Sentimental Analysis On Hotel Reviews Using Classic Approaches And A Smaller Data Set

Raghav Sehgal

ABSTRACT: There are large no of customers visiting hotels in various cities daily. Various online forums, websites , social networks etc having details of all these hotels consist of text information in the form of comments. These comments are not captured and utilised properly. Using sentimental Analysis on this information can help in enhancing customer experience to a great extent. It helps in choosing the right hotel to tourist/visitor in any city. Estrela Do Mar Beach Resort,Calangute, Goa is the hotel which is chosen for our sentimental analysis endeavours. TripAdvisor.com lists down more than 1000 mixed bag of reviews on this hotel online which is taken for our dataset. My manual crawler crawls the reviews and feeds them to Apache Lucene software taking review and polarity(1,0,-1) based on the star marked rating of the review. The index matrix is given to K Nearest Neighbours (KNN) classifier (taking k=5) for it to be trained on. Another set of 100 reviews extracted act as test data. I obtained an accuracy of 62% via the classifier. Another classification methodology gave a score to each of the review extracted using sentiwordnet. We find that accuracy increases to 64 %. Tripadvisor.com does not bridge the gap between the star rating user assigns and what the user actually means by the review. The SentiWordnet bag of words approach using adjectives aims to bridge the gap between this by giving a unique score to each review.The sentiwordnet assigns polarity based on dictionary valuation and so can also act as a classifier in itself assigning a polarity to each review by using four separate mechanisms of assigning polarity. The accuracy in each of these cases was also noted and the maximum was obtained by using only adjectives, taking the neutral value as negative. It is also noted that the KNN Accuracy increases a substantial amount to about 10% increase on doubling the trainee set. We are thus able to compare both the supervised as well as unsupervised learning approaches for text mining.

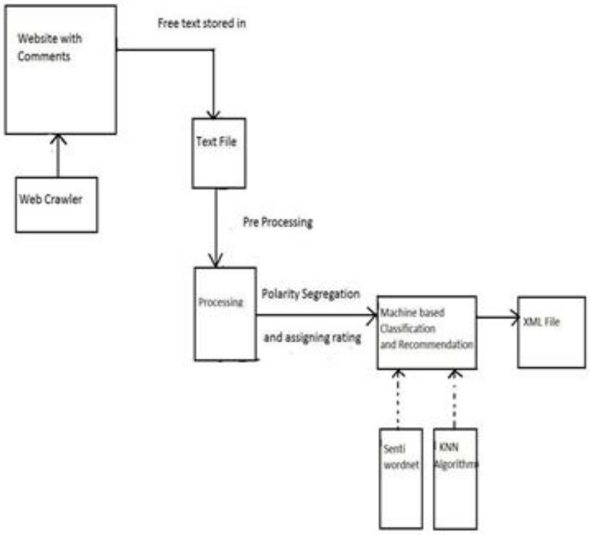
INDEX TERMS: Sentimental analysis, KNN, Manual crawler, Tripadvisor, sentiwordnet, java, opinion mining.

I. INTRODUCTION

Sentimental analysis or opinion mining is an application of text mining and NLP is being widely used in the world today. The purpose is to extract user sentiments from a given textual data and obtaining some useful information from it. It is widely being used in social media applications, e-commerce websites, tourism industry, medical fields etc. to enhance the customer experience. The idea is to extract whatever user has tried to communicate. It has been used in the past to extract user data from twitter. Every such analysis is based upon a certain classifier to be used and implemented. The process may be as simple as doing the analysis for a simple sentence to doing the analysis for multiple documents. There are a few problems which may arise and hamper the accuracy of results which need to be taken care of. The reviews or any such textual data may contain some non required elements such as emoticons, hash tags, smiley, sentence terminators or even non required words and bits such as stop words like conjunctions-and, or , but etc. which if used cause decrease in the accuracy of our results. All such non essential data needs to be removed before we continue on with our analysis part. Hence, we first filter our text from the stop words and also stem the words to get only the root words. This system makes use of sentimental analysis to understand the customer experience based on their online comments. The comments are categorised based on whether they are positive or negative or neutral and accuracy of the machine is predicted based on the classification algorithm. Also an actual rating predicted by the machine rather than as given by the customer on the online forum is obtained by the machine.

I have taken customer reviews about hotel Estrela Do Mar Beach Resort,Calangute,Goa on Tripadvisor.com website as the dataset. Both trainee and testing data is obtained and accuracy is determined. This methodology involves using three classes for the purpose of classification. These classes are based on the star marked rating as assigned by the user online. The rating is assigned to each review and is in the range of 1 to 5. The experiment was done using a smaller dataset of about 1400 reviews and the hotel for our research was chosen judiciously and accordingly. This was done to have a dataset comprising of almost equal amounts of positive, negative as well neutral reviews. Tripadvisor website allows any user to give a review on any hotel in a particular city. A star marked rating out of 5 is specified. It also stores information like the name of user, Date of writing and URL of the review. All this information is extracted by my web crawler and used for the analysis. The web crawler and the entire process has been carried out in java, JDK 8 using NetBeans API. The crawler extracts data based on regular expressions formed by looking at the html tags used on the website and needs a high speed internet connection if the extraction is to be carried out expiditely. The data related to the reviews is crawled and stored in a text file from where we have stored it in an Array List for further processing.Each review is thus stored in an array list of objects along with the information associated with each review. The process as carried out is mentioned in the architecture diagram provided below:

- *Raghav Sehgal, Analyst System Development, Verizon, raghav.sehgal.2512@gmail.com, raghavsehgal77@gmail.com, +91 8754551856*



II. METHODOLOGY

Classification is done using a step by step process as follows:

- Crawler extracted 1400 reviews from TripAdvisor.com website about hotel Estrela Do Mar Beach Resort, Calangute,Goa.
- Reviews along with star marked rating,polarity assigned ,review id are stored in a text file.
- This information is then stored into an array list
- Preprocessing and stemming is done over the data.
- First 1300 reviews are taken as trainee data. This data is fed to Apache Lucene Software for indexing which stores reviews and their associated polarities in the matrix.
- The indexed reviews are then trained over a classifier using KNN algorithm K=5. 100 more reviews are taken as test data to obtain the polarity as predicted by the classifier. Using this I obtained an accuracy of 62.1%.
- Another parallel approach uses SentiWordnet for assigning a score to each review and uses dictionary based approach for classification.
- SentiWordnet yielded a better accuracy than the above classification model.
- Both Trainee as well as Test reviews, their polarity and other details along with those predicted by machine were passed in an XML file and displayed. Also a bar graph for each review assigning a score against the star marked rating and assigned polarity is displayed.

III. STEP BY STEP MODULES

Now we go about the various modules that were used in carrying out this methodology:

A. Crawler Module

It creates a web crawler for extracting customer reviews from website ,Tripadvisor.com about hotel Estrela Do Mar Beach Resort, Calangute,Goa . The information extracted for each hotel is put separately in file which is used for further processing. The crawler uses Regular Expressions to extract the specific data required for processing. The URLs to be crawled are put in a separate text file and all the links are crawled sequentially.

B. Extract Module

Extracts reviews and their associated properties from the file and puts them in an array list. The polarities are also assigned based on star marked rating in this manner:

Star Rating	Polarity Allotted
>3	1
=3	0
<3	-1

C. Pre-processing Module

In this module first tokenization of the text is done. First stop words are removed. After this stemming process is carried out for getting the root words. The text obtained can now be used for further processing. We have used porter stemmer for carrying out the stemming.

D. Indexing Module

Uses Apache Lucene software for indexing. Reviews and polarities are stored as fields in matrix format. The indexing is done for trainee data. Lucene jar file is easily available in java.

E. Classification Module

a) KNN Algorithm Approach

The indexed matrix is given as input to the classifier. We take K=5 here.The machine is trained over this data. Later test data is given to the classifier for predicting its polarity and the accuracy is obtained.

b) SentiWordNet Approach

We use SentiWordnet 3.0.0 for analysing each review, assigning a score to it based only on adjectives and the net score count of each sentence is calculated and its polarity is predicted. The accuracy increases in this approach. I have used 4 different mechanisms for finding out 4 different observations using this approach.Senti wordnet uses 4 functions:

1) classifyAllPOSY()- this uses adjective, noun, adverb, and verb as pos tags to assign polarity and classify the review text. We take 0 as +ve in one case, creating 2 separate classes and in the other it is taken as neutral case creating 3 separate classes.

2) classifyAllPOSN()- this uses adjective, noun, adverb, and verb as pos tags to assign polarity and classify the review text. We take 0 as -ve in one case, creating 2 separate classes and in the other it is taken as neutral case creating 3 separate classes.

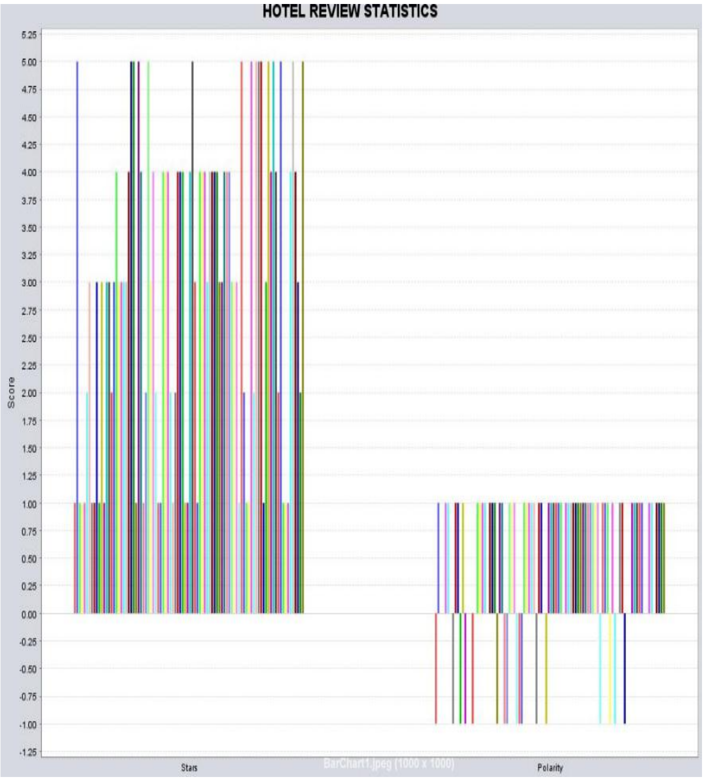
3) classifyADJY()-this uses only adjectives as pos tags to assign polarity and classify the review text. We take 0 as +ve in one case, creating 2 separate classes and in the other it is taken as neutral case creating 3 separate classes.

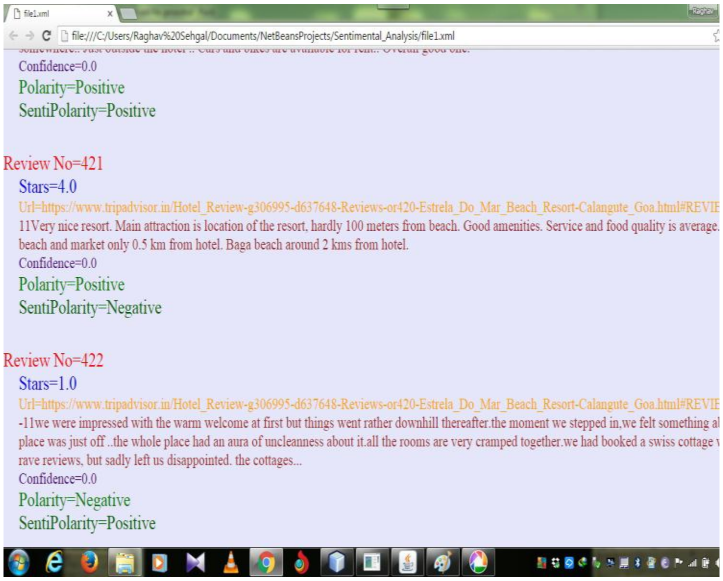
4) classifyADJN()-this uses only adjectives as pos tags to assign polarity and classify the review text. We take 0 as -

ve in one case, creating 2 separate classes and in the other it is taken as neutral case creating 3 separate classes.

F. Display Module

All the extracted reviews along with their details including polarity as assigned by the man and the machine is displayed in a stylised XML file using document object model.Also a bar graph for each review assigning a score against the star marked rating and assigned polarity is displayed.




file1.xml
Raghav
file:///C:/Users/Raghav%20Sehgal/Documents/NetBeansProjects/Sentimental_Analysis/file1.xml

somewhere.. Just outside the hotel .. Cars and bikes are available for rent.. Overall good one.
Confidence=0.0
Polarity=Positive
SentiPolarity=Positive

Review No=421
Stars=4.0
Url=https://www.tripadvisor.in/Hotel_Review-g306995-d637648-Reviews-or420-Estrela_Do_Mar_Beach_Resort-Calangute_Goa.html#REVIE
11Very nice resort. Main attraction is location of the resort, hardly 100 meters from beach. Good amenities. Service and food quality is average.
beach and market only 0.5 km from hotel. Baga beach around 2 kms from hotel.
Confidence=0.0
Polarity=Positive
SentiPolarity=Negative

Review No=422
Stars=1.0
Url=https://www.tripadvisor.in/Hotel_Review-g306995-d637648-Reviews-or420-Estrela_Do_Mar_Beach_Resort-Calangute_Goa.html#REVIE
-11we were impressed with the warm welcome at first but things went rather downhill thereafter.the moment we stepped in,we felt something al
place was just off ..the whole place had an aura of uncleanness about it.all the rooms are very cramped together.we had booked a swiss cottage v
rave reviews, but sadly left us disappointed. the cottages...
Confidence=0.0
Polarity=Negative
SentiPolarity=Positive

IV. RESULTS

Both the classifiers work on the extracted reviews from hotel Estrela Do Mar Beach Resort,Calangute,Goa and we obtain some interesting findings:

1) For the KNN algorithm approach I had taken in 1300 reviews as the trainee data. I took about 100 reviews as the test set. This approach gave an accuracy of 62.1052%.Now I tested the system by increasing the trainee set by doubling the reviews crawled. Interestingly the results were better and the accuracy increased by about 10%. Here, we got an accuracy of 71.6%.
2) Now I took to the non-supervised method of SentiWordnet using bag of words(unigram) approach.Here 2 different methodologies are:

   a) We first take the same 100 reviews as given to KNN approach and classified based on each of the methods mentioned above. Following results are obtained:

Algorithm	Accuracy(%)
classifyAllPOSY()	50
classifyAllPOSN()	52.12765
classifyADJY()	52.12765
classifyADJN()	51.06382

Interestingly, also in the same algorithms when 0 was taken as neutral, the results obtained are a bit contrasting. For instance, classifyADJY() gives an accuracy of 50% but classifyAllPOSY() in the same scenario gives an accuracy of 51.063%.

   b) Now taking the entire length of 1400 reviews extracted from the Estrela Do Mar Beach Resort,Goa we get the following results:

Algorithm	Accuracy(%)
classifyAllPOSY()	62.33859
classifyAllPOSN()	62.91248
classifyADJY()	63.7733
classifyADJN()	63.84505

The SentiWordnet approach uses a dictionary based classification mechanism and increases the accuracy to a substancial amount.The accuracy found in this case is the deviation of the user assigned star marked rating to that found on the basis of the SentiWordnet.

Accuracy= (Number of times both polarities match/Total reviews)*100

V. CONCLUSION

Opinion Mining is an emerging field in computing and this research upholds the fact. It is seen through my model that performing k nearest number algorithm to classification on about 1300 reviews and testing on 100 reviews we obtain accuracy of about 60% .On doubling the amount the same accuracy went to upto 72%.Thus classification works best on increasing the number of reviews. My SentiWordnet based polarity assigner compares the star marked rating assigned polarity to itself to find deviation and on about 1400 reviews we got access of about 64%. on using only adjectives classification and taking positive for three marked rating as well. Thus it was found that dictionary based approach(monogram) was better than a machine based KNN classification and the accuracy increases on increasing number of reviews. There has been a lot of work on sentimental analysis before. This model which trains currently on a single hotel on Trip advisor (Estrela Do Mar Beach) is further working on other classification algorithms to find which works the best. I haven't used WEKA tool for the classification process to check the accuracy based on only manually applied algorithms but plan to use it in the future for further research. I even plan to incorporate the Hybridization techniques of classification on the same dataset and try to increase the accuracy as desired even on a smaller dataset.

ACKNOWLEDGMENT

A project is a golden opportunity for learning and self-development. I consider ourselves myself lucky and honoured to have so many wonderful people leading us through in the completion of this project.I would first like to thank my parents for their constant support. My grateful thanks to Dr. C. Muthamizhchelvan (Director E&T, SRM University) , who has provided me with this opportunity to practically apply the skills that I have acquired. Dr. G. Vadivu, our Head of the Department,Infomation Technology(SRM) who monitored our programs and arranged all facilities.I choose this moment to acknowledge her contribution gratefully. I would also like to thank my teammate Vaibhav Sachdeva who helped me through documentation and with various inputs in the project. I would like to express my deepest gratitude to my guide, Mr N. Arivazhagan (Asst Professor) for his valuable guidance, consistent encouragement, personal caring, timely help and providing me with an excellent atmosphere for doing research. All through the work, in spite of his busy schedule, he has extended cheerful and cordial support to me for completing this project work.I would also like to thank Ms. E. Tamil Selvi(Asst Professor), Mrs.K.Nimala(Asst Professor) and Dr. G. Maragatham(Asst Professor) for their valuable inputs ,help, guidance and support all through the project. Last but not the least, I would like to thank all the faculties of my department who have made efforts to hone my skills over these years.

REFERENCES

[1]. H. Song, Y. Fan, X. Liu and D. Tao, "Extracting product features from online reviews for sentimental analysis," Computer Sciences and Convergence Information Technology (ICCIT), 2011 6th International Conference on, Seogwipo, 2011, pp. 745-750.

[2]. Wang and X. Luo, "Sentimental Space Based Analysis of User Personalized Sentiments," Semantics, Knowledge and Grids (SKG), 2013 Ninth International Conference on, Beijing, 2013, pp. 151-156.

[3]. A. Celikyilmaz, D. Hakkani-TÃijr and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," Spoken Language Technology Workshop (SLT), 2010 IEEE, Berkeley, CA, 2010, pp. 79-84.

[4]. Z. Kechaou, M. Ben Ammar and A. M. Alimi, "Improving e-learning with sentiment analysis of users' opinions," 2011 IEEE Global Engineering Education Conference (EDUCON), Amman, 2011, pp. 1032-1038.

[5]. P. Kherwa, A. Sachdeva, D. Mahajan, N. Pande and P. K. Singh, "An approach towards comprehensive sentimental data analysis and opinion mining," Advance Computing Conference (IACC), 2014 IEEE International, Gurgaon, 2014, pp. 606-612.