Journal of Engineering and Applied Sciences 12 (24): 7529-7533, 2017
ISSN: 1816-949X

Automatic Rating for Services of Tourism Industry by using Opinion Mining

[1]Rupesh Kumar Mishra, [1]Meghana Bhradwaj, [2]A.K. Awasthi, [3]Rafael Berlanga Llavori and [4]Kannan Srinathan
[1]Department of Computer Science and Engineering,
[2]Departments of Mathematics, Manav Rachna University, Faridabad, India
[3]Department of Computer Science and Engineering, Universitat of Jaume I, Castellon, Spain
[4]Department of Computer Science and Engineering, International Institute of Technology, Hyderabad, Andhra Pradesh, India

Abstract: In this study, we have proposed a platform for extracting and summarizing users opinions about the services offered by the tourism industry. Perspectives extracted from the public generated content regarding aspects specific to services provided at various tourist spots and adventure spots are useful to both people who want to visit that place as well as the tourism industry to help them in improving their services. Here, using Naive Bayes classifier is applied to classify the data of a tourism industry based upon either positive and negative Tweets and posts on the given social media. In that approach the interconnection between of the post and public opinion has successfully managed by using the score of that post has to exist in the opinion or not. Also defined the threshold for the given Tweets or post on social media and we cannot take those post or Tweets which of score less than the measure threshold. The proposed system uses a hybrid approach mixing lexical and supervised learning methods. Three types of data we have taken for the experiment for this problem are first all the online news related to tourism data and non tourism data and its public opinion, second one the facebook post of public opinion and the third one is taken a Twitter data and TripAdvisor datasets.

Key words: Opinion mining, tourism industry, lexical analysis, Naive Bayes, sentiwordnet, threshold

INTRODUCTION

Tourism industry is one of the world's largest private commercial sectors whose advancement, economic importance and potential are phenomenal across the globe. Due to the rapid growth in tourism people who travel to different places, want to gather information regarding that particular place. Also, people usually share their experience about the places they visit on social networks and specialized opinion web sites. As a consequence, travelers check the opinions published by other travellers opinions on separate web platforms before planning their own vacation. However, while trying to analyze these opinions travelers usually get confused because of the vast amount of reviews and mixed opinions. Therefore, there is a requirements of the automatic machine that will be retrieve different opinions from this complex information, summarize the reviews and offer an overall perspective about how the services provided at a particular place are taken by the visitors to that place. This will not only ease the decision process of the people who wish to visit the place but will also help the tourism industry to know what services are liked or disliked by tourists. Therefore, helping them in improving their services.

In this research, the opinions and reviews posted by the travelers on social websites and tourism websites were collected. Then, a lexical-based method was used to pre-process and score social data which are in turn processed by a Naive Bayes learning algorithm. The final result is an accurate assessment of the different services according to the users perspectives.

Requirement of automatic system for the ranking of tourism industry services: Public based evaluation for determining the efficiency of the reviews with respect to the services provided by the tourism industry is a complex and durable techniques. Therefore, to determine the efficacy of the reviews and its consequent impact on the tourism industry as well as on its functional stakeholders such as airlines, travel intermediaries and hotel industry which are expected to gain in the whole process. This problem statement focuses upon automatically process development which determines on the influence of reviews posted by the tourists on social websites.

Corresponding Author: Rupesh Kumar Mishra, Department of Computer Science and Engineering, Manav Rachna University, Faridabad, India

Steps to following the appropriate approach: In this research tourism related data was collected from Twitter, using its API and also from TripAdvisor website. These data were then manually labeled into two datasets, positive and negative datasets, according to their impact in the service reputation. Then, a lexical-based approach is used in the initial stage which included pre-processing and data scoring. Scoring was done with the help of AFINN word list. Naive Bayes algorithm is then applied on this data in order to improve its performance by predicting the final assessment. The system finally generates a graphical representation related to the accuracy of the training data and also its effectiveness. That will minimize the time of decision for every tourist and also it will give impact upon the tourism industry. Also, public can show his/her opinions about tourism either like or disliked. Definitely improv their services which will enhance the growth and development of the tourism industry.

Our contribution: In this research, a hybrid approach has been used to find out the accuracy and effectiveness of tourism services based on the reviews and opinions given by the tourists who have visited these places and utilized the services provided there. The approach is divided into two modules: the data collection module which collects data from different web sources and the classification module which processes data and implements supervised technique.

Literature review: Akehurst (2009) presented the work on how the User Generated Content (UGC) can help to inform tourism and marketing managers, educate tourism consumers and facilitate tourism transactions. The conclusion being that if an automated tracker system is successfully completed, perhaps utilizing artificial intelligence, tourism organizations, enterprises and tourists will have a potentially powerful new weapon in their decision-making armories.

Mishra and Saini (2014) in their research presented the insights into opinion mining at different levels. The precise method for predicting opinions enable us, to extract sentiments from the web and foretell online customer's preferences which could prove valuable for marketing research.

Bucur (2014) proposed a platform for extracting and summarizing of opinions expressed by users in tourism related online platforms. The system could extract opinions from user generated reviews, regarding aspects specific to hotel services which proved to be useful both to clients looking for accommodation and also hotels trying to improve their services.

Sharma *et al.* (2014) proposed a document based opinion mining system that classified the documents as positive, negative and neutral. Experimental results using reviews of movies showed the effectiveness of the system.

Mishra and Jha (2012) proposed sentimental analysis based social issued through social media and do it with opinion mining. Experimental result we can see through the table which we have displayed in the resultant part. Kim *et al.* (2009) an unsupervised approach to domain-specific term extraction. In Australasian Language Technology Association Workshop 2009.

MATERIALS AND METHODS

Algorithmic approach for linking tourism industry related threads: In this part Mishra *et al.* (2016), we are discussing about the linking for tourism algorithm which will be related to various interconnected information based on the tourism and non tourism data based on feedback of persons. In the last, we will analyze the impact of the tourism comments based on the tourism industry to predefined evaluation criteria. There are following the steps:

- Select the social media based post and comments related to tourism industry
- Linking of all different tourism industry related post and comments and non tourism post and comments

We are applying the following approach (Mishra *et al.*, 2016) to identify and matching words related to all the above discussed elements:

$$\text{domain\_score}_{imp}(W_i) = \text{tfidf}(w_i) \times \text{pol\_score}(w_i) \quad (1)$$

Here, $\text{domain\_score}_{imp}(w_i)$ is the domain wise valuable score of the word W_i. Here, $\text{tfidf}(w_i)$ is the tf-idf score for different words of the tourism industry but here we are using the term frequency of tourism data and idf is of non tourism data in which we have calculated the score for that data to tourism sentiment based on tourism data. Because we have taken data for both the tourism industry and non tourism industry because if we have taken that tf-idf for the same domain than in this noisy data will be contained in the collected purified data. We need to remove all the noisy data and contained those data which will be useful for our problem area, so, we are calculating both tourism and non tourism data here. This will provide the accuracy of the tf-idf because the term frequency is for the same domain but inverse domain frequency is taken for different domains. pol_score_i is the highest polarity score of words of the taken word W_i.

According to Kim *et al.* (2009), given tf-idf is different than the traditional tf-idf, because this will give the specific domain oriented result for the tourism data. Here, term frequency will count from the tourism domain and idf will be taken from non tourism domain data. We have randomly selected 15000 Wikipedia document which are not related to tourism or any historical place or person. The entire idf is calculated on the same 15000 Wikipedia documents sets. The tf-idf score of different words are calculated by using the this equation in which term frequency will be of same tourism related data and idf is for non tourism data which is taken online or from the Twitter data:

$$\text{tfidf}(w_i) = \text{tf}(w_i) \times \log_2\left(\frac{n}{\text{idf\_count}(w_i)}\right) \quad (2)$$

Where:

n = 15000 (Wikipedia document sets)

idf_count(w_i) = Number of Wikipedia document which contain w_i. This is unsupervised way to give higher preference to domain specific term (Kim *et al.*, 2009)

tf(w_i) = The number of repetition of word w_i in the tourism domain data

This is enhancement of previous (Mishra *et al.*, 2016) way of calculation of sentimental weight of term. Calculate the impactful work based upon the given different opinions by different persons.

Now, we have to find out sentimental user's feedback and all the negative polarity score from online data and Twitter or Facebook data for the tourism industry and non tourism industry. The goal of this is to identify different sources through the given Eq. 3 and 4:

$$T_{pos\_Tfeed.com} = \text{mod}\left(\sum_{j=1}^{k} \text{domain\_score}_{imp}(w_j)/Tp\right) \quad (3)$$

$$T_{neg\_Tfeed.com} = \text{mod}\left(\sum_{j=1}^{k} \text{domain\_score}_{imp}(w_j)/Tr\right) \quad (4)$$

Here, $T_{pos\_Tfeed.com}$ and $T_{neg\_Tfeed.com}$ are the total positive and negative feedback after taking modulus of the total domain score which is calculated for tourism domain and non tourism domain scores.. 'Tp' and 'Tr' is the total number of tourism based positive and negative sentiment polarity score words for non tourism domain data.

Here, the Naive Bayes approach has been applied to classify and interlinked between public feedback and his comments related to the tourism industry:

P(tourist_feed.com|sent) = P(sent|tourist_feed.com)
*P(tourist_feed.com) / p(sent))

Where:

$$P(\text{tourist\_feed.com}) = T_{pos\_Tfeed-com} + T_{neg\_Tfeed-com}$$

Here, P(tourist_feed.com/sent) and P(sent|tourist_feed.com) are posterior and prior probability for all the positive as well as negative comments which is related to tourisn based feedback and comments from different persons.

The baseline format will be given by the public comments and feedback with respect to tourism domain related feedback and non tourism domain related feedback and comments to find out the effectiveness of the tourism industry related post as well as to determine how efficiently the effort has been applied in giving the rating on tourism industry. In Naive based approach, we have to find out the Euclidean distance between feedback (P) and comment (Q):

$$ED = \sqrt{((\min(\text{neg\_feed}) - \min(\text{pos\_feed}))^2 - (\max(\text{neg}_c\text{om} - \text{pos}_c\text{om})^2)}$$

Here, the perpous of Euclidean distance is that which post is closer than the other post and the distance is showing and calculating the accuracy between the post and comments. If the distance will be more closer then we can analyze the data up to 85% which would be considered a good score and based upon that it will give the best rating on the tourism industry.

Pseudo code of the tourism linked thread

Input: The input is a set of labeled training and testing data. The data is automatically collected from social websites (Twitter) and also manually from tourism websites (TripAdvisor). The dataset contains positive and negative reviews given by the users about the tourist spots they visited and other services provided by the tourism industry.

Output: The output generated displays how many reviews are positive and how many are negative. It also displays how accurate the result is. A confusion matrix displays the performance of the classification algorithm.

Steps: We divide the entire system into two process.

Data collection process: Data is collected from sites such as Twitter, using the Twitter API and manually from tourism website such as TripAdvisor. The data from these sources is then combined, into a single file and labelled to form a training data and test data.

Classification process: Pre-processing and input cleaning is applied. In this, the data is first split into sentences. Then with the help of AFINN wordlist the sentence contents is analysed. AFINN wordlist which contains about 2475 words and phrases which are rated from very positive (5) to very negative (-5).

Some more words related to tourism were also added to the wordlist. Opinion mining is performed on each sentence. Each sentence is split into component words through a tokenization process. With AFINNN wordlist the words are rated as very positive, positive, very negative and negative. In order to classify the sentences into positive and negative Naive Bayes algorithm is implemented (ignoring the manually assigned sentiment).

RESULTS AND DISCUSSION

Evaluation: The method has been applied on a dataset of 2475 Tweets per target place and online tourism, we sites which were automatically captured through the Twitter API. Feedback and comments were captured from facebook and online tourism web pages related to the different services related to the tourism industry.

Table 1: Shows the feedback of the travelers who visited the place

Places	Feedback	Scores
Bangkok	Thread 1	0.96
Rajasthan	Thread 2	0.87
Nanital	Thread 3	0.90

Table 2: Suggestions or comments the traveler wants to add after visiting

Places	Suggestions	Scores
Bangkok	Dataset 1	0.96
Rajasthan	Dataset 2	0.87
Nanital	Dataset 3	0.90

Table 3: Threshold table

Effectiveness of social media	Workdone (%)	Threshold range
Worst	Below 25	F<0.25
Bad	25-45	0.25≤F<0.45
Average	45-65	0.45≤F<0.65
Good	65-85	0.65≤F<0.85
Best	85-100	F>0.85

Table 4: Interconnecting score of post and comment

Places	Comments/Suggestion	Scores
Bangkok	Dataset 1	0.96
Rajasthan	Dataset 2	0.87
Nanital	Dataset 3	0.90

Table 5: Evaluation results

Places/Evaluation matrix	Devised system	Baseline
Bangkok		
Precision	0.78	0.56
Recall	0.68	1.00
F-means	0.73	0.72
Rajasthan		
Precision	0.84	0.54
Recall	0.54	0.74
F-means	0.66	0.62
Nanital		
Precision	0.85	0.56
Recall	0.75	0.94
F-means	0.79	0.70

Table 1 and 2 show the feedback score according to the target place. After visiting one place, users express their experience through comments. Here, a thread represents the reviews scores about a given place.

Table 3 shows the relation between the posts scores and the given feedbacks. If some comments about particular Tweets or post are not related to the domain, they are directly rejected according to the post scores.

In Table 4, a threshold table has been defined in order to account for the efffectivenes of social media information according to the defined parameters over the lexical-based scores. From this Table 4, we define the Baseline with the decisions taken by the lexical-based component with F>0.2.

Finally, Table 5 shows the precision, recalland F-measure of the post/tweets according to the manual feedback. The devised system corresponds to our proposal which combines lexical and Naive Bayes components. Results shows improvements on the three evaluated target places.

CONCLUSION

The given study have presented for an opinion mining technique to extracting and classify the reviews and opinions related to tourist locations and provided good services by the tourism industry. The system used a hybrid approach in which lexical approach and supervised learning technique were used. According to Bucur (2014), the given approach has an acceptable accuracy and has the good advantages that are domain independent and does not need expensive resources to operate. After analysis of the review, it can be concluded that in the domain of tourism an aspect oriented analysis would improve the performance of the platform, due to the multitude of aspects of users express opinions about and the mixed sentiments that are present in reviews.

RECOMMENDATIONS

A future direction for improving the performance could be the point of linked data (Llavori *et al*., 2015), oriented to tourism domain. The proposed architecture could be a very useful background tool for summarizing the opinions of tourism oriented web platforms. In

forthcoming work, we also plan to perform much larger experiments and perform a more comprehensive evaluation over alternative classifiers.

REFERENCES

Akehurst, G., 2009. User generated content: The use of blogs for tourism organizations and tourism consumers. Serv. Bus., 3: 51-61.

Bucur, C., 2014. Using opinion mining techniques in tourism. Proceedings of the 2nd Global Conference on Business, Economics, Management and Tourism, October 30-31, 2014, Elsevier, Prague, Czech Republic, ISBN:9781510809680, pp: 1-685.

Kim, S.N., T. Baldwin and K. Min-Yen, 2009. An unsupervised approach to domain-specific term extraction. Proceedings of the Workshop on Australasian Language Technology Association, December 3-4, 2009, University of New South Wales Sydney, Kensington, New South Wales, Australia, pp: 94-98.

Llavori, R.B., L.G. Moya, V. Nebot, M.J. Aramburu and I. Sanz *et al*., 2015. SLOD-BI: An open data infrastructure for enabling social business intelligence. Intl. J. Data Warehous E Data Min., 11: 1-28.

Mishra, N. and C.K. Jha, 2012. Classification of opining mining techniques. Intl. J. Comput. Appl., 56: 1-6.

Mishra, R., K. Saini and K.B. Sumreen, 2016. Social media based linking approach for finding the effectiveness of the social issues. Proceedings of the International Conference on Computing, Communication and Automation (ICCCA), April 29-30, 2016, IEEE, Noida, India, ISBN:978-1-5090-1667-9, pp: 7-12.

Mishra, R.K. and K. Saini, 2014. Automatic detection of interlinked events for better disaster management. Proceedings of the IEEE International Conference on Advance Computing (IACC), February 21-22, 2014, IEEE, Gurgaon, India, ISBN:978-1-4799-2573-5, pp: 595-600.

Sharma, R., S. Nigam and R. Jain, 2014. Opinion mining of movie reviews at document level. Intl. J. Inf. Theor., 3: 13-21.