REVIEWS ANALYSIS USING GAUSSIAN NAÏVEBAYES IN MACHINE LEARNING

Kalakonda Shashank[1], Anumandla Sahithya[2], Shaik Shakeel[3], Dr. R. Lakshmipriya[4]

[1,2,3] B. TECH Scholars, Dept. of Computer Science and Engineering Hyderabad-501301, India
[4] Associate Professor, Dept. of Computer Science and Engineering, SNIST, Hyderabad-501301, India

---***---

Abstract - *In our day-to-day life reviews plays a crucial role in every business aspect. To increase business scale, the productmust have positive reviews from the users. There are hundreds of similar products in every business. To avoid his/her difficulty customer always checks the reviews whether it satisfies them or not, and they buy the products only after it satisfies them. So, owner always should observe their reviews and do changes according to them. If there are below hundred reviews, owner had possibility to read those reviews and determine whether they are positive or not. If there are thousands of reviews owner cannot read all of them. So, to overcome this problem, we are using Machine Learning concept Gaussian naïve bayes and predict the positive or negative reviews. A dataset consisting of reviews taken as inputs and preprocessing the data and applying the algorithm. A spreadsheet consisting of reviews with their respective values (0/1 1 for positive and 0 for negative) is the output.*

***Key Words*: Sentimental Analysis, Machine Learning, Gaussian Naïve Bayes, Restaurant reviews.**

1. INTRODUCTION

Platforms such as Swiggy, Zomato, Dineout, allows people to comments, outlooks, passions, judgements on many topics like from food taste to ambience. These platforms contain the huge quantum of the data in the form of textbook, blogs, and updates on the status, posts, etc. Sentiment Analysis aims to determine the opposition of feelings like happiness, anguish, grief, abomination, wrathfulness and affection and opinions from the textbook,reviews, posts which are available online on these platforms. Opinion Mining finds the sentiment of the textbook with respect to a given source of content. Emotion analysis is complicated because of the shoptalk words, misspellings, short forms, repeated characters, use of indigenous language and new forthcoming emoticons. So it's a significant task to identify applicable sentiment of each word. Sentiment Analysis is one of the most active exploration areas and is also Extensively studied in data mining. Sentiment analysis is applied in nearly every business and social sphere because opinions are central to utmost mortal conditioning & actions. Sentiment analysis is veritably popular because of its effectiveness. Thousands of documents can be reprocessed for sentiment analysis. Since it's an effective process which provides good accurateness, thus it has colorful operations Purchasing Merchandise or Service while picking up a goods or service we must take a right decision which isn't a tough task presently. By sentiment analysis, people can effortlessly estimate reviews and opinions of any object or service and can painlessly compare the racing brands.Quality enhancement in Product or Service By Opinion mining, the directors can collect the stoner's opinion whether favorable or not about their product or service and also they can enhance and upgrade the quality of their product or service. Recommendation Systems By breakingdown and grading the people's opinion according to their preferences and interests, the system can forecast which item should be recommended and which one should not be recommended. Decision Making People's sentiments, ideas, passions are veritably important factor to make a decision. While buying any item be it book or clothes orelectronic particulars stoner's first to read the opinions and reviews of that particular product and those reviews have a great impact on stoner's mind. Marketing exploration, the result of sentiment analysis ways can be employed in marketing exploration. By this Fashion, the station of consumers about some product or services or any new government policy can be anatomized. Discovery of honey, the monitoring of newsgroups, blogs and social media is fluently possible by sentiment analysis. This fashion can determine bold, arrogant, over heated words used in tweets, posts or forums and blogs on the internet.

There are following phases of Sentiment Analysis:

Pre-Processing Phase: The data is to be cleaned first to reduce noise.

Feature Extraction: A token is given to the keywords and this token is now put under analysis.

Classification Phase: Based on different algorithms these keywords are put under certain category.

2. BACKGROUND STUDY (LITERATURE)

Customer satisfaction is an essential concern in the field of marketing and examination in terms of consumer actions. As in the habits of hostel consumers when they get excellent service, they will transmit to others mouth to mouth. Text mining or retrieval of data from a collection of documents stores constantly with the help of analysis tools or manuals. Through the analysis process of several text mining perspectives, information can be produced that can be used to increase gains and services. Sentiment analysis is used to find opinions from the author about a specified reality. Sentiment analysis of a review is an opinion investigation of a product. The base of sentiment analysis is using Natural Language Processing (NLP), text analysis and some computational portions to extract or forget gratuitous corridor to see the pattern of the judgment negative or positive. In the 18th century, Reverend Thomas Bayes developed a method known as Naive Bayes that used probability and opportunity approaches. Naive Bayes calculates future probability prognostications from data or gests that have been given, grounded on the occasion point of view. One characteristic of the Naive Bayes Bracket is the existence of independentinput variables which assume the presence of an articular point from a class that's mutually independent to other features. The major step involved in determining the sentiment of a textbook. In our approach, we've resolved the preprocessing part into three major ways. The first step involves removing the punctuation in the sentences. All special characters like exclamatory mark and quotations are removed by designing applicable regular expression. The attendant data would be containing only alphabetical characters. Removing the stop- words from collected reviews can be done in second step. Stop- words are the words which aren't used to express any emotion or sentiment but used as connectors or papers in the English language. This includes words like and, with, of, the. Natural language processing (NLP) ways like verbal analysis, syntactic analysis, semantic analysis, exposure integration, and realistic analysis are applied on the datasetto identify and remove stop- words. Step words are not well removed by semantic analysis. But, in opinion mining, the presence/ absence of the word not plays an important part. For illustration, the review says the crust isn't good. By removing of stop words, it will result the sentences into crust good, a negative opinion will turn to positive. To avoid this problem, we've modified the semantic analysis step in NLP and made sure that similar stop- words aren't being removed in the process. The third step in preprocessing is to convert the original words to their root words (no prefix or suffix). For illustration, love is the root word for the words loving, loved, loves, etc. As we're interested only in factual opinion sentiment rather than English alphabet, Similar conversion eases the job. The Porter Stemmer algorithm is applied for converting all words in the dataset into root words.

3. METHODOLOGY

OBJECTIVE

This design focusing on the estimation of the contrariness of the sentiment evoked by an text through input box. To apply an algorithm for automatic classification of text into positive, negative or neutral. Sentiment Analysis is determined to check the reviews positive or negative or neutral.

PROBLEM STATEMENT

To give a Sentiment Analysis system for customers review classification, that may be helpful to analyze the information where opinions are largely unstructured and are either positive or negative.

EXISTING SYSTEM

The content of user generated opinions in the social media such as face book, twitter, review spots, etc are growing in large volume. These opinions can be tapped and used as business intelligence for various uses similar as marketing, prediction, etc. Generally, sentiment analysis is used for finding out the aptitude of the author considering some content. But in our social network spots not implemented Sentiment analysis. Some survey depends on the static transferred word dataset to find the sentiment analysis but we require finding a proper result to find the polarity of the micro blogs.

PROPOSED SYSTEM

We'll collect the unstructured data through the text box. With that data covert the data to lower case and data is reused as follow.

Preprocessing:

Before the feature extractor can use the reviews to make feature vector, the review text goes through preprocessing step where the following steps are taken. These stepsconvert plain text of the review into process suitable rudiments with further information added that can be employed by feature extractor. For all these way, third partytools were used that were specialized to handle unique nature of review text. PyCharm is an integrated development environment (IDE) which is used in computer programming. Specifically for the Python language. It's developed by the Czech company JetBrains.

Step 1 Tokenization:

Converting of large string into smallest individual elements is called Tokenization. In the environment of a review, these elements can be words, emoticons, url links, hashtags or punctuations “an insanely awsum” Text was broken into “an”, “insanely”, “awsum”. These elements are frequently separated by separated by a space. On the other hand, hash markers with ”#” preceding the label needs to be retained since a word as a hash label may have different sentiment value than a word used regularly in thetext.

Step 2 Parts of Speech Tags:

Parts of Speech (POS) tags are characteristics of a word in a judgment based on grammatical orders of words of language. This information is essential for sentiment analysis as words may have different sentiment value depending on their POS label. For illustration, word like “good” as a noun contains no sentiment whereas “good” as an adjective positive sentiment.

Step 3 Dependency Parsing:

For our purposes, dependence parsing is embedding the relationship between words in a judgment. This can be useful in relating relationship between "not" and "good" in expressions like "not really good" where the relationship isn't always with the adjacent word.

4. ALGORITHM

Gaussian Naive Bayes is a probabilistic algorithm based onclassification with strong independence hypotheticals.

In this classification, independence refers to the idea that the presence of one value of a point does not impact the presence of another (unlike independence in probability proposition). Naive refers to the use of a supposition that the features of an object are independent of one another.

In the terrain of machine literacy, naive Bayes classifiers are known to be largely suggestive, scalable, and nicely accurate, but their performance deteriorates swiftly with the growth of the training set. Utmost especially, they do not bear any tuning of the parameters of the type model, they gauge well with the size of the training data set, and they can easily handle continuous features.

$$P(X|Y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

5. ARCHITECTURE

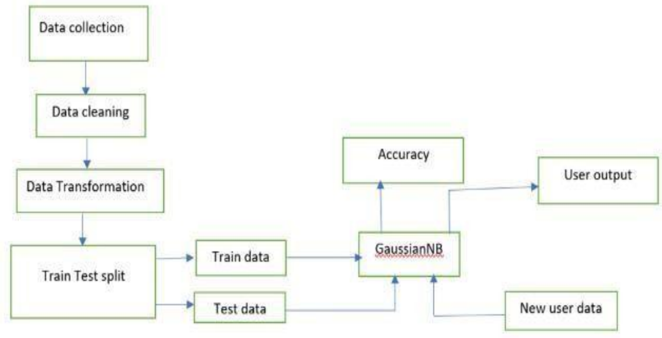


Fig 5.1. Architecture diagram

Step 1: Collecting the data from restaurants.

Step 2: Cleaning the data by removing special characters, stop words .

Step 3: Converting the data into Lower case.

Step 4: splitting the data into training and testing data and label the training data.

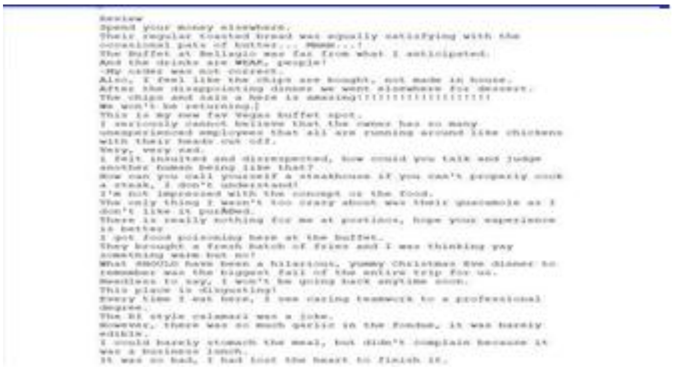
Step 5: Apply train data into GaussianNB. Now the modelis trained. Apply the test data into the model.

Step 6: A new user data is applied into model and theoutput is displayed.

Step 7: Accuracy of model is to be find using accuracy score() method.

6. IMPLEMENTATION

6.1 Collecting data from restaurants about their reviews they have got from customers.



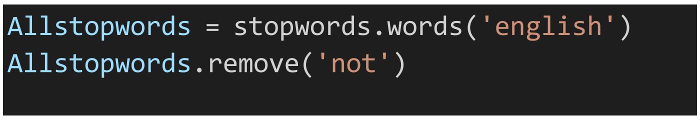
```
Review
Spend your money elsewhere.
Their regular toasted bread was equally satisfying with the
occasional pats of butter... Mmmm...!
The Buffet at Bellagio was far from what I anticipated.
And the drinks are WEAK, people!
-My order was not correct.
Also, I feel like the chips are bought, not made in house.
After the disappointing dinner we went elsewhere for dessert.
The chips and sals a here is amazing!!!!!!!!!!!!!!!!!!!
We won't be returning.
This is my new fav Vegas buffet spot.
I seriously cannot believe that the owner has so many
unexperienced employees that all are running around like chickens
with their heads cut off.
Very, very sad.
i felt insulted and disrespected, how could you talk and judge
another human being like that?
How can you call yourself a steakhouse if you can't properly cook
a steak, I don't understand!
I'm not impressed with the concept or the food.
The only thing I wasn't too crazy about was their guacamole as I
don't like it puréed.
There is really nothing for me at postinos, hope your experience
is better
I got food poisoning here at the buffet.
They brought a fresh batch of fries and I was thinking yay
something warm but no!
What SHOULD have been a hilarious, yummy Christmas Eve dinner to
remember was the biggest fail of the entire trip for us.
Needless to say, I won't be going back anytime soon.
This place is disgusting!
Every time I eat here, I see caring teamwork to a professional
degree.
The RI style calamari was a joke.
However, there was so much garlic in the fondue, it was barely
edible.
I could barely stomach the meal, but didn't complain because it
was a business lunch.
It was so bad, I had lost the heart to finish it.
```

Fig 6.1 Dataset

6.2 Remove all special characters like @, #, &, $, etc., for better understanding.

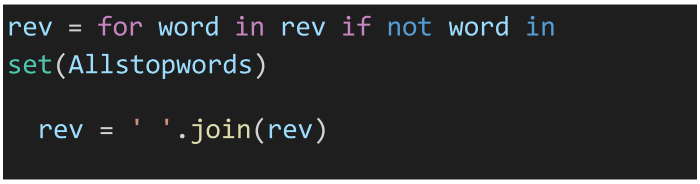
6.3 Convert all reviews into same case (lower/upper) by using reviews.lower() /reviews.upper() method.

6.4 Remove all stopwords from reviews except 'not'.



```
Allstopwords = stopwords.words('english')
Allstopwords.remove('not')
```

It downloads the package of all stopwords in English .



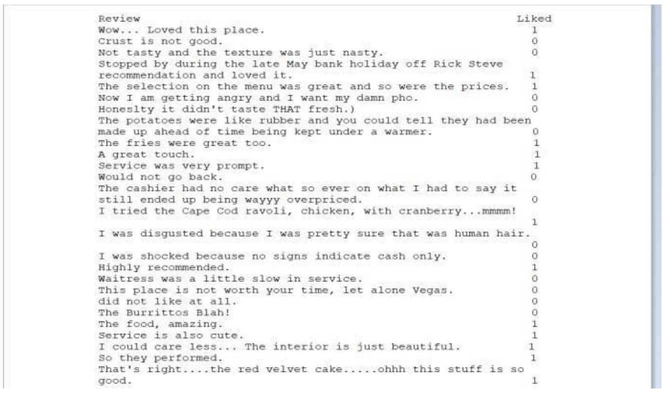
```
rev = for word in rev if not word in
set(Allstopwords)

  rev = ' '.join(rev)
```

We get the all the reviews without stopwords.

6.5 Now divide the reviews as training data and testingdata using train test split function.

6.6 For training data add one more column (+/-) for each review.



```
Review                                                          Liked
Wow... Loved this place.                                          1
Crust is not good.                                                0
Not tasty and the texture was just nasty.                         0
Stopped by during the late May bank holiday off Rick Steve
recommendation and loved it.                                      1
The selection on the menu was great and so were the prices.       1
Now I am getting angry and I want my damn pho.                    0
Honeslty it didn't taste THAT fresh.)                             0
The potatoes were like rubber and you could tell they had been
made up ahead of time being kept under a warmer.                  0
The fries were great too.                                         1
A great touch.                                                    1
Service was very prompt.                                          1
Would not go back.                                                0
The cashier had no care what so ever on what I had to say it
still ended up being wayyy overpriced.                            0
I tried the Cape Cod ravoli, chicken, with cranberry...mmmm!
                                                                  1
I was disgusted because I was pretty sure that was human hair.
                                                                  0
I was shocked because no signs indicate cash only.                0
Highly recommended.                                               1
Waitress was a little slow in service.                            0
This place is not worth your time, let alone Vegas.               0
did not like at all.                                              0
The Burrittos Blah!                                               0
The food, amazing.                                                1
Service is also cute.                                             1
I could care less... The interior is just beautiful.              1
So they performed.                                                1
That's right....the red velvet cake.....ohhh this stuff is so
good.                                                             1
```

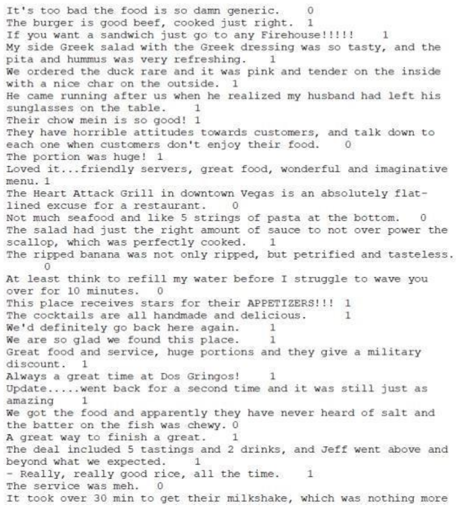
6.7 Now import Gaussian Naïve Bayes model form sklearn.

```
from sklearn.naive_bayes import
GaussianNBmodel = GaussianNB()

    model.fit(X_train, y_train)
```

6.8 Add training data into model to train model .

6.9 Now testing data is applied on model to get the outputof testing data.



```
It's too bad the food is so damn generic.    0
The burger is good beef, cooked just right.  1
If you want a sandwich just go to any Firehouse!!!!!    1
My side Greek salad with the Greek dressing was so tasty, and the
pita and hummus was very refreshing.    1
We ordered the duck rare and it was pink and tender on the inside
with a nice char on the outside.  1
He came running after us when he realized my husband had left his
sunglasses on the table.    1
Their chow mein is so good! 1
They have horrible attitudes towards customers, and talk down to
each one when customers don't enjoy their food.    0
The portion was huge!  1
Loved it...friendly servers, great food, wonderful and imaginative
menu. 1
The Heart Attack Grill in downtown Vegas is an absolutely flat-
lined excuse for a restaurant.    0
Not much seafood and like 5 strings of pasta at the bottom.   0
The salad had just the right amount of sauce to not over power the
scallop, which was perfectly cooked.    1
The ripped banana was not only ripped, but petrified and tasteless.
     0
At least think to refill my water before I struggle to wave you
over for 10 minutes.   0
This place receives stars for their APPETIZERS!!! 1
The cocktails are all handmade and delicious.     1
We'd definitely go back here again.     1
We are so glad we found this place.     1
Great food and service, huge portions and they give a military
discount.  1
Always a great time at Dos Gringos!     1
Update.....went back for a second time and it was still just as
amazing    1
We got the food and apparently they have never heard of salt and
the batter on the fish was chewy. 0
A great way to finish a great.    1
The deal included 5 tastings and 2 drinks, and Jeff went above and
beyond what we expected.    1
- Really, really good rice, all the time.    1
The service was meh.   0
It took over 30 min to get their milkshake, which was nothing more
```

Fig 6.3. Output

6.10 Accuracy of model can be find usingaccuracy_score() method.

```
pr= classifier.predict(xtst)


from sklearn.metrics import
confusion_matrix,accuracy_score

output =
confusion_matrix(xtst,pr)
print(output)



accuracy_score(xtst,pr)
```

72.777777

7. Usecase diagram

Use case diagrams helps in summarizing the system user's data and their interactions with system. Use case are used to represent system and user interactions. It is used in defining and organizing the functional requirements in a system.

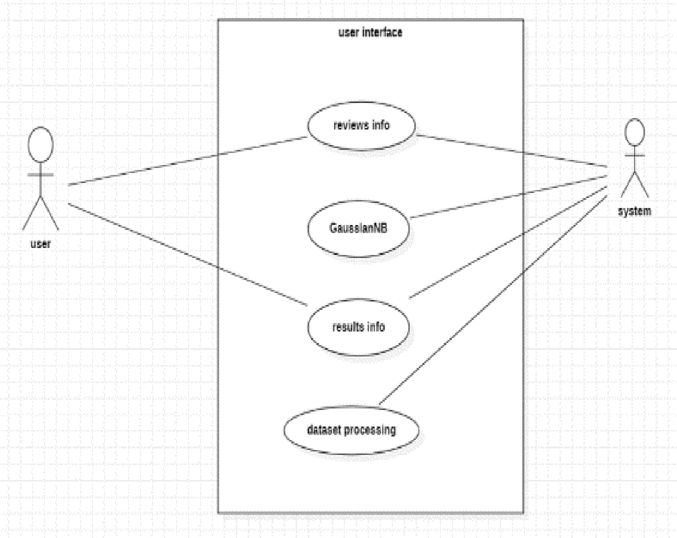


Fig 7.1. Use case diagram

Sequence diagram

Sequential diagrams are one of the interaction diagrams that deals thith how operations are carry out. The interaction between objects are captured in the context of collaboration.

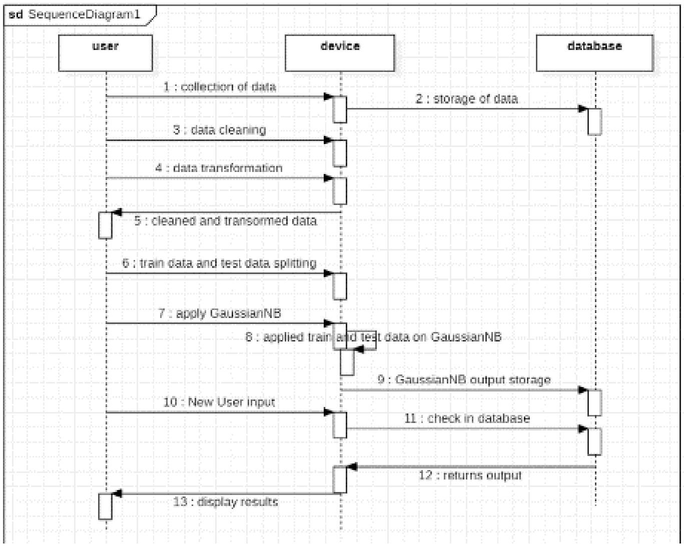


Fig 7.2 . Sequence diagram

8. ADVANTAGES OF SYSTEM

The main theme of this project is to make owner of specific businesses ease of finding the reviews positive or negative.This will help your owner find negative reviewsfast (indicates 0 binary value) and can overcome those reviews.

9. CONCLUSION

By observing the results of this paper that has done on reviews taken from restaurant by using Gaussian Naive bayes we found we can able to understand the sentiment from customers. The accuracy given this model is 72.77%.Further research can be done by taking more number of reviews or variety of data or by using different models to increase the accuracy.

10. REFERENCES

[1] R. Sarno and M. Fikri," A relative Study of Sentiment Analysis using SVM and SentiWordNet," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 13,no., 2019.

[2] B. Rintyarna, R. Sarno, and C. Fatichah," Enhancing the performance of sentiment analysis task on product reviews by handling both original and global terrain," International Journal of Information and Decision lores, vol., 2018.

[3] D. L. Gupta, A. K Malviya, and S. Singh," Performance analysis of classification tree learning algorithms," International Journal of Computer operations, vol. 55,no., 2012.

[4] B. Agarwal, N. Mittal and P. Bansal," novelettish analysis environmental information," Computational intelligence and neuro wisdom, vol. 2015, p. 30, 2015.

[5] K. Ilieska," customer satisfaction index-- as a base for strategic marketing operation," TEM Journal, vol. 2, no. 4,p. 327, 2013.

[6] Mikel Joaristi, Edoardo Serra, Francesca Spezzano “assessing the Impact of Social Media in Detecting the caffs Violating the Health morals ” In 2016 IEEE/ ACM international Conference on Advances in Social Networks Analysis and Mining.

[7] HumaParveen and ShikhaPandey “ novelettish analysis on Twitter Dataset using Naive Bayes algorithm ” 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology(iCATccT) Runner 416-419@article{ Parveen2016SentimentAO}.

[8] Niu, Zhen, Zelong Yin, and Xiangyu Kong." novelettish type for microblog by machine knowledge." In 2012 Fourth International Conference on Computational and Information lores,pp. 286- 289. Ieee, 2012.

[9] Meena Rambocas, João Gama, “ Marketing disquisition The part of novelettish Analysis ”, April 2013, ISSN 0870-8541.

[10] P.Kalaivani, “ Sentiment type of Movie Reviews by supervised machine knowledge approaches ”et.al, Indian Journal of Computer Science and Engineering(IJCSE) ISSN 0976- 5166 Aug- Sep 2013.