Machine learning Techniques for Hotel Online Reputation

Pankaj Chaudhary, Anurag Aeron, Sandeep Vijay

Abstract: ***Now days when someone decide to book a hotel, previous online reviews of the hotels play a major role in determining the best hotel within the budget of the customer. Previous Online reviews are the most important motivation for the information that are used to analyse public opinion. Because of the high impact of the reviews on business, hotel owners are always highly concerned and focused about the customer feedback and past online reviews. But all reviews are not true and trustworthy, sometime few people may intentionally generate the fake reviews to make some hotel famous of to defame. Therefore it is essential to develop and propose the techniques for analysis of reviews. With the help of various machine learning techniques viz. Supervised machine learning technique, Text mining, Unsupervised machine learning technique, Semi-supervised learning, Reinforcement learning etc we may detect the fake reviews. This paper gives some notions of using machine learning techniques in analysis of past online reviews of hotels, Based on the observation it also suggest the optimal machine learning technique for a particular situation.***

Keywords: ***Unsupervised machine learning technique, Text mining, Supervised machine learning technique, Semi-supervised machine learning technique, Reinforcement machine learning technique, Hype, Quantification, collision, manipulation, machine learning, mining, deep learning etc***

I. ANALYSIS OF ONLINE REVIEWS.

To make an analysis of online reviews, we need to gather the available reviews from various hotel websites such as Treebo, Goibibo, Yelp, Makemytrip, Booking.com, Yelp etc. After collecting various types of reviews we need to follow the following steps:

A. TECHNICAL UNDERSTANDING OF DATA

We must clearly understand what all kind of hotel reviews are available within that budget. Analysis algorithms that we want to implement must also be chosen keeping the variety of the data in our mind. It must also be understood that so may be implemented on smaller sample sets while others require larger samples. Some algorithms work with certain types of data and some can be implemented on any type of data.

We must able able to find out the Percentiles, average, medians, correlations, regressions , central tendency, strong relationships etc.

Revised Manuscript Received on July 11, 2019.

Pankaj Chaudhary, Research Scholar, FST, ICFAI University, Dehradun, India.

Dr. Anurag Aeron, Associate Professor, FST, ICFAI University, Dehradun, India.

Dr. Sandeep Vijay, Director, Shivalik College of ngineering, Dehradun, India

B. STATISTICALLY VISUALIZATION OF THE DATA

There should be the provisions for Box plots, outliers, density plots, histograms, scatter plots can describe bivariate relationships.

C. CLEANING OF THE DATA

1. We must able to deal with the missing value.
2. What to do with outliers must be clearly defined, especially in multidimensional data.
3. Do we need the data needs to aggregated

D. AUGMENT THE DATA

1. Mechanisms for conversion of raw data into the data form that is ready to be used for the modeling must be clearly specified.

   It can serve several purposes such as:

- Models can be easily interpreted.
- More complex relationships can be modeled.
- It can Reduce the data redundancy and also dimensionality rescale the variables
- Different models may have some built in feature engineering implementations.

E. CATEGORIZE THE PROBLEM

Data need to be categorised based on the input or output.

1. **Categorize by input:**

- If labelled data is available, better to implement supervised learning problem.
- If data is unlabelled we want to find structures, it's then better to implement unsupervised learning problem.
- If purpose is optimization of the objective function after interacting with the related environment, then better to implement reinforcement learning problem.

2. **Categorize by output**.

- If the output is a number of the model, it can be specified as regression problem.
- If the output is a class of the model, it can be specified as classification problem.

- If the output is a set of input groups, of the model then it can be specified as a clustering problem.
- If purpose is to detect an anomaly? Then it's anomaly detection.

Understand the constraints

- What is the data storage capacity?
- Does the prediction have to meet the deadlines?
- Does the learning have to meet the deadlines?

Find the available algorithms

Now it can be identified that the applicable algorithms and practicals to implement the tools will be at our disposal. Some factors that may affect the choice may be:

- Does model can meet the business objectives?
- What amount of pre processing of the model is needed?
- What mathematical concepts are available to measure the accuracy of the model?
- Upto what extent model can be expended?
- How model can be scalable?
- On how many features model learns and predicts?
- Up to what extent it relies on more complex feature engineering/
- Up to what extent it has complex computational overhead.
- Which regression model is utilized?
- How the decision tree concept is utilized?

Developing the same complex algorithm will surely increases the chance of over fitting.

1I. MACHINE LEARNING TECHNIQUES

By analysing the reviews and extracting meaningful features of text by implementing (NLP) Natural Language Processing it is easy to implement spam detection using various machine learning techniques.

There are mainly 4 types of machine learning techniques:

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

A. SUPERVISED LEARNING

Supervised learning refers to the task of inferring a function from labelled set of data.
Equations are used to fitting Labelled training data; and the purpose is to find the most optimal model parameters to predict unknown labels on other objects.
In case label is a real number, then this task is termed as *regression*. If the label is in the from of the limited number of un-ordered values, then this is called as *classification* [2].

B UNSUPERVISED LEARNING

Unsupervised learning is implemented when we have less information about objects; in this case the training set of data is unlabeled.
In this case our goal is to observe some similarities among the groups of the objects and also to include these within appropriate clusters. Few objects may differ drastically from all other clusters; these objects are declared as the anomalies [3].

C. SEMI-SUPERVISED LEARNING

Semi-supervised learning algorithm includes both labelled and unlabeled data. This method allows to improve accuracy significantly, because one can use unlabeled data in the train set with a small amount of labelled data also.

D. REINFORCEMENT LEARNING

Reinforcement learning is an important area of machine learning which concerned with how software agents ought to implement the actions in some given specific environment to maximize and optimize some notion of cumulative reward.

1II. EXPERIMENT AND ANALYSIS

A questionnaire was distributed among 276 persons who have taken decision in past based on the reviews to know their opinion about the factors that they keep in mind while deciding any review and good bad or neutral.

Which of the following are the major criteria to decide the good or bad review?

SN	Parameters	First Choice by
A	By seeing the infrastructure Customer keep in mind what should be the rating. If review is not matching the expectations it means something is wrong.	72/276**[26%]**
B	Customers pre-assume the categories of the reviews and try to fit in the next review in a particular category.	81/276**[29%]**
C	Some Clusters are identified and predict the review good or bad with rating based on the parameters matching with the clusters.	71/276**[26%]**
D	Individual reviews are analyzed and by making the groups we try to conclude and verify the type.	52/276**[19%]**

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

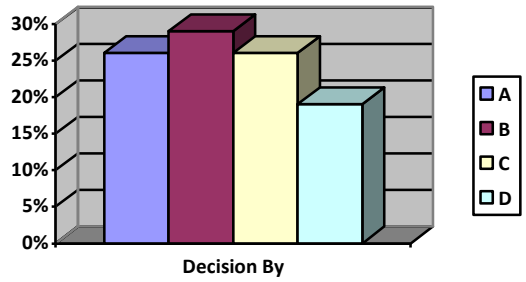

FIG 1: REVIEW CATEGORISATION CHOICE

IV DEVELOPMENT OF OPTIMAL MACHINE LEARNING ALGORITHM

In case of hotel reviews we may analyse the reviews given by the groups and can predict the suggested reviews by another groups. By analysing the difference between predicted and suggested reviews, genuineness may be analysed. This is the case of supervised learning.

In case of reviews some pre-defined clusters may be identified and hotel reviews may be fit in some group. The reviews that did not fit any group will be termed as the spam. This is the case of unsupervised learning.

In hotel reviews some clusters and some predictions both may be defined as per the user requirements, and analysis of reviews may be made. This is the case of semi-supervised learning. In case we are considering and analyzing the environment and outcomes of hotel reviews may generate specific re-actions on specific actions then re-enforcement learning must be used for hotel reviews.

CONCLUSION AND FUTURE WORK

It is concluded that various machine learning techniques can be used for analysis of online reviews of the hotels.

Step No 1: Collect maximum available review and based on the quantity mathematically decide whether quantity of reviews is big.

Step No 2: Analyse whether all expectations of the customer can be identified from the reviews.

Step No 3: If step no 1 and 2 are positive then supervised learning can be used. Otherwise it is better to analyse the hotel reviews based on the group theory using unsupervised machine learning.

Step No 4: Analyse whether we have labelled and unlabelled defined clusters for genuinity analysis, we must use semi supervised learning technique.

Step No 5: In case we are also considering and analyzing the environment also then Re-enforcement learning must be used.

REFERENCES

1. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection,"IEEE Com-munications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, Oct.2015.
2. Huang, C. L. Gutterman, P. Samadi, P. B. Cho, W. Samoud, C. Ware,M. Lourdiane, G. Zussman, and K. Bergman, "Dynamic mitigationof EDFA power excursions with machine learning,"Optics Express,vol. 25, no. 3, pp. 2245–2258, Feb. 2017.
3. A. Marinescu, I. Macaluso, and L. A. DaSilva, "A Multi-Agent NeuralNetwork for Dynamic Frequency Reuse in LTE Networks," inIEEEInternational Conference on Communications (ICC), 2018.
4. Rout J.K., Dalmia A. and Choo K. K. R., (2017), "Revisiting Semi-supervised Learning for Online Deceptive Review Detection", IEEE Access, 2169-3536, pp. 11-19.
5. Deng X. and Chen R., (2014), "Sentiment Analysis Based Online Restaurants Fake Reviews Hype Detection", proceedings of APWeb Workshops, Springer International Publishing Switzerland, pp. 1–10.
6. Ruchansky N., Seo S. And Liu Y., (2017), "A Hybrid Deep Model for Fake News Detection", Computer Society of India ACM. 978-1-4503-4918-5/17/11, pp. 17- 27.
7. Mukherjee A., Venkataraman V., Liu B., Glance N., (2013), "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews", Technical Report, Department of Computer Science (UIC-CS-2013-03).University of Illinois at Chicago, pp. 268- 279.
8. Kokate S., Tidke B., (2015),"Fake Review and Brand Spam Detection using J48 Classifier", International Journal of Computer Science and Information Technologies ISSN: 0975-9646, Vol. 6 (4), pp. 3523-3526.
9. Kolhe N.M., Joshi M.M., Jadhav A.B., Abhang P.D., (2014)," Fake Reviewer Groups' Detection System", IOSR Journal of Computer Engineering (IOSR-JCE). e-ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 16, Issue 1, Ver. V (Jan. 2014), pp. 06-09.
10. Chaitanya Kale, Dadasaheb Jadhav., Tushar Pawar., (2016), "Fake Spam review detection using natural language processing techniques", International journal of innovations engineering research and technology ISSN: 2394-3696, Vol. 3, Issue 1, Jan.-2016, pp. 31-37.
11. Adike R.G., Reddy V., (2016), "Detection of Fake Review and Brand Spam Using Data Mining Technique", International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 02, Issue 07; July - 2016 [ISSN: 2455-1457], pp. 251-256.
12. Bonde Y.P., Kharabi K.L., Sabale A.N., (2017), "Detection and Elimination of Fake Review from Real- Time Data using Cloud Computing", International Journal of Advance Engineering and Research Development ISSN: 2348-6406 Volume 4, Issue 5, May-2017, pp. 187-194.
13. Crawford M., Khoshgoftaar T.M., Prusa J.D., Richter A.N. and Najadain H.A., (2015), "Survey of review spam detection using machine learning techniques", Springer Journal of Big Data Machine Learning Methods Crawford et al, pp. 17- 39.
14. Elmurngi E. and Gherbi A.,(2017),"An Empirical Study on Detecting Fake Reviews", proceeding of IEEE The Seventh International Conference on Innovative Computing Technology, pp. 107-114.
15. Fontanarava J., Pasi G. and Viviani M., (2017), "Feature Analysis for Fake Review Detection through Supervised Classification", proceedings of IEEE International Conference on Data Science and Advanced Analytics, pp. 658-666.
16. Wahyuni E. D. And Djunaidy A., (2017), "Fake review detection from a product review using modified method of iterative computation framework", proceedings of MATEC Web of Conferences, pp. 121-127.
17. Lin Y., Zhu T., WuI H., Zhangl J., Wang X., Zhou A., (2014), "Towards Online Anti-Opinion Spam: Spotting Fake Reviews from the Review Sequence", proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 261-264.
18. Li Y., Feng X., Zhang S., Li Y., (2016), "Detecting Fake Reviews Utilizing Semantic and Emotion Model", proceedings of 3rd IEEE-International Conference on Information Science and Control Engineering, pp. 317-320.

19. Yin R., Wang H., Liu L. , (2015), "Research of Integrated Algorithm", proceedings of 4th IEEE International Conference on Computer Science and Network Technology, pp. 584- 589.
20. Rajamohana S.P., Umamaheswari K., Dharani M., Vedackshya R., (2017), "A Survey on online review spam detection techniques", proceedings of IEEE International Conference on Innovations in Green Energy and Healthcare Technologies, pp. 8- 13.
21. Liu P., Xu Z., Ai J. , Wang F., (2017), **"**Identifying Indicators of Fake Reviews Based on Spammer's Behavior Features**",** proceedings of IEEE International Conference on Software Quality, Reliability and Security, pp. 396- 403.
22. Chauhan S.K., Goel A., Goel P., Chauhan A. and Gurve M.K., (2017), "Research on Product Review Analysis and Spam Review Detection", proceedings of IEEE 4th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 399- 393.
23. Christopher S.L. and Rahulnath H. A., (2016), "Review authenticity
24. verification using supervised learning and reviewer personality traits", proceedings of IEEE International Conference on Emerging Technological Trends, pp. 16- 23.
25. Fei G., Mukherjee A., Liu1 B., Hsu M., Castellanos M., Ghosh R., (2013), "Exploiting Burstiness in Reviews for Review Spammer Detection" Proceedings of the Seventh International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Weblogs and Social Media, pp. 175- 185.
26. Shojaee S., Azman A., Murad M., Sharef N. and Sulaiman N., (2017), "A Framework for Fake Review Annotation", proceedings of 17th IEEE Computer Society UKSIM-AMSS International Conference on Modelling and Simulation, pp. 153- 159.
27. Visani C. and Jadeja N., (2017), "A Study on Different Machine Learning Techniques for Spam Review Detection", IEEE Transaction 978-1-5386, pp. 1887-1892.
28. Xue H., Li F., Seo H. and Pluretti R., (2015), "Trust-Aware Review Spam Detection", IEEE Computer Society Trustcom/BigDataSE/ISPA, pp. 726-733.
29. Jiang M. and Cui P., (2016),"Suspicious Behaviour Detection: Current Trends and Future Directions", IEEE Intelligent systems/1541-1672/16, Computer Society, pp. 31-39.
30. Ahsan M.N.I., Nahian T., Kafi A.A., Hossain I., Shah F.M., (2017), "An Ensemble approach to detect Review Spam using hybrid Machine Learning Technique", IEEE 19th International Conference on Computer and Information Technology, pp. 381- 388.

AUTHORS PROFILE

Pankaj Chaudhary has completed his B.Tech and M.Tech, he is Ph.D(CSE) reasearch scholer at ICFAI University, Dehradun . He has published 16 National and International research papers in jounrals of repute. He has also attended several conferences.Currently he is doing his reasearch in analysing the genuinity of the online reviews of hotels.

Dr. Anurag Aeron is Associate professor(CSE) at ICFAI University, Dehradun, He has completed his Ph.D from IIT Roorkee**.** His research areas are Remote Sensing and GIS, Open Source Systems, Disaster Management, AI, Android Operating System, Machine Learning, IOT, NLP.

Dr. Sandeep Vijay, Working as Director at Shivalik College of Engineering, Dehradun, He has completed his Ph.D from IIT Roorkee. He has Proficiency in spearheading overall strategic research and developments projects, right from planning, cost controls,, resource mobilization, structured communications to final reviews, within cost & time parameters.