Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 115 (2017) 563–571

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India

Aspect based Sentiment Oriented Summarization of Hotel Reviews

Nadeem Akhtar[a]*, Nashez Zubair[a], Abhishek Kumar[a], Tameem Ahmad[a]

[a]*Department of Computer Engineering, Zakir Husain College of Engineering & Technology, AMU, Aligarh 202002, India*

Abstract

Hotel booking websites use online ratings and customer feedback to help the customer's decision making process but reviews provide a better insight about the hotel but most travellers don't have the time or patience to read all reviews. This study analyzes the hotel reviews and gives information that ratings might overlook. The reviews and metadata are crawled from website and classified into predefined classes as per some of the common aspects. Then Topic modelling technique (LDA) is applied to identify hidden information and aspects, followed by sentiment analysis on classified sentences and summarization. Finally we discuss results and future work, ultimately building towards Hotel Recommender System.

Keywords: Review; sentiment analysis; topic modeling; classification; lda; MALLET; nltk;opinion mining; wordnet; sentiwordnet

1. Introduction

The Internet, in more ways than one has been a boon to many people around the world. The users tend to move towards convenient methods to find what they want and with the advent of the Internet, there has been an influx of web surfers searching for a good place to stay during their trip. The Hotel Industry, therefore is in a completely new phase now, and relies on the web for advertisement and publicity, for which there are tremendous websites available these days. Some of the prominent ones in India are TripAdvisor, GoIbibo, Yatra, etc. Each of these provide preview of what the user should expect from a certain hotel when he/she visits it or whether or not it would suit the budget.

* Corresponding author. Tel.: +91-9450658150.
E-mail address: nadeemalakhtar@gmail.com

10.1016/j.procs.2017.09.115

Mainly, what these sites rank the hotels according to are the ‘Ratings’ that some of the previous travelers provide from the feedback. Even though some of the websites might provide with ratings based on some aspect, but still research shows us that for proper customer satisfaction, interaction with service providers and other customers is necessary[1]. Also the reliability of these ratings is questionable [2]. For this, the reviews of users mentioned on the website are necessary and do provide a good insight into what the other customers experienced while visiting this particular hotel. Customers are more sensitive to personalized information that they find in the reviews and more often than not use it as a basis for decision making. Although the traditional review set is helpful but it fails when a customer’s interests fall outside of the information provided by the reviews[3].

The complications in this kind of study are plenty. There are a number of factors that make the data analysis difficult [4]. For example, the requirements of one user may vastly differ from the others; one may be on to a business trip, another on a vacation. Also the writing style of each reviewer is different from the other. Different aspects of the reviewer also affect his decision making, like their living standards, native place, lifestyle, budget, etc. Another important factor is the size of review set. Most of the customers are likely to read the first fifty or the most up voted reviews which may be missing out on some of the features described by the others.

This study analyses the data from hotel reviews and applies various different natural language processing techniques to reveal some important information that is apparently not visible to the viewer. Some predefined aspects were used to classify customer reviews into various categories and then ran over to topic modeling techniques which revealed some hidden topics over the raw classified data.

The study target for this was the website TripAdvisor (www.TripAdvisor.com) and since the dataset wasn’t freely available so a custom scrapper was built in Python to crawl customer reviews and their metadata.Discussion of a little bit of background is followed by delving into the brief concepts of text analytics. Then some related works are discussed that have been done in this field, building towards the proposed approach of handling the problem. A discussion of the experiments conducted and the received results follows, concluding towards the future scope of improvement.

2. Background

2.1. Modeling Customer Experience

Guest Experience and satisfaction is the backbone of the Hotel industry. It is a complex human experience within a hospitality setting. Customer satisfaction can be more commonly defined as interaction between his expectation and post purchase evaluation.

Modeling guest experience is a complex and challenging task because the demands and expectations of all users are not alike. Moreover depends on writing style of user. Although consumer surveys can be used to gather the data and it is very efficient but more often the customers are uninterested in writing about their experience so this method suffers from poor response rates[5].

2.2. Text Analytics

Due to unpredictable size of review sets and customer generated content, various text analysis techniques like sentiment analysis, opinion mining, topic modeling, classification, etc. play an important role. Classification is an important step and is a supervised method and is used to classify the text into various classes. In our study, we use it to classify the sentences of the text into predefined aspects that we already know are present in hotel reviews. Topic Modeling is a type of machine learning technique that can be used as a statistical model to discover the abstract topics in a collection of documents. It is a valuable tool to identify hidden semantic structures in the textual data and performs this function over our dataset[6].

Sentiment Analysis can be used easily for extracting opinions from the data about a certain aspect[7]. It is particularly useful for unstructured human authored documents and is a very important factor in business intelligence. It has become the central part of Information Retrieval process. The approaches towards short text summarization have improved sentiment analysis techniques [8].

3. Related Work

OpeNR[10] is an NLP platform applied to the hospitality domain to automatically process customer-generated reviews and extract important information from it. It consists of a set of OpenSource and free NLP tools to analyze text based on a modular architecture to simplify its modification and extension. It basically takes in social media generated content to perform text analysis on reviews[11].

Then there is system which is a part of the BESAHOT[16] project[15]. It is meant to be targeted at hotel customers who wish to know the information and actual overviews and summaries of the textual content about their hotels on the web. It handles only German reviews from German websites as of the time of this writing. It is based on the GWT framework and is an interactive web application. The core system handles data acquisition, analysis and storage while the user interface provides various types of summaries of the analyzed data. Data retrieved from the web by the acquisition system are checked by analysis system for language check to filter out reviews only in German. Then it segments the review text into sentences, such that to perform statistical polarity classification and linguistic information extraction on the text to get even better results of polarity values and the topics of the text. These values are then combined to get global polarity value and stored in a Result DB.

Other notable works include the Sumview system[13] although it is not for hotel but product reviews but provides a good insight into the review summarization model, the work by Hsiao-Wei Hu[9] that shows to summarize hotel reviews as per the aspect input provided by the user, personalized review and rating based summarization[6] emphasizes on integrating two topic models and use them to predict hotel ratings, etc.

4. Proposed Method

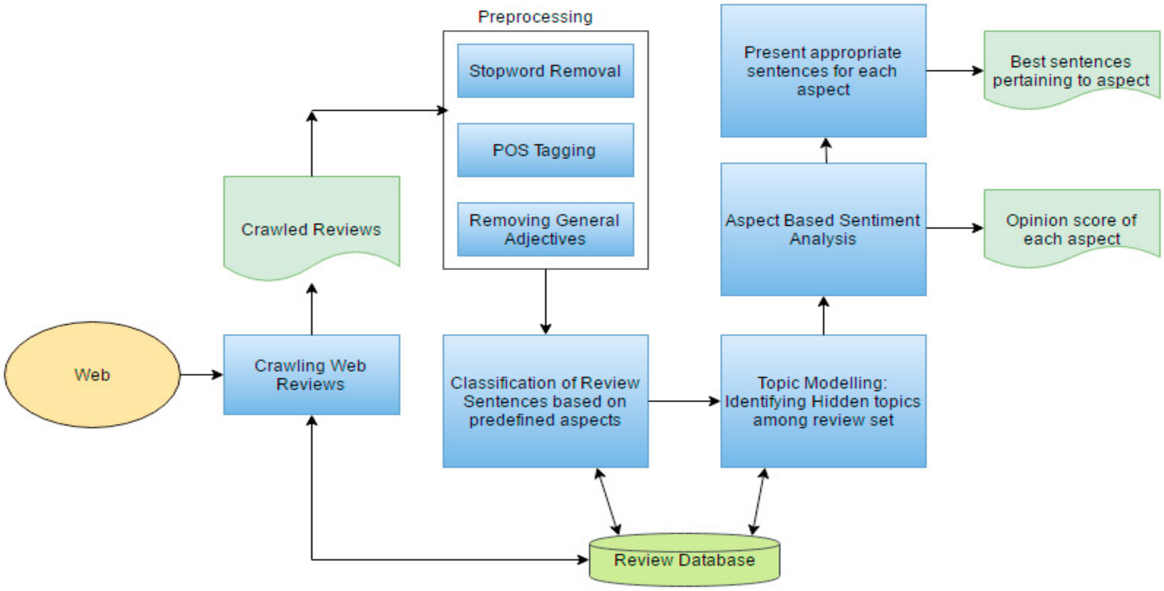

Preprocessing
Stopword Removal
POS Tagging
Removing General Adjectives
Present appropriate sentences for each aspect
Best sentences pertaining to aspect
Crawled Reviews
Aspect Based Sentiment Analysis
Opinion score of each aspect
Web
Crawling Web Reviews
Classification of Review Sentences based on predefined aspects
Topic Modelling: Identifying Hidden topics among review set
Review Database

Fig 1. Basic Underlying Architecture of our Review Processing system.

TripAdvisor website was chosen as the target for this study and analysis because it is one of the most popular hotel review website in the country and contains major data on most of the hotels. First a custom scrapper was built using python's libraries 'beautifulsoup' and 'urllib'. It takes as input the url of a hotel's webpage on the TripAdvisor site and then crawls the review one at a time. It creates two files per review, one that contains the review text only and the other that contains the review and metadata and assigns a unique id to each review file as per the review number crawled. These are placed into two different folders one that contains only review and the other that contains review + metadata. The metadata includes the following:

- Review Title
- Review date
- Reviewer’s Username on the TripAdvisor website
- Contribution Level of the reviewer
- Total Reviews
- Number of Hotel Reviews given by this user
- Travel Style of the reviewer if provided in the reviewer profile

The next step is classifying reviews into some of the predefined categories. These categories are nothing but the aspects that are frequently recurring in the review data set. Upon manual inspection of around 1200 reviews, and 5 hotels the following aspects and there commonly recurring words were defined.

Table 1. Manually Defined Aspects and their frequently occurring keywords

Value	Location	Service	Meal	Facility	Room	Quality	Staff	Surrounding
Price	Railway	Desk	Drink	Pool	Bed	Satisfactory	Good	Landmark
Amount	View	Check-in	Breakfast	Spa	Dirty	Ample	Polite	Monument
Rate	Station	Check-out	Spicy	Wi-fi	Clean	Hygienic	Helpful	Temple
Cheap	Airport	Reliable	Food	Gymnasium	Toilet	Proper	Friendly	Mosque
Worth	Distance	Fast	Tasty	Gym	Bathroom	Ambience	Reliable	Church
Low	Far	Convenient	Tea	Internet	Shower	Odour	Quick	Restaurant
Money	Close		Buffet	Ample	Dryer	Smell		Diner
Economical	Convenient		Bar	Parking	Fridge			Mall
Reasonable	Train		Restaurant	Wireless	View			Market
Fee	Metro		Dinner	Broken				
Expensive			Lunch					
			Brunch					
			Delicious					

Classifier.py (*Python program for classification of sentences under one of the above headings*)

```
1. Begin Procedure
2. Hardcode list of aspects and their keywords. And create the empty files with aspect names.
3. Repeat unless all files in the directory are read
     a. Repeat for each sentence in the review
          i. Repeat for all words in the sentence
               1. Compare the word with each and every keyword in aspect list
               2. Find out the similarity scores with the synset of these words
               3. Calculate average of similarity score of all aspect keywords
          ii Find out which of the aspects have the maximum similarity to the sentence
          iii Assign this sentence to the file of this aspect
4. End procedure
```

After running the classifier python script on the only review data set crawled (meta data excluded) ten different files are created that are named as per the aspect given in the *Table 1* and they contain the review sentences that are most similar to their aspect in the *txt* file format. For topic modeling, the MALLET tool is used. The number of topics is assumed to be *15* (The topics for LDA are not given because this is not labelled LDA).

After Topic Modelling, Sentiment Analysis is performed. For this feature’s implementation again the Senti-Word net corpus is utilised. The focus is on carrying out Sentiment Analysis on the topics that we have previously classified sentences on, i.e. the presented method reads the files created by the classifier and calculates its aggregate sentiment score to classify it as a positive or a negative aspect for the hotel. The most polarized sentences per aspect in this method are also presented. It works on the following algorithm:

Senti.py

```
1) Begin Procedure
2) Repeat for each file unless the all files are read
   a) Initialize two variables, one for total positive score and other for total negative score both to be 0.0
   b) Repeat for each sentence in the file unless all sentences of the file are read
      i) Tokenize each sentence
      ii) Create tagged element sets with each word
      iii) Keep only the words with common noun tags, adjectives tag and adverbs tag and discard all others.
      iv) Calculate the positive as well as negative score of this word and add it to the sentence.
      v) If the sentence contains a 'not' word or any negation word like that, interchange the positive and
         negative scores.
      vi) Calculate overall score of the sentence, positive or negative and store it along with sentences.
   c) Output and Write this aspect's positive score and negative score
   d) Display the three most positive, negative and neutral sentiment sentences.
   e) Calculate overall score of this aspect
 3) End Procedure.
```

5. Experiments, Results and Discussion

5.1 Tools Used

As the datasets were not available, a custom scrapper in Python 2.7 was built using two of Python's popular and freely available libraries, the 'BeautifulSoup' library and the 'urllib' library.For the textual processing this study turns to NLTK - Natural Language Tool Kit 3.0 which is a library of Python and contains many of text processing functions. Integrating it with other tools such as topic modeling tool MALLET or python's very own 'Gensim' library provides a very good framework to carry out the natural language processing tasks. Senti-Word Net corpus is also used to generate score of similarity using its synsets.

For topic modeling, the freely available tool called MALLET is used. It is a Java based tool for statistical natural language processing and includes various tools that perform complex classification, clustering, topic modeling, information extraction, and various other natural language processing tasks. It can also perform sequence tagging and numerical optimization on the textual data. This method uses it for topic modeling to identify various hidden topics and latent information that can be present in the reviews but is potentially overlooked by other techniques.

Sentiment Analysis has been carried out using Python and the SentiWord Net corpus and calculates opinion score of each sentence based on that aspect under which the sentence is classified.

5.2 Dataset

The procedure was run for a hotel *Orchid Residency*, the url for this hotel is:

https://www.TripAdvisor.in/Hotel_Review-g297634-d736092-Reviews-Orchid_Residency-Kottayam_Kerala.html

The data set that was crawled consists of 78 reviews. Each review consists of the following components:

- title of the review
- actual review text
- username of the reviewer
- date of review
- location of the reviewer
- contribution level of the reviewer
- total reviews reviewer has given
- helpful reviews, the ones that got up votes
- travel style

Then the reviews were separated from metadata to extract crude review text and put it to another directory.

5.3 Experiments

Term frequency analysis on the crawled reviews after removing the stop-words shows the following results:

We can see that ignoring the name of the city ‘*Kottayam*’ and hotel name ‘*Orchid*’ we get mostly those terms that we defined in our classifier.

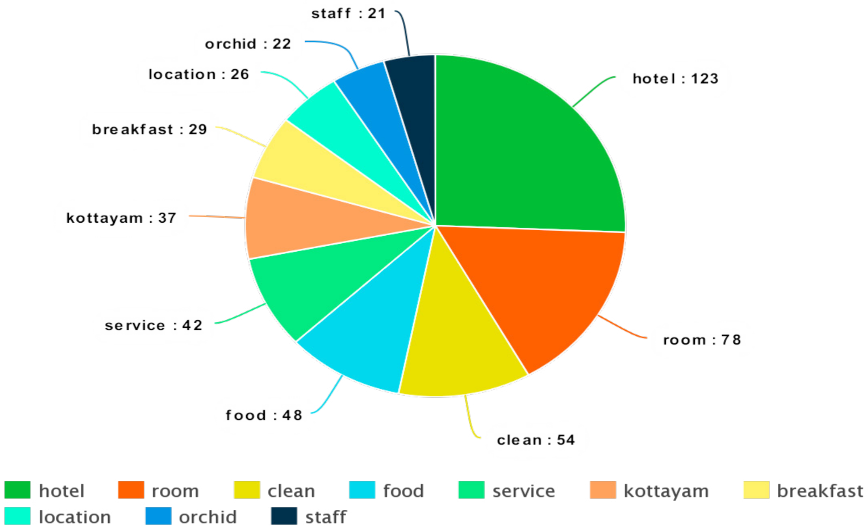


Fig 2. A word distribution of our dataset, showing most frequent 10 words.

Applying classification the following ten files are created:

value.txt	staff.txt	location.txt	meals.txt	quality.txt	facilities.txt
nhood.txt	room.txt	service.txt	others.txt		

Each of these ten files contains the sentences that are relevant to that aspect after which the file is named. Here the proposed method writes these files in such a way that each line contains one sentence that belongs to one of the reviews. One important point worth mentioning is the fact that each sentence will be put into only one of the files, otherwise in topic modeling we would be getting overlapping hidden topics but this decision has alternate consequences that will be discussed later. A snapshot of *meals.txt* looks like this:

```
chef delivers tastefully prepared food in good ambience
stayed in this hotel for 1 night in december while attending the wedding of a family member
staff were helpful and polite and the breakfast in the restaurant quite tasty too
the restaurant in aida serves very good food
the thali meals they offer is not that much appreciated but other foods are of good taste
my wife, daughter and myself stayed at aida in 2 deluxe rooms on 3rd and 4th jan 2016, as we
had a wedding to attend there food was very good
good budget hotel with clean rooms and good breakfast
hi,i have been to aida hotel many times hope this review will help the people who are planning
to kottayam and have a nice stay and wonderful food
i had been to aida most of the times when i am in kottayam(as i am from kottayam)
i like the food the ambience in this restaurant
 the food i tried was butter chicken with chicken fried rice
that was a super dish as the other items like beef chilly was also good but it was overcooked
so try butter chicken with chicken fried rice and a beer
 i had visited this hotel many times for lunch and last week also i went there for lunch
great food and nice ambiance guys at room service too were well mannered
i did receive 2 calls from the hotel's pr dept after the stay
 potato soup,mutton biriyani ,egg plant in garlic sauce ,aapams,egg roast all were good
```

As can be seen, some of the sentences that contain mixed aspects like “good budget hotel with clean rooms and good breakfast” also make it here in this file even though they also contain some relevance to the aspect of value or rooms. But still our classifier works correctly while taking into account the proper nouns like names of dishes.

Applying topic modeling using MALLET, certain hidden topics and their frequent words are discovered. Since the LDA model does not provide a heading to the hidden topic it provides and the number of topics it takes should also be predefined, so we take them to be 15. A key file of the meals.txt file after topic modeling looks like this:

Table 2. A *key.txt* file for *meals.txt*

Topic No.	Frequent Words
0	nice office waiter sites exploring kerala a/c bring filtered reply highlight maker tea/coffee overpriced taxi priced firstly cheaper replacement behave front
1	offer housekeeping orchid entered items beer/coctails hang paid water enjoyed myself.a extremely doesnt keeping lizards helpless comfy alternative
2	pathetic booked disappointing.refill page hotel,rooms poorly chef alternate assist representatives team visiting families restaurant orchid maintained,food heavenly super buffet draw
3	days veg back supervise feat suggested carry alternative.hotel tariff coming bcoz orchid send dingy night hotel.a making number atmosphere wanted
4	staff family slow uninspiring warm loud order advance bowl city star.great share skip respond.u restaurant.bugs.house expensive big assistance show rate
5	checkin cleartrip orchid location told rooms town smooth exhausted loud dont husband electric bathroom indian comfortable.food parking cockroach volunteered months
6	worth christian nite service.i diners asks hotelat residency.good doors guest apprehensive time attitude card gave simple important neatly window night
7	sambar food directly expedia roughly mainaintans claims hygienic extremely dated TripAdvisor sick hot alternative supposed passed maintained promised night good
8	service food heart behaviour average midnight good.but base trough okay.we ample bath comfortable vacant conscious castle complained cockroaches courteous executive
9	turned stay business dinner restaurant guard clean curry booked picked breakfast money's worked kerala bit alcohol showing rajesh charged
10	windsor middle chicken visited waiters backwaters terrible sir kettle mineral south specious provided closed.i slightly better right bhavans money castle improvements
11	hot found booked reach service there reservation printed site system operate heat rank mediocre drop tired provide oki call poor lighting reminder
12	hotel good rooms food room breakfast stayed water spacious called residency stay buffet service star finger bring experience charge
13	checked march beds manager pick thirty fantastic caramel open affordable explain
14	maintained guests located food reached today the shave bowls!!no categorized smile the guy prices shocked saviour unapologetic sun hot waiting provide out-of-business

The Sentiment Analysis part of each aspect file gives the aggregate score of each aspect, i.e. their positive and negative scores on the aspect and writes the output of this to a file ‘sentianalysis.txt’. It can be helpful in many ways. The scores in this file show normalized as well as un-normalized scores. This is important as it can help to identify the weaker points / aspects of the hotels. The results of this file in normalized format have been plotted in form of bar chart.

The results of sentiment scores of aspects show something of peculiar interest. It can be seen that the score of ‘staff’ attribute is coming out to be negative only and shows no score on the positive side. This was because this aspect was not mentioned much in our review sentences. In fact the classifier puts only three sentences under this aspect’s file, all of them being negative. This happens even though there are some sentences mentioning ‘staff’ because of the fact that one sentence is placed under exactly one aspect. Also because of the fact that people only mention things like this if there is something to complain about or something very good to praise. The other results show the hotel's aspect scores.

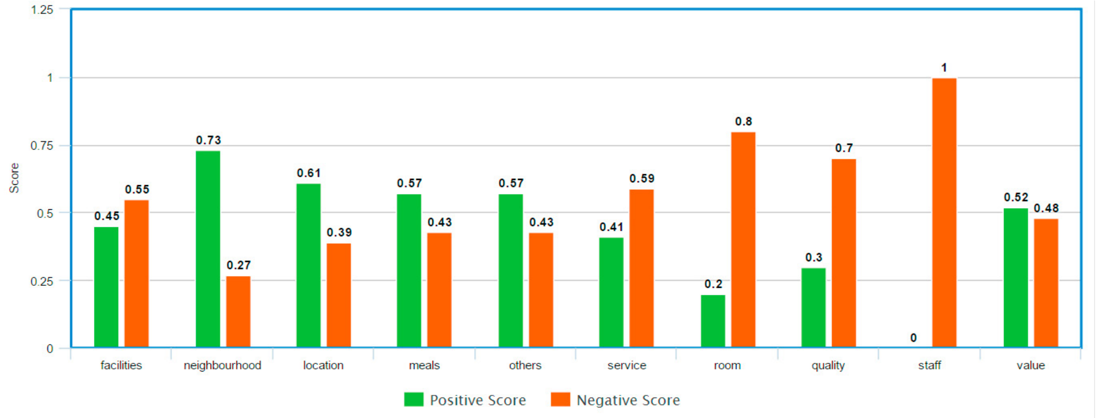
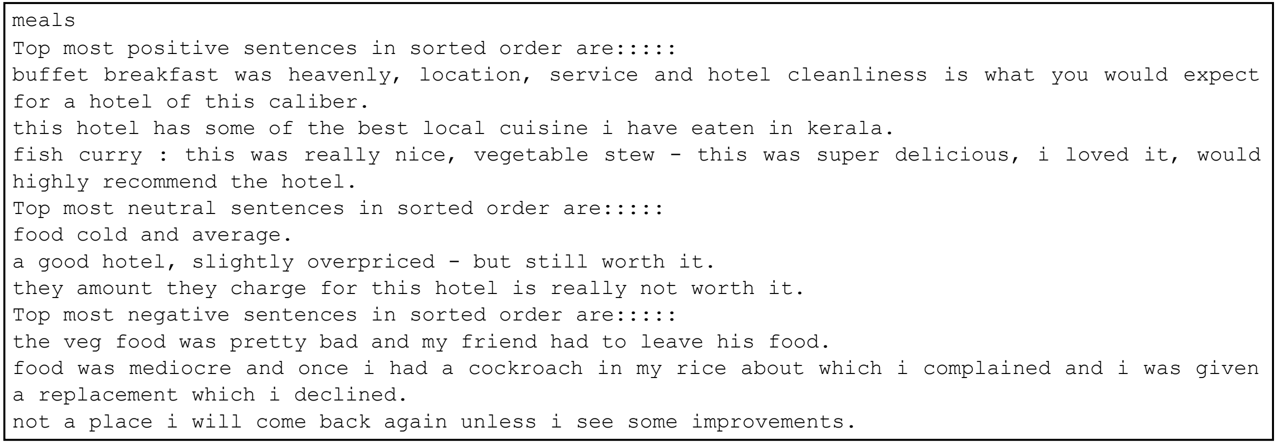


Fig 3. Sentiment Analysis Results

The summarization part proceeds as given in the method before. It presents the most polarized, positive negative as well as neutral sentences about each aspect in a descending sort order of their scores. The top three sentences have been displayed. A snapshot looks like this:



```
meals
Top most positive sentences in sorted order are:::::
buffet breakfast was heavenly, location, service and hotel cleanliness is what you would expect
for a hotel of this caliber.
this hotel has some of the best local cuisine i have eaten in kerala.
fish curry : this was really nice, vegetable stew - this was super delicious, i loved it, would
highly recommend the hotel.
Top most neutral sentences in sorted order are:::::
food cold and average.
a good hotel, slightly overpriced - but still worth it.
they amount they charge for this hotel is really not worth it.
Top most negative sentences in sorted order are:::::
the veg food was pretty bad and my friend had to leave his food.
food was mediocre and once i had a cockroach in my rice about which i complained and i was given
a replacement which i declined.
not a place i will come back again unless i see some improvements.
```

6. Conclusion and Future Work

This study focuses on the analysis and summarization of the user reviews of hotels mentioned on the TripAdvisor website. Review summarization provides a better understanding to the users about what they are looking for in the hotel. The analysis on the reviews, providing the hidden topics and frequent words can be further of importance for finding out valuable information. The study also analyses the sentiment scores of the hotels on the basis of their aspects, which gives a better understanding as to which of the aspects of the hotel under study are better than the others as per the user comments and on which of these aspects more improvement needs to be done. The hotel review summary this paper provides and sentiment score are separate from the ranking provided by the website and can be more realistic than the one provided by the website.

There is a variety of directions that this work can take. We can also rate the hotels as per the aspects crawled[14], i.e. how is the hotel in terms of the location for example. This can better help the customer in his decision making process of which hotel to choose as per his/her requirement. This can also be of help to the hotel management because they will now be aware of what areas they need improvement in and what their strong points are.

As of now the metadata that we crawled has not been used to much effect. It can be used to further improve our results. The review date can be used to find out some other useful patterns such as the seasonal patterns when a hotel is visited most, e.g. a hotel in Shimla is more likely to be visited during snowfall when most of the tourists are attracted to hill stations. Another potential application that can be of important consequence is including the metadata to account for the author's credibility. Many reviewers can just to defame the hotel can leave the negative aspects of the comments and discredit the hotel rating. Hence using the data like number of 'helpful reviews' or the number of up-votes that this author has gained can be used to further strengthen the credibility of this review and increase its weight-age on the final result.

An important direction that this work can take is to ask the user the aspects they want in the hotels, e.g. economical, for family, good food, etc. This can be used to effectively build a Personalized Hotel Recommender System that would suggest the user which hotel to choose while looking for their preferences. Although effectiveness of such personalization requires instructiveness with the consumers[17][18], but still we can effectively remove all user intervention by crawling the entire user profile, the hotels he/she has visited, travel history, travel style, etc. and then suggesting him/her the hotel that suits him the most by just inputting the travel schedule and giving a brief summary of the strong points and the weak points as retrieved from user reviews. The use of social media information if freely available can also be incorporated to better improve on the recommendations as is denoted by the work of SY Wang[6]. On the management's side, this recommender system can be used to know the potential travelers that can visit a hotel and improve on the weaker points.

References

[1] Potgieter, Marius, Johan W. de Jager, and Neels H. van Heerden. An innovative marketing information system: A management tool for South African tour operators. *Procedia-Social and Behavioral Sciences* 2013; **99**: 733-741.

[2] Wan, Yun, and Makoto Nakayama. The reliability of online review helpfulness. *Journal of Electronic Commerce Research* 2014; **15.3**: 179.

[3] Raut VB, Londhe DD. Survey on opinion mining and summarization of user reviews on web. *International Journal of Computer Science and Information Technologies*. 2014; **5(2)**: 1026-30.

[4] Bansal P, Ahmad T. Methods and Techniques of Intrusion Detection: A Review. *In International Conference on Smart Trends for Information Technology and Computer Communications, Springer, Singapore* 2016; 518-529.

[5] Xiang Z, Schwartz Z, Gerdes JH, Uysal M. What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International Journal of Hospitality Management* 2015; **44**: 120-130.

[6] Zhang W, Wang J. Integrating Topic and Latent Factors for Scalable Personalized Review-based Rating Prediction. *IEEE Transactions on Knowledge and Data Engineering* 2016; **28(11)**: 3013-3027.

[7] Tsytsarau M, Palpanas T. Managing Diverse Sentiments at Large Scale. *IEEE Transactions on Knowledge and Data Engineering* 2016; **28(11)**: 3028-3040.

[8] Philander K, Zhong Y. Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management* 2016; **55(2016)**: 16-24.

[9] Hu HW, Chen YL, Hsu PT. A Novel Approach to Rate and Summarize Online Reviews According to User-Specified Aspects. *Journal of Electronic Commerce Research* 2016; **17(2)**:132.

[10] García-Pablos A, Cuadros M, Linaza MT. OpeNER: Open Tools to Perform Natural Language Processing on Accommodation Reviews. *InInformation and Communication Technologies in Tourism Springer, Cham.* 2015; 125-137.

[11] García-Pablos A, Cuadros M, Linaza MT. Automatic analysis of textual hotel reviews. *Information Technology and Tourism* 2016; **16(1)**: 45-69.

[12] A Study of Recommendation System Combining Social Information. *Master's Thesis 2015.*

[13] Dingding Wang, Shenghou Zhu, Tao Li. SumView: A Web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications 2013; **40**: 27–33.*

[14] Dim Nyaung DE, Thein TL. Feature-Based Summarizing and Ranking from Customer Reviews. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 2015; **9(3)**.

[15] Steffen J. N-Gram Language Modeling for Robust Multi-Lingual Document Classification. InLREC 2004; : 731–734.

[16] Walter Kasper, Mihaela Vela. Sentiment Analysis for Hotel Reviews. *Proceedings of theComputational Linguistics-Applications Conference* 2011; : 45-52.

[17] Ho SY, Ho KK. THE EFFECTS OF WEB PERSONALIZATION ON INFLUENCING USERS'SWITCHING DECISIONS TO A NEW WEBSITE. *PACIS 2008 Proceedings* 2008: 67.

[18] Xu Y, Yu Q, Lam W, Lin T. Exploiting interactions of review text, hidden user communities and item groups, and time for collaborative filtering. *Knowledge and Information Systems* 2017; **52(1)**: 221-254.