Customer Review Rating Analysis Using Opinion Mining

K. Sowmya, K. Monika, M. Radha, V. Vijay Kumar

Abstract- The customer ratings and reviews is very important to the service providers. The customer rating will act as a feedback to the service provider. Sometimes, the customer may give the good review but he may give the bad rating to the service. So, the service provider will be in a confusion. So, we should predict the rating with the help of customer review. It can be done with the help of optional mining. We used logistic regression, Naive Bayes, SVM algorithms. We applied these algorithms on the data set containing of 1500 reviews and ratings of the customer. When we see above three algorithms logistic regression is giving 80.82% accuracy, Naive Bayes is giving 67.6% accuracy, where asSVM is giving 80.80% accuracy. When we compare the above classification algorithms accuracy logistic regression and SVM are having good accuracy and better performance.

Key Words: Opinion mining, Stop words, Reviews, Positive words, Negative words.

I. INTRODUCTION

As days passing the technology is increasing day by day. And the no of people using the technology is also increasing day by day. Due to this lot of data is generating abnormally. In olden days if any person wants to know about any particular place or thing, they used to ask some of the experienced persons. They used to have very less feedback. But now we are having lot of feedback in the form of comments and ratings. But this is not the exact solution of find the correct feedback or information about the required aspect. The correctness is the main required thing that people of this generation is needed.

If a person wants to know get the overall ideal or to know the overall range of the required products, he/she need to see the reviews and ratings of the customer who already experienced the product. But the real obstacle comes here, the customer needs to read each and every review and ratings of the previous customer and some of the reviews doesn't match with the ratings.

This can be dealt with an example, one of the customers commented "the rooms are very clean and neat". However, the customer given only 4.5 out of 10 which is practically not fair. Here there is another example where the customer commented "I am not satisfied with food and accommodation" and given 9 out of 10.

Revised Manuscript Received on May 06, 2019

K. Sowmya, is pursuing B.Tech in Computer Science and Engineering at Koneru Lakshmaiah educational.

K. Monika, is pursuing B.Tech in Computer Science and Engineering at Koneru Lakshmaiah Educational Foundation.

M. Radha, M. Tech, is an Assistant Professor at KoneruLakshmaiah Educational Foundation in Computer Science and Engineering Department.

V. Vijay Kumar, M. Tech, is an Assistant Professor at KoneruLakshmaiah Educational Foundation in Computer Science and Engineering Department.

Any person who read the comment will understand the bias. So, with the help of machine learning algorithms we can find the customer ratings from the reviews.

II. PROPOSED METHODOLOGY

The proposed methodology used reviews from many customers who visits different hotels and book rooms and order food. We have to mainly concentrate on the reviews and ratings that are given by the customer. Here we have taken around 1500 customer reviews from the hotels. In [1] they have also calculated the accuracy with the help of different classification algorithm like decision tree and naïve bayes. But they have only taken 400 data sets as their inputs. So we get better results than [1]. To find the rating from the customer review we have several steps that involves here. First of all we have to collect the data from required hotels. The data set should have the customer review, customer rating, details of the hotel etc. Next, we have to pre-process the dataset. While data pre-processing the data, we will remove the stop words. Means, we collect the data in the raw format and this raw data is not suitable for the data analysis. So, we do pre-processing to convert the raw data into clean data. Now, we will calculate the frequency of the frequently occurring words. Then with the help of machine learning algorithms we will access the customer review as positive or negative.

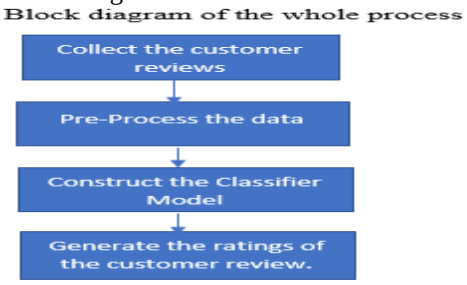

Fig 1. Block diagram of whole process

Flow chart of the construction of the model classifier

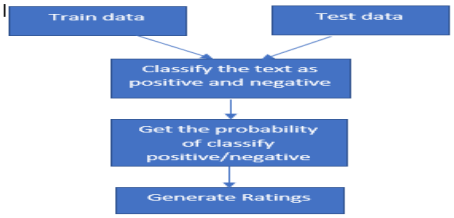

Fig 2. Classification block diagram.

Pre-processing

In pre-processing we will remove the unwanted data like the, a, for etc. these kind of words are called as stop words. In pre-processing we will remove the punctuations, numbers, stop words, whitespaces etc. Now we have to grab the most frequent words that are used in the reviews. We have to calculate the individual word frequency. We have to see the top 200 words including both positive and negative.

```
          word.freq[1:15]
the                  1652
food                 1197
good                 1007
place                 938
great                 694
service               632
time                  526
back                  478
ordered               361
restaurant            353
this                  337
order                 334
chicken               331
dont                  320
menu                  303
```

Fig 3, list of individual word frequency.

Below Fig 4, are the top 200 terms with highest correlation magnitudes those includes both positive and negative.

```
  [1] 121 110 268  80  76 172 203 586 453 741 587 492  69 242 673 526 606 560 489 127 506 300 598 309 656 596 604 290 710 742  57 646 539 177 781 661 561 143 262  14
 [41] 454 313 295 156  72 241 328 250  28 570 118 752 112 424  51 138 223 333 563 638 716 678 797 176  44 782 726  81 519 135  39 214 304 115 113 267 480 583 162 614
 [81] 294  84 552 418 452  31  95 768 103 307  61 767 578 648 279  24 740 755 610  89 315  30 306 466  46 438 298 738 374 245 590 224 386 591 653 147 650  91 566 403
[121] 579 621 723  82 419 680  48 582 145 213  92 744 173  79 617 736 749 349 441 721 527 289 473 568  78 330 663 417 675 329 327 285  86 803 576 228 537 261 416   7
[161] 608 530 771 404 332 750 355 628 715  98 544 531 373 667 735 711 554 437 696 191 247 387 380 426 108 654 639 233 343 240 801 737 144 371 146  70 525 253 148 676
> top200words = colnames(absCorr)[top200]
> top200words
  [1] "clean"        "chicken"      "fries"        "burger"       "brunch"       "dessert"      "eggs"         "sandwich"     "night"        "vegas"
 [11] "sandwiches"   "party"        "bowl"         "filling"      "table"        "pork"         "service"      "reservation"  "pancakes"     "coffee"
 [21] "pho"          "gyro"         "seated"       "hash"         "steak"        "seasoned"     "server"       "great"        "toast"        "vegetables"
 [31] "birthday"     "split"        "pulled"       "dinner"       "wine"         "strip"        "reservations" "counter"      "fresh"        "ambiance"
 [41] "noodles"      "healthy"      "group"        "curry"        "breakfast"    "filet"        "hot"          "flavor"       "atmosphere"   "rice"
 [51] "choose"       "waiter"       "chinese"      "menu"         "beer"         "cool"         "expensive"    "however"      "restaurant"   "sour"
 [61] "top"          "talk"         "year"         "dining"       "bbq"          "wings"        "turkey"       "burgers"      "plate"        "completely"
 [71] "bar"          "evening"      "hands"        "chocolate"    "chipotle"     "friends"      "our"          "salsa"        "date"         "show"
 [81] "grilled"      "busy"         "reasonable"   "meats"        "nice"         "authentic"    "cant"         "weekend"      "charlotte"    "happy"
 [91] "bland"        "week"         "run"          "staff"        "generous"     "arrived"      "vegan"        "walk"         "shared"       "cake"
[101] "heard"        "attentive"    "happened"     "onion"        "beans"        "mom"          "guy"          "valley"       "left"         "fine"
[111] "saturday"     "experience"   "live"         "sauce"        "start"        "cream"        "standard"     "california"   "review"       "mac"
[121] "sad"          "simply"       "true"         "burrito"      "mediocre"     "taste"        "beautiful"    "salmon"       "crab"         "even"
[131] "call"         "veggie"       "desserts"     "bun"          "sides"        "upscale"      "visiting"     "italian"      "months"       "traditional"
[141] "portion"      "greasy"       "option"       "rib"          "buffet"       "hour"         "stuff"        "meat"         "taco"         "hotel"
[151] "hostess"      "glass"        "butter"       "youll"        "room"         "fact"         "prime"        "french"       "means"        "alcohol"
[161] "set"          "potatoes"     "went"         "made"         "house"        "wait"         "job"          "slices"       "tonight"      "chain"
[171] "rare"         "prepared"     "leave"        "super"        "unique"       "recently"     "mixed"        "there"        "drinks"
[181] "finished"     "lived"        "light"        "met"          "cheese"       "started"      "special"      "fancy"        "including"    "felt"
[191] "you"          "usual"        "couple"       "las"          "craving"      "boyfriend"    "popular"      "floor"        "crisp"        "tacos"
>
```

```
> top.pos # Top 15 Positive Words
  counter     even happened  noodles  filling    clean charlotte   valley   months   simply  recently    meats    fancy including vegetables
3.314698 3.002619 2.881763 2.823072 2.574088 2.531568 2.299976 2.166588 1.969091 1.939380 1.838152 1.771822 1.758483 1.695223  1.668968
```

Fig 5, Top 15 frequent positive words

```
> top.neg # Top 15 Negative Words
     split     brunch reservation      toast   pancakes      hands completely california     slices       date        mom    weekend       rare
-17.640835 -17.563211  -17.221507 -17.220389 -15.004308  -7.832615  -5.376735  -4.444019  -3.942995 -3.053012 -2.939471 -2.565880 -2.400620
     filet       cant
-2.364994  -2.343717
```

Fig 6, Top 15 frequent negative words

Now we have to take 15 top frequent positive words and 15 top frequent negative words as above.

III. Model Construction

Here we have used 3 classifier model constructors. They are logistic regression, naïve bayes, SVM. Here we have trained data set and tested data set. Here, we have 80 percent of the data as trained data and 20 percentage of the data as tested data.

The above mention classifiers are described below in detail.

1 Logistic regression

Logistic regression is one of the machine learning algorithms which is used to predict a value based on the previous data. It is one of the classification models. By analysing the relationship between one or more independent variables it predicta dependent data.

The logistic regression outcome will have only two values. Those two values are 0 or

1. The main idea behind is to estimate the probability of the outcome will be either 0 or 1. If we consider the probability of a particular situation occurrence is p. then the opposite situation will have 1-p probability

2 Naïve bayes

Naïvebayes algorithm is one of the classificationalgorithm where it will classify the posterior probability where we have prior probability, likelihood, predictor prior probability.

Naive bayes posterior probability equation is given below.

$$P(\frac{A}{B}) = \frac{P\left(\frac{B}{A}\right).P(A)}{P(B)}$$

Here, P(A/B) is the posterior probability of (A, target) given predictor (B, attributes).
P(A) is the prior probability.
P(B/A) is the likelihood
P(B) is predictor probability.
Naïve bayes takes the training data set and classify the data. Naïve bayes have several steps to implement.Inthe naïve bayes algorithm first we will convert all the data set into a frequency
table. Then with the help of probabilities we will create likelihood table. Then posterior probability is calculated with the help of Naïve Bayesin equation.
3 Support Vector Machine
Support Vector Machine is one of the supervised learning algorithms. Supervised learning algorithm will come under the context of artificial intelligence and machine learning algorithm concepts. SVM will come under this context. This Support vector machine algorithm will analyse data used for classification and regression analysis. Mainly SVM algorithm will analyse weather the model is working correctly or not.
The main theme of the Support Vector Machine is to find a hyper plane in an n-dimensional space which classifies the data points differently.

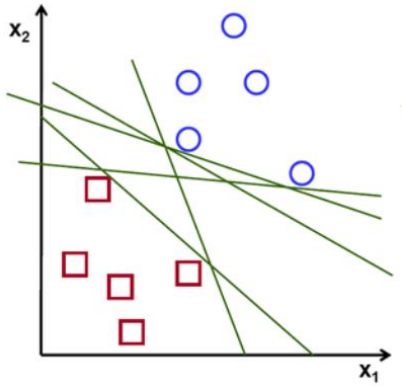

Fig 7, possible hyper plane 1

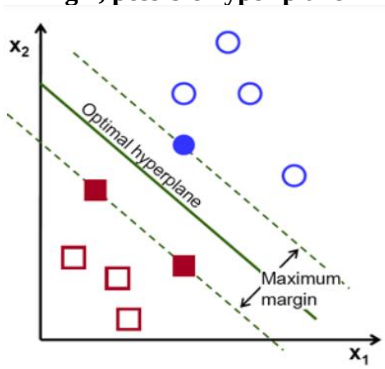

Fig 8. Possible hyper plane 2

The SVM classification algorithm is to find a plane which have maximum margin. Here, maximum margin means the maximum distance between the data points of both classes.

IV. EXPERIMENTAL RESULTS

These results are tested with review texts which are taken from 1500 service utilizers in different hotels and these are taken from their respective websites. Logistic regression which is having 80.82% accuracy and SVM which is having 80.80% accuracy are having better accuracywhen we compare it with Naïve Bayes which is having 67.6%.
Here we are representing the top frequently used words in the form of graphical representation.

Fig 9, graphical representation of frequent words.

Here, in the above figure the frequency varies with the size of the word present in the output. The font size of the word is directly proportional to the frequency of the word in the reives.

Attribute	Logistic Regression	Naïve Bayes	SVM
Accuracy	80.82%	67.6%	80.80%

Table 1, Accuracy % table

V. Conclusion

Here, We calculated the accuracy of the three different machine learning classification algorithms. Every algorithm will have its own benefits but they may vary according to the attributes which we have taken. So here, logistic regression is preforming well. To get these required outputs we have first collected data from online hotel websites, we pre-processed the data, then classification algorithms are used. And finally, we calculated the accuracy.

REFERENCES

1. WararatSongpan, “The Analysis and Prediction of Customer Reivew Rating Using Opinion Mining”, 2017 IEEE SERA 2017, June 7-9, London, UK.
2. I. K. C. U. Perera ; H.A. Caldera, “Aspect based opinion mining on restaurant reviews”, 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)
3. Chi Cho Hnin, NawNaw, Aung Win, “Aspect Level Opinion Mining for Hotel reviews in Myanmar Language”, 2018 IEEE International Conference on Agents (ICA)
4. K L Santhosh Kumar, Jayanti Desai, Jharna Majumdar, “Opinion mining and sentiment analysis on online customer review”, 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)
5. MaruthiPrithivirajan, Vivian Lai, Kyong Jin Shim, Koo Ping Shung, “Analysis of star ratings in consumer reviews: A case study of Yelp”, 2015 IEEE International Conference on Big Data.
6. Humaira G. Momin, Chandrabhaga S. Kondhawele, Rubina E Shikh, Kiran G. Gawandhe, “Feature based evaluation of hotel reviews and ratings for differently abled people”, 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)

7. Vijay B. Raut, D.D. Londhe, “Opinion Mining and Summarization of Hotel Reviews”, 2014 International Conference on Computation Intelligence and Communication Networks.
8. A. Angelpreethi, S. Britto Ramesh Kumar, “An Enhanced Architecture for Feature Based Opinion mining from product reviews”, 2017 World Congress on Computing and Communication Technologies(WCCCT).
9. Santhosh Kumar K L, Jayanthi Desai, Jharna Majumdar, “Opinion Mining and Sentiment Analysis on Online Custome review”, 2016 IEEE International Conference on Computational Intelligence and Computing Research.
10. K. M. Azharul Hasan, Mir Shahriar Sabuj, Zakia Afrin, “Opinion Mining using Naïve Bayes”, 2015 IEEE International WIE Conference on Electrical and Computer Engineering(WIECON-ECE)

AUTHORS PROFILE

K. Sowmya, is pursuing B.Tech in Computer Science and Engineering at KoneruLakshmaiah educational.

K. Monika, is pursuing B.Tech in Computer Science and Engineering at Koneru Lakshmaiah Educational Foundation.

M. Radha, M. Tech, is an Assistant Professor at KoneruLakshmaiah Educational Foundation in Computer Science and Engineering Department.

V. Vijay Kumar, M. Tech, is an Assistant Professor at KoneruLakshmaiah Educational Foundation in Computer Science and Engineering Department.