Enhanced Lexical Based Technique For Opinion Mining In Tourism

Dr. Meenakshi Bansal, Ashok Bathla

Abstract: Today is the era of online shopping. Most of the people do online shopping for their convenience. It includes both M-Commerce and E-Commerce. Important part of the online shopping are the reviews given by the customer which gives the rating to the product they have purchased. In addition to the product reviews, customer also give reviews about company and post purchase experience. These reviews effects the promotion of the product and also helps others to take wise decisions regarding the purchase of goods. This research is focused on studying the reviews on tourism. In which people gives reviews regarding Hotels, Recreation places etc. They also reviews regarding Food, Room, Room Service, Parking, Pick up facility etc. In this research lexical based approach is used to identify the collective analysis for positive and negative reviews. So that people who are trying to take the services following the reviews can see the collective scenario. Current lexical based approach is better than the previous research which was based on sentiwordnet. In sentiwordnet the marks or grade given to the positive, negative, and neutral word. With the help of lexical approach reviews have been tokenized. Positive, negatives and neutral reviews are compared with established Ontology. This mined information will provide better view regarding the total reviews rather than studying all the individual reviews. Experimental results shows much better performance to have collective analysis by automated tool in terms of accuracy and time taken.

Key Term: Ontology, Reviews, Opinion Mining, sentiwordnet, lexical, tokens.

1 INTRODUCTION

1.1 Web Review
Web is a repository of data which large amount of information is added each day through online users in different ways. This information grows rapidly as the number of users increase day by day. Large chunk of this information is the product reviews, that people had purchased through e-commerce or M-commerce. The analysis of this review data will provide expertise view point of the results. Any person before his or her purchase wants to checks these reviews. Main issue in this type of data is its processing in positive or in Negative sense. Among E-Commerce and M-Commerce one main category which is most commonly used is online booking of tour and packages. Online booking provides economical and convenient booking solutions. Once people get the services they provide the expert reviews about the services. Later on machine based processing and analysis of these reviews is done which gives valuable results in lesser time. These machine controlled tools provide easy classification of the data [1].

1.2 Stages of Text Mining Methodology
Text Mining is used in various fields, starting from information retrieval to the language processing applications. The Text based mining application is processed in multiple steps as shown below:

1. Knowledge Retrieval System establish the document from the retrieved reviews from different websites. The collected reviews must be highly relevant reviews. For example tourism related reviews.
2. Language method is one of oldest and most hard way of information processing. It is human like behavior. Where large information is understood as person speaks. This system has many problems like sentence boundaries, Part of speech tags, parsing
3. Knowledge Modelling (KM) first retrieve the relevant reviews data. Classify the retrieved data in useful classification. Keep the classified data in to database. Later on provide the graphical interface through which processed information can be retrieved with.
4. Knowledge Extraction is that way in which information is retrieved in structured way from unstructured language document[2].

1.3 Machine Learning based Text Classification
Machine learning primarily based text classification contains quantitative approaches Natural language Processing (NLP) that uses machine learning algorithm. Most commonly used supervised learning techniques for text classification are:

- Rocchio Algorithmic rule(RAR)
- Decision Trees and Support Vector Machine
- Artificial Neural Networks

1.4 Ontology based Text Classification
Ontology or primary keyword based extraction is considered as the powerful solution of the problems. It benefits from the domain knowledge and provides good accuracy as compared to traditional text classification techniques. Large data base of positive and negative words is prepared. This data base belongs to the domain knowledge of the related field. For ex. Ontology related to tourism. The word used is specifically related to tourism in positive and negative sense [1].

2 LITERATURE SURVEY
K. Amarouche et al. (2016) [1] numerous corporations in today’s atmosphere square measure supported numerous natural philosophy market places. These market places as like net, and TV tec. They shares their contents through social networking sites. folks sell their merchandise on the web. therefore folks opinion concerning their product matters. as a result of negative
review can enforce merchandiser to seem at their
merchandise therefore the product improvement. Even to
took back their product from the market. They collects the
review knowledge, analysis it and so soon method this data to extract helpful analysis.
C. Bucur et al. (2016) [2] in line with this paper merchandiser or analyzer is to analysis the opinion
expressed by the folks on touristy. The collected knowledge is to be summarized and classified in
pre outlined dataset classification. Within the paper the potency of algorithms square measure analyzed victimization text mining domain
specific measures.
A. Severyn et al. (2016) [3]
There needs sturdy model mistreatment that great deal of
reviews knowledge by users will be sub set and summarized
for higher form of analysis. This study model is specified it will handle great deal of knowledge,
this knowledge even will have noise. This analysis is
predicated on opinion mining on the information created in
YouTube connected videos. This analysis is predicated on
proposing a strong shallow grammar structure (STRUCT)
that adapts well once tested across domains.
as sentiment analysis, refers to distinctive, Extracting knowledge from
the source by mistreatment tongue process, text analysis
and process linguistic. options extraction is predicated on
extracting those options solely onto that opinion is
predicated on. It is effective approach of this IEDR technique
for the aim of feature extraction.
S. Saquib Sohail et al. (2015)
[5] wearable devices has nice positive response from the
general public, therefore numerous analyzers has inclined
themselves for the research within the field
of wearable devices. they need steered user feedback
model for wearable devices sweetening. it'll helps in
enhancing the standard of the merchandise associated
[6] scientist has stress on the method of distinctive
the method of opinion mining to trace the mood of
the folks from the massive dataset. They has specify the
seven classes like sturdy positive, positive, weak-positive,
negative, weak-negative, neutral and powerful negative
words for the opinion mining task. They need used the
technique like Naïve Bayes, SVM and Multilayer perception.
Dhivya Bino et al. (2016) [7] Opinion mining uses
supervised learning algorithms to seek out the polarity
of the coed feed back knowledge. This study includes the
mix of technique like machine learning and tongue process.
This paper has conjointly provide the technique
Supported SVM, Naïve mathematician, K Nearest
Neighbors and neural network process.

3 TECHNIQUE
Ontology of positive and negative words will be prepared. This ontology is specially related to tourism domain. By using
lexical based approach these positive and negative words will be matched to extract general point of view. The results
generated will be highly useful for system representation for
reviews wants to purchase some thing from the shopping
website.

4 ALGORITHM
Step 1 Collect the reviews related to Tourism especially
related to Hotels.
Step 2 Collect the ontology of various positive and negative
words.
Step 3 Convert the reviews data into individual words using
lexical based approach.
Step 4 Identifying the number of positive and negative words
into the reviews data.
Step 5 Categories the reviews into six sub categories like
Room service, Rooms, Pickup Facility, Food, Parking etc.
Step 6 Identify the collective analysis for all the six sub
categories.
Fig. 1 show the flowchart of the above specified steps in the
algorithm. It explains how the ontology related to tourism
works to accurately analyze the tourism related reviews.

6 RESULT ANALYSIS
Current research is to detect the sentiments of the people
reviews regarding tourism. It is the field of opinion mining.
Where large number of people reviews are collected from
different sources. Each source provides the reviews related
to tourism. A large ontology of positive and negative words
will be generated. With the help of lexical approach positives
and negative words are matched in lexical divided reviews.
So that percentage of positive and negative sentences can
be found. This approach may be suitable for those people
who cannot find the positive and negativity in the reviews by
reading each single reviews. Current research approach will
provides the collective approach for both positive and
negative reviews.

6.1 Sample Data
Various reviews are collected from TripAdvisor.com about
hotels in New Delhi. It includes the reviews regarding
different aspects related to tourism. So that a general point
of view can be developed with current research approach.
Aspects that are collected from the reviews are shown in Table 1.

**Table 1: Different Aspects for Reviews**

<table>
<thead>
<tr>
<th>S no.</th>
<th>Aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Room Service</td>
</tr>
<tr>
<td>2</td>
<td>Food</td>
</tr>
<tr>
<td>3</td>
<td>Cost</td>
</tr>
<tr>
<td>4</td>
<td>Parking in Hotel</td>
</tr>
<tr>
<td>5</td>
<td>Room Furnishing</td>
</tr>
<tr>
<td>6</td>
<td>Hotel pickup services</td>
</tr>
</tbody>
</table>

**6.2 Sample Review**

1. Nov 2 2016: Good hotel for an overnight stop. We spent two separate nights at this hotel as part of our tour of India. As others have said, it is heavily used by and geared to tour groups. There are a lot of arrivals and departures in the night and early morning, which may be disturbing to individual travelers. This is a good hotel though. Our room was pleasant and the bathroom clean. The food was good and the service very friendly and welcoming, except at the front desk where staff were a little abrupt. The gardens are very well maintained, quiet and peaceful. On our second visit our enjoyment of them was increased by the beautiful decorations and fairy lights for a wedding reception to be held that evening. We would not hesitate to return for an overnight stay.

2. Nov 12 2016: Getting there is rather odd, you have to drive down a dirt track and the hotel is at the end - we were in transit, so it didn't matter much, but you wouldn't want to stay there if you were on a trip to see the sights of Delhi. The hotel and rooms (although large) are characterless. Service in the restaurant was OK, as was the food. When we arrived at midnight we were unable to get a beer - bar closed and no beers in the mini-bar! Reception did however find us a couple of beers which were delivered to our room.

3. Sep 10 2016: away from the bustle of Delhi, useful stop over if visiting Delhi. Good pool and excellent food. Road works when I was there so getting to and from was a bit of a problem but was over come. Taxi ride away from nightlife in Delhi.

4. Sep 5 2016: Old but comfortable. On a five star scale rating this hotel would qualify as a high end 3 star hotel. It's an older resort style hotel but has few amenities. The building and the grounds are surrounded by a tall wall which is what separates the front of the building from the busy street. We stayed in a room facing away from the street and did not have any complaints about the noise. The bathrooms were clean and so were the linens. There are nice big flat panel televisions with tons of local cables and news stations to watch. The grounds are manicured and well kept and the pool was clean. Since it was the winter season when we were there we opted not to go for a swim but the pool itself seemed nice. The gym is a bit cramped and dark and equipment seriously out of date but if you are just looking for descent cardio equipment it should get the job done. The breakfast served was minimal but all you can eat. Overall our stay was pleasant and would consider staying again. Such reviews are taken 3006 in number. Each review is in sentences given by different people in different years. The reviews taken are from year 2006 to 2016. It provides the collected scenario for hotels in tourism on different aspects. Each aspect is carefully studied and only those are considered which has large contribution to the reviews analysis in tourism related field.

**6.3 Dataset for Review**

**Table 2 Dataset for Review Collection**

<table>
<thead>
<tr>
<th>S no.</th>
<th>Number of Reviews</th>
<th>Number of Sentences</th>
<th>Number of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>153</td>
<td>987</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>1017</td>
<td>9100</td>
</tr>
<tr>
<td>3</td>
<td>2000</td>
<td>2087</td>
<td>19450</td>
</tr>
<tr>
<td>4</td>
<td>3000</td>
<td>3007</td>
<td>32021</td>
</tr>
</tbody>
</table>

In Table 2 reviews are subdivided into smaller parts of 100, 1000, 2000 and 3000 reviews. In this dataset we counted the number of sentences and number of words with the help of lexical approach.

**6.4 Dataset for Positive and Negative Review**

**Table 3 Dataset**

<table>
<thead>
<tr>
<th>Positive</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>4783</td>
</tr>
</tbody>
</table>

We prepared ontology of positive words 2007 and negative words 4783 as shown in table 3. This ontology includes those words which commonly occurs in the reviews data. These ontology words are matched against the reviews data. So that the number of positive and negative words percentage can be counted. It helps in providing the systematic reviews to the people which makes the purchasing more convenient after reading the reviews positive or negative.

**6.5 Parameters Taken**

**6.5.1 Accuracy**

Accuracy is the total amount of accurate results that has been generated in terms of positive or negative. 

\[ \text{Accuracy} = \frac{\sum p + \sum n}{\sum p \times 100} \]

**6.5.2 Time**

It is total time taken by the procedure to perform the activity. That means total time taken for comparing the reviews with Ontology.

\[ T = \sum t \]

**6.6 Evaluated Accuracy and Time for Review Data**

**Table 3 Accuracy**

<table>
<thead>
<tr>
<th>Reviews Count</th>
<th>Accuracy Percentage for Positive</th>
<th>Accuracy Percentage for Negative</th>
<th>Average Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>75.83%</td>
<td>16.66%</td>
<td>131 sec</td>
</tr>
<tr>
<td>1000</td>
<td>77.66%</td>
<td>15.66%</td>
<td>231 sec</td>
</tr>
<tr>
<td>2000</td>
<td>74.33%</td>
<td>15.66%</td>
<td>300 sec</td>
</tr>
<tr>
<td>3000</td>
<td>80%</td>
<td>17.66%</td>
<td>377 sec</td>
</tr>
</tbody>
</table>
Table 3 shows the Accuracy percentage of positive and negative reviews and time taken to analyze whether the review is positive or negative. This result has been calculated for 100, 1000, 2000 and 3000 reviews. The accuracy percentage grows as number of reviews grows. Because in aspect based analysis using lexical approach results will improves as amount of data grows. Addition to it as compare to positive reviews negative percentage has shown the improvement. But the negative percentage growth is less compared to the positive percentage growth. But time taken has shown the linear growth with little percentage.

6.6 Graph For Accuracy Percentage and Review Count

This graph shows the growth of accuracy percentage for positive and in comparison to the growth in the Reviews count.

6.7 Graph For Accuracy Percentage And Review Count

This graph shows the accuracy growth of negative count percentage against growth in the reviews count.

6.8 Graph For Time And Reviews Count

This graph shows the growth of average time and the growth in reviews count. Time grows as the number of reviews grows.

6.9 Comparison Accuracy of Current and Previous Research

<table>
<thead>
<tr>
<th>Accuracy based on Lexical based Approach</th>
<th>Accuracy based on SentiWordNet based Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.83</td>
<td>72</td>
</tr>
<tr>
<td>77.83</td>
<td>76</td>
</tr>
<tr>
<td>74.33</td>
<td>74</td>
</tr>
<tr>
<td>80</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 4 Comparison of accuracy for positive

This graph shows the accuracy of previous and current approach. Values represented shows there is a improvement in accuracy in current lexical based approach.

6.10 Comparison Graph of Accuracy of Current and Previous Research

This graph shows the accuracy percentage comparison. In current lexical based approach there is improvement in the percentage of accuracy as the number of reviews grows. But in previous research there is decrease in accuracy as the number of reviews grows.

6.11 Comparison Time of Current and Previous Research

<table>
<thead>
<tr>
<th>Time based on Lexical based Approach</th>
<th>Time based on SentiWordnet based Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>131</td>
<td>539.24</td>
</tr>
<tr>
<td>231</td>
<td>5985.11</td>
</tr>
<tr>
<td>300</td>
<td>12371.8</td>
</tr>
<tr>
<td>377</td>
<td>17987.58</td>
</tr>
</tbody>
</table>

Table 5 time comparison of current and previous research
This table shows the time comparison of both lexical based approach and the sentiwordnet based approach. There is a
marked improvement in the processing time for current research compared to the previous research.

6.12 Comparison Graph of Time of Current and Previous Research

![Graph showing time comparison of current and base approach]

Graph 6 shows the time comparison of current and base research.

This graph shows the improvement in the time spent in both current and previous research. Current research which is based on lexical based approach takes less time in comparison to sentiwordnet based approach.

7 CONCLUSION

Today’s world is more emphasized on internet based shopping. Where people reviews regarding their experience about the product matters. Each time people post reviews about their experience about the product they have purchased new users read the reviews. But it is difficult to read each and every review for identifying the positive and negative reviews. So Current research based on lexical based approach is to identify the collective analysis for positive and negative reviews. So that people who sees to the reviews can see the collective scenario. Current lexical based approach has outperform than the previous research based on sentiwordnet. Both the parameters like accuracy and time has shown the improvement than the previous research.

8 FUTURE WORK

Current research is focused on tourism specially on hotels. This research can further be enhanced in other fields related to tourism. So that collective system can be developed to analysis the reviews data related to tourism.

9 REFERENCES