Andrei P. Kirilenko
University of Florida

Svetlana O. Stepchenkova
University of Florida

Automated Topic Modeling of Negative Tourist Reviews

Abstract: Automated content analysis of online travel reviews allows analysis of topics of travelers' satisfaction and dissatisfaction, yet its domain is not well researched. We suggest that the "Anna Karenina principle" positing a greater variability of the factors leading to business failure as opposed to those leading to its success results in limitations of topic modeling applied to dissatisfied visitor reviews. We test our hypothesis using TripAdvisor reviews of the Museum of Terracotta Warriors (China). Our findings confirm the hypothesis; we also report the main themes of satisfaction and dissatisfaction of visitors from mainland China and industrialized English speaking countries.

Key words: Traveler satisfaction, Latent Dirichlet allocation, Anna Karenina effect, TripAdvisor, Social media.

Andrei P. Kirilenko, Ph.D.
Department of Tourism, Hospitality, and Event Management
College of Health and Human Performance
P.O. Box 118208
Gainesville, FL 32611-8208, USA
Email: andrei.kirilenko@ufl.edu

Svetlana Stepchenkova, Ph.D.
Department of Tourism, Hospitality, and Event Management
College of Health and Human Performance
University of Florida
P.O. Box 118208
Gainesville, FL 32611-8208, USA
Email: svetlana.step@ufl.edu

Dr. Andrei P. Kirilenko (andrei.kirilenko@ufl.edu) is Associate Professor in the Department of Tourism, Hospitality and Event Management at the University of Florida. He received his Ph.D. in Computer Science and held positions at the Center for Ecology & Forest Productivity, Russia, European Forest Institute, Finland, U.S. Environmental Protection Agency laboratory, OR, Purdue University and University of North Dakota. His research interests include big data analysis, data mining, tourism analytics, climate change impacts, and sustainability issues.

Dr. Svetlana Stepchenkova (svetlana.step@ufl.edu) is Associate Professor at the Department of Tourism, Hospitality and Event Management at the University of Florida. Her research interests are in the area of marketing communications, branding, and positive image building. She studies tourism behavior and the effectiveness of destination promotion in situations of strained bilateral relations between nations. She is also interested in usability of user-generated content for managerial decision making in destination management.

Introduction

Recent Web 2.0 growth has provided ample opportunities to research tourist opinions of attractions and provided services based on reviews tourists share online. In the literature, two main directions of such analysis are clearly distinguished: (1) sentiment analysis (or polarity analysis), focused on extraction of the sentiment (positive, negative, or neutral) expressed in a review of a particular attraction, hotel, or destination and (2) topical analysis (or topic modeling), targeted at extraction of the review's meaning. In sentiment analysis, a variety of tools have been applied to various types of user-generated content and applicability of those tools has been evaluated (Kirilenko et al. 2018). The applicability of automated topical analysis, however, is not researched well.

The goal of topical analysis is to assist in understanding, classification, and generalization of the meaning of a collection of documents (a corpus). While the first developments of automated topical analysis can be traced back to the text indexing research published in the early 1960s (Borko & Bernick, 1962), very few tourism papers used topic modeling prior to 2016 and all of them were published in non-tourism journals. It is only as late as 2019 when the leading tourism journals have published more than a handful papers that used topic modeling. Apparently, this development is owed to a recent crop of user-friendly software based on the Latent Dirichlet allocation (LDA) method (Blei, Ng, & Jordan, 2003).

Recently, LDA applications in tourism are gaining popularity: e.g., two major tourism journals, *Tourism Management* and *Journal of Travel Research* published two LDA based papers in 2017 (Guo, Barnes, & Jia, 2017; Xiang et al., 2017), none in 2018, and seven in the first nine months of 2019; (the third major journal, Annals of Tourism Research, published none). In comparison, another popular in text mining topic analysis method based on the singular vector

decomposition (SVD) was used only in three articles published in these tourism journals. While multiple other topic modeling methods vary from factor analysis (Borko & Bernick, 1962) to hierarchical clustering to ontology-based semantic clustering (Vincent & Moreno, 2015), their acceptance in tourism applications is very limited. Meanwhile, the limitations of LDA method are largely unknown (Tang et al. 2014).

The experimental research of LDA based classification of multiple datasets (Tang et al. 2014) suggested the following informal guidelines: (1) The number of documents should be large enough; (2) The length of documents should be large enough; (3) when either the number of documents or their length are above a certain threshold, topical analysis obtained from a sample are similar to the one obtained from the entire corpus; (4) extracting overly large number of topics should be avoided; and (5) For LDA success, the topics should be well-separated, that is, the topics should be concentrated at a small number of words. In travel review analysis, LDA has been applied to datasets of radically different sizes, ranging from as few as 50 (Putri & Kusumaningrum 2017) to over 250,000 (Guo, Barnes, & Jia, 2017) reviews. The latter article regressed the review star rating on 19 review topics extracted with LDA with the purpose of finding the most important dimensions of tourist satisfaction. The dataset size (266,544 reviews of 25,670 hotels located in 16 countries) clearly fits within the guidelines suggested in Guo et al. (2017) research which operated with only up to a few thousand document long corpora. Restricting LDA analysis to large cumulative datasets is however impractical: a single popular destination typically contributes several hundred to several thousand reviews annually as evidenced from TripAdvisor pages. Dividing the collected dataset into several groups e.g. by the expressed sentiment may further reduce subset sizes. In their analysis of Indonesia tourist reviews, (Putri & Kusumaningrum, 2017) collected 100 reviews, which were further divided into positive and negative review classes. It is

not entirely clear if LDA analysis applied to the small datasets is still capable of returning coherent topics.

The opening lines of Lev Tolstoy's novel Anna Karenina state: "All happy families resemble one another; each unhappy family is unhappy in its own way". The overarching principle stating that while "no feature guarantees success, many guarantee failure" (Shugan, 2007, 146) was famously popularized by Jared Diamond (1997) in his bestselling book "Guns, germs, and steel: the fates of human societies". Since that time, the principle has been applied in many areas of science as diverse as statistics, geotectonic, pest control, and genetics. In tourism, however, The Anna Karenina principle (TAK) was used in a single study (Tasci, Croes, & Villanueva, 2014).

Shugan (2007) suggested that an important outcome of the TAK is that "the most revealing variables might exhibit negligible variation among survivors because survivors are necessarily alike. Perhaps variability is inversely related to the variable's importance for survival" (pg. 145). In application to topic modeling, we might argue that same principle suggests that the reviews of the satisfied customers are "more alike" as compared to reviews of the dissatisfied customers. That does not mean that there are fewer topics in positive reviews as compared to the negative ones; rather, the positive topics are more likely to be shared among multiple reviews. In practice, that should result in lesser separability of the topics in the negative reviews. Hence, this research investigates the following proposition:

> Reviews of the customers with low satisfaction of the attractions have low topic separability as compared with reviews left by the satisfied customers, which results in poorly interpretable topics. This proposition is invariant for visitors from different countries.

Data

The data was collected from TripAdvisor reviews of the Museum of Qin Terracotta Warriors and Horses popular known as the Terracotta Army in Xi'an, China. The museum exhibits cir. 8,000 life-size figurines of the imperial guards of Qin Shi Huang, the first emperor of a unified China. The excavated tomb is a UNESCO World Heritage Site and is believed one of the most significant archeological findings of the 20th century. The historical significance of the site attracts over 50,000 tourists at peak days (Tianzhu, 2019), bringing numerous problems related to transportation, crowdedness, logistics, food, and crime.

We collected 14,273 visitor reviews from Tripadvisor site of the museum. The collected reviews were referenced to the reviewers' home countries in the following way. First, the text describing location of the reviewer was resolved into the latitude and longitude using Google geolocational API. Next, the geographical coordinates were reverse-geolocated into countries using OpenCage API (OpenCage.com). In total, the location was found for 11,242 reviews left by the visitors from 129 countries. The most of the reviewers came from the USA (19.7%), followed by the mainland China excluding Hong Kong and Taiwan (15.0%) and the UK (11.6). For further processing, two subsets were selected: "en" subset (5859 reviews) representing English language reviews coming from visitors from five industrialized and culturally close countries (USA, UK, Australia, Canada, and New Zealand) and "zh" subset (1792 reviews) representing Simplified Chinese reviews by visitors from the hosting nation (mainland China excluding Hong Kong and Taiwan). Table 1 shows reviews rating distribution.

Notice that distribution of the reviews from two samples over the ratings is significantly different (Chi square = 571, df=4, $p<0.001$; Mann-Whitney U = 3,842,409, $p<0.001$). The zh sample contains significantly lesser percentage of 5-star reviews and significantly higher

percentage of 3 and 4-star reviews while the percentage of 1 and 2-star reviews are same. The English and Chinese subsamples were then divided into subsamples of 1– 3-star reviews and 4 – 5-star reviews for a total of four subsamples.

Table 1. Distribution of collected reviews from the English speaking (en) countries and China (zh) over the attraction point ratings.

Stars	N en	% en	N zh	% zh
1	11	0.2	5	0.3
2	38	0.6	13	0.7
3	149	2.5	156	8.7
4	785	13.4	605	33.8
5	4876	83.2	1013	56.5
Mean	4.8		4.5	

Method

Main topics discussed by the visitors to the Terracotta Army site were extracted in the following way. First, the zh dataset was translated from Simplified Chinese to English using Google Translate. Second, all collected reviews were pre-processed using the standard steps in computer-assisted content analysis:

1. Tokenization with removal of short (below 4 symbols) and long (above 25 symbols) tokens;
2. Filtering English stopwords (from Page Analyzer, https://www.ranks.nl/stopwords);
3. Filtering tokens by part of speech (POS filtering retaining nouns and adjectives);
4. Stemming (reducing inflected words to their word stems using Porter stemmer);
5. Transformation of words to low case;
6. Transformation of words with variant spellings (e.g., terracotta and terra cotta);
7. Filtering the most frequent and infrequent words (those encountered in over 70% and lesser than 1.5% of documents, accordingly).

Third, the main topics appearing in the pre-processed reviews were extracted with LDA as implemented in MALLET package with optimized topic density/words density parameters. Two indices were used to compare performance of LDA topic models: perplexity and coherence. The perplexity measures how well the word distribution predicted by the LDA model matches the actual word distribution. That is, perplexity measures how well the outcomes of the theoretical model built on the underlying LDA assumption matches the observations. The coherence measures the semantic similarity between the high-ranking words in a topic. Specifically, the coherence evaluates the frequencies with which the high-ranking and lower-ranking words composing same topic tend to appear together in documents. Note that the perplexity and coherence both depend on the number of topics K in the LDA model and as such are suitable for comparison of performance of different models, but not as substitutes for an absolute measure of model performance.

The latter point is important in deciding upon the values of LDA parameters. The LDA algorithms optimize distribution of the words in topics and topics in documents given a pre-set number of topics K, hence the value of K affects LDA results. The automated methods based on perplexity minimization that allow to find the "best" K are known to return very large K values roughly equal the 0.05 of the number of documents, frequently resulting in thousands of poorly interpretable topics. Hence, the majority of practical applications employ a manual method to determine the value of K that returns a small number of topics which are well interpretable. In our research we used a combination of both approaches known as the "elbow method". As K increases, topic coherence tend to improve; this improvement is fast when K is small and slow when K is large. The elbow method evaluates K as the number of topics approximately corresponding to the "elbow" on the coherence(K) function plot (Figure 1). Mathematically this elbow corresponds to the maximum of the second derivative of the function. The precise value of K is then determined

based on the topic interpretability. Following this procedure, first we determined the optimal value of K to be between 10 and 15 from the elbow analysis, then manually interpreted results of LDA models with K=10…15, and finally selected the K value corresponding to the best topic interpretability. Hence, the optimal K value was set at K=14 for en reviews and 11 for zh reviews.

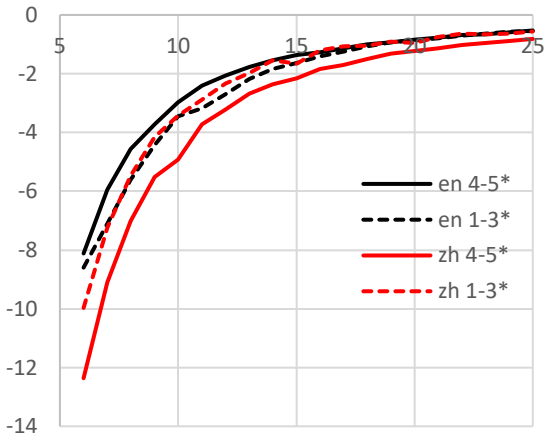

Figure 1. The change in topic coherence as a function of the number of topics for English (en) and Chinese (zh) datasets; 1 – 3 and 4 – 5 star reviews are shown separately.

Results

Table 2A and 2B show the topics identified from the en and zh datasets, respectively. While the topics arising from the positive (4 – 5 star) reviews are easily interpretable based on words comprising those topics, the majority of topics arising from negative (1 – 3 star) reviews are hard to interpret based on the words alone. Manual analysis of the negative reviews which load highly on those topics found multiple complaints about a large variety of service failures. In the en, but not the zh dataset, the most common shared complain was crowdedness of the site. Indeed, "crowd" was the most frequent negative word in en 1 – 3 star reviews and ranked 13th in the overall word frequency distribution; the same word was also frequent in the positive reviews and ranked 20th. Connected to the crowdedness issue, en dataset also contained multiple complaints about tourist behavior: pushing, rudeness, cutting into the lines, and similar. The majority of other

negative review topics were similar in the en and zh datasets. The most common shared complains were about local entrepreneurs, including soliciting, pushing sales, dishonesty, seedy business street leading from the station to museum and similar. Interestingly, many reviewers mentioned a local entrepreneur positioning himself as a farmer who discovered the site in 1974 selling autographed books; some reviewers seemed to believe the farmer was genuine, while many others commented he was fake. Finally, English and Chinese visitors alike complained about prices; in addition, en students noticed that while the Chinese student tickets were half-priced, they did not share this discount.

Other complain topics, however, were dispersed. Those negative reviews were related to prices for various services, management problems (tour guide confusion, bad software, dirty toilets, half of the museum closed during the travel peak, ticket lines and ticket scalpers at gates), unmanaged photography such as dazzling flashes, and many others. Many reviewers disputed authenticity of the place, ranging from its commercialization ("Disney attraction rather than a archeological treasure") to its genuineness as a whole ("a ploy for Communist China to gain tourism"). Similarly, some negative reviewers from both data sets believed that the museum artefacts were either replicas or entirely forged. Another criticism was connected with interpretations of the site history by the museum guides misaligned with the one found in historical books. Complains about travelling to museum site were also diverse and included highway traffic, bus drivers miscommunicating the route to get more passengers, scary taxi rides, and a ride from the train station in a golf cart. Finally, many reviews expressed general disappointment: boring, just a dirty pit, broken pottery, missing figurines, and similar (see comments to Figures 2A and 2B).

Table 2A. Topics extracted from the English sample. The tokens represent the highest ranking words.

4 – 5 star reviews	**Tokens**
Culture, 8th Wonder of the World	site world china wonder history archeology great ancient visitor culture
Farmer signs the book, gifts	farmer book picture warrior shop sign site gift terracotta souvenir
The place is worth to visit	place visit amazing worth time history picture museum site guide great
Travel logistics	city Beijing wall Xian time hour china trip airport hotel train Shanghai
Excellent private tour	guide tour museum informative private group time history knowledge excel
Amazing original excavation	warrior work pit archeology piece site restore excavate horse amaze original
Area logistics, shops and food	shop walk museum souvenir restaurant food price area good park gift
Clay army description	emperor soldier terracotta warrior hors army tomb chariot life china face
Impressive museum	warrior museum site terracotta visit area pit impressive hour attract main
Main excavated pit of museum	warrior hall exhibit picture pit chariot display walk view good glass main
Busy place, crowds esp. mornings	crowd people time busy place holiday picture good site morning front
Getting there locally	station ticket train yuan entrance bus hour taxi stop cost museum guide
Amazing life-size warrior army	warrior size amazing detail terracotta scale mind sheer incredible life army
Great experience worth a travel	warrior china terracotta amazing trip great experience visit worth highlight
1 – 3 star reviews	**Tokens**
Disorganized place	clear lack background organize admission work constant empire commission
Do not trust local bus operators	operator problem terracotta cash town bus larger green term resident
Hope for future improvement	nice expect complex quality hangar future opinion actual type unexcavated
Multiple criticisms[1]	visitor disappoint moment camera vendor fact actual fake room massive
Fantastic place, but[2] …	figurine queue hawker hanger extra Quin theatre rude box pottery fantastic
Multiple criticisms [3]	guide museum tour terracotta Xian information shop army driver soldier
How to get there. Overrated.	walk ticket station train yuan cost park number sign price stop wall
Multiple criticisms[4]	hall color weapon mausoleum task film dynasty example shot cool trap
No discount for foreign students	student charge government store price country factory replica pick amount
Multiple criticisms[5]	preserve light famous control commerce able safety extreme advance pity
Multiple criticisms[6]	warrior site china people terracotta picture visit tourist crowd place pit time
Multiple criticisms[7]	cart ride golf highlight video seller floor movie course corridor plan extra
Multiple criticisms[8]	wall world total book guess food easy Shanghai other wonder skeptic
Local businesses[9]	warehouse vast copies giant awesome event card heritage hold bronze

1: Sketchy story, likely fake, waste of time, noise, crowds.
2: Crowded, busy, rude people, pushing, lack of organization, no toilet paper, fake, etc.
3: Most information in Chinese, too much information from guides, terrible guides, overrated, long drive.
4: Probably fake, crowded, poor documentary movie, best colored figurines are missing, etc.
5: Uncontrolled local commerce, seedy street from the station, bad display inside, crowds, rude tourists
6: Distance to figurines, photos misrepresent the site (figurines not colored, site size overstated etc.), crowds, local tourists cutting into lines, etc.
7: Poor management: gold cart rides to museum, information video, missing plans, dump, hot, crowds.
8: Crowds, pickpockets, little to see, broken movie theater, local commerce, fake. Tiring trip from city.
9: Rude and overpriced local businesses, real/fake farmer who discovered the site selling overpriced autographed books, expectations to pay for everything. Many of displayed artefacts are copies.

Table 2B. Topics extracted from the Chinese sample. The tokens represent the highest ranking words.

4 – 5 star reviews	Tokens
Description of the figurines	figurine color unearth face thousand expressive look amazing life
Guided tour: Pros and cons	tour guide people good time price ticket group student look money talk
Museum and excavation pits	museum hall exhibit archeology hole amazing walk site pit travel antique
Tourism environment and services	good area attract beautiful scenic tourism environment place spot regret
Pride, 8th wonder of the world	people ancient china world wonder great history wisdom culture eighth
Emperor's tomb, historic place	history tomb soldier sense emperor general imperial front king face
Descriptions of terracotta warriors	terracotta warrior horse spectacular feel time look museum figurine
Visit w/friends, children; foreigners	time foreign summer school travel love child friend door experience
Shock and emotions	feel people history ancient dynasty shock terracotta place momentum kind
How to get there	station railway train ticket hour yuan convenient line direct area morning
Worth a visit	visit worth attract history tourist feel good look place foreign interest
1 – 3 star reviews	**Tokens**
Multiple criticisms[1]	admire scenery play dirt distance wave ancient native book similar
Expensive, nothing special[2]	price terracotta face loess visitor help communist fine pile okay work
Multiple criticisms[3]	scenic care capital driver phenomenon travel carriage copper period
Great place but something is not right	history people china ancient great place culture worth world foreign site
Multiple criticisms[4]	rain weapon guide hand vehicle hour result Xiang went environment
Spectacular, but hard to see, costly	terracotta warrior horse look feel time spectacular figurine people attract
Problems with accessibility	station tourist convenient individual train store railway sale tour dollar
Spectacular, but undeveloped	exhibit technology undeveloped mausoleum spectacular spring vote area
Management problems	lack management effect protect visitor majestic downtown software
Tour dissatisfaction	good guide tour visit regret people feel ticket expensive place museum
Tickets: lines, price, scalpers	ticket time office free dollar scalper local hole wood obvious opportunity

1: The documents loading to this topic while expressing some admiration also contain multiple criticism points, including: just ancillary services, many wild tour guides, soliciting, toilets are flawed, during the New Year there are many people, but only half of the site open, boring, crowded, violent people, place probably fake, just a dirt pit, and others.
2: An example of a loading review: the place is ok, but I just go for a local craft shop.
3: Touching on all issues but in various ways: wanted to see copper carriage but was pushed away.
4: Expensive tickets, rain, site destroyed by Xiang [Yu rebellion] - all sorts of complaints

Discussion

The results of LDA analysis showed that 4 – 5 star topics are easily interpreted with the visitors commenting on the historical significance of the place, a feeling of amazement observing huge clay army, logistics of the travel, and local services. In addition, many Chinese visitors commented on the feelings of pride for their history and people.

Meanwhile, the majority of the dissatisfaction topics in the 1 – 3 star reviews were left unidentified in the LDA analysis and required additional consulting with the original reviews (Tables 2A and 2B). Note that the objective measures of topic coherence (Figure 1) did not show significant differences in the coherence of the topics arising from positive and negative reviews, suggesting the LDA topical modeling was similarly successful for either of four datasets. To confirm that this effect is due to the nature of negative reviews, as compared to the positive reviews, and not sample size differences, we randomly selected two samples from 4 - 5 star en and zh datasets with sizes matching the 1 - 3 star samples. We run LDA analyses on those samples and found the topics arising from those samples interpretable and similar to those in the full datasets.

Overall, we confirmed our initial hypothesis that the results of the LDA analysis on the low customer satisfaction datasets are significantly less interpretable compared with the analysis of the positive reviews. We suggest that this effect is due to the observed diversity of negative opinions leading to a much greater diversity of the negative comments as compared with the positive ones. More research is required to confirm this result. If, however, confirmed, it would mean that the Anna Karenina effect limits application of the automated topic modeling to the analysis of the main topics of customer dissatisfaction to very large datasets where the sheer volume of reviews would warrant keeping much greater number of topics arising from the LDA analysis.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3(1), 993–1022.

Borko, H. & Bernick, M. D. (1962). *Automatic Document Classification*. System Development Corp., Santa Monica, California.

Tasci, A. D. A., Croes, R., & Villanueva, J. B. (2014). Rise and Fall of Community-Based Tourism–Facilitators, Inhibitors and Outcomes. *Worldwide Hospitality and Tourism Themes*, 6 (3), 261–276.

Diamond, J. (1997). *Guns, Germs, and Steel: The Fates of Human Societies*. NY: WW Norton & Company.

Guo, Y., Barnes, S. J., & Jia, O. (2017). Mining Meaning from Online Ratings and Reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation. *Tourism Management*, 59, 467–483.

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated Sentiment Analysis in Tourism: Comparison of Approaches. *Journal of Travel Research*, 57 (8), 1012–1025.

Putri, I. R. & Kusumaningrum, R. (2017). Latent Dirichlet Allocation (LDA) for Sentiment Analysis toward Tourism Review in Indonesia. *Journal of Physics: Conference Series*, 801:012073.

Shugan, S. M. (2007). The Anna Karenina Bias: Which Variables to Observe? *Marketing Science*, 26(2), 145–148.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. *International Conference on Machine Learning*, 190–198.

Tianzhu, Z. (2019). Record Number of Tourists Visit Terracotta Warriors in New Year's Day Holiday. *People's Daily Online*. Accessed September 12, 2019. http://en.people.cn/n3/2016/0104/c98649-8998790.html.

Vicient, C. & Moreno, A. 2015. Unsupervised topic discovery in micro-blogging networks. *Expert Systems with Applications*, 42(17-18), 6472–6485.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A Comparative Analysis of Major Online Review Platforms: Implications for Social Media Analytics in Hospitality and Tourism. *Tourism Management*, 58, 51–65.