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

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.

**II. 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.

**III. 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?**

<table>
<thead>
<tr>
<th>SN</th>
<th>Parameters</th>
<th>First Choice by</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>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.</td>
<td>73/276[26%]</td>
</tr>
<tr>
<td>B</td>
<td>Customers pre-assume the categories of the reviews and try to fit in the next review in a particular category.</td>
<td>81/276[29%]</td>
</tr>
<tr>
<td>C</td>
<td>Some clusters are identified and predict the review good or bad with rating based on the parameters matching with the clusters.</td>
<td>71/276[26%]</td>
</tr>
<tr>
<td>D</td>
<td>Individual reviews are analyzed and by making the groups we try to conclude and verify the type.</td>
<td>52/276[19%]</td>
</tr>
</tbody>
</table>
Detecting Fake Reviews

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 genuineness 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

Machine learning Techniques for Hotel Online Reputation


AUTHORS PROFILE

Pankaj Chaudhary has completed his B.Tech and M.Tech, he is Ph.D(CSE) research scholar at ICFAI University, Dehradun. He has published 16 National and International research papers in journals of repute. He has also attended several conferences. Currently he is doing his research in analyzing 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.