]. They noticed the order of fake review which followed some pattern. Based on reviews written and activity of reviewers, they proposed a supervised model which showed promising results with high accuracy.

Luca, M., & Zervas, G. studied the business incentives of posting fake review using yelp dataset [9]. Their findings include different circumstances when fake reviews of an organization or a product are increased. They claimed that a restaurant posts fake reviews when their reputation is weak or when there is an increased competition among restaurants. They found that fake reviews are mostly positive reviews which must have been posted by the restaurant itself. They further revealed the financial profit to organization who are doing business of posting fake reviews.

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R did their study on Amazon dataset and found some bursting pattern when spam reviews are posted [10]. They claimed that fake reviews tend to appear together and genuine reviews tend to appear together. They exploited this behavior of spammers and proposed a model to detect a spammer. They also proposed another supervised classification system to classify genuine and fake reviews. Both of the proposed model showed good results.

Mukherjee, A., Liu, B., & Glance, N studied the behavior of groups of spammers who are employed to write fake reviews rather than studying individual spammers behavior [11]. They claimed that fake reviewer group tend to work together to promote or demote a specific product and hence can be classified. They proposed a model to study such behavioral pattern and it outperformed the existing state of art baseline model.

IV. MODULES OF THE PROPOSED SYSTEM

The implementation of the proposed work can be divided into 2 parts. In the 1st part, we get the data from yelp.com about the restaurant reviews which is achieved by performing website scraping using python script. In the 2nd step, we used the obtained dataset to train and test the classifier.

In 1st step, we are fetching reviews of 150 restaurants of New York found in and around Pin Code 10001. A total of 46,266 reviews has been fetched from 150 restaurants. Out of 46,266 reviews, 11,103 reviews are fake and 35,163 reviews are genuine. This categorization of data into fake and genuine has been done on yelp.com which we will use as ground truth for machine learning algorithm to learn. After fetching dataset from the website, it is saved in .csv format file so that it can be read during training phase of a classifier. CSV (comma separated value) file can be considered as central data repository as it consists of both fake and genuine labelled data. It has been made sure that no duplicate reviews are picked from the website. Reducing the duplicate reviews decreases the size of dataset which takes a lot of space and it increases the quality of dataset.

2nd step is to test and train the classifier. Before training the classifier, the dataset has to be pre-processed and must be converted to tf-idf vector so that textual data is converted into statistical measure. In pre-processing, all the stop words have to be removed. Stop words removal has been taken care during construction of tf-idf vector. Removal of stop words decreases the space consumed by dataset and increases the quality of dataset. Tf-idf vector is basically words in data along with their frequency in corresponding

dataset. Once tf-idf vector is ready, it is passed to classifier along with their category to which it belongs. Train data and test data is divided in the ratio of 75:25. 75 correspond to training dataset and 25 corresponds to testing dataset. After the classifier has been trained, it is tested and the resultant accuracy and confusion matrix of every classifier used are noted. Different classifiers used are K nearest neighbor (KNN), Logistic Regression, Support Vector Machines (SVM), Decision Tree, Random Forest.

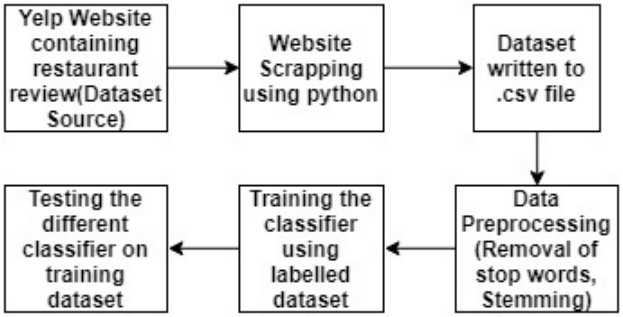

Fig. 1. Modules of the system

V. RESULTS AND DISCUSSION

The implementation is done using python script. Dataset of 150 restaurant present in New York in and around pin code 10001 was collected from yelp.com and is written in .csv file. The website has reviews about restaurants of New York along with the label of genuine or fake with every review (which we considered as a ground truth for training different classifiers). We fetched the author of review, id of the author, review written by the author, rating given by the author to restaurant, date of review and category of review (fake or genuine review). In the classifier, review comment and category of the review is passed. The dataset is divided into training and testing dataset in the ratio of 75:25 respectively. The training set is mostly kept large so that classifier can be trained properly. Yelp.com does not delete fake or biased review. Instead it keeps them under 'not recommended reviews' category which is publicly available.

Pre-processing of dataset was performed which involved removal of stop words. Stop words consume a lot of space in the dataset. Stop words like 'a', 'an', 'the' are removed because they don't hold any importance in training the classifier. Therefore, the stop words are removed so as to increase the quality of dataset.

Textual data has to be converted into a proper format so that classifier can understand the data and train itself. So, Tf-idf vector was constructed from reviews. Tf-idf vector is a statistic which reflects the importance of each and every term in a document. The importance of each term is measured by its frequency in the document. Tf (Term frequency) in simple words is count of every unique word in the document. Idf (Inverse Document frequency) is a measure which takes into account the frequency of very frequent words which are not important for classification. High frequency of such useless words will outweigh the use of much more important and less frequent terms. Hence Idf is a way of decreasing the weight of such useless words in documents and increasing the weight of other useful but less frequent term. TF-IDF vector is simple the product of these two measures.

$$\text{tf-idf}_{t,d,D} = \text{tf}_{t,d} \times \text{idf}_{t,D} \quad (1)$$

$$\text{tf}_{t,d} = 0.5 + 0.5 \times (f_{t,d} / \max \{f_{t'd} : t' \in d\}) \quad (2)$$

$$\text{idf}_{t,D} = \log (N / 1+|d \in D: t \in d|) \quad (3)$$

where N: total number of documents in corpus N=|D|

The tf-idf vector and the category of the review was passed in different classifiers during training phase. During testing phase only tf-idf vector is passed to classifier and it's predicted category is noted. Its predicted category is compared with actual category and the accuracy of different classifier is calculated. Classifiers used are K Nearest Neighbor (KNN), Logistic Regression, Support Vector Machine (SVM), Decision Tree and Random Forest classifier. All the above-mentioned classifiers are supervised leaning classifier. All the classifier used in the model are compared in the table below on the basis of their accuracy.

Table 1: Performance Comparison of classifier

S.No.	Classifier	Accuracy (%)
1.	KNN	81.15 %
2.	Logistic Regression	87.02 %
3.	SVM	87.87 %
4.	Decision Tree	86.49 %
5.	Random Forest	88.12 %

From the table, it can be concluded that SVM and random forest works better than other classifier for this particular model with an accuracy of 87.87% and 88.12% respectively.

VI. CONCLUSION, LIMITATIONS AND SCOPE FOR FUTURE WORK

Through this project, it can be concluded that fake and genuine review can be identified through supervised machine learning algorithm. From the series of experiments, we conducted on yelp dataset through different classifier, it can be said that Support Vector Machine classifier (SVM) and Random Forest classifier works better than other classifiers and were able to identify fake reviews with an accuracy of 87.87 % and 88.12 % respectively. Pre-processing of data is necessary to improve the quality of dataset which in turn improves the accuracy of classifiers. This model can be used by any website which takes online reviews such as e-commerce websites to

detect and delete biased and fake reviews. Hence a transparent and actual reputation of an organization or a product/service supplier can be reported to the customer who wants to know the reputation of an organization or product/service provider.

Limitation of the model is that a huge labeled dataset for any product, service or organization is difficult to acquire. Manually reading each and every review and then assigning it manually a categorical is a very laborious and time-consuming work. Hence manually labelling of reviews/comments is not feasible.

Another limitation of any classifier and hence this model is that if a very huge data set (in GB) are passed in any classifier, then it won't be able to predict the category of new test data point very accurately and accuracy of certain threshold (say 60%) is not acceptable.

This model can further be improved in future if more information from review writer is obtained. One of such information is the IP address of reviewer. From the IP address, we can triangulate a person's identity. If too many reviews are coming from a particular IP address, then reviews/comments written from that IP address can be removed automatically because that reviewer is certainly writing fake reviews. Another information which can be fetched from the reviewer is the number of reviews one person has written and how frequently a person writes a review. If frequency and number of reviews are too high, then the reviews written by that particular reviewer can be analysed and deleted so that it does not affect the overall accuracy of a certain organization/product/service.

ACKNOWLEDGMENT

We are thankful to yelp.com which helped us to acquire dataset of reviews along with their category (fake or genuine) of multiple restaurants in New York.

REFERENCES

1. Mukherjee A, Venkataraman V, Liu B, Glance N. Fake review detection: Classification and analysis of real and pseudo reviews. UIC-CS-03-2013. Technical Report. 2013.
2. Bajaj, Simran, Niharika Garg, and Sandeep Kumar Singh. "A novel user-based spam review detection." Procedia computer science 122 (2017): 1009-1015.
3. Li, Huayi, Zhiyuan Chen, Bing Liu, Xiaokai Wei, and Jidong Shao. "Spotting fake reviews via collective positive-unlabeled learning." In 2014 IEEE International Conference on Data Mining, pp. 899-904. IEEE, 2014.
4. Dellarocas, C.: Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In Proceedings of the 2nd ACM conference on Electronic commerce (pp. 150-157). ACM. (2000, October)
5. Li, H., Chen, Z., Mukherjee, A., Liu, B., & Shao, J.: Analyzing and Detecting Opinion Spam on a Large-scale Dataset via Temporal and Spatial Patterns. In ICWSM (pp. 634-637). (2015, May).
6. Li, H., Fei, G., Wang, S., Liu, B., Shao, W., Mukherjee, A., & Shao, J.: Bimodal distribution and co-bursting in review spam detection. In Proceedings of the 26th International Conference on World Wide Web (pp. 1063-1072). International World Wide Web Conferences Steering Committee (2017, April).
7. Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., & Ghosh, R.:Spotting opinion spammers using behavioral footprints. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 632-640). ACM. (2013, August)
8. Lin, Y., Zhu, T., Wu, H., Zhang, J., Wang, X., & Zhou, A.: Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on (pp. 261-264). IEEE. (2014, August)
9. Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. Management Science, 62(12), 3412-3427
10. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R.: Exploiting Burstiness in Reviews for Review Spammer Detection. Icwsm, 13, 175-184. (2013)
11. Mukherjee, Arjun, Bing Liu, and Natalie Glance. "Spotting fake reviewer groups in consumer reviews." In Proceedings of the 21st international conference on World Wide Web, pp. 191-200. ACM, 2012.

AUTHORS PROFILE

Divyanshu Jalther completed his B.Tech in Computer Science and Engineering from Vellore Institute of Technology, Vellore in 2018. His area of interest is Machine leaning, Test based classification and Reputation system.

Priya Govindaraj is an Associate professor in School of computer science and Engineering, Vellore Institute of Technology, Vellore. She completed her B. E in computer science and Engineering under Madras university, M.Tech Computer science and Engineering and Ph.D in VIT. She published more than 30+ research papers in reputed journals. Her area of interest is Trust management, cloud computing, IoT and Deep learning.