Psychographic Segmentation Based on Online Reviews

Prof. Lalit K Ojha, IMT Nagpur

ABSTRACT

Online customer reviews are great insights into customer thought process. Prior studied have used online reviews to understand sentiment etc. We looked at TripAdvisor reviews for two properties of same hotel chain to confirm whether the heritage and business travelers accord varying priorities to services. Training and Validation was performed using three different statistical analysis techniques were used to ascertain suitability of one technique over another. The results show that SVM has better prediction capability for classification of results. We also validate the non-overlapping aspect of heritage and business hospitality segments.

INTRODUCTION

This paper is divided into four parts. In first part, we will discuss the nature of hospitality industry and some unique characteristics that are of interest to us for study topic. In second part, we will discuss user-generated content (UGC) w.r.t its features that will allow us to move beyond simple sentiments but obtained deeper understanding of human personality and behavior. In third part, we specify the elements that allow us to extract more psychographic traits from UGC. Lastly, we compare some of most commonly used techniques for Natural Language Processing (NLP) for classification problems to obtain best results.

HOSPITALITY INDUSTRY

The hospitality industry is unique because it relies so heavily on discretionary income and free time (Chan & Ni, 2011), Lee K. & Ha I(2012), (Canina & Carvell 2005). One of the most defining aspects of this industry is that it focuses on customer satisfaction (Kandampully & Suhartanto,

2000), (Pizam & Ellis, 1999). This is because these businesses are based on providing luxury services rather than basic needs. The industry is on the verge of an evolutionary leap where the interaction between customer and company becomes real-time and relevant. Hence, Industry is now shift towards data centric personalization.

Data centric personalization in Hospitality Industry

Researchers had envisaged an information management system that would be integral to deliver service to customer satisfaction (Minghetti, 2003). (Crotts et al, 2009) established the use of online blogs as one of the means to estimate customer satisfaction as well as competitive positioning. A completely new framework for this goal for suggested by (Neuhofer et al.2015). This framework highlights that instead of traditional B2C and C2B information flow, now hospitality interactions are significantly mediated by technology. The data generated from these sources is instrumental in personalization of service for customers. This key initiative indicates drive of a business to go extra mile to provide their customers with experiences and services designed according to individual needs and preferences. It can be frequent business flyer being offered their favorite drink without asking for it or being addressed by name on boarding. Or a hotel guest finding that room temperature is already set to their preference even before they enter. Data centric personalization becomes enabler for employees to provide better service while also providing right promotions and message to right set of customers.

CHARACTERSTICS OF UGC

As a social being, people generally want to be part of something and have meaningful connection. This concept in social psychology is known as of sense of community. There are four elements that encouraged people to feel like part of a community. "*Membership* is the feeling of belonging or of sharing a sense of personal relatedness. *Influence* is a sense of mattering, of making a difference to a group and of the group mattering to its members. *Integration and fulfillment of needs* is the feeling that members' needs will be met by the resources received through their membership in the group. *Shared emotional connection,* the commitment and belief that members

have shared and will share history, common places, time together, and similar experiences", (McMillan and Chavis, 1986).

While all of the points are relevant to online UGC, influence and emotional connect are specifically useful to understand people involvement with UGC. This sense of community enables customer to navigate thru information overload. Multiple domains like psychology, marketing, information science etc. have studied information load form respective perspectives (Meyer, 1998), (Jacoby, 1984). The research also focused on impact of information overload on consumer decision making (Speier et. Al. 1999), (Malhotra, 1984). With the advent of internet as a popular source of information, this phenomena was observed in online environment also (Lee & Lee, 2004). Growth in smartphones further fueled a distinct phase of UGC (Furner & Zinko, 2017). When we look at UGC from this perspective, it serves the purpose of easing decision process by promoting sense of community and reducing information overload. Individual can use UGC for decision making when they trust the source of information (Chari et al, 2016).

UGC and Personality

Traditional methods of measuring personality characteristics were derived from earlier research from psychology perspective. This mandated use of survey methodology limiting the possibility of taking large samples. But in recent times there have been studies that used text data to obtain personality elements. (Fast & Funder, 2008) derived elements of personality from self-reported and acquaintance-reported personality characteristics in text format. On the other hand, (Hirsh & Peterson, 2009) required subject to write about past experiences and future scenarios. The text elements used were used to arrive at personality elements (Yarkoni 2010). (Pennebaker & King, 1999) demonstrated that linguistic styles differ for segments. In their study, substance abuse inpatients, students and scholars demonstrated varying use of language. We are furthering this observation by demonstrating that heritage tourism customers and business customers have different perspective to same luxury experience. .

Computational linguistics and allied areas have been at forefront of using machine learning to explore the personality traits from social media text. (Mairesse & Walker, 2006)a detailed that text based personality identification would can work better than self-reported data. (Mairesse &

Walker, 2006)b validated text elements that confirm personality type. Hence, using linguistic psychology and data-mining algorithms and automatic detection of personality characteristics is very much possible. We can now use readily available unstructured data from digital platforms and conduct large scale analysis.

LINGUISTIC ELEMENTS OF UGC

Various researcher have used different elements of language and approaches for extracting personality elements from text data. Below table summarizes text elements used by earlier studies.

Table 1: Platforms and linguistic elements used in study of text data for personality elements

S.no.	Article	Platform/s	Text elements
1	Fast & Funder, 2008	LIWC (Pennbaker et al. 2001)	66 of 87 Linguistic categories (e.g. pronouns, preposition, number, affects etc) offered by LIWC
2	Hirsh & Peterson, 2009	LIWC (Pennbaker et al. 2007)	Categorized LIWC dimensions (e.g. Affect words, Cognitive process, Relativity etc.)
3	Yarkoni, 2010	LIWC (Pennbaker et al. 2007)	Categorized LIWC dimensions supplemented with individual word correlation with Personality traits
4	Mairesse & Walker, 2006	LIWC (Pennbaker et al. 2001) and MRC Psycholinguistic database (Coltheart, 1981)	LIWC word categories supplemented with MRC features (e.g. Inclusive words, past tense verbs etc.)

This is a representative table to illustrate the elements of text that have been utilized to extract personality data. For tagging the parts of text, there are advanced tools present today but we would be limiting the discussion to personality type related studies. Secondly, having established that text information is indicator of personality type, we would move to core discussion topic of using text data for identifying psychological elements that differentiate Business vs Heritage travelers.

In the next section we will look into data collection and methodology used to arrive at suitable classification models for this purpose.

DATA COLLECTION AND ANALYSIS

Data Extraction

Past studies have indicated that heritage value of hotels have impact on pricing of hotels even under the same brand of hotels. About 80% of the articles in top 6 travel and hospitality journals on online reviews have used TripAdvisor as the sources. We extracted all reviews for two hotel properties i.e. Oberois Bangalore and Oberois Jodhpur. Oberoi Bangalore is new age modern business hotel while Oberoi Jodhpur is a Heritage property. A scrapper was design to extract all TripAdvisor reviews will June2017. The reviews were screened for language and brevity. Non-english reviews were removed from data, about 2% of total reviews. We also used a minimum word count criteria so exclude review that did not present enough information for analysis. In end, we worked with 1200 reviews for Oberoi Bangalore and 1400 reviews for property Oberoi Jodhour.

In their study, (Mathur & Dewani, 2016) estimated that cultural heritage itself has economic value and has impact on pricing and profit. In this preliminary study, we want to establish that consumers of heritage tourism are inherently different from business traveler in psychographic profile and personality characteristics. Hence, they constitute a different segment altogether. In this direction, we would primarily address this is classification problem based on text analysis of online reviews posted by consumers of these services.

Most of the past literature has been focused on using text analytics for sentiment analysis of the content (Pang et al, 2002). Primarily two types of techniques are used for this purpose, machine learning and semantic orientation. Semantic orientation depends on interpretation and context of words and statement used in reviews. (Chaovalit and Zhao, 2005) have suggested that supervised machine learning approach works better in case of online reviews. Naive Bayes, SVM and N-gram

are the three most important approaches in text mining and sentiment classification (Joachims, 1998; McCallum & Nigam, 1998; Pang et al., 2002; Yang & Liu, 1999)

SVM: SVM is a supervised machine learning algorithm, primarily used for classification problems. The underlying technique involves finding a hyperplane that best divides the dataset into two classes. It derives its name from data points called support vectors which if removed will change the position of hyperplane. We applied SVM classifier with information gain (IG) as a feature selection method. In the experiments, we chose the word frequency to present a document rather than word presence for probability estimation.

Naives Bayes: Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

N Gram Model: An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n − 1)–order Markov model. Two benefits of n-gram models are simplicity and scalability – with larger n, a model can store more context with a well-understood space–time tradeoff, enabling small experiments to scale up efficiently. In addition, because of the open nature of language, it is common to group words unknown to the language model together.

RESULTS

Results confirms that the drivers of value perceptions for customers using heritage hospitality brands are significantly different from those of business hospitality consumers. We also put forth a new approach to identify segments from customer population based on their key motivational drivers. The insights from methods selected suggest that key motivational drivers for drivers of heritage products and services consumers in hospitality industry can be obtained from the text

analytics methods. The results also show that SVM has best prediction ability this type of classification. Segmentation methods were held adequate in cross validation techniques.

S.No.	Technique	Precision (%)
1	SVM	85.07
2	Naives Bayes	83.71
3	N Gram	82.23

LIMITATIONS AND NEXT STEPS

Primary aim of this study was to propose and validate methodology for segmentation using online reviews. As the next step, role of individual element of personality needs to be in context of service products. Apart from business and heritage traveler context, this study needs to be extended to other industries to further validate the outcomes.

REFERENCES

Wilco W. Chan & Shanshan Ni (2011) Growth of Budget Hotels in China: Antecedents and Future, *Asia Pacific Journal of Tourism Research*, 16:3, 249- 262.

Lee K. & Ha I(2012): Exploring the Impacts of Key Economic Indicators and Economic Recessions in the Restaurant Industry, *Journal of Hospitality Marketing & Management*, 21:3, 330-343

Canina, L., & Carvell, S. (2005). Lodging demand for urban hotels in major metropolitan markets. *Journal of Hospitality & Tourism Research*, *29*(3), 291-311.

Kandampully J. & Suhartanto, D. (2000) "Customer loyalty in the hotel industry: the role of customer satisfaction and image", *International Journal of Contemporary Hospitality Management*, Vol. 12 Issue: 6, pp.346-351,

Pizam A. & Ellis T. (1999) "Customer satisfaction and its measurement in hospitality enterprises", *International Journal of Contemporary Hospitality Management*, Vol. 11 Issue: 7, pp.326-339,

Crotts, J. C., Mason, P. R., & Davis, B. (2009). Measuring guest satisfaction and competitive position in the hospitality and tourism industry: An application of stance-shift analysis to travel blog narratives. *Journal of Travel Research*, *48*(2), 139-151.

Minghetti, V. (2003). Building customer value in the hospitality industry: towards the definition of a customer-centric information system. *Information Technology & Tourism*, *6*(2), 141-152.

Neuhofer, B., Buhalis, D., & Ladkin, A. (2015). Smart technologies for personalized experiences: a case study in the hospitality domain. *Electronic Markets*, *25*(3), 243-254.

McMillan, D. W., & Chavis, D. M. (1986). Sense of community: A definition and theory. *Journal of community psychology*, *14*(1), 6-23.

Meyer, J. A. (1998). Information overload in marketing management. *Marketing Intelligence & Planning*, *16*(3), 200-209.

Jacoby, J. (1984). Perspectives on information overload. *Journal of consumer research*, *10*(4), 432-435.

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, *30*(2), 337-360.

Malhotra, N. K. (1984). Reflections on the information overload paradigm in consumer decision making. *Journal of consumer research*, *10*(4), 436-440.

Lee, B. K., & Lee, W. N. (2004). The effect of information overload on consumer choice quality in an on-line environment. *Psychology & Marketing*, *21*(3), 159-183.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, *47*, 140-151.

Furner, C. P., & Zinko, R. A. (2017). The influence of information overload on the development of trust and purchase intention based on online product reviews in a mobile vs. web environment: an empirical investigation. *Electronic Markets*, *27*(3), 211-224.

Chari, S., Christodoulides, G., Presi, C., Wenhold, J., & Casaletto, J. P. (2016). Consumer Trust in User-Generated Brand Recommendations on Facebook. *Psychology & Marketing*, *33*(12), 1071-1081.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology*, *94*(2), 334.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of research in personality*, *43*(3), 524-527.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Mairesse, F., & Walker, M. (2006, January)a. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 28, No. 28).

Mathur, S., & Dewani, P. P. (2016). Influence of cultural heritage on hotel prices, occupancy and profit: Theory and evidence. *Tourism Economics*, *22*(5), 1014-1032.

Mairesse, F., & Walker, M. (2006, June)a. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (pp. 85-88). Association for Computational Linguistics.

Ghose, A. and Ipeirotis, P.G., 2007, August. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce* (pp. 303-310). ACM.

B. Pang, L. Lee, S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques P. Isabelle (Ed.), *Proceeding of 2002 conference on empirical methods in natural language, Philadelphia, US, Association for Computational Linguistics* (2002), pp. 79-86

Chaovalit, P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. In R. R. Sprague (Ed.), *Proceedings of the 38th Hawaii international conference on system sciences, Big Island Hawaii* (pp. 1–9). IEEE.

Yang, Y. M., & Liu, X. (1999). A re-examination of text categorization methods. In F.Gey, M. Hearst, & R. Tong (Eds.), *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, California, United States* (pp. 42–49). ACM Press

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. *In AAAI-98 workshop on learning for text categorization* (pp. 41–48). AAAI Press.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of 10th European Conference on Machine Learning (ECML-98), Chemnitz, Germany, 1998* (pp. 137–142). Springer