Annals of Tourism Research xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Annals of Tourism Research

journal homepage: www.elsevier.com/locate/atoures

Research Note

Application of text mining in tourism: Case of Croatia

Uroš Godnov[a], Tjaša Redek[b,*]

[a] University of Primorska, Slovenia
[b] University of Ljubljana, Slovenia

Ninety percent of travelers use on-line reviews in travel decision and planning (Simms & Gretzel, 2013; TripAdvisor, 2013). On-line platforms typically provide both numerical and textual evaluations and are especially important in decision-making (Gretzel, Yoo, & Purifoy, 2007), as they are considered objective and trust-worthy. This paper highlights how text mining, accompanied by numerical evaluations' analysis, provides efficient review summations, valuable decision-making input for travelers and management, and facilitates comparative analysis.

Several text-mining methods (Table 1) were applied to 18 thousand reviews of 87 Croatian hotels[1] (Table 2). Sentiment analysis numerically captures the text's overall "feel," and can be progressed with an analysis of the review's prevailing emotion and polarity. Content analysis methods reveal the main topics discussed: key-words identify the most common words and capture the essential idea, while more advanced methods investigate correlations between words and identify topics using probabilistic topic models.

The average numerical evaluation was 4.28 (of 5); average sentiment was roughly 20. Mildly positive words (primarily adjectives) prevail (good, great, nice, clean, friendly, helpful, recommend, sentiment value +2 and 3), followed by the words "free" (no charge), big (usually with "room"), fresh, easy, and huge. The most common negative word is "problem."

Text analysis shows that numerical evaluation is inadequate. The correlation between numerical evaluation (1–5) and sentiment, which reflects hotel quality as "read between the lines," is significant but weak (0.28, $p < 0.001$). Numerical evaluation is negatively related to review length (dissatisfied customers write more). If accounting for length (sentiment per word; longer reviews have higher absolute sentiment value), the correlation is of medium strength (0.45, $p < 0.001$).

Emotional analysis based on the NRC lexicon (Mohammad, 2015) showed that trust, anticipation, and joy were the predominant emotions (71% of reviews). Polarization showed that 62.8% of reviews were positive.

* Corresponding author at: Faculty of Economics, University of Ljubljana, Kardeljeva ploščad 17, 1000 Ljubljana, Slovenia. Tel.: +386 15892400; fax: +386 15892698.

E-mail addresses: uros.godnov@fm-kp.si (U. Godnov), tjasa.redek@ef.uni-lj.si (T. Redek).

[1] Data were gathered from the biggest traveler platform during 2015. Selected hotels cover all major Croatian seaside destinations. Other accommodation types have few reviews. Since paper focuses on selected methodology and not on comparative analysis, the focus on hotels should not present a problem.

http://dx.doi.org/10.1016/j.annals.2016.02.005

Table 1
Systematization of the methodology used.

Method	Selected references	Research output	Implications for destination image analysis
Sentiment analysis (a) Comment's sentiment value (b) Polarization analysis (c) Emotional analysis (NRC)	Nielsen (2011), Hansen, Arvidsson, Nielsen, Colleoni, and Etter (2011), Mohammad (2015)	Numerical evaluation provides a significantly different picture of the tourist destination image than the textual evaluation Identification of the prevailing emotion in the text (disgust, fear, anger, sadness, surprise, anticipation, trust, joy) or identification of whether the review as a whole is positive, negative, or neutral List of the most positive and negative terms	Important to evaluate also the text of reviews from emotional value, it reveals how the reader might understand the review from the emotion/impression it conveys Helps in the competitiveness analysis; useful to strategic behavior of firms and decision-making of potential travelers The selection of positive/negative words used in reviews points to the deficiencies and strengths of a location
Key-words analysis Correlation between words	Zhang et al. (2008), Feinerer, Hornik, and Meyer (2008)	The first step in topic analysis Identification of most discussed aspects Word-network	Keywords capture the essential storyline and points of discussion in the document/review Word-net reveals the strongest relationships between words: what appears together. Helps identify main topics Both reveal the most important elements of a destination's image determination and what was good or bad. Useful for travelers and management
Probabilistic topic models (e.g. Latent Dirichlet Allocation)	Blei (2012)	Identification of key topics discussed in the text and the most common terms associated with each of those topics. Several approaches can be used	Identification of key topics discussed is an upgrade to key-words as it reveals the aspects (topics) most important to travelers, as well as the most common opinion about a specific topic or what is being discussed within a specific topic Identification of key areas (positive, negative) of discussion and what is being discussed. Useful for travelers and managers, similarly as above, and for travel choice or strategic management decisions

Table 2
Descriptive statistics.

	N	Minimum	Maximum	Mean	Std. deviation
Numeric evaluation given by reviewer	18,288	1	5	4.28	.904
Number of words in textual evaluation	18,288	7	3468	186.10	171.205
Sentiment	18,288	−49	309	19.89	15.456
Sentiment per word	18,288	−0.440	1.375	0.144	0.108

Regarding content, key-words analysis relying on word frequencies (stems) revealed the most common nouns: hotel, room, staff, food, pool, and breakfast (Fig. 1). Prevailing adjectives included: good, great, nice, clean, and friendly. Results signal that guests mostly appreciate room quality, staff friendliness, hotel facilities (pool), and food (Fig. 1).

Network of words (Fig. 2) shows that reviews revolve primarily around the hotel: room, food, pool, staff, and their qualities (and not other resort facilities).

Latent Dirichlet Allocation was used to identify topics discussed in texts (Chen, 2011) (Table 3). An iterative procedure (5000 iterations) identified the 'nodes' of each word cluster and accompanying