



## Client Audits through Sentiment Analysis

Bhavin Kumar<sup>1</sup>, Dr. Dayanand Lal.N<sup>2</sup>, Deepak D.M<sup>3</sup>, Swasthika Jain.T.J<sup>4</sup>, Anusha N<sup>5</sup>

<sup>1</sup>Assistant Professor, GITAM University, India, bkumar@gitam.edu

<sup>2</sup>Assistant Professor, GITAM University, India, dnarayan@gitam.edu

<sup>3</sup>Assistant Professor, GITAM University, India, dmanjuna@gitam.edu

<sup>4</sup>Assistant Professor, GITAM University, India, sjain@gitam.edu

<sup>5</sup>Assistant Professor, Sambhram Institute of Technology, India, anushagpn089@gmail.com

### ABSTRACT

The purpose of this study is to demonstrate how Natural Language processes and Artificial Intelligence helps us to understand customers' emotions, feelings, behavior and their requirements. This study plays a major role in the field of business, Ecommerce, politics, education and technology to understand the customers feedbacks. In this paper, we present our primer tests on client surveys on restaurant information. This examination is intended to extricate notion dependent on surveys that exist in restaurant information. It distinguishes the assessment that alludes to the client surveys utilizing Machine learning, Artificial Intelligence and Natural Language Processing. To arrange supposition, our analysis comprises of Classification algorithms, NLP techniques like tokenization, Stemming, Stopwords, WordCloud and assessment of execution by utilizing ROC bend. The examination uses wordcloud to recognize positive and negative client audits on cafe dataset. Test results showing that the Naïve Bayes calculation gives better precision and execution as contrast to remaining all classification algorithms.

**Key words :** WordCloud, Stopwords, NLP, Stemming, Sentiment Analysis, Classification Algorithm.

### 1. INTRODUCTION

Today, an immense measure of data is accessible in online records, for example, website pages, postings of newsgroup, twitter and Facebook, on-line news databases etc. Among these sorts of data accessible, one valuable sort is the supposition, or feeling individuals express towards a subject. (A subject is either a theme of intrigue or a component of the topic.) [1]. For instance, knowing the notoriety of their own or their rival's eatery/restaurant is important for improvement, showcasing and client relationship the board. Generally, organizations lead customer overviews for this reason. Natural language processing[11] is best way to train and improve the words extraction from the given client surveys.

The prepared sets can undoubtedly channel the qualities according to the client necessities. From Natural Language Processing we can undoubtedly pre-process every one of the words from the given sentences.

Classification algorithm helps to classify the customers based on their reviews and ROC graphs helps to determine relationship between true positive rate versus false positive rate. By using Confusion matrix, we can easily summarize the result and we can predict accuracy of model.

Classification algorithm includes Logistic regression, Decision tree, random forest, Support Vector Machine, Naïve baye's[12] and KNearest neighbour algorithm. For applying machine learning algorithm, we required packages like pandas, numpy, sklearn, seaborn, matplotlib, NLTK, wordcloud, porterstemmer etc.

Pandas majorly used for handling dataset operations like extracting Excel, csv, text, reading dataset, selecting particular rows and columns, converting dataset into data frame, removing unwanted data, Checking with null values etc. Numpy package used for performing numerical, calculus, integration, Array operation etc. Matplotlib and seaborn packages helps to diagrammatic representation of data like pie chart, Bar plot, correlation graph etc. Sklearn contains all the regression and classification algorithm supporting packages, accuracy, confusion matrix packages. Sklearn also contain data feature selection packages like binary encoder, one hot encoder, standardization, normalization etc.

With the help of NLTK we can process the feedback and extract the important features from customer reviews. Wordcloud package helps us to clean the dataset and extracting only keywords words from dataset.

#### 1.1 Literature Review

Z.Callejas et al[1], proposed a different roads for the mix of assessment investigation in up close and personal human agent communications. To start with, it is important to pick a sufficient psycho-phonetic model to portray human-specialist full of feeling exchanges. Need to shuffle with the utilization of semantic standards and machine learning strategies so as to incorporate: the multimodal nature of supposition related wonders, the inconstancy of worldly and choice casings,

shifting levels of multifaceted nature required by the planning limitation of the connection, and the heterogeneity of the logical data. Another enormous issue is to manage ASR yields. Need to explore supposition examination strategies as an information and a yield of both short and long haul procedures.

Guixian Xu et al[2], proposed an improved word portrayal technique is proposed, which incorporates the commitment of slant data into the conventional TF-IDF calculation and creates weighted word vectors. The weighted word vectors are contribution to BiLSTM (Bidirectional Long Short Term Memory) to catch the setting data viably, and the remark vectors are better spoken to. The assessment propensity of the remark is acquired by feedforward neural system classifier. Under similar conditions, the proposed slant examination strategy is contrasted and the slant investigation strategies for RNN, CNN, LSTM, NB. The test results show that the proposed assessment examination technique has higher exactness, review and F1 score. The technique is end up being successful with high exactness on remarks. In this paper, an assessment investigation strategy for remarks dependent on BiLSTM is proposed and applied to the remark assessment investigation task. As indicated by the lack of the word portrayal strategy in the flow inquiries about, the opinion data commitment degree is incorporated into the TF-IDF calculation of the term weight calculation, and another portrayal strategy for word vector dependent on the improved term weight calculation is proposed.

Mondher Bouazizi et al[3], proposed a novel approach that, notwithstanding the previously mentioned errands of double and ternary classifications, goes further in the classification of writings gathered from Twitter and classifies these writings into various estimation classes. Here they limit degree to seven diverse supposition classes, this approach is adaptable and can be rushed to characterize writings into more classes. First present SENTA, apparatus worked to assist clients with choosing out of a wide assortment of highlights the ones that at the most for their application, to run the classification, through a simple to-utilize graphical UI. At that point use SENTA to run our own investigations of multi-class classification. Trials show that the proposed approach can reach up to 60.2% precision on the multi-class classification. All things considered, the methodology ends up being exact in double classification also, ternary classification: in the previous case, we arrive at an exactness of 81.3% for similar informational index utilized after evacuating impartial tweets, and in the last case, arrived at an exactness of classification equivalent to 70.1%.

Zhao Jianqiang et al[4], proposed a word embeddings got by Zhaoid learning on huge twitter corpora that utilizes inactive logical semantic connections and co-event measurable attributes between words in tweets. These word embeddings are joined with n-grams highlights and word notion extremity score highlights to frame a feeling highlight

set of tweets. The list of capabilities is incorporated into a profound convolution neural system for preparing and foreseeing feeling grouping names. Tentatively contrast the presentation of our model and the pattern model that is a word n-grams model on five Twitter datasets, the outcomes demonstrate that our model performs better on the precision and F1-Measure for Twitter estimation arrangement.

Kim Schouten et al[5], proposed an objective is to discover and total feeling on elements referenced inside reports or parts of them. A top to bottom outline of the current best in class is given, indicating the enormous advancement that has just been made in finding both the objective, which can be an element accordingly, or some part of it, and the comparing supposition. Perspective level supposition investigation yields very fine grained slant data which can be helpful for applications in different areas. Current arrangements are classified based on whether they give a technique to perspective discovery, supposition examination, or both. Besides, a breakdown dependent on the kind of calculation utilized is given. For each talked about examination, the announced exhibition is incorporated. To encourage the quantitative assessment of the different proposed techniques, a call is made for the institutionalization of the assessment system that incorporates the utilization of shared informational collections. Semantically-rich idea driven perspective level conclusion examination is talked about and distinguished as one of the most encouraging future research course.

## 2. METHODOLOGY

### 2.1 Work Flow

The Flow have been shown how Natural language process and classification algorithm is being demonstrated:

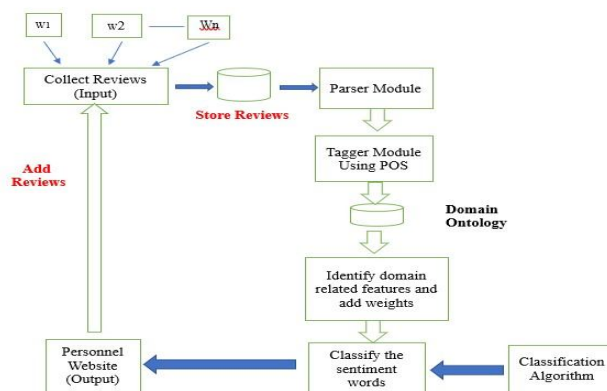


Figure 1: Flow Chart for NLP and Classification

- 1- Input:** Collecting required data from different web resources and to convert it into dataset format.
- 2- Parser Module:** It helps to identify grammatical structure of each and every sentence and words.

**3- Tagger module:** The POS tagger used to relegate POS labels to words in a sentence (such as: labels for things, action words, and descriptor)

**4- Domain Ontology:** These modules separate sentences and then it helps to extract pronoun, noun and respective adjectives, verbs and adverbs.

**Step 5- Classification algorithm:** Here we train the machine with different classification algorithm for finding accuracy.

## 2.2 Procedure for dataset analysis

**i. Sentiment Analysis:** Sentiment analysis helps to understand the human emotions, feelings, likes and dislikes. So that it directly helps the business mans, investors, Ecommerce, education, politics etc.

**ii. Natural Language processing:** It is a technique for train the PC so that it needs to break down, perceive and comprehend the human language. NLP can be utilized for discourse acknowledgment, understanding client surveys, news arrangement and so forth.

### 2.2.1.NLP processing has following steps.

**i. Tokenization:** It is a strategy for breaking the sections into little lumps or words. Sentence tokenizer breaks the sections into sentences. Word tokenizer breaks the sections into words.

**ii. Stemming:** It keeps only root words. Ex: From the words like Interested, Interesting is reduced to common word like interest.

**iii. Parts of Speech tagging (POS):** Is utilized to distinguish syntactic gathering of words. POS is utilized to distinguish thing, pronoun, intensifier, and action word and so on., in light of setting. POS is utilized to distinguish the relationship with in the sentences and appoint the labels to comparing words.

**iv. Stop words:** It can be noticed as noise in the text and it has to be removed compulsory. E.g. is ,as, was, at, it ,for etc. IN NLTK for stop words packages identifies the matching stop words and these stop words can be removed while processing textual data.

#### v. Word cloud:

Word Cloud is a Data perception method utilized predominantly for content portrayal where size of each word speak to a recurrence or significance of an each word. Word mists are fundamentally utilized for dissecting information from interpersonal organizations, client surveys, grievances of a client's and so on.

**vi. Count Vectorization:** Count vectorization package converts all the words int 0's and 1's and it will be represented as count vector matrix. It also identifies the frequency of each and every word.

Once all NLP process is done then dataset has to split into 80% of training dataset and 20% of testing dataset. By using training dataset, we can train the machine and by using testing dataset we can test the machines in order to find the accuracy of an model.

## 2.3 Choosing algorithms

**2.3.1 Classification Algorithms-** It is used to train the machine by using past data and then uses this learning methods to classify new observation. Here we using Decision tree, Random forest, KNN Logistic regression, SVM and Naïve Bayes algorithm in order to train the machine.

**i.KNearest Neighbors algorithm (KNN)** is an supervised classification algorithm and it classify the customers(data) based on the number of nearest neighbors. Here K represents the number of nearest neighbors. KNearest Neighbors algorithm is also called as lazy algorithm and it is best suited if dataset is small and noise free. This algorithm uses Euclidean Distance formula to find the distance of all neighbors.

Euclidean distance formula:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

where d=Euclidian Distance  
x1,x2,y1,y2 are coordinates.

**ii)Decision tree** is also called as Classification and Regression Tree (CART) . It is an supervised algorithm used for both classification and regression purpose. It represents the data in the form of tree, where we can select root node which is having less entropy, branches represent all possible choices and leaf node represents the final decision. This model is best suited for the less noisy dataset.

For calculating entropy we are using formula:

$$E(S) = \sum_{i=1}^c - p_i \log_2 p_i \quad (2)$$

Where E=Entropy, P=Probability

#### iii)Naive Bayes algorithm

It is supervised classification algorithm works based on the principle of conditional probability of the occurrence of an event based on past knowledge that will be helpful for correlating to an event.

It is majorly used for news classification, sentiment analysis, spam filtering etc.

Conditional Probability formula is

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad (3)$$

Where,  $P(B|A)$ = Probability B on A  
 $P(A|B)$  =Probability A on B  
 $P(A)$ =Probability of A  
 $P(B)$ =Probability of B

**iv)Random Forest algorithm** is applicable for both regression and classification problems. It is an ensemble method, for continuous dataset it will consider average of an all the trees and for classification dataset it will choose an result in which maximum trees are voting.

Random forest algorithm works based on below formula:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (4)$$

Where  $b=1..B$  and  $B$  is an free parameter  $x'$  is an unseen sample can be predicted by taking average of an all prediction from regression tree or by considering majority votes for classification problems.

**v)Logistic Regression:** It is an classification and supervised algorithm. It has dependent variable 'y' and independent variable 'x' and it will consider only discrete values. This algorithm uses sigmoid function, that will map the results between 0 and 1(Based on threshold value)

**Logistic regression works based on below formula:**

$$y = \frac{1}{1 + e^{-z}}$$

$$\text{Or, } y = \frac{1}{1 + e^{-(ax+b)}} \quad (5)$$

Y=Sigmoid function  
a=slope  
x and b is coordinates.

### 3.4 Implementation code(snippet)

#### i)Code snippet 1

```
import re
from nltk.stem.porter import PorterStemmer
corpus = []
for i in range(0, 1000):
    review = re.sub('[^a-zA-Z]', ' ', data['Review'][i])
    review = review.lower()
    review = review.split()
    ps = PorterStemmer()
    review = [ps.stem(word) for word in review if not word
              in set(stopwords.words('english'))]
    review = ' '.join(review)
    corpus.append(review)
```

Here initially sentences are divided into words by using word tokenizer and then stopwords removed by using stopwords package and then we can keep only root words by using porterstemmer. Finally we got cleaned data. All cleaned data kept in corpus.

#### ii) Code snippet 2

```
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer()
X=cv.fit_transform(corpus).toarray()
y=data.iloc[:,1].values
```

CountVectorizer converts all the words into vector matrix

#### iii)Code Snippet3

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.20, random_state = 0)
```

This code splits dataset into training dataset(80%) and of test dataset(20%).

#### iii) Code snippet3

```
from sklearn.naive_bayes import GaussianNB
classifier=GaussianNB()
classifier.fit(X_train,y_train)
y_pred=classifier.predict(X_test)
```

Dataset can be trained by using Naïve Bayes algorithm and then we have to predict accuracy of an model.

#### iv)Code snippet4

```
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,y_pred)
```

This code used to print confusion matrix.

### 3. RESULTS

In every model, the accuracy and the cost analysis plays an important role in the acceptance of that model for the application. The result of Restaurant data set is being displayed using confusion matrix and ROC curve.

In each model, the exactness and the cost examination assumes a significant job in the acknowledgment of that model for the application. The consequence of Restaurant informational index is being shown by utilizing ROC curve and Confusion Matrix .In this model Initially, word Tokenizer divides sentences into words then Stopword package removes the stopwords from all sentences. By using porterstemmer we can keep only root elements in an sentences and then this cleaned words saved in corpus. Wordcloud displays positive words and negative words separately. These words intern converted into vector matrix and finally this vector matrix giving as an input to an machine.

**i. carpus: After removing Stop words**

Carpus contain cleaned data where special symbols, stopwords will be removed and it contains only root elements. The carpus results shown below.

Index	Type	Size	
0	str	1	wow love place
1	str	1	crust good
2	str	1	tasti textur nasti
3	str	1	stop late may bank holiday rick steve recommend love
4	str	1	select menu great price
5	str	1	get angri want damn pho
6	str	1	honeslti tast fresh
7	str	1	potato like rubber could tell made ahead time kept warmer
8	str	1	fri great
9	str	1	great touch
10	str	1	servic prompt
11	str	1	would go back
12	str	1	cashier care ever say still end wayyy overpr

Snapshot-1: Carpus result

**ii. Vectorization:**

Converting all the cleaned words into vector matrix. Vector matrix result shown below.

	22	23	24	25	26	27
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	1	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	1
12	0	0	1	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0

Sanpshot-2: Vector Matrix

**iii. Word Cloud:**

Word Cloud is a Data perception method utilized primarily for content portrayal where size of each word speak to a recurrence or significance of an each word. Word mists are principally utilized for breaking down information from interpersonal organizations, client audits, protests of a clients and so on.

**Positive Review**



Snapshot- 3: Positive review wordcloud

**Negative review**



Snapshot-4: Negative Review wordcloud

**iv. Confusion Matrix:**

It is used for predicting accuracy of an model.

**a) True Positives (TP):** For this situation anticipated is additionally yes and genuine likewise yes

Ex: There will be fire and anticipated additionally is fire.

**b) True negatives (TN):** For this situation anticipated is additionally no and genuine likewise no.

Ex: There will be no fire and anticipated as additionally no fire.

**c) False Positives (FP):** For this situation genuine is no yet anticipated as yes

Ex: There will be no fire however anticipated as Fire

**d) False negatives (FN) :** For this situation real is yes however anticipated as no.

Ex: There will be fire yet anticipated as no fire,

**Table 1:** Confusion Matrix

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

**Accuracy:** It is the proportion of number of redress expectations to the overall number of input tests  
 $Accuracy = (TP+TN)/(TP+FN+FP+TN)$

**v. ROC (Receiver Operating Characteristics) curve.** It is an important metrics to evaluate the performance of a classification model. The ROC curve is plotted with True Positive Rate(TPR) against the False Positive Rate(FPR), where TPR is on y-axis and FPR is on the x-axis.

$TPR = TP/(TP+FN)$

$FPR = FP/(TN+FP)$

**3.1. Metrics used**

**Cost , difference with other efficiency algorithm used features**

**ROC Curve:**

$TPR = TP/(TP+FN)$

$FPR = FP/(TN+FP)$

$Roc\ Curve = TPR/FPR$

$Accuracy = (TP+TN)/(TP+FN+FP+TN)$

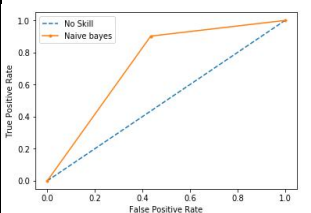
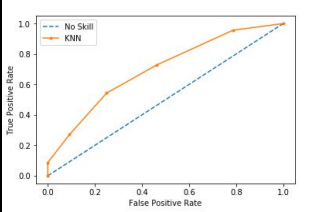
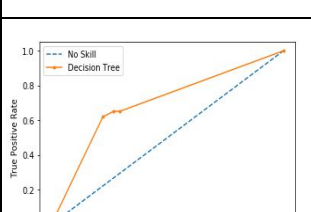
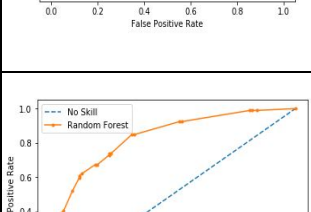
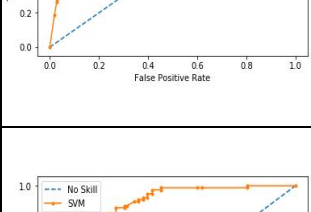
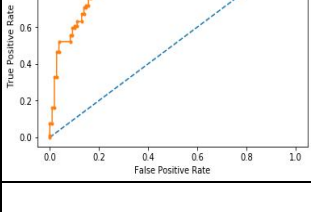
**3.2 Comparison tables**

**Comparison of Confusion Matrix ,accuracy and ROC curve for all Classification algorithms**

**Table 2:** Confusion Matrix for classification algorithms

Sl no	Classification Algorithm	Confusion Matrix
1	Naïve baye's algorithm	
2	KNN algorithm:	
3	Decision Tree	
4	Random Forest	
5	SVM	
6	Logistic Regression	

**Table 3:**ROC Curve for Classification Algorithms

Sr no	Classification Algorithm	ROC Curve
1	Naïve baye’s algorithm	
2	KNN algorithm:	
3	Decision Tree	
4	Random Forest	
5	SVM	
6	Logistic Regression	

**Table 4:** Accuracy Score for Classification Algorithms

Sr no	Classification Algorithm	Accuracy
1	Naïve baye’s algorithm	0.73
2	KNN algorithm:	0.61
3	Decision Tree	0.71
4	Random Forest	0.72
5	SVM	0.72
6	Logistic Regression	0.71

**4.CONCLUSION**

We applied NLP procedures to comprehend the client's surveys of restaurant dataset. From NLP we effectively process the client's audits and by utilizing word cloud we can distinguish positive and negative surveys of clients. With the help of ROC and confusion matrix we can summarize results of all classification algorithms and finally we got better accuracy from Naïve Bayes algorithm.

**REFERENCES**

1. Chloe Clavel et al, "Sentiment Analysis :from opinion mining to human-agent interaction " 1949-3045/2015 IEEE.
2. Guixian Yu and Yueting Meng , "Sentiment Analysis of Comment Texts Based on BiLSTM" 2169-3536/2018.
3. Mondher Bouazizi, and Jay Bal "A Pattern Based Approach for Multi-Class Sentiment Analysis in Twitter" 2169-3536/2017 IEEE. Translations and content mining are permitted for academic research only. <https://doi.org/10.1109/ACCESS.2017.2740982>
4. Zhao Jianqiang et al "Deep Convolution Neural Networks for Twitter Sentiment Analysis" 2169-3536/2017 IEEE. Translations and content mining are permitted for academic research only.
5. Kim Schouten et al " Survey on Aspect-Level Sentiment Analysis"1041-4347/2018 IEEE.

6. XIANGHUA FU et al, “**Lexicon-Enhanced LSTM With Attention for General Sentiment Analysis**” 2169-3536 2018 IEEE.
7. Zhao Jianqiang et al, “**Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis**” 2169-3536 (c) 2016 IEEE.  
<https://doi.org/10.1109/SmartCity.2015.158>
8. Liang-Chih Yu, K. Robert Lai , and Xuejie Zhang “**Refining Word Embeddings Using Intensity Scores for Sentiment Analysis**” 2329-9290 © 2018 IEEE.
9. Rui Xia, Feng Xu et al, “**Dual Sentiment Analysis: Considering Two Sides of One Review**” 1041-4347 (c) 2015 IEEE.
10. Zubair Md. Fadlullah et al, “**State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow’s Intelligent Network Traffic Control Systems**” 1553-877X (c) 2016 IEEE.
11. Maganti Syamala; N.J.Nalini ”**A Deep Analysis on Aspect based Sentiment Text Classification Approaches**” Vol.8, No.5 2019-10-15 IJATCSE.
12. Rein Rachman Putra Monika Evelin Johan; Emil Robert Kaburuan “**A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia**” Vol.8, No.5 2019-10-15 IJATCSE