A Combinatorial Approach for Predicting Online Review Helpfulness of Indian Online Travel Agencies

Hari Bhaskar Sankaranarayanan[1] and Viral Rathod[1]

[1]Amadeus Labs
Bangalore, India
{hari.sankaranarayanan, viral.rathod}@amadeus.com

Abstract. Online user reviews are quite popular in social media, e-commerce and review websites. It is commonly referred as word of mouth which provides positive and negative messages from users about products and services. It helps users to get insights through review ratings and subjective feedback. As the volumes of reviews are high, it makes it harder for users to identify the helpfulness upfront. In general helpfulness rating is provided by the users who read the review, but many reviews still stay unrated. In this paper, we propose an approach of predicting helpfulness of such reviews from mouthshut.com using a combinatorial approach of empirical analysis and naïve Bayes machine learning method. The data set is chosen for Indian Online Travel Agencies (OTA) namely Makemytrip, Cleartrip, Yatra, Goibibo, and Expedia India. A detailed experiment is conducted and results are discussed by analyzing review metadata characteristics.

Keywords: travel, reviews, Bayesian, prediction

1 Introduction

Online reviews are extensively used for inspiration, validation, and decision making for buying products and services online. Social media, e-commerce and review websites provide valuable insights like pros, cons, ratings, subjective feedback from users that can help other users make the right choice among products and service providers [1]. On the flip side, online reviews are abundant in nature and it is hard to sift through the large volumes to identify helpfulness. It is hard to figure out which would be a helpful review since the perception of helpfulness varies based on user needs, reviewer quality, review content, and type of product or service. To solve this problem, the review sites provide an option for users to tag the reviews as helpful or not. Users can filter the reviews based on the usefulness rating. India OTAs form an interesting case study for the analysis since travel market in India is booming due to raising in disposable income and travel propensity for middle-class [2]. Many travelers prefer online trip planning and booking tools for their trip [3]. Online reviews for OTAs are quite relevant forum to make purchase decisions and evaluate the travel services offered by them. In this paper, we propose a novel approach of analyzing the dataset for online reviews of India OTAs using a combination of empirical analysis and supervised machine learning techniques to predict the helpfulness rating without dwelling into the review content itself. Mouthshut.com is a very popular online review site for Indian consumers and it is used by researchers for understanding Indian consumer behavior.

The paper is organized into following sections. Section 2 highlights the related work done in this area, section 3 discusses the data set information, section 4 discusses the empirical analysis, section 5 mentions the attribute classification, section 6 discusses the assumptions, section 7 elaborate the analysis method, section 8 summarizes the experiment results, section 9 discusses the limitations, section 10 provides directions for future work and section 11 concludes the paper.

2 Related Work

In our literature analysis, there are various scholarly articles are available on sentiment analysis, opinion mining and natural language processing methods for online reviews. For travel and hospitality and tourism online reviews, there is a detailed survey paper studying 50 papers on recent trends and future directions. This paper is quite extensive and provides valuable insights like online buying and reviews, satisfaction & management, motivation, and the role of reviews. There is also a mention about the complexity of this problem space and lack of widespread adoption of scientific research in practice due to time and cost implications. Online reviews are also analyzed from the economy and customer service perspective that can help to improve the overall quality of product and services [1]. Online review helpfulness is analyzed from psychological and behavior science perspective like associating user emotions on anxiety or anger [4]. Online reviews helpfulness is assessed from the content point of view using sentiment analysis techniques like topic mining, feature identification, and content structure [5]. Research work is carried out for effectively identifying fake and manipulation of online reviews [6]. On the methodology, Support Vector Model (SVM) is mentioned on research articles for predicting review helpfulness [7] [8]. A research paper identifies product type influences the review helpfulness which makes a case for us to study travel products and services from OTAs [9]. A paper on hypothesis testing of reviewer characteristics like self-described expertise and personal information had a positive effect on helpfulness and reviewer demographics doesn't have a much significant effect [10]. This paper provided us a strong motivation to pursue a new approach of prediction based on additional characteristics like reviewer contribution and past usefulness rating. Also from Indian context we intend to test the influence of reviewer demographics once again with a different dataset.

3 Dataset Information

The section covers information about collection and cleansing of online review data sets for analysis collected from mouthshut.com website. OTAs are selected in the search box of the mouthshut web page.

3.1 Data Extraction

Post the selection of OTA, the reviews are listed. The reviews are extracted using screen scraping technique using an online based tool named "import.io". The output is generated as a comma separated value file from the tool. Table 1 highlights the sample size of the online review data extracted for various OTAs

OTA	Number of Review Samples
Makemytrip	263
Cleartrip	247

Goibibo	399
Yatra	400
Expedia.co.in	146

Table 1 Online Review sample size

3.2 Data Cleansing

The raw data which is screen-scraped available in comma separated value format is cleansed further in excel spreadsheet on the following areas:

1. Location information is cleansed with common city codes in upper case (Ex: BANGALORE, KOLKATA, and CHENNAI)
2. For the ease of classification, Review Contribution and Read count are classified as Very Low, Low, Medium, High, and Very High based on the range of numeric values. Table 2 & 3 below highlights the classification.
3. Normalized review contribution numbers to avoid any duplicates.

Classification	**Reviewer Contribution Range**
VL (Very Low)	1-10
L (Low)	11-30
M (Medium)	31-100
H (High)	101-500
VH (Very High)	500+

Table 2 Review Contribution Classification

Classification	**Review Read Count Range**
VL (Very Low)	1-500
L (Low)	501-3000
M (Medium)	3001-8000
H (High)	8001-15000
VH (Very High)	15000+

Table 3 Review Read Count Classification

4 Empirical Analysis Phase

The goal of the empirical analysis phase is to achieve attribute classification that would serve as an input to machine learning phase. The data set is analyzed in a spreadsheet pivot table and the key findings are listed below:

1. *Reviewer Location* has significant influence in the data set in OTA reviews. For instance, significant reviews are written by reviewers from metro or big cities like Bangalore, New Delhi, Mumbai, Hyderabad and Kolkata contributing 53.78% of total reviews. Table 4 lists the top 5 reviewer locations where reviews are posted.

Location	**Number of Reviews**	**% Contribution**
New Delhi	4891	16.15%
Bangalore	4596	15.18%
Mumbai	4176	13.79%

Hyderabad	1313	4.34%
Kolkata	1310	4.33%

Table 4 Top 5 review contribution location wise

2. The dataset consists of *usefulness rating.* In this experiment, we consider only ‘Useful’ and ‘Very useful’ as a measure of helpfulness of reviews. ‘Somewhat useful’ rating is excluded since it doesn’t provide a strong endorsement on usefulness. Table 5 lists the % of Helpful reviews of the Top 5 reviewer locations.

Location	% Helpful Reviews
New Delhi	63.4%
Bangalore	62.4%
Mumbai	60.93%
Hyderabad	75%
Kolkata	58.82%

Table 5 Review helpfulness of Top 5 reviewer’s location

3. *Read count* of every review which provides access characteristics. Table 6 lists the read count of Top 5 reviewer’s location.

Location	Review Read Count
New Delhi	474057
Bangalore	1023446
Mumbai	478802
Hyderabad	187192
Kolkata	122804

Table 6 Review read count of Top 5 reviewer’s location

From the observation of the dataset, good *customer service recovery* processes and practices have high review score and helpfulness rating. In this dataset, Makemytrip has a team that responds to complaints and queries in mouthshut website. The data is validated with mouthshut.com website as they archive issue resolutions database, especially for Makemytrip. There is no specific evidence for other OTAs in our study of the serviceability process in mouthshut. Table 7 lists the OTA’s overall rating.

OTA	Overall Rating
Makemytrip	4.14
Cleartrip	1.33
Goibibo	1.65
Yatra	1.44
Expedia	1.18

Table 7 OTA overall rating

5 Attribute Classification

Based on the above empirical analysis, we have derived the following attribute categories that potentially influence the helpfulness of a particular review. Fig 1 depicts the attribute class which is an input to the next phase.

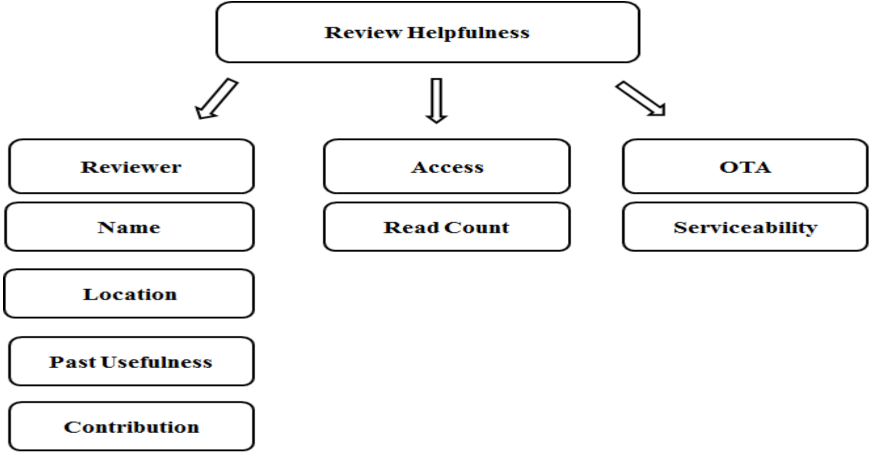


Figure 1. Attribute Classification

6 Assumptions

The following assumptions are made for the dataset:

1. We didn’t do any text analysis for opinion mining like sentiment analysis considering the limitations of accuracy, especially on natural language processing.
2. Review content itself is excluded since the content is subjective for interpretation and doesn’t provide any direct measure of helpfulness.
3. Review rating on Services & Support, Information Depth, Content Timeliness, Website Load time and Design/Usability are not taken into account. OTA reviews are classified under website category but there is a heavy inclination towards posting reviews about Services & Support issues than website characteristics itself.
4. Reviewer contribution is across the mouthshut.com site and not specific to OTA review only.
5. For certain reviews, there is no location information available (around 7.63% review contributions) hence we need to discount the same.

7 Analysis Method

In the current problem context, naïve Bayes classifier is used for predicting the helpfulness of reviews. The reasons are as follows:

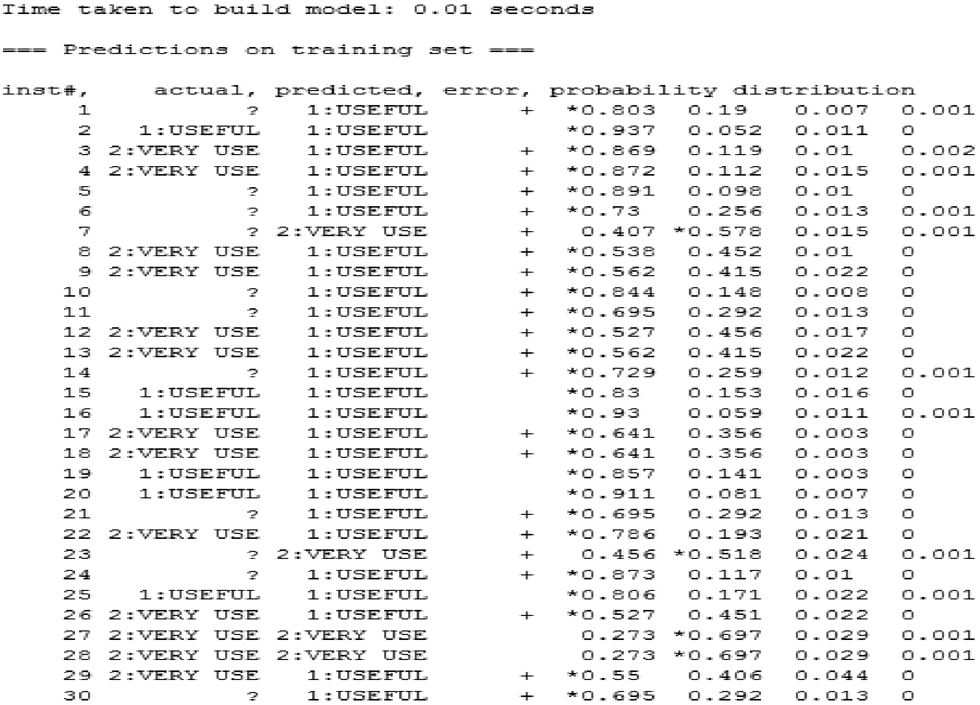
1. Naïve Bayes is suitable for probabilistic modeling of attributes which can independently influence the outcome.
2. Naïve Bayes has several advantages over other machine learning methods as it classifies based on prior probability. It is simple to use and easy to understand. In the current OTA review data set, prior data is considered for review helpfulness. This naturally converges into a probability based approach with heavy reliance on the prior probability distribution of attributes.

3. The attributes can be added opportunistically where dependency is yet to be ascertained.

Weka tool for machine learning is used for running the naïve Bayes algorithm. It consists of inbuilt packages that support the classification. The data is classified into training and testing sets. Training dataset consists of review rating already marked by users. The testing dataset consists of ‘Not Rated’ ones.

8 Results and Discussion

Fig 2 shows the output of Bayesian classification run in Weka tool. The data highlighted as “?” are the “Not rated” from the original data set. The run predicts the usefulness rating based on the training dataset. First 30 records are shown for illustration.



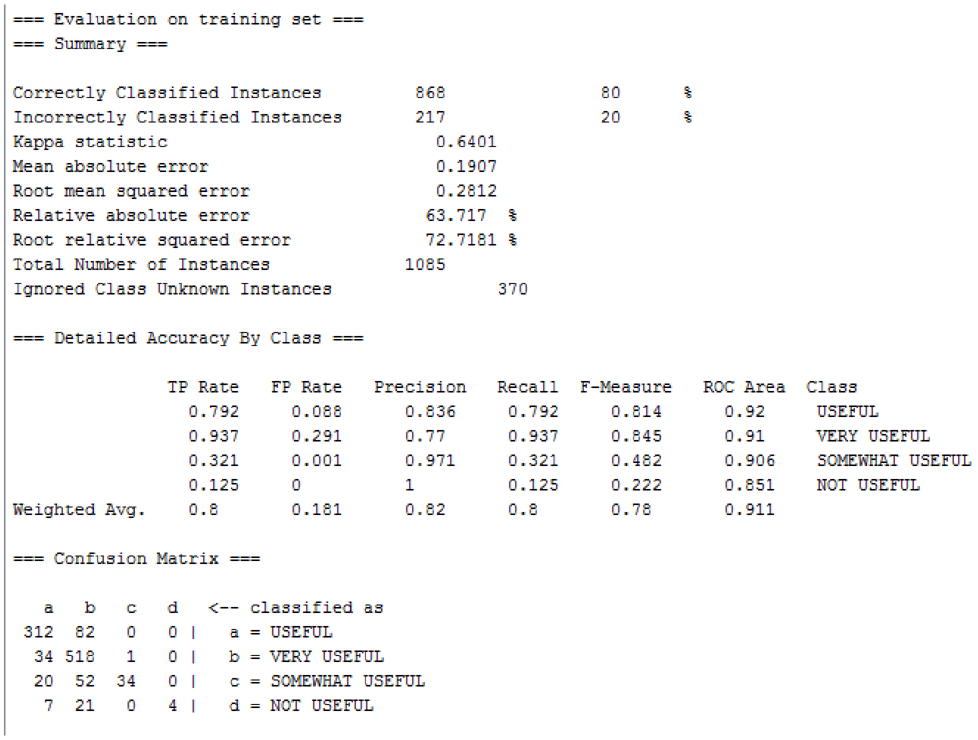
```
Time taken to build model: 0.01 seconds

=== Predictions on training set ===

inst#,     actual, predicted, error, probability distribution
     1          ?   1:USEFUL      +  *0.803  0.19    0.007  0.001
     2   1:USEFUL   1:USEFUL         *0.937  0.052   0.011  0
     3 2:VERY USE   1:USEFUL      +  *0.869  0.119   0.01   0.002
     4 2:VERY USE   1:USEFUL      +  *0.872  0.112   0.015  0.001
     5          ?   1:USEFUL      +  *0.891  0.098   0.01   0
     6          ?   1:USEFUL      +  *0.73   0.256   0.013  0.001
     7          ? 2:VERY USE      +   0.407 *0.578   0.015  0.001
     8 2:VERY USE   1:USEFUL      +  *0.538  0.452   0.01   0
     9 2:VERY USE   1:USEFUL      +  *0.562  0.415   0.022  0
    10          ?   1:USEFUL      +  *0.844  0.148   0.008  0
    11          ?   1:USEFUL      +  *0.695  0.292   0.013  0
    12 2:VERY USE   1:USEFUL      +  *0.527  0.456   0.017  0
    13 2:VERY USE   1:USEFUL      +  *0.562  0.415   0.022  0
    14          ?   1:USEFUL      +  *0.729  0.259   0.012  0.001
    15   1:USEFUL   1:USEFUL         *0.83   0.153   0.016  0
    16   1:USEFUL   1:USEFUL         *0.93   0.059   0.011  0.001
    17 2:VERY USE   1:USEFUL      +  *0.641  0.356   0.003  0
    18 2:VERY USE   1:USEFUL      +  *0.641  0.356   0.003  0
    19   1:USEFUL   1:USEFUL         *0.857  0.141   0.003  0
    20   1:USEFUL   1:USEFUL         *0.911  0.081   0.007  0
    21          ?   1:USEFUL      +  *0.695  0.292   0.013  0
    22 2:VERY USE   1:USEFUL      +  *0.786  0.193   0.021  0
    23          ? 2:VERY USE      +   0.456 *0.518   0.024  0.001
    24          ?   1:USEFUL      +  *0.873  0.117   0.01   0
    25   1:USEFUL   1:USEFUL         *0.806  0.171   0.022  0.001
    26 2:VERY USE   1:USEFUL      +  *0.527  0.451   0.022  0
    27 2:VERY USE 2:VERY USE          0.273 *0.697   0.029  0.001
    28 2:VERY USE 2:VERY USE          0.273 *0.697   0.029  0.001
    29 2:VERY USE   1:USEFUL      +  *0.55   0.406   0.044  0
    30          ?   1:USEFUL      +  *0.695  0.292   0.013  0
```

Figure 2. Prediction Output

Fig 3 shows the summary of the run with correct classification of 80% and incorrect classification of 20%. 370 “Not Rated” test data instances are classified based on learning from training data set. Thus, naïve Bayes technique classifies the data with a reasonable accuracy and provides an encouraging outcome of the experiment.



```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         868               80      %
Incorrectly Classified Instances       217               20      %
Kappa statistic                          0.6401
Mean absolute error                      0.1907
Root mean squared error                  0.2812
Relative absolute error                 63.717  %
Root relative squared error             72.7181 %
Total Number of Instances             1085
Ignored Class Unknown Instances                370

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.792     0.088     0.836     0.792     0.814      0.92     USEFUL
                 0.937     0.291     0.77      0.937     0.845      0.91     VERY USEFUL
                 0.321     0.001     0.971     0.321     0.482      0.906    SOMEWHAT USEFUL
                 0.125     0         1         0.125     0.222      0.851    NOT USEFUL
Weighted Avg.    0.8       0.181     0.82      0.8       0.78       0.911

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 312  82   0   0 |   a = USEFUL
  34 518   1   0 |   b = VERY USEFUL
  20  52  34   0 |   c = SOMEWHAT USEFUL
   7  21   0   4 |   d = NOT USEFUL
```

Figure 3. Results Summary

9 Limitations

Our current approach to this problem has certain limitations. The limitations are as follows:

1. Mouthshut provides review read statistics which may not be prevalent across other review sites. The approach cannot fit directly if the read statistics are not available.
2. Serviceability score is based on the issue resolution process and overall review rating from mouthshut which does discount other means of issue resolution. For example, OTAs might have different mitigation like dedicated customer service and recovery.
3. The analysis considers location as an attribute since reviews are extensively biased towards metro or big Indian cities considering review contribution. This model may not work for all geographies if there are no distinction on number of reviews across cities.

10 Directions for Future Work

The current work can be extended to e-commerce product and services reviews in India like flipkart, amazon, and snapdeal. Review content can be added as a parameter by looking at simplification options like overall sentiment of review and nature of product, services mentioned (Ex: Offers, discounts). The OTA dataset can be extended from other data sources where users provide reviews like social media tools like Facebook, Twitter. In terms of machine learning techniques, further correlation of reviews like trusted reviewer circles (mouthshut term for reviewer credibility) and its influence on the outcome can be evaluated.

11 Conclusion

Online reviews is a great way to know more about the feedback from the user community on making an online purchase decision. The nature and complexity of reviews make it an interesting and evolving space in terms of solving the problem of scale. As each review is unique it is a challenge to provide a generalized approach for rating helpfulness. Our analysis shows that naïve bayes can come handy to predict helpfulness considering its simplicity and prior probability based approach. In the Indian context, it is important to understand the demographic characteristics since online consumer behavior varies drastically across cities. Naïve bayes helps to classify such problems by adding the attributes which are independent in nature yet influence the outcome of review helpfulness. The experiment shows encouraging results for Indian OTAs for reliably predicting the usefulness of reviews.

References

[1] Markus Schuckert, Xianwei Liu & Rob Law (2015), Hospitality and Tourism Online Reviews: Recent Trends and Future Directions, Journal of Travel & Tourism Marketing, 32:5, 608-621, DOI: 10.1080/10548408.2014.933154

[2] KPMG, Travel and tourism sector: Potential, opportunities and enabling framework for sustainable growth, http://www.kpmg.com/IN/en/IssuesAndInsights/ArticlesPublications/Documents/KPMG-CII-Travel-Tourism-sector-Report.pdf

[3] Yourstory.com, The big picture of online travel space in India that is set to get bigger, http://yourstory.com/2014/07/online-travel-india/

[4] Yin, Dezhi, Samuel Bond, and Han Zhang. "Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews." Mis Quarterly 38.2 (2014): 539-560.

[5] Schindler, Robert M., and Barbara Bickart. "Perceived helpfulness of online consumer reviews: the role of message content and style." Journal of Consumer Behaviour 11.3 (2012): 234-243.

[6] Hu, Nan, et al. "Manipulation of online reviews: An analysis of ratings, readability, and sentiments." Decision Support Systems 52.3 (2012): 674-684.

[7] Zhang, Yadong, and Du Zhang. "Automatically predicting the helpfulness of online reviews." Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on. IEEE, 2014.

[8] Kim, Soo-Min, et al. "Automatically assessing review helpfulness."Proceedings of the 2006 Conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006.

[9] Mudambi, Susan M., and David Schuff. "What makes a helpful review? A study of customer reviews on Amazon.com." MIS quarterly 34.1 (2010): 185-200.

[10] Connors, Laura, Susan M. Mudambi, and David Schuff. "Is it the review or the reviewer? A multi-method approach to determine the antecedents of online review helpfulness." System Sciences (HICSS), 2011 44th Hawaii International Conference on. IEEE, 2011.