Sentiment Analysis on Customer Reviews in Tourism - A Text Mining Approach

Arun Rajan
Amrita School of Business,
Amrita Vishwa Vidyapeetham University
Coimbatore
arun160890@gmail.com

Prof. Shyam A V
Amrita School of Business,
Amrita Vishwa Vidyapeetham University
Coimbatore

***Abstract*—Tourism is a rapidly-growing practice of travelling across international and national borders to obtain various objectives. Tourism is growing to be an important source of income across globe. Online customer reviews on tourism destined hotels and restaurants play a vital role in the decision making while choosing destination hotel and restaurants. Travel blogs and text reviews in the travel website play a vital role while choosing destination as well as hotel for a given destination. Tourists subsequently use the same information to match their preference. Similarly, hotel authority uses the same information to improve customer service. A better and more effortless inference could be used if all the reviews are analysed for a given location, rather reading sample reviews. Scraping the online customer reviews from travel websites and sentiment analysis on hotel reviews by customers as well as hotel authorities could be used to improve the service to be offered by the hotel executives through meticulous decision making. Sentiment classification is a supervised learning performed on text based data to mine the sentiment of a particular keyword present in the text.**

Keywords*—Sentiment; Tourism; Text Mining; Online Customer Reviews; Decision Tree; Scraping;Semi Automation; Supervised Learning

I.INTRODUCTION

Tourism sector is one of the key factors that contribute heavily for the economic growth and for integration of the world. US$7.2 trillion was generated by tourism and travel and it is 9.8% of the global GDF. Tourism also created 284 million jobs in 2015, which is 1 in 11 jobs created, thus stressing on the fact on how important tourism is for the world economy.

This study aims to explore and demonstrate the utility of text analytics through sentiment analysis to better understand important hospitality issues, mainly the relationship between hotel guest experience, and their satisfaction rating. Specifically, this study applies a text mining and analyze a large quantity of consumer reviews extracted from Tripadvisor.com to deconstruct hotel guest experience via the guest generated reviews and examine its association with their satisfaction ratings. Text mining performed on the website is automated through R programming. The findings reveal several aspects of guest experience that carries varying weights and, more importantly, have novel, meaningful semantic compositions. The association between guest experience and satisfaction appears correlated, suggesting that these two domains of consumer behavior are coherently interconnected. This study will reveal that sentiment analysis can generate new insights into keywords that have been extensively studied in existing tourism and hospitality literature. In addition, implications for theory and practice as well as directions for future research are discussed. It also aims at mining text of written reviews from customers for certain products or services, and classifying the reviews into positive or negative opinions.

Special challenges are associated with text mining on tourist reviews. In this subject specific area of text analytics, keyword semantics in a particular review could contradict with the overall semantic direction (good or bad) of that review. For instance, an "unpredictable" camera implores negative sentiment to the review; whereas a tour with an "unpredictable" experience is positive to explorers.

II.LITERATURE REVIEW

The tremendous social media growth and online customer reviews on the internet has inspired the development of the so-called text analytics to understand and solve real-time problems. No systematic application of text analytic techniques has been implemented in hospitality based studies. However Important research based problems in hospitality and tourism could be resolved using these real time techniques. [1]

Social networking websites have recently become a valid resource for mining sentiments in fields as diverse as customer relationship management and public opinion tracking. Many marketing research companies get ample sentiment insights and knowledge from social networking websites such as Twitter and Facebook based on opinion mining performed on the text data. Public opinion and other text mining entities derived from social networking websites have become very significance in research. However, both at the lexical and the syntactic levels in web texts, imparts considerable problems and hence many a times these data are classified as noisy. [2]

Customer generated contents have helped consumers to know about the strengths and weaknesses of different products and services, and find the ones that best fit their needs. But customer generated contents because of their volume, variety, velocity and veracity, introduce enormous challenge for businesses as well as for consumers in analyzing and deriving insights from them. [3]

Stock market across globe has been significantly influenced by the applications of Internet technologies. Barriers between brokers and geographical location has been eliminated because of the wide usage of the Internet, which in turn, has rendered investors to buy and sell their shares by accessing the stock market status from anywhere at any real time. Wide application of internet has helped investors to predict the trend followed by the stock market at a given point of time before making any transaction, that is, buy or sell stocks. Current Internet based technologies in digital world such as opinion mining, text analytics and Cloud Computing, have significantly changed the way people do business. Sentiment analysis or opinion mining makes use of extensive text mining as well as natural language processing (NLP) and, in order to identify and extract the subjective content and opinion, evaluation on user's attitudes, opinion, sentiments and emotions is required. [4]

Sentiment analysis has also marked its significance in sports domain. Predicting the outcomes of sporting events has a long and rich tradition across the globe. Since ancient times people have designed methods to predict natural and physical sports events. Successful prediction of outcome of such game events still enthralls academicians and gamblers alike. Prediction being more depending on probability, now is considered a complex and intriguing science. Most perplexing dimension of prediction is filtering out noise of the game event from the relevant pattern and parameters. Identifying the relevant pattern is the second most intriguing process while classification of data. Critical parameters or outlier datasets are always difficult to be identified or measured. So need to extract data, which is subjected to constant changes, intermittently, to avoid real time interpretation. Various supervised classification is implemented in sports domain to classify the outcome of the game event. The best methodology is adopted based on minimum percentage error or maximum accuracy rate. [5]

The complexity of the data extractable from social media and the dearth of marketing analyses, in many cases, have posed hindrances in deriving meaningful insights. Bias in data extraction has always resulted in crippled system or erred interpretation. This could be avoided by either by extracting relevant data based on probabilistic sampling which includes all relevant set of diverse dataset, or extracting the entire set of data. [6]

Opinion mining on reviews of advertisement has become revolutionary in digital marketing world due to attribution feature of Internet. Reviews being real time expression of the customer emotions, customer written text data are comparatively more authentic and reliable than any other source of text data.[7]

Air transportation, infrastructure development and technological advancement has direct impact on tourism and hospitality industry of a nation.

III.Data and Methodology

The data is collected from the tourism website Tripadvisor.com for multiple destinations- Goa and Shimla. The motive behind choosing these two location is to compare and identify significance difference in the customer sentiment towards hotels at a beach side destination and hill top destination. Customer reviews were collected from top 10 most reviewed hotels in each destination. 100 reviews were collected from each hotel. To inculcate supervised learning, overall satisfaction rating also was fetched from the review. Overall rating was rated on a scale of 5, 5 being the highest rating and 1 being the lowest. To minimize the effect of biasing, from each hotel 25 reviews were taken each set for 1, 2 4 and 5 rating out of 5. Based on mean split, 4 and 5 rating out of 5 are considered to be positive sentiment. Similarly, 1 and 2 rating out of 5 are attributed as negative sentiment. Out of 1000 reviews, 500 reviews were attributed as positive sentiment, and remaining 500 reviews are attributed as negative sentiment. The same data structure is maintained in hotel in the destination – Shimla.

Customer reviews from Tripadvisor.com are devoid of fraud entries due to their algorithm which internally detects and filters fraud reviews. Hence Tripadvisor.com was chosen to be the source of data. Text reviews were scraped using a semi-automated methodology using R programming. Matching HTML tags are pointed out and customer text reviews are extracted based on matched tags. Following are the steps followed while semi-automated scraping of data:

- *URL parsing*: Fetches the hardcoded URL and navigates to the webpage based on the URL
- *HTML parsing*: Finding the relevant tags which holds the customer generated text reviews Find also the relevant HTML tags which holds the overall satisfaction rating.
- *Extracting the data*: Fetch the text reviews and the overall satisfaction rating from the relevant tags.
- *Saving data*: Save the data into excel or comma separated variable format to improve understanding.

After each iteration of fetching the data, manual entry of the URL which redirects to the next hotel is required. This entire semi-automated mechanism could be improved by complete automated fetching of the text reviews.

Matrix with high frequency keywords needs to be created from the scraped text reviews. Process followed for the creation of such a matrix involves the following steps

- *Punctuation removal*: All punctuations from the text reviews are removed
- *Stop words removal and stemming*: Top 10 stop words by usage like "I", "Me", "My", "Myself", "We", "Our", "Ours", "Ourselves", "You", and "Your" are removed from the text reviews. Similarly, certain other keywords

like “poor”, “never”, “great” etc. are also removed. This removal is done because above mentioned semantic keywords has no significance while decision making to improve customer service

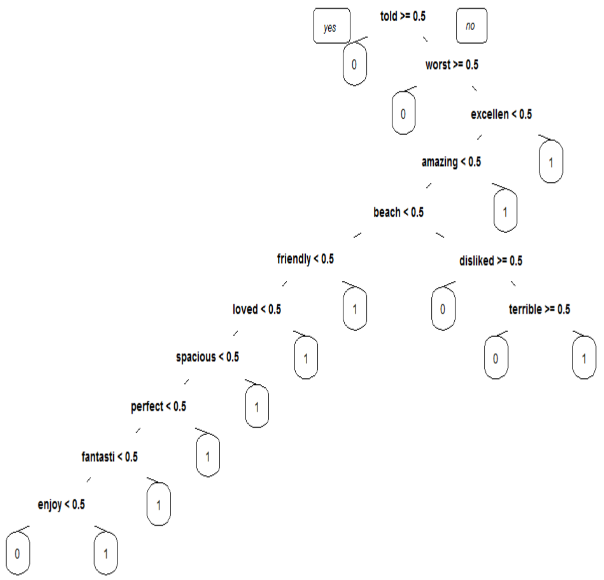
- *Matrix building*: The remaining semantic keywords after stop words removal and stemming, are separated and are treated as different entities. Matrix would look like a 2 dimensional array, with occurrence of each keyword in a text review mapped to its overall sentiment (derived from overall satisfaction rating). Stemming method used in this context is porter stemming. Porter stemming is a semantic based stemming. For instance, keywords like “customer” and “customers” are stemmed into “custom”, whereas keyword “customize” is stemmed into “custo”.
- *Removal of low frequency fields*: Keywords which occur more than 20 times across all the text reviews are only taken for processing.

Supervised learning of thus created matrix is possible due to the inclusion of overall sentiment of the text review. A tree like graph which best portray the algorithm of keyword sentiment are derived from the matrix- decision tree. These decision tree shows each significant keyword, and sentiment related to it, with all the text reviews taken into account. Following are the facts which needs to be understood prior analyzing the decision tree

- Digit ‘0’ represents negative sentiment
- Digit ‘1’ represents positive sentiment
- When the statement holds verity, next statement down left is checked for verity.
- When the statement does not hold verity, next statement down right is checked for verity.

Separate decision tree is created for different destinations of Goa and Shimla. Goa based decision tree on keyword sentiment is as follows:

FIGURE I DECISION TREE ON KEYWORDS DERIVED FROM GOA BASED TEXT REVIEWS

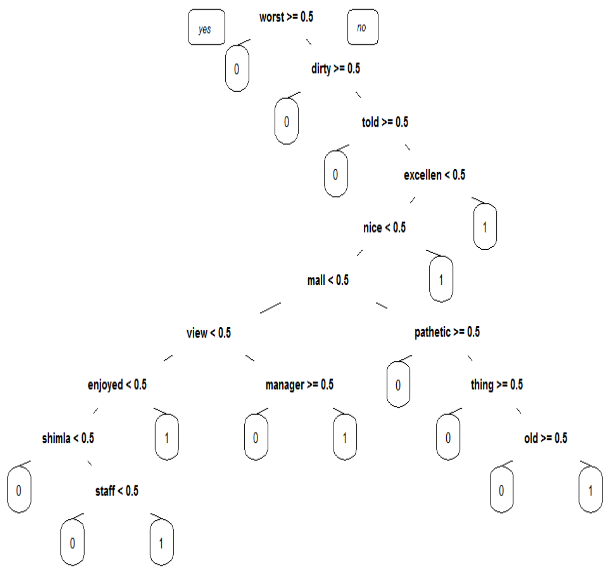


Decision tree arrives at conclusion on sentiment of text reviews, whenever a particular keyword is present or is absent. Based on the Goa based decision tree, following stemmed keywords have significance on sentiment of the review, subsequently: “told”, “worst”, “excellent”, “amazing”, “beach”, “disliked”, “terrible”, “friendly”, “loved”, “spacious”, “perfect”, “fantasti” and “enjoy”.

Keywords have subsequent effect on sentiment. For instance, presence of keyword “told” in the text reviews have always resulted in a negative sentiment (told>=0.5). Similarly, subsequent keywords down the decision tree are verified with subsequent effect. For instance, presence of keyword “worst” (worst>=0.5) is checked whenever the keyword “told” is absent. So the analysis derived from the combination of keyword is – absence of the keyword “told” and presence of keyword “worst” always result in a negative sentiment. Similarly, absence of keywords “told” and “worst” along with presence of keyword “excellent” always result in positive sentiment. Keywords at the top of the decision tree plays more vital role in the determination of sentiment, than keywords at the bottom of the tree.

Shimla being more mature and established tourism destination, the reviews are taken from wider time range. Shimla based decision tree on keyword sentiment is as follows:

FIGURE II DECISION TREE ON KEYWORDS DERIVED FROM SHIMLA BASED TEXT REVIEWS



Based on the Shimla based decision tree, following stemmed keywords have significance on sentiment of the review, subsequently: “worst”, “dirty”, “told”, “excellen”, “nice”, “mall”, “pathetic”, “thing”, “old”, “view”, “manager”, “enjoyed”, “shimla” and “staff”.

As mentioned in Goa based decision tree, stemmed keyword has subsequent effect on sentiment. For instance, presence of keyword “worst” in the text reviews have always resulted in a negative sentiment (told>=0.5). Similarly, subsequent keywords down the decision tree are verified with subsequent effect. For instance, absence of keyword “worst” and presence of keyword “dirty” in the text reviews always impart negative sentiment. Similarly, absence of stemmed keywords “worst”, “dirty”, “told” and presence of keyword “excellen” in the text review has always resulted in a positive sentiment of the review.

IV.CONCLUSION

Sentiment analysis on hotels in both destinations – Goa and Shimla have been performed. Set of stemmed keywords in both the destination show significant differences. Inferences vary according to the stemmed keywords. For instance, Goa based decision tree has “told” at the top of the tree. This indicates the pattern exhibited by majority of the text reviews. Most of the Goa based reviews which has been marked negative sentiment, includes the keyword “told”. In most of the cases, hotel management failed to provide what they told the customer. So the hotel authorities could be cautious on what they promise to the customer.

Similarly, stemmed keyword “worst” and “dirty” are evident from majority of the text reviews of Shimla based hotels. So hotel authorities could improve their services as well as hygiene of the hotel to improve customer satisfaction.

V.LIMITATION AND DIRECTION FOR FURTHER RESEARCH

Since the text mining of customer reviews are semi-automated, there is a scope of improvement with complete automation of text review scraping. This would eliminate the excess time consumed while fetching the online customer reviews from tourism website. Complete automation requires expertise in various programming languages like R and Java(Selenium framework). A more generic automation tool which could automate online text review scraping from multiple tourism based website would provide better and accurate insights.

REFERENCES

[1] Salvador Anton Clave´ Planeta ()“Lessons on tourism:the challenge of reinvesting destinations”,‖ Annals of Tourism Research, vol. 41, pp.249-250, 2013.

[2] Zheng Xianga,∗, Zvi Schwartzb, John H. Gerdes Jr.c, Muzaffer Uysala ()“What can big data and text analytics tell us about hotel guestexperience and satisfaction?” ,| International Journal of Hospitality Management, vol. 44, pp.120-130, 2015.

[3] Mohamed M. Mostafa ()“More than words: Social networks’ text mining for consumer brand sentiments” ,| Expert Systems with Applications, vol.40,pp.4241-4251, 2013.

[4] Mohammad Salehan a, Dan J. Kimb ()“Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics”,| Decision Support Systems, vol. 81, pp.30-40, 2016.

[5] Aditya Bhardwaja*, Yogendra Narayanb, Vanrajc, Pawana, Maitreyee Duttaa ()”Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty", Procedia Computer Science, vol.70, pp.85-91, 2015.

[6] Robert P. Schumaker a, A. Tomasz Jarmoszkob, Chester S. Labedz Jr. cComputer ()”Predicting wins and spread in the Premier League using a sentiment analysis of twitter”,| Decision Support Systems, vol.88, pp.76-84, 2016.

[7] Haeng-Jin Jang a,1, Jaemoon Sim b,2, Yonnim Lee b,2, Ohbyung Kwonc ()”Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media”

[8] Syed Abdul Rehman Khan, Dong Qianli, Wei SongBo, Khalid Zaman, and Yu Zhang () “Travel and tourism competitiveness index: The impact of air transportation, railways transportation, travel and transport services on international inbound and outbound tourism”, ‖ Journal of Air Transport Management, vol. 58, pp. 125 – 134, 2016

[9] Garín-Muñoz, Teresa; Pérez-Amaral, Teodosio ()“Internet Usage for Travel and Tourism. The Case of Spain”, 21st European Regional ITS Conference, Copenhagen 2010

[10] Suresh M, Kranmoy Bid, Sangeetha Gunashekhar, “Inbound international tourism development in India: A panel data analysis on its affecting factors”, IEEE International Conference on Computational Intelligence and Computing Research, Vickram College of Engineering 2015