# SENTIMET ANALYSIS: REVIEW EXTRACTION FOR CHECKING QUALITYOF PRODUCT

[1]Ankit Maurya, [2]Anamika Sharma, [3] Nandani Sharma
[1]Research Scholar, [2]Assistant Professor, [3]Research Scholar
Department Of Computer Science and Engineering, Goel Institute of Technology and Management, Dr. A. P. J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India.

***Abstract:*** Sentiment Analysis and Opinion Mining has become an investigation trouble spot with the speedy improvement of online business destinations. Amazon is a normal web business application with a large number of customers imparting their assumption consistently for a thing experience. In this work, we examined completely the procedures applied in feeling gathering over thing review data : word reference based, rule-based and AI put together methodologies thus with respect to Our instructive assortment is crawled and cleaned with the rule of AI, colossal data and ordinary language taking care of system.

For word reference based strategies, we attempted various things with the Simple Word Count approach and Feature Scoring approach using most renowned appraisal word references and semantic resources, explicitly TF - IDF vectorization, Sentiment jargon and General Inquirer. We built changed end word references, arranged remembering scores and considered ten classifiers for authentic overview data.

Further, we arranged Linguistic Inference Rules(LIR) to additionally foster jargon based classifiers. LIR intends to manage nullification, valence shift and separation conjunctions in normal language. For AI based techniques, we used state of the art oversaw learning models: Naive Bayes, Maximum Entropy and Support Vector Machines. The computations that we differentiated are Naiave Bayes and XGBoost Model.

***Index Terms*** **– Sentiment Analysis, Opinion Mining, Review Extraction, Natural Language Processing, Machine Learning.**

## INTRODUCTION

We are getting millions of views, opinions and emotions in a single subject or product everyday through social media, blog, forums, shopping sites, etc. as tweets, announcements, sites posts, audits and so forth To oblige this different uses of assessment examination have arisen in a few areas like notion investigation over Product Review, Political Exit Poll, Financial News and Healthcare etc.
In this study we are focusing on Amazon product review. We break each review/rating in 3 poles like :

1. Positive
2. Negative
3. Neutral

| S. No. | Review | Rating | Pole |
|---|---|---|---|
| **1.** | My experience is fantastic with this product. |  | **Positive** |
| **2.** | Product is Ok I guess. |  | **Neutral** |
| **3.** | Product is useless. |  | **Negative** |

**Table : Pole classification.**

““Sentiment analysis is a process of mining attitudes, opinions, views and emotions from text speech, tweets and database sources through natural language processing (NLP). It involves classifying opinion in the text into categories like positive or negative or neutral. It also refers to subjectivity analysis, Opinion mining and Appraisal extraction.”

Sentiment Analysis is a type of text classification based on sentimental orientation (SO), below are the general steps that used for SA :

**Firstly** : Evaluative terms are extracted through a review dataset.
**Secondly** : Determining SO (polarity) of opinion.
**Thirdly** : Opinion strength or opinion intensity determined.
**Finally** : Classify the reviews with respect to sentiment class, positive, negative or neutral.

## 2. A BRIEF DISCUSSION ON PREVIOUS WORKS OVER SENTIMENT ANALYSIS

Unavoidable genuine applications are just basic for the motivation driving why thought evaluation is a remarkable examination issue. It is in like way particularly testing as a NLP research theme, and covers different novel subproblems as we will see later. Furthermore, there was little examination before the year 2000 out of one or the other NLP or in semantics. A piece of the explanation is that before then there was little examination text open in modernized developments. Since the year 2000, the field has developed quickly to get conceivably the most remarkable evaluation regions in NLP. It is additionally widely explored in information mining, Web mining, and data recovery. In all honesty, it has spread from PC programming to the board sciences (Archak, Ghose and Ipeirotis, 2007; Chen and Xie, 2008; Das and Chen, 2007; Dellarocas, Zhang and Awad, 2007; Ghose, Ipeirotis and Sundararajan, 2007; Hu, Pavlou and Zhang, 2006; Park, Lee and Han, 2007).
Some latest research works are described below:

- **SEMEVAL 2013, Team IITB :**
  "Team IITB used the Bag of Words model with Discourse Information and could achieve an accuracy of 39.80%."

- **SEMEVAL 2013, Team Uottawa :**
  "Team uottawa used NaÃŕve Bayesian Classifier were able to achieve an accuracy of 42.51%."

- **SEMEVAL 2013, Team NRC-Canada :**
  "Team NRC-Canada used SVM model with unigram, bigram, OS tags, negation etc as features and were able to achieve an accuracyof 69.02%, securing the first position."

## 3. Data Description

The following 3 dataset used for the carried work:


All these datasets are publicly available for research work. We can describe the attribute in the dataset into the following:

**Fig. 1 : Data set entity**

We give load to review and rating column of dataset as we perform sentiment analysis on product review of dataset.

| | reviews.text | reviews.rating |
|---|---|---|
| 0 | This product so far has not disappointed. My c... | 5.0 |
| 1 | great for beginner or experienced person. Boug... | 5.0 |
| 2 | Inexpensive tablet for him to use and learn on... | 5.0 |
| 3 | I've had my Fire HD 8 two weeks now and I love... | 4.0 |
| 4 | I bought this for my grand daughter when she c... | 5.0 |

**Fig. 2 : Cleaned data**

## 4. COMPONENTS AND BACKGROUND

Notion can be isolated into various parts: holder, target, extremity and perspective. Every part compares to explicit errands in a framework.

- **Holder :** It indicates a substance that means the notion.
- **Target :** It distinguishes the substance that will be chosen with the point of assumption.
- **Polarity :** Sentiment can be isolated into various segments: holder, target, viewpoint and extremity. Every segment compares to explicit undertakings in a framework.
- **Aspect :** It characterizes the specific part or highlight of the objective that the assumption is communicated toward.

### 4.1 TASKS

Most assumption investigation errands are characterized by the conclusion part it concerns. With the assistance of new innovation, specialists have been meagerly led going from holder/target location, feeling arrangement, perspective extraction, assessment spam identification and so on In this analysis, we were centered around report level notion order. In particular, given an audit dataset, we will investigate various strategies for allotting an extremity name.

### 4.2 LEXICON

Vocabularies convert sentences into tokens. It peruses the audit sentences and breaks into a little substance called a token.

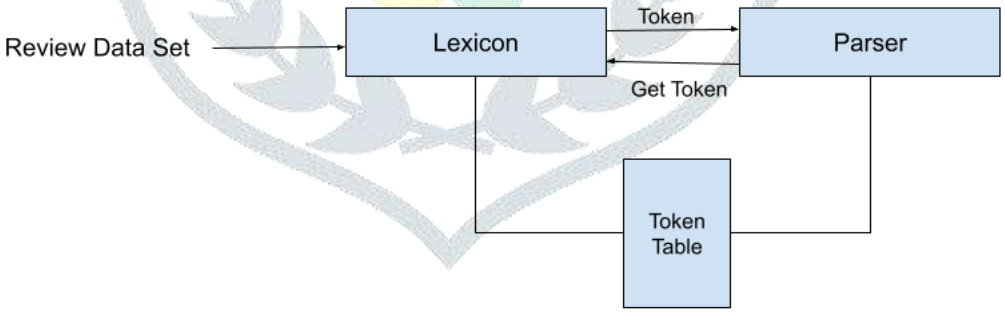



**Fig. 3 : Lexicon functionality**

The sentiment Lexicons are a rundown of words or expressions that passes positive or negative extremity data. Dictionaries are vital asset in conclusion examination. It gives estimation data towards about the littlest semantic unit of the sentence. Indeed, even AI put together techniques can depend with respect to opinion vocabulary in highlight designing.

### LEXICON BASED ALGORITHMS

The vocabulary based grouping calculations are bring assumption of a record is controlled by the it's prevailing parts (words or expressions). The essential plans incorporate greater part casting a ballot, record scoring with thresholding and basic word tallying.

Dictionary based technique gives us a pattern to additional investigation. As of late there has been a pattern of utilizing outfit learning with different frail vocabulary based classifiers. Augustyniak et. al. utilize an assortment of vocabulary based powerless classifiers and a C3.4 choice tree as solid classifiers. The vocabulary extraction technique is called Frequentiment and it is demonstrated three to multiple times quicker than administered AI. While this is useful and prominant, no comparative realized work has been directed to test its adequacy in English language text.

In our work, we would apply two methodologies on the audit informational index, Simple Word Count and Feature Scoring (F-Score).

### 4.3 MACHINE LEARGING BASED ALGORITHM

Supposition grouping is two way classification of undertaking. Text arrangement that generally characterizes information into some predefined classes. It is an all around considered field with exceptionally ideal arrangements and applications. Most of examination in both content arrangement and supposition investigation fall into AI based philosophy.
Machine learning methods such as Linear Regression, Decision Tree, K-Nearest Neighbour, Random Forest, Support Vector Machine, Navie Bias, XGBoost etc.

## 5. EVALUATION MATRICES

Let we have a set of classification results, in each class of n documents, Cij ($0 \le i \le n-1$, $0 \le i \le n-1$) indicates the number of instances where a document in the ith class is categorized as belonging to the jth class. The per-class measures can be calculated as follows:

- **F-Score**
  The main evaluation metrics for the project is F-Score that is the average polar of the dataset. It is known as Feature Score.

Fscore =[FI(-ve) + FI(+ve)]/2

Given a sentiment lexicon l, the scoring function of a feature f maps a feature to a real-valued number where

Score l, f > 0 if Positive
Score l, f = 0 if Neutral
Score l, f < 0 if Negative

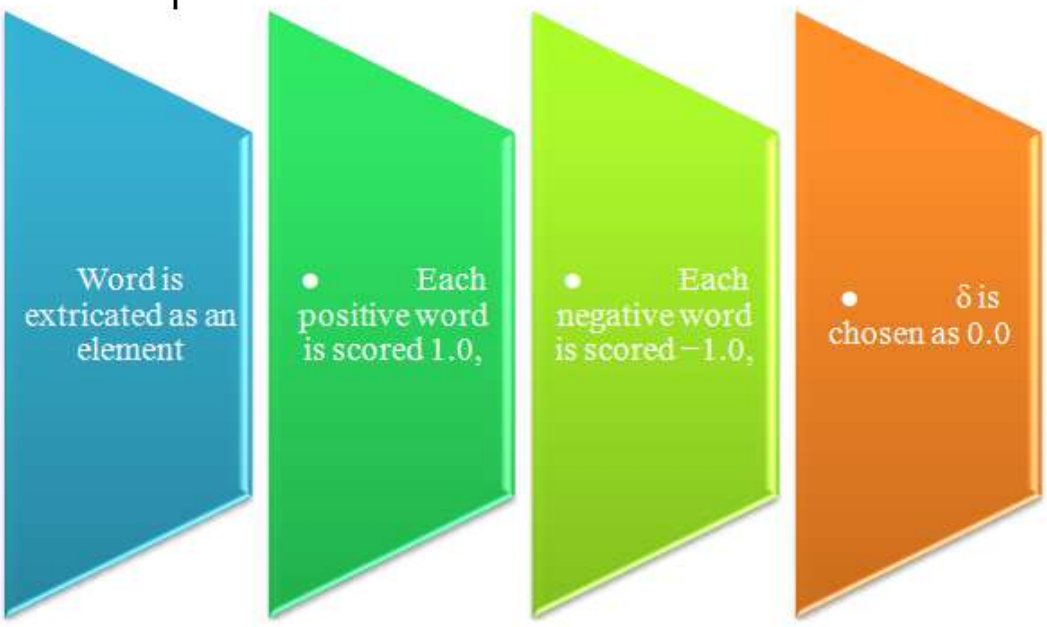
Given document d = {f1, f2, ..., fn} where fi($1 \le i \le n$) represents the ith feature in d, the overall sentiment sum of d can be calculated as:

$$Sum\ (l,\ d)\ =\ \sum_{i=1}^{n} Score\ (l,\ f_i)$$

By selecting a threshold δ = 0, the sentiment orientation of d can be defined as:

$$s_{fc}(d) = \begin{cases} 1, sum(l,d) > \delta, \\ RC(d), -|\delta| \le sum(l,d) \le |\delta|, \\ -1, sum(l,d) < \delta. \end{cases}$$

To fit in our concern, we relegate an irregular name to d when the total worth is between edge stretches. Basic Word Count is an exceptional instance of Feature scoring where:




For the Simple Word Count method, the key is to create a lexicon with polarity attached to each word entry from

dataset. For the Feature Scoring (F- Score) method, the key is to extract features, define as an effective scoring function and find an accurate threshold δ.

- **Precision**
  Analysis of what fraction of instances is correctly classified.

$$p = \frac{c_{ii}}{\sum_j c_{ji}}$$

- **Recall**
  Analysis of what fraction of correct instances is classified.

$$r = \frac{c_{ii}}{\sum_j c_{ij}}$$

- **Accuracy**
  Analysis of what fraction of instances is correctly classified across all classes of review dataset.

## 6. EXPERIMENT

This work describes the implementation of different machine learning and natural language processing algorithms and shows which algorithm accurately predicts the sentiment with optimum accuracy. Step for enumeration of proposed methodology are as follow:

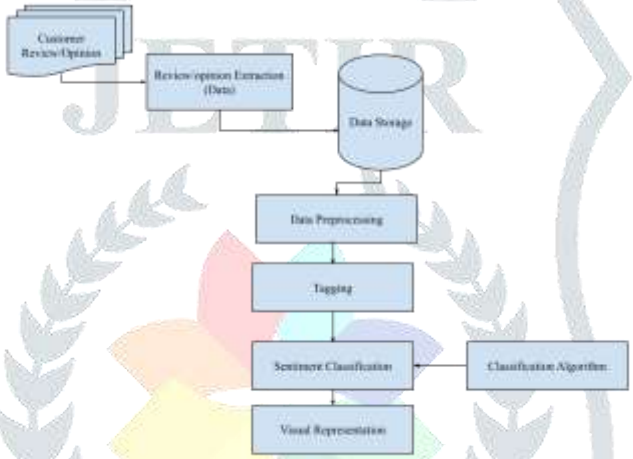



**Fig. 4 : System Architecture.**

### 6.1 Flowchart for the proposed model

The process contains 5 main steps i. E. importing package and loading dataset, exploratory data analysis, text preprocessing, sentiment engineering and modeling with machine learning algorithms.

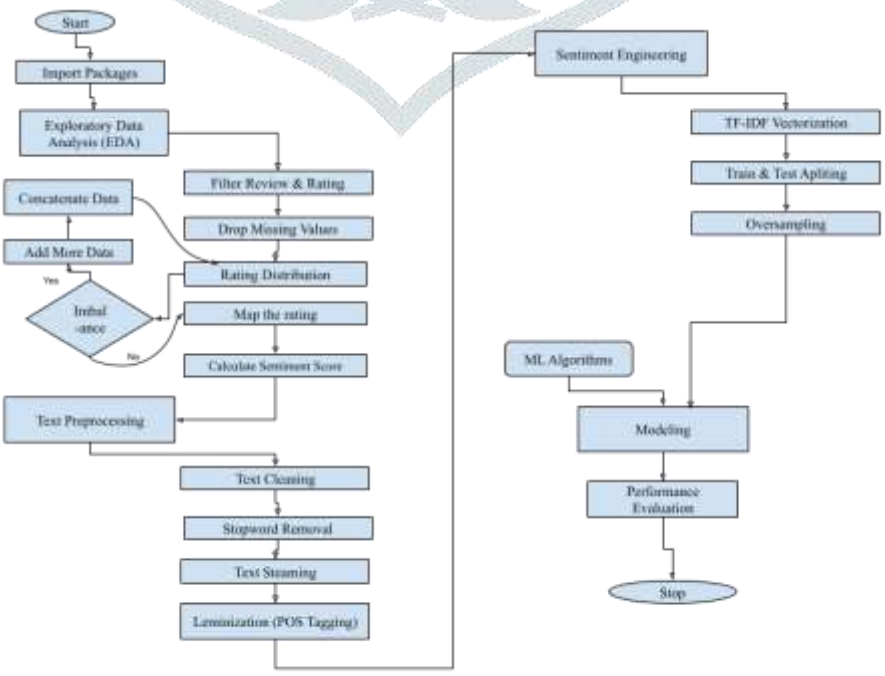



**Fig. 5 : Flowchart for proposed model.**

- **Import Packages :**
  In these steps we import all the required packages and for the environment.

- **Exploratory Data Analysis :**
  Since we are interested in sentiment analysis, we will only use review text and review rating, So we give load to selected columns.

  - **Filter review and rating :** Filters only review and rating of the product.
  - **Drop missing values :** Some columns contain missing review and rating, with the aim of fast execution we drop these rows.
  - **Rating distribution :** In this step we find distribution of rating. It can be 1- star, 2 - start, 3- star, 4 - star and 5 - star. We need to calculate the weight of the distribution.
    If the weight seems imbalanced then add more data until it seems balanced, we are doing it to analyze that our system is performing well for all the review and rating stars.

- **Mapping of rating :** In this step we give the sentiment score for the mapping value. Here is the mapping score given as :

  0 : Negative
  1 : Positive

- **Sentiment Score :** Sentiment score is given according to the rating map.

```
Sentiment score =  {1: 0,
                    2: 0,
                    3: 0,
                    4: 1,
                    5: 1}

Sentiment   =      {0: 'NEGATIVE',
                    1: 'POSITIVE'}
```

- **Text Preprocessing :**
  Text processing performs following sub steps :

  - **Text cleaning:** We will go over some simple techniques to clean and prepare text data for modeling with machine learning.
    It will return cleaned text:
    - lowercase
    - remove whitespaces
    - remove HTML tags
    - replace digit with spaces
    - replace punctuations with spaces
    - remove extra spaces and tabs

  - **Stop word Removal:** There can be some words in our sentences that occur very frequently and don't contribute too much to the overall meaning of the sentences. We usually have a list of these words and remove them from each of our sentences. For example: "a", "an", "the", "this", "that", "is", "it", "to", "and".

  - **Text Steaming:** Stemming is a rule-based system to convert words into their root form. It removes suffixes from words. This helps us enhance similarities (if any) between sentences.

    For example :
    "Happied" , " Happies" converts to "Happy"

  - **Leminization:** If we are not satisfied with the result of stemming, we can use the Lemmatization instead. It usually requires more work, but gives better results. As mentioned in the class, lemmatization needs to know the correct word position tags such as "noun", "verb", "adjective", etc.

- **Sentiment Engineering :**
  It has the following steps to analyze the sentiment and apply natural language processing and we will apply machine learning algorithm over it's resultant.

  - **TF - IDF Vectorization :** Advance ML environments does not supports strings so we need to convert it into integers and floating datatypes. For this we use a machine learning algorithm called feature extraction or (Vectorization).

  - **Train & Test Splicit :** In this step we split our data set into train and test subset for applying machine learning algorithms.

- **Oversampling :** As the data may be imbalanced so we use oversampling technique to optimize sentiment score.
- **Model :** Model gives us an environment to apply different machine learning algorithms over splitted data.

- **Machine learning algorithms :** Here we apply different machine learning algorithms with aim of achieving better accuracy of sentiment analysis.

- **Performance Evaluation :** In this step we evaluate the performance of proposed work with score metrics.

### 6.2 Proposed Algorithm

In this algorithm we take amazon product review data set and machine learning algorithms as input and find sentimental values for review data set in positive and negative data set plot with its accuracy.

**Algorithm:**

1. Environment setup.
2. Importing package and loading data.
3. Exploratory data analysis.
    1. Filter review text and review rating from loaded data.
    2. Drop the missing values.
    3. Calculate the distribution of ratings.
        1. If rating distribution imbalance.
            1. Load more data.
            2. Concatenate it with previous data.
            3. Go to 3.3.
    4. Map the rating
        1. 0 : for Negative.
        2. 1 : for Positive. {1, 2, 3 star → 0 (NEGATIVE), 4,5 → 1 (POSITIVE)}
    5. Calculate the sentiment score.
4. Perform Text preprocessing
    1. Text cleaning.
        1. Lowercase conversion.
        2. Remove white spaces.
        3. Remove tags.
        4. Replace digits with spaces.
        5. Remove extra spaces & tabs.
    2. Remove stop words.
    3. Steaming of text.
    4. Lemmatization.
5. Sentiment vectorization
    1. Create a TF-IDF vector.
    2. Split the dataset into a train and test subset.
    3. Perform oversampling of final data.
6. Modeling of data and algorithm
    1. Apply vectorizer with ML algorithms .
        1. Filter classification to training set.
        2. Predict test set result.
        3. Assign F1 score to variable.
        4. Calculate and plot confusion matrix .
    2. Performed modeling with different ML algorithms.

## 7. ANALYSIS

- **Rating Distribution : Unbalanced Data**

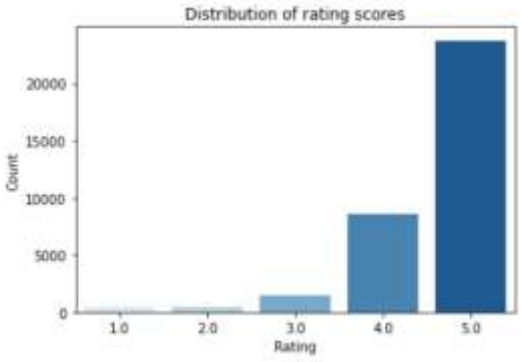



**Fig. 6 : Unbalanced rating distribution.**

As we see the distribution of 4,5 star rating is very high, and 1,2,3 star distribution is very low hence the overall analysis will move towards positive analysis only. To balance it we added more datasets.

- **Rating Distribution : Balanced Data**

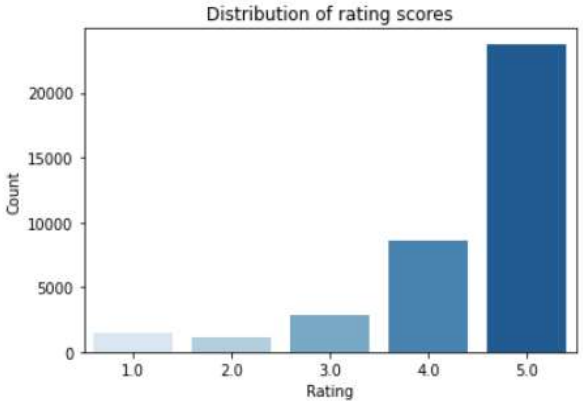



**Fig. 7 : Rating distribution (Balanced data).**

- **Sentiment Distribution**

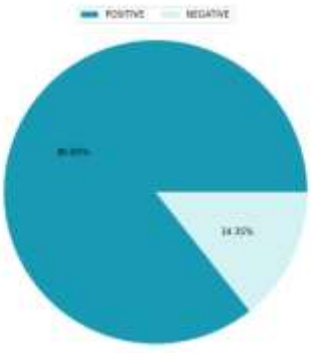


**Fig. 8 : Sentiment Distribution**

- **Modeling with Naiave Bayes Algorithm**

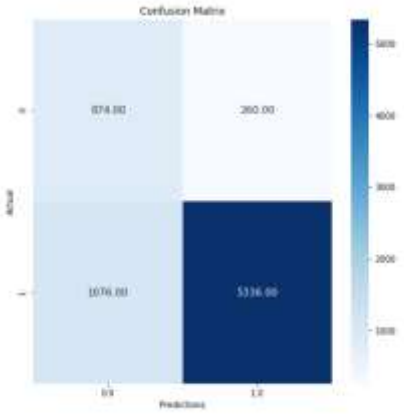



**Fig. 9 : Confusion Matrix – Naiave Bayes**

- **Modeling with XGBoost Model**

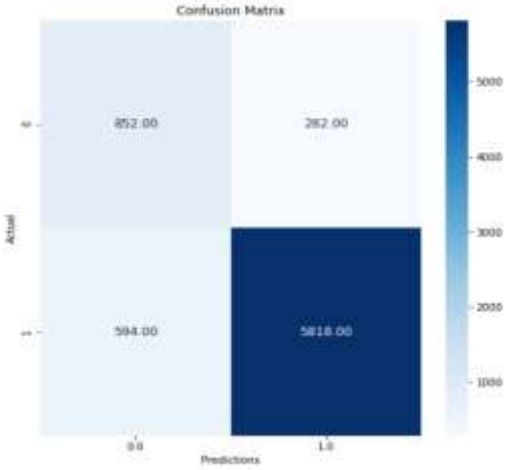



**Fig. 10 : Confusion Matrix – XGBoost Model**

## 8. CONCLUSION

We proposed a bunch of procedures for mining and summing up item audits dependent on AI and normal language handling strategies. To give a highlights based outline of an enormous number of client audits and trial results shows that the proposed procedures are promising in playing out their undertakings. We accept that observing will be especially valuable to item in new certain or negative remarks on there.

We got XGBoost Model is more accurate than Naiave Bayes model. Feature score for XGBoost model is 82.88 % while for Naiave Bayes model we analyze feature score 80%.