Sentiment Analysis for Hotel Rating using Machine Learning Algorithms

P. Sanjay Bhargav, G. Nagarjuna Reddy, R.V. Ravi Chand, K.Pujitha, Anjali Mathur

Abstract*: *As the technology is growing rapidly everyone is expressing their views in many websites. The rating of any online commercial site is highly dependent on opinion of its users. Not only commercial sites now a days every type of business-like hotels, banks, shops, malls etc are available on internet. They all seek online rating from their consumer. Consumers put their opinion according they got the service. Consumers sentiments are reflected in the form of 'opinion dataset' on internet. The proposed research work is for the sentiment analysis for hotel reviews. Consumers of certain hotel put their opinion on website, on the basis of that opinion we can rate the hotel and analyze it with other hotels. The proposed research work is implemented through Machine learning algorithmsusing Naïve Bayes Algorithm and opinion Mining Algorithms on the basis of Natural Language Processing.

Index Terms:Machine Learning, Naive Bayes, Natural Language Processing, Recommender Systems, Sentiment Analysis.

I. INTRODUCTION

In present days a company, a business organization or a service-based companies which requires feedback from its customers. Increase in expansion of company will be providing more number of services and products. So, as to increase this the organization must bother about the reviews and ratings given by its users. The service-consumers can mention their feelings and reviews on online-portals. By performing the opinion mining and sentiment analysis on these details we can predict the rating of that organization. One recommender system is required for generating the ratings inprecise and accurate manner. For a hotel business, reviews about various aspects like Maintenance, Food, Hospitality, Room neatness,Response from the staff ofaHotel, etc. plays a major role for recommender system.

The Customer's sentiments regarding to a hotel depends upon the facilities he/she got from that hotel, just like cleanliness, location of the hotel, services provided by the hotel like free wi-fi, multilingual staff, bar/lounge, babysitting rooms, wheel chair etc. The sentiments can be expressed in the form of excellent, good, average, poor, terribleetc.Generally, customers want to express their feelings also with these rating and review values.

Revised Manuscript Received on April 06, 2019.

P.Sanjay Bhargav, Computer Science, Engineering Koneru Lakshmaiah Education Foundation , Vaddeswaram, India.
G.Nagarjuna Reddy, Computer Science Engineering Koneru Lakshmaiah Education Foundation , Vaddeswaram, India.
R.V.Ravi Chand, Computer Science, Engineering Koneru Lakshmaiah Education Foundation ,Vaddeswaram,India.
K.Pujitha,Computer Science, Engineering Koneru Lakshmaiah Education Foundation , Vaddeswaram, India.
Dr.AnjaliMathur, AssociateProfessor , Computer Science, Engineering Koneru Lakshmaiah Education Foundation , Vaddeswaram, India.

A recommender system can provide suggestions for a single item or a sequence of items. A single item is generally used at e-commerce websites to find out the sale and product of one single item while the sequence recommender system for sequence of items is used to predict the business of any organization such as hotel business, cleanliness, services, food , these values are used. The recommender system is considered as a good one that can identify the appropriate users, their motivation, expectations and the goals. The recommender system can understand properly the properties of the item set such as volume, distribution, nature and the user's rating.

The recommendation technologies should make use of "trust metrics" to be considered as a strong recommender system. It has been proved by research that the people tend to rely more on recommendations from people they trust just like their friends rather than the online recommender systems. The trust enhanced recommender system must utilize the knowledge that originates from the trust networks like friends or the persons of same choice or person belongs to same city etc,. For a strong and reliable recommender system we need to apply group recommendation to individual users. Group recommendation techniques can be useful in three ways: (1) to aggregate multiple criteria, (2) to solve the so-called cold-start problem, (3) to take into account opinions of others.

II. LITERATURE SURVEY

[1] Nibedita Panigrahi and Asha. T. in 2018 proposed a method based on aspect level Sentiment analysis for rating the hotels, they used RHALSA (Ranking Hotels Using Aspect Level Sentiment Analysis) algorithm. The data was taken from some trip advisor's reviews. The proposed work was for two major aspects, i.e.cleanliness and hotel service. The Standard Core-NLP sentiment approach was used in the proposed method forsentiment analysis togenerate scores. The Sentiment Score is calculated based on the Stanford Core-NLP Sentiment Levels and depicted as: 0 for "Very Negative", 1 for Negative, 2 for Neutral, 3 for "Positive" and 4 for "Very Positive". The result shows with respect to each of the aspects that were considered on the same hotel are of mixed opinions. The RHALSA presented was limited to handle negative comments but can be extended to handle discourse relation which may change the orientation of the sentence.

[2] In this paper the authors created a sentiment classifier that uses Support Vector Machine algorithm. Using Unigram language model, a machine learning algorithm is trained which calculates the frequency of individual words. The TF-IDF was applied as weighing scheme and also for finding the occurrence of selected polarity words. The

tenfold cross validation was applied on dataset that divided the data into ten equal size folds that contain 180 reviews each. The cross-validation process was applied then. The proposed work excludes Extremely short (i.e. less than 30 words) and very lengthy (i.e. more than 250 words) reviews from the corpus. The authors got accuracy of TF-IDF as 95.78% and accuracy of TO-approach as 71.76%.22.

[3] People give reviews and rate hotels based on their opinions on a numerical score. There lie many aspects behind their reviews. People generally see the overall rating of the hotel but do not read all the reviews. So, aspect-specific sentiment analysis provides a good solution. A paper was presented on this concept, that used ILDA (Interdependent Latent Dirichlet Allocation) algorithm. The proposed method separates the vocabulary of reviews into header and modifier terms. For example, if a review is given as nice service in this service is the header term and nice is the modifier term. Header terms does not have sentiment polarity. The modifier terms lead to the sentiment polarity. The modifier terms can be changed by using the terms like nice, worst, good and so on but the header term cannot be changed. Modifier term gives the polarity while the header term gives the aspects.

The algorithm used is ARIH (Aspect and Rating Inference Using Hotel Specific Aspect Rating Priors).

Formally, we tend to outline an information corpus of reviewed N documents denoted by $D=\{x_1,x_2,x_D\}$. Every reviewed document x_D is created of a sequence of twenty sixWWW(2017) 20:23–37 of tokens. Each review x_d is associated with an overall rating r_d, which takes an integer value from 1 to S (S = 5). An aspect is a predefined property of a hotel, such as value, room, location, and service. A text review says opinions of the reviewers based on several aspects. The review comments on aspect value is indicated by the occurrence of the word price is an example. For every review there are several integer scores identified by$\{l_1, l_2, l_K\}$, where K is the number of aspects. Phrase We assume each review is a set of opinion p_h.

[4] In this paper author approach was classifying sentence as dependency tree sub trees. Classifying the sentence polarity based on sub opinions in thedependency tree with different aspects like sum-product belief propagation, sub-opinion relations, word granularity. This paper uses methodology of sentiment classification system based on dependency sub-trees. Corpus is dataset used andcorpus contain id and overall score of comment. Further sentence istokenized and parsed for analysis. Therefore, calculating recall and precision will give the results.

[5] one paper proposed sentiment classification is divided in to two approaches. One is machine learning approach and other approach is semantic orientation. In machine learning approach used bag of words which is predefined dataset and used for two steps of training and prediction. Semantic orientation approach deals with identifying the key words such that words describes the sentiment words in the document. As machine learning and semantic approach as different computation approaches in finding the sentiment in document or sentence.

[6] one of the research paper approach was text mining and topic modelling. Topic modelling is a type of statistical model is frequently used text mining approach for discovery of hidden semantic structures of text body. This paper uses topic modelling for finding frequency of words. The authors propose using topic modelling for sentiment classification as a possible future research direction to explore.

[7] one of paper proposed approach was classification using machine learning algorithms. Opinion mining is a method of extracting information from text processing by making analysis. For this we use two machine learning classification algorithms like Naïve Bayes and Decision tree. Although, Naïve Bayes is more reliable classification algorithm than decision tree, because when the size of data increases decision tree may not give accurate results as Naïve Bayes. For machine learning algorithms data set is divided into training set and testing set by making further analysis on this sets will produce results.

[8] one of research paper offered research as sentiment classification is divided into two ways. First approach is lexicon based which used for calculating orientation of a document of words or phrases and second approach is machine learning techniques such as clustering. For this applying TF-IDF gives the term frequency of words in a document. By calculating sentiment strength calculation then we will get certain values and those values clustering technique will be applied. By calculating Euclidean distance, we classify the words/text into positive, negative and neutral.

[9] In this paper the authors used a method called contradiction detection. the contradiction detection is identified through data analysis and pre-processing. They explained how numerical mismatch leads to contradiction by normalizing dates and time formats into a range of t and h. If a text is not in the range of t and h it is contradiction. The proposed contradiction focuses on data pre-processing and feature selection for accuracy.

III. PROPOSED SYSTEM

In the proposed work we are using machine learning algorithms for sentiment analysis for hotel business. The machine learning techniques can do work on large amount of data sets. The real-world data can contain noisy values in it, for that some cleaning process is required which is known as pre-processing step. That reduces undesirable data impact present in information which augmenting its data. The real-world data always requires to be clean and transform inorder to be usedby machine learning techniques. The pre-processing step includes sampling also to reduce big population intoprecise number of data-set. Somecommon pre-processing techniques are collaborative filtering, sampling, reducing dimensionalities, principal component analysis, singular value decomposition etc.,

A. Machine Learning

The goal of Machine Learning is to grasp the structure of the data and convert that knowledge into models which will be understood and utilized. Machine learning is the sub branch of artificial intelligence and the advance concept of machine learning is deep learning. Machine learning will be able to predict the future based on the past or historical data. The most commonly used Machine Learning algorithms are decision tree, k-means clustering, support vector machine, random forest, neural network.

B. Classification

In Machine Learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a trainingset of data containing user reviews. Examples are assigning a rating to the hotel based on positive and negative reviews. Classification is an example of patternrecognition. Classification comes under supervised learning, which learns a function and maps an input to the output. some of the classification algorithms are Naïve Bayes, decision trees etc.,

IV. METHODOLOGY

A. Naive Bayes:

Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability and the probabilities of observing various data.

P(h/D) = (P(D/h) * P(h) (1)

- A concept learning algorithm considers a finite hypothesis space 'h' defined over an instance space X.
- What can we do if our data *d* has several attributes?
- Naïve Bayes assumption: - Attributes that describe data instances are conditionally independent given the classification hypothesis.

$$P(d|h) = P(a_1,\ldots,a_T|h) = \prod_t P(a_t|h) \quad (2)$$

- It is a simplifying assumption, obviously it may be violated in reality
- In spite of that, it works well in practice
- The Bayesian classifier that uses the Naïve Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier
- One of the most practical learning methods

B. Natural Language Processing (NLP):

Natural Language Processing, or NLP for short, is extensively characterized as the programmed control of natural language, similar to speech and text, by programming. NLP is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to method and analyse massive amounts of language information. It is the branch of machine learning which is about analyzing any text and handling predictive analysis.

NLP is useful in hotel rating for dividing the sentences and determines whether the sentence is Positive, Negative and Neutral and also, NLP will also be useful as translator in case translation of one language to required language.

V. EXPERIMENTAL RESULTS

In this work we implemented machine learning algorithms On the data collected from various websites such as trip advisor, trivago etc., The Naïve Bayes classifier is used to classify the sentiments into two broad categories i.e., positive and negative. The dataset contains sentimental statements from various reviews.

Steps:

1. Create a dictionary that has the data with labels(positive/negative).
2. Create the dictionary, feature set = { }, that contains the count of occurrences of word under each label.
3. Split the sentences with respect to non-characters and key words.
4. Use Naïve Bayes classifier with respect to non-characters and key words.

Weuse 60% data as training set and 40% data as testing set. The classifier was perfectly able to find out positive and negative classes on the basis of lexicon values. In lexicon we used 45 key values for positive and 45 for negative.

Figure 1 Naive Bayes results

In the figure 1 shows the classification of the reviews as positive or negative. For this the input takes five sentences and classify each of the review. Finally it shows the total number of positive ratings and negative ratings.

Figure 2 NLP result1

For figure 2 input is a single sentence and it calculates the overall sentiment dictionary score for the positive, negative, and neutral. It calculates the rating of positive, negative, neutral with the help of compound rate. Finally it gives the review overall rating as positive or negative or neutral.

Figure 3NLP result2

As mentioned in figure 2 here in figure 3 we have given different reviews as input and calculated the overall review rating.

VI. CONCLUSION

The content-based recommender implies matching of attributes from a user profile in which preferences and interest are stored with attributes of content object. If a string, or some morphological variant, is found in both the profile and the document, a match is made and the document is considered as relevant. The major limitation is matching the keywords and the overall vocabulary. The Naïve Bayes classification algorithm is good for scaling the dataset and implement the linear equation on features and predicators. The classifier got some mismatched values for neutral results.

REFERENCES

1. Nibedita Panigrahi and Asha T, ”RHALSA: Ranking Hotels using Aspect Level Sentiment Analysis”,
2. George Markopoulos, George K. Mikros, Anastasia Iliadi “Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features Article”, pg:-373-383
3. by Wei Xue, Tao Li, Naphtali Rishe, “Aspect identification and ratings inference for hotel reviews”, pg:-24-37
4. Tran Sy BANG and Virach SORNLERTLAMVANICH, “Sentiment Classification for Hotel Booking Review Based on Sentence Dependency Structure and Sub-Opinion Analysis”, pg:-910-916
5. Youngseok Choi & Habin Lee, “Data properties and the performance of sentiment classification for electronic commerce applications”,pg:-994-1012
6. Sentiment classification of consumer generated online reviews using topic modeling
7. Wararat Songpan, “The Analysis and Prediction ofCustomer Review Rating Using Opinion Mining”, pg:-71-77
8. Sumbal Riaz, Mehvish Fatima, M. Kamran, M. Wasif Nisar, “Opinion mining on large scale data using sentiment analysis and k-means clustering”
9. Siti Nuradilah Azman et al, “Towards an Enhanced Aspect-based Contradiction Detection Approach for Online Review Content”
10. Francesco Ricci · Lior Rokach · Bracha Shapira · Paul B. Kantor(Editors), Recommender systems handbook, springer, 2011.

AUTHORS PROFILE

P.Sanjay Bhargav, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. He is doing his research work in knowledge engineering.

G.Nagarjuna Reddy, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. He is doing his research work in knowledge engineering.

R.V.Ravi Chand, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. He is doing his research work in knowledge engineering.

K. Pujitha, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. She is doing his research work in knowledge engineering.

Dr.Anjali Mathur,she is an Associate Professor of Computer Science and Engineering Department of KL University. Her research interest are Artificial Intelligence, Machine Learning and Big Data Analytics. She completed her Ph.D. on Artificial Neural Networks.