Tourist Attraction Sentiment Analysis with TripAdvisor Reviews

Yu-Ting Chiu
Dept. of Computer Science,
New York University
New York City, USA

Jane Liu
Dept. of Computer Science,
New York University
New York City, USA

***Abstract*—The travel industry has seen significant changes over the past decade as travelers have become more sophisticated and have begun to demand more unique, one-of-a-kind experiences. Increasingly, travel consumers refer to online user reviews to make decisions about where to go, where to stay, and where to eat. The rapid growth of web 2.0 applications has generated an enormous number of online user-generated content (UGC) for travel-related information. An intrinsic problem of the internet is overwhelming amounts of data. While search engines can help users identify the reviews they need, search results still contain information beyond the visual capacity of human beings. For our research, we explored the application of supervised machine learning together with a character-based N-gram model for sentiment classification of online travel reviews. This predictive model can be applied to travel reviews to generate a sentiment score and rank tourist attractions based on the polarity of their score (from 1.0 to 5.0). Scores for various attractions can be compared to help travel industry professionals make the best recommendations to their customers.**

***Keywords*—tourist attractions, sightseeing, unique travel.**

I. Introduction

With over 1 billion international tourist arrivals worldwide in 2018, international tourism has recovered strongly from the economic downturn of the last decade. In 2011 the United Nations World Tourism Organization predicted that international arrivals will reach 1.4 billion by 2020, but this number has already been surpassed in 2018.[1] As global travelers become more sophisticated and demand more unique experiences, travel agents, tour operators, hotel concierges, and other travel industry professionals will need to stay competitive by creating unique, compelling itineraries for their guests. Sentiment analysis provides an excellent opportunity to analyze large quantities of data from travel websites, blogs, social media, and other sources to identify the best, most offbeat, and unique tourist attractions that will appeal to increasingly sophisticated globetrotters.

The importance of word-of-mouth in influencing consumer choice is well-documented in the literature and increasingly word-of-mouth is being replaced by online user-generated content (UGC).[2] User-generated content has become widely used by individuals and organizations to make purchase and business decisions. Positive opinions can result in significant financial gains and fame for businesses and individuals. The rapid growth of web 2.0 applications has generated an enormous number of online content about travel destinations, sightseeing attractions, hotels, and other travel services. Travelers, like other consumers, increasingly depend on online word-of-mouth recommendations in traveler reviews, blogs, and social media posts to make travel purchase decisions.

Aside from the benefits to travelers online word-of-mouth can have important implications for managers to consider their brand building, product development, and quality assurance. In brief, the internet can serve as a useful platform for personal communications for sharing information about ownership, particularly goods and services, the suppliers, and the place of supply.[3]

Because user-generated content has increasingly replaced word-of-mouth recommendations there is an urgent need for new techniques that can automatically analyze attitudes of customers in their reviews. Sentiment analysis (or opinion mining) can perform the task of extracting useful information from online reviews automatically. To assist with the current effort to develop new techniques for analyzing traveler sentiment we developed a predictive model that can analyze website reviews for a tourist attraction and rank those attractions based the overall polarity of the reviews (positive,

[1] (2019, January 21). International Tourist Arrivals Reach 1.4 billion Two Years Ahead of .... Retrieved August 11, 2019, from http://www2.unwto.org/press-release/2019-01-21/international-tourist-arrivals-reach-14-billion-two-years-ahead-forecasts

[2] (n.d.). Talk of the Network: A Complex Systems Look at the Underlying .... Retrieved August 11, 2019, from https://link.springer.com/article/10.1023/A:1011122126881

[3] (n.d.). The Digitization of Word of Mouth: Promise and Challenges of Online .... Retrieved August 12, 2019, from https://pdfs.semanticscholar.org/a920/066b9dfa968625f76e8738e1b9a19a52bf23.pdf

negative, or neutral). It is our hope that our predictive model can suggest interesting alternative destinations for seasoned travelers looking for experiences outside of the mainstream. To that effect, our model can be used by travel-related businesses to craft a more unique, one-of-a-kind experience for their guests.

II. Motivation

Word-of-mouth is generally defined as the communication between individuals regarding their perceptions of goods and services. The importance of word-of-mouth in influencing consumer purchase decisions has been well-documented in the existing literature. Since the development of web 2.0 online online travel reviews have become another type of word-of-mouth recommendation, providing information for tourists to make decisions about where and when they will travel, where to stay, and at which restaurants to dine.

Online search engines like Google and Bing offer millions of search results for travel-related keyword searches. As stated this amount of information is impossible for a human being to process; even a single travel blog, social media newsfeed, YouTube video, or a list restaurant reviews on Yelp can take days for a person to read, digest, and process. Nonetheless, many travelers search online for information about their preferred travel destinations before planning their trip. As the size of travel-related data on the web increases it becomes more important than ever to find ways to analyze that data and extract from it the information that we need.

Sentiment classification is the mining of written text reviews with the goal of identifying the polarity of its prevailing opinion (favorable or unfavorable).[4] It has been mainly applied to the computing fields of information retrieval and natural language processing. A model builder often has three sources of information available--a small collection of labeled documents, a large collection of unlabeled documents, and human understanding of language. Special challenges are associated with mining tourists reviews. Depending on context the semantics of a word can have good or bad meaning. For example, an "unpredictable airplane ride" implies a scary, unpleasant experience while an "unpredictable walking tour" could imply a tour with unexpected and exciting elements. Although researchers have started to investigate content analysis of travel-related websites, blogs, and tweets, more sophisticated web mining techniques need to be developed for tourist review analysis.

To contribute to the existing tourism literature and the need for new methods to analyze large quantities of travel-related data, our study attempts to perform automatic classifications of tourist attractions by opinion mining user reviews on TripAdvisor.com. Data from TripAdvisor was chosen due to the availability of pre-labeled data to train our supervised learning model (a different user rating ranging from 1 to 5 bubbles accompanies each review). This study constructs a predictive model by combining the natural language processing technique of term frequency-inverse document frequency (tf-idf) with a multinomial logistic regression algorithm to generate a ranked list of the best sightseeing attractions for three popular cities--London, UK, Paris, France and New York City, USA. Performance is measured by calculating the number of accurate predictions divided by the total number of predicted reviews.

III. Related Work

We present and contrast research paper summaries prepared by each author.

Sentiment Analysis of TripAdvisor Reviews

Web 2.0 has brought a variety of new travel-related content from blogs, social networks (e.g., Facebook, Twitter), collaboration tools (e.g., Dropbox, Google Docs), media content (e.g., YouTube, Flickr). This rapid growth in user-generated content indicates that customers often check other users' opinions before buying a product or service. Included in this growth is TripAdvisor, a website containing nearly a billion user reviews about destinations, hotels, restaurants, activites, and other travel-related services around the world.[5]

Many studies on TripAdvisor exist but there is currently no complete analysis of the sentiments expressed by users who use the TripAdvisor site. This paper attempts to develop a more thorough understanding of how user sentiments are matched to sentiment-detection algorithms on TripAdvisor and discusses the challenges of analyzing user sentiment.

Sentiment analysis is performed using classification and is divided into three steps–polarity detection (label the sentiment of the text as positive, negative, or neutral), aspect selection/extraction (obtain the features for structuring the text), and classification (apply machine learning or lexicon approaches to classify the text).

Some popular approaches include aspect-based sentiment analysis, where attributes of a business, such as cleanliness, staff behavior, or service at a restaurant, are assessed for positive and negative emotions. The opinions expressed by customers about these attributes are analyzed to determine how customers feel about that business. Another approach is to ascertain irony and sarcasm in user reviews, which has remained an ongoing and intractable problem. Being able to

[4] (n.d.). The Sentimental Factor: Improving Review Classification via Human .... Retrieved August 11, 2019, from https://web.stanford.edu/~hastie/Papers/ACL2004.pdf

[5] (n.d.). Sentiment Analysis in TripAdvisor - IEEE Computer Society. Retrieved August 12, 2019, from http://www.computer.org/csdl/mags/ex/2017/04/mex2017040072-abs.html

determine irony and sarcasm will undoubtedly help extract more valuable information from TripAdvisor reviews.

Detection of Fraudulent TripAdvisor Reviews

Consumer-created reviews of products and services are a critical driver of everyday decision-making. However, the credibility of these reviews is damaged when businesses commit review fraud, either creating fake positive reviews for themselves (boosting) or negative reviews for their competitors (vandalizing). So the authors analyzed millions of hotel reviews on TripAdvisor, where no verification of an actual hotel stay is made before a review is posted Agoda and Booking.com, where only reviewers who booked through their website can review properties. The paper tries to identify patterns and indications of fake reviews.[6]

They stated that there are three primary types of information related to a hotel review: (1) the review content, (2) the reviewer who wrote the review, and (3) the product (in this case a hotel room) being reviewed. So they analyzed the reviews form these 3 aspects.

They found that independent hotels and hotel chains comprised of primarily franchises are more likely to display signs of review fraud than corporate-run chains. Properties that had a large number of reviewers who contributed only a single TripAdvisor review are more likely to leave reviews that are shorter and use fewer words.

TripAdvisor reviews contained different term frequency and different sentiment than Agoda/Booking.com reviews. When TripAdvisor reviewers of our 103 suspected properties are examined separately, the language used was markedly different from reviewers who left TripAdvisor reviews on properties we did not find suspicious.

Learn the method and tool and used to perform analyzation on Tripadvisor reviews. Their finding of what does fake reviews look like? What kind of hotels tend to have more fake reviews? What are important aspects to analyze a review?

Comparison of the Summaries

Both summaries discuss methods of detecting characteristics of TripAdvisor reviews but differ in what is being investigated. The first summary examines the current state of sentiment analysis for TripAdvisor reviews while the second summary examines techniques for the detection of fraudulent user reviews. Both sentiment analysis and spam review detection use similar techniques in NLP and machine learning. Fraudsters sometimes leave semantic patterns in their reviews that can be detected using methods that are similar to detecting users' sentiments about products.

[6] (2019, April 25). Decomposing TripAdvisor: Detecting Potentially Fraudulent Hotel .... Retrieved August 12, 2019, from https://www.researchgate.net/publication/329956728_Decomposing_TripAdvisor_Detecting_Potentially_Fraudulent_Hotel_Reviews_in_the_Era_of_Big_Data

IV. Datasets

Because of the large number of tourist reviews (over 1 billion as of 2019) we chose to use reviews at the TripAdvisor website for our training and testing datasets. On TripAdvisor users can read the accumulated opinions of millions of tourists. They can also check the popularity index, which is computed using an algorithm that accounts for user reviews and other published sources such as guidebooks and newspaper articles. This index runs from number 1 to the overall total number of tourist attractions, restaurants, or hotels within the city. In addition the bubble rating (user rating), a 1–5 scale where one bubble represents a terrible experience and five bubbles an excellent experience. Because the number of bubbles given by a user sometimes differs from how they truly feel about a sightseeing attraction, restaurant, or hotel we propose a model to perform opinion mining on reviews to determine travelers' true feelings about a tourist attraction. We plan to offer this information to travel industry professionals to help them craft the best possible travel experience for their guests.

The data was obtained by web scraping user reviews from the TripAdvisor website for the cities of London, Paris, and New York. The Beautiful Soup library provided the tools for scraping. It is a Python library for pulling data out of HTML and XML files. It works with Python's html.parser to provide idiomatic ways of navigating, searching, and modifying the parse tree.

For each city we identified less well-known tourist attractions, i.e., rank 11 to rank 30, and obtained their user reviews. We also attempted to scrape reviews from top 10 attractions to be training data. However, changes to the TripAdvisor website introduced errors in the scraped data and Spark was unable to process those later datasets. As a result, we limited our training data to the original data that we were able to obtain.

Our dataset contains the following schema: *user name*, *review rating* (the number of bubbles given by the user), *review text*, and *date of review*. Although TripAdvisor review is realtime data, we decided to scrape only once because based on the number of data, we assumed that the newly coming data in a short time is not going to affect the positivity too much.

V. Description of Analytic

We used a combination of feature engineering, natural language processing, and supervised machine learning to build a model that could predict which less well-known tourist attractions would appeal to the largest number of travelers.

First, we created a balanced training set containing equal numbers of reviews for each category. For example, if there are *n* reviews with a 2-star label and 2-star reviews appear the least often in the entire corpus, the training set would have *n* number of reviews for every label. This helped address the problem with data skew, which had dramatically lowered our accuracy score during initial testing.

Feature engineering is the process of constructing or extracting features from data. In this section, we describe some commonly used feature extraction techniques for user review sentiment analysis. One of the most common types of features that can be extracted from reviews are the words found in the review's text. This is typically implemented using the bag-of-words approach, where features for each review consist of either individual words or small groups of words known as n-grams, found in the corpus.[7]

Natural Language Processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. An NLP technique that we utilized in our model is term frequency-inverse document frequency (tf-idf). The tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a corpus of documents. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. We present a mathematical explanation of the tf-idf algorithm.

The tf–idf is the product of two statistics, term frequency and inverse document frequency. There are various ways for determining the exact values of both statistics.

Term frequency measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length as a way of normalization:

TF(t) = (Number of times a term t appears in a document) / (Total number of terms in the document).

Inverse Document Frequency measures how important a term is. While computing *TF*, all terms are considered equally important. However it is known that certain terms, such as "tour", "bus", and "hotel", may appear many times but have little importance in the context of a travel-related review. Thus we need to weigh down the frequent terms while scaling up the rare ones, by computing the following:

IDF(t) = log_e (Total number of documents / Number of documents with term t in it)

For our project we applied tf-idf to the set of all user reviews for a tourist attraction. After cleaning the corpus by tokenizing and removing numbers, punctuation, and common stopwords, a document term matrix was constructed to hold each word of the corpus. Tf-idf was applied to the matrix to vectorize the features.

Supervised machine learning can be used to determine sentiment by looking at it as a classification problem where reviews are separated into multiple classes. In our project the reviews were separated into 1-star, 2-star, 3-star, 4-star, and 5-star reviews. A smaller number indicates a very negative review and a larger number indicates a very positive review.

For our model we chose *multinomial logistic regression*, a type of regression that returns the probability of the occurrence of an event by fitting the data to a mathematical function called a *logit function*. It is a member of a class of models called generalized linear models. Unlike linear regression, logistic regression can directly predict probabilities—specifically, values that are restricted to the (0,1) interval. The coefficients of the model can provide some hint of the relative importance of each input variable.

We begin with an explanation of the standard logistic function. The logistic function is a sigmoid function, which takes any real input , and outputs a value between zero and one; for the logit, this is interpreted as taking input log-odds and having output probability. The logistic function is defined as follows:

[7] (n.d.). Text mining and probabilistic language modeling for online review .... Retrieved August 12, 2019, from https://www.semanticscholar.org/paper/Text-mining-and-probabilistic-language-modeling-for-Lau-Liao/eaead47da83d850b2f58e0c60177db36e7cbc07b

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

A graph of the logistic function on the t-interval (−6, 6) is shown in Figure 2 below.

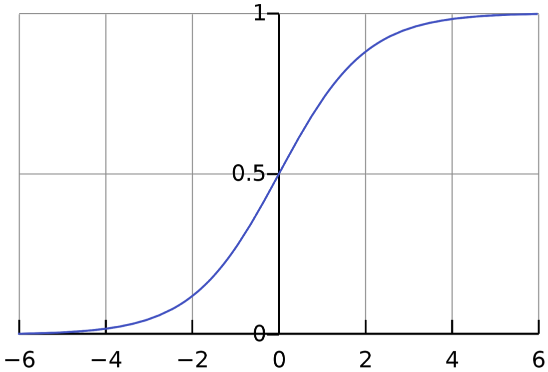


Figure 1. The standard logistic function $\sigma(t)$, where $\sigma(t) \in (0,\ 1)$ for all t.

Assume that t is a linear function of a single explanatory variable x. We can then express t as follows:

$$t = \beta_0 + \beta_1 x$$

The logistic function can now be written as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Note that is interpreted as the probability of the dependent variable equaling a “success” or “case” rather than a failure or non-case.

We can now define the inverse of the logistic function, , the logit (log odds):

$$g(p(x)) = \operatorname{logit} p(x) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x,$$

After exponentiating both sides:

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

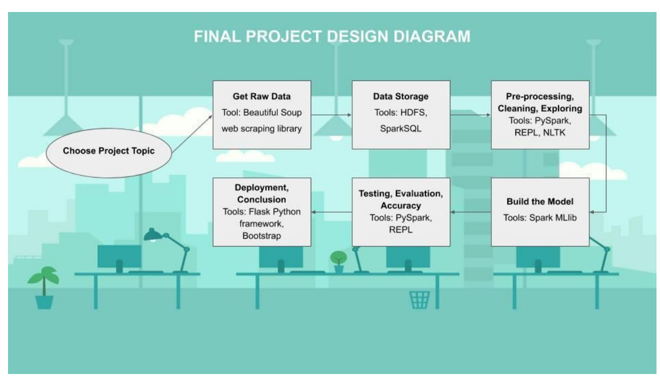
Model Implementation

To implement the multinomial logistic regression learning algorithm we used the PySpark MLlib package. We used the MLlib Pipeline functionality to pass a sequence of commands to calculate tf-idf, generate the label column (*StringIndexer*), and prepare the training set. Next, a predictive model was created using *pipelineFit* and used to train on the training set. The predictive model was then applied to the datasets of other tourist attractions to generate their predicted label values.

Interestingly, MLlib assigns a different set of labels to the star ratings every time *StringIndexer* is applied to a dataframe. It seems that star ratings that are least represented in the corpus are assigned the smallest label value, while star ratings that are over-represented in the corpus are assigned higher label values. This is a guess and needs to be researched further.

Due to constraints in the data size and challenges with scraping data from the TripAdvisor website, we used data from two cities to make predictions for the third city. For example, data from London and Paris were used to make predictions for New York City. The training data for the new city is trained again and a new predictive model is created. User reviews from each tourist attraction is cleaned, preprocessed, fitted with the pipeline and run through the predictive model, which outputs a dataframe with the predicted results. A user-defined *getScore* function is applied that identifies the label that MLlib has generated for each star rating (again, this varies every time the model is run) and obtains the sentiment score for the tourist attraction. All the sentiment scores are collected and output to a text file in descending order. This is peformed for each city and the output files are written to the local directory in the folder /website/output. Finally, the website displays the newly ranked tourist attractions.

VI. Application Design



VII. Actuation or Remediation

We gained deep insight into how to build machine learning applications in the Spark environment. After initial testing using binary classification with a logistic regression model we attempted to create a more accurate predictor using multinomial logistic regression. While the accuracy of the multiple classifier model varied according to the training set, we were able to achieve an accuracy score of 67% on predictions for New York City. Next, we will attempt to improve the accuracy by applying transfer learning.

Based on the set of sentiment scores output for each city we created a web application that displays a list of less well-known, offbeat tourist attractions ranked according to their score.

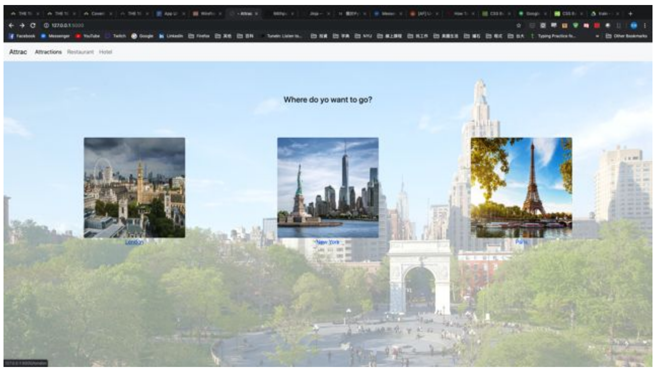


Figure 2. The landing page of the web application.

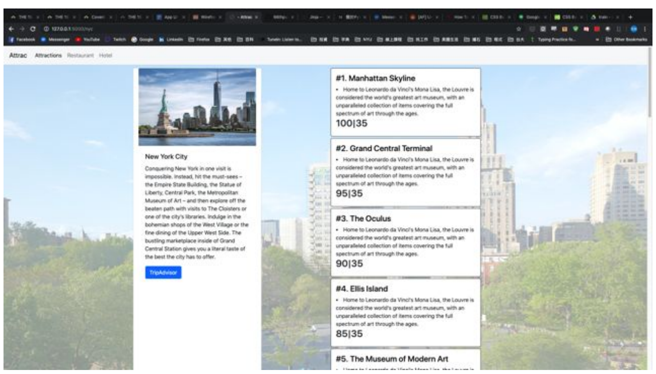


Figure 3. A ranked list of tourist attractions for New York City

VIII. Analysis

The sentiment scores generated from our model offered a number of surprises. For example, the Wallace Collection in London, which contains a large collection of English and French decorative arts from the 15th to the 19th centuries, received the highest sentiment score of all tourist attractions in London. For New York City, whose model had the highest accuracy score (67%), the Manhattan Skyline received the most positive sentiment (TripAdvisor had ranked it 15th on their list). This insight led us to conclude that tourists do not always enjoy visiting famous sightseeing attractions compared to less well-known attractions. Many go because the attraction is famous and is a "must-visit" location, but tourists do not always enjoy themselves during the visit. This can be confirmed from first-hand observations of tourist behavior at overcrowded locations like the Statue of Liberty. In contrast, the best time to view the Manhattan skyline is at night from across the Hudson or East River. At that time of day viewing locations are relatively empty and deserted and may offer some peace and quiet for a tourist after a busy day navigating crowded spaces.

IX. Conclusion

NLP and machine learning are powerful tools that can distill useful information from large amounts of data. The travel industry has the good fortunate of having access to vast amounts of useful information contained in travel blogs, travel-related websites, social media, and even YouTube videos filled with consumers' likes, dislikes, and other opinions about destinations, hotels, restaurants, tour operators, and more. There is a growing need to be able to extract useful information from all this data. We hope that our predictive model for tourist attractions has contributed to that effort.

X. Future Work

Anecdotally, logistic regression is an algorithm that requires significant amounts of data to achieve accurate results. The varying accuracy of our models may be due to the limited data that we were able to scrape from the TripAdvisor website. In the future we plan to train our predictive model on a larger dataset of user reviews, blog posts, and social media posts and apply them to more cities. Applying additional NLP techniques, such as lemmatization, N-grams, and parts-of-speech tagging (obtaining only adjectives and adverbs for sentiment) may also improve precision and accuracy. Other ideas include re-building the model with a different library, such as scikit-learn and applying transfer learning.

In the future we plan to generate a list of more obscure, rarely visited places to create a truly unique, one-of-a-kind list of tourist attractions for seasoned travelers.

Acknowledgment

We would like to thank the HPC support staff for their advice and recommendations and for keeping the servers running even when significant resources were requested. We would also like to thank Professor McIntosh for helping us troubleshoot some very difficult challenges when trying to get the application to run. And lastly, we would like to thank TripAdvisor for providing us with the data sets.

References

1. (2019, January 21). International Tourist Arrivals Reach 1.4 billion Two Years Ahead of Forecasts. Retrieved August 11, 2019, from http://www2.unwto.org/press-release/2019-01-21/international-tourist-arrivals-reach-14-billion-two-years-ahead-forecasts

2. Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters,*211-223. Retrieved August 11, 2019, from https://www.ama.org.
3. Dellarocas, C. N. (2003). The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback Mechanisms. *SSRN Electronic Journal*. doi:10.2139/ssrn.393042
4. Beineke, P., Hastie, T., & Vaithyanathan, S. (2004). The sentimental factor. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL 04*. doi:10.3115/1218955.1218989
5. Valdivia, A., Luzon, M. V., Herrera, F. (2017). Sentiment Analysis in TripAdvisor. IEEE Intelligent Systems, 32(4), 72-77. doi:10.1109/mis.2017.3121555
6. Christopher G. Harris. (2018). Decomposing TripAdvisor: Detecting Potentially Fraudulent Hotel Reviews in the Era of Big Data doi:10.1109/ICBK.2018.00040
7. Lau RY, Liao SY, Kwok RCW, Xu K, Xia Y, Li Y (2011) Text mining and probabilistic language modeling for online review spam detecting. ACM Trans Manage Inf Syst 2(4):1–30
8. Liu, B. (2013). *Web data mining: Exploring hyperlinks, contents, and usage data*. Berlin: Springer.
9. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In P. Isabelle (Ed.), Proceeding of 2002 conference on empirical methods in natural language, Philadelphia, US (pp. 79–86). Association for Computational Linguistics.
10. Li, H., Ye, Q., Law, R. (2013). Determinants of Customer Satisfaction in the Hotel Industry: An Application of Online Review Analysis. Asia Pacific Journal of Tourism Research, 18(7), 784-802. doi:10.1080/10941665.2012.708351
11. Raimbault, T., Chareyron, G., Krzyzanowski-Guillot, C. (2014). Cognitive map of tourist behavior based on Tripadvisor. 2014 IEEE International Conference on Big Data (Big Data). doi:10.1109/bigdata.2014.7004492
12. Subramaniyaswamy, V., Logesh, R., Vijayakumar, V., Keerthana, K., Rakini, K., Swarnamalya, B. (2018). Dynamo: Dynamic Multimodal Route and Travel Recommendation System. 2018 International Conference on Recent Trends in Advance Computing (ICRTAC). doi:10.1109/icrtac.2018.8679443
13. Ye, Q., Law, R., Gu, B., Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human Behavior, 27(2), 634-639. doi:10.1016/j.chb.2010.04.014
14. Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, Veselin Stoyanov. (2016). Sentiment Analysis in Twitter. doi:10.18653/v1/S16-1001
15. Snehasish Banerjee, and Alton Y. K. Chua. (2016). In search of patterns among travellers' hotel ratings in TripAdvisor. doi: 10.1016/j.tourman.2015.09.020
16. Hee “Andy” Lee , Rob Law & Jamie Murphy (2011) Helpful Reviewers in TripAdvisor, an Online Travel Community, Journal of Travel & Tourism Marketing, 28:7, 675-688, doi: 10.1080/10548408.2011.611739