ABSTRACT

With the fast growth of e-commerce, a large number of products are sold online, and a lot more people are purchasing products online. People while buying also give feedback of products purchased in the form of reviews. The user generated reviews for products and services are largely available on the internet. Since information available on the internet is so widespread we need to extract the needful information for which we make use of sentimental analysis. Sentiment analysis extracts abstract and to the point information required for source materials by applying the concept of Natural language processing. It is used to deal with identification and aggregation of the opinions given by the customers. These reviews play a vital role in determining potential customers for the products as well as market trends for the product. Another important use of sentiment analysis is Customer segmentation. Focusing marketing efforts on those groups of people who are happy about your business and can naturally embrace your products is the key to a successful business. This technique is also useful for people who always talk negatively about your business. Once you identify them you can take necessary measures to protect the image of your brand. This report provides a summary of reviews for products by classifying these reviews as positive, negative or neutral. Information on internet is highly unstructured, machine learning approaches are applied including naïve Bayes and support vector machine algorithms by first taking inputs as unstructured product reviews, performs preprocessing, calculates polarity of reviews, extracts features on to which comments are made and also plots graph for the result. The algorithms precision, recall and accuracy are measured finally.

Keywords: Sentiment Analysis, Customer reviews, Classification, Product reviews, Amazon

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Online shopping has been growing for 20 years and many e-commerce websites such as Amazon, have been created to meet the increasing demand. Consequently, a specific product can be bought on several websites and the prices may vary. As customers usually want the best quality for the lowest price but can't directly check it, reviews from other customers seem to be the most reliable way to decide whether to buy the product or not. Therefore, sentiment analysis has proven essential to understand a product's popularity among the buyers all over the world. Buyers and sellers in the e-commerce platforms get a lot of ease of life facilities because of the automated machine learning algorithms which continuously run in the background to optimise the overall e-commerce process. Buyers get better product recommendation, less price, and trends in the market which help them make better purchase decisions. At the same time, sellers get information about the market trend, buyer interest and optimal price per product using which they can increase their profit.

A computer has no concept of naturally spoken language i.e it cannot simply deduce whether the sentiment is joy, anger, frustration or something else.The field of sentiment analysis solves this problem by using the process of natural language processing and computational linguistics to analyse and even predict human sentiments in various situations. It recognizes necessary keywords and phrases which helps the algorithms to determine the emotional state of the text. It has become widely popular among large corporations which rely on more and more customer interaction and customer satisfaction. Machine learning solutions to review the customer behaviour analysis problems provide a hands-off approach in which these algorithms once developed continue providing the market information. These require very little manpower compared to traditional methods while being more accurate than ever. Hence this is beneficial to both the corporate and the customer.

CHAPTER 2

LITERATURE SURVEY

Reference Paper (Name,Author,Year)	Dataset	Techniques/ Algorithm	Advantages	Disadvantages
1)Sentiment analysis of product based reviews - Manvee Chauhan, DivakarYadav. (December 2015)	Collected from different sites like consumerreview.com, cnet.com, download.com, zdnet.com. It consists of 13094 product reviews where 12094 are of training and 1000 for testing	To develop an interface Microsoft visual studio is used .It is possible to test and train datasets, extract features out of it. Either naïve bayes or support vectors are used to work upon the data and predict polarity of opinions.	Naïve Bayes gives better accuracy at 84.02% .	SVM is less accurate compared to naive bayes that is 80.2%. Large text files take a long time for computation.
2)Sentiment Analysis and Sentiment Classification using NLP- G.Divya, R.Suresh (July-2016)	1)The Multi Domain Sentiment dataset:contains product reviews from Amazon.com that includes books,DVDs, Electronics and Kitchen appliances with 1000 positive and 1000 negative reviews for each domain.(http://www.cs.jhu.edu/mdr	Machine learning techniques like Maximum entropy,Naïve Bayes and Support vector machines for text categorization. Other well known machine learning algorithms in NLP are K-Nearest Neighborhood,ID3 ,C5,centroid classifier,winnow classifier and Ngram model.	Maximum entropy,Naïve Bayes and SVM help achieve great success in text categorization.	

	edze/data sets/sentiment) 2)Another review data is available. http://www.cs.uic.edu/liub/FBS/CustomerReviewData.zip: Consists of reviews of 5 electronic products downloaded from Amazon and Cnet. 3)Movie review data is available. (http://www.cs.cornell.edu/People/pabo/movie-review-data) 4)Blogs,reviewers data collected from ecommerce websites like Amazon,yelp,CNET,dpreview,zdnet	Basic idea is estimating probabilities of categories using joint probabilities of words and categories.		
3)Sentiment Analysis for Amazon.com reviews-LeventGuner,Emilie Coyne,JimSmit (March 2019)	Dataset containing 60,000 product reviews from Amazon.com which are randomly selected from a dataset available from Kaggle containing 4 million	The performance of 3 different algorithms were compared: Multinomial Naïve Bayes(MNB), Linear Support Vector Machine(LSVM), Long short-term memory network(LSTM)	With tokenization numerical data represents the whole sentence which is convenient for a recurrent neural network like LSTM.Hence LSTM networks are most suitable for binary	In this study the assumption that the amount of stars corresponds with the sentiment in the review is made.However it could be the case that a very

	reviews.	. For classification with MNB and LSVM,a TF-IDF vectorizer is used.Whereas for classification with LSTM a method called tokenization is applied.	sentiment analysis on Amazon.com product reviews.	positive review is given one star or vice versa,polluting the dataset.
4)Twitter Sentiment Analysis Based on Ordinal RegressionShihabElbagir,Jing Yang (October 2019)	Data is collected from Twitter using Twitter API which contains 5000 positive tweets and 5000 negative tweets	Count vectorizer and term frequency-inverse document frequency is used for feature extraction.Machine learning techniques such as Multinomial logistic regression,Support Vector Regression,Decision Trees and Random Forest used to build and study machine learning classifier.	Experimental results conclude that the proposed model can detect ordinal regression with a good accuracy result.	----------
5) Sentiment analysis on online product reviewRaheesasafrin ,K.R. sharmila,T.S.shrisubangi ,E.A. vimal. (April 2017)	The dataset contains online product reviews along with their associated binary sentiment	Novel incremental diffusive algorithm is being used to extract features from online product descriptions, and then	The POS tagging is used to extract the most relevant features to get better results in classifying the sentence as positive or negative. This	----------

	polarity labels comments are taken as review and it is considered as a dataset for our project. The number of entries in the dataset is 3100	employ association rule mining and the k nearest neighbour	positive and negative separation of comments is used to analyze the quality of the online products. ANN is used to predict the comments. ANN provides more accuracy than the support vector algorithm	
6) Machine Learning-Based Sentiment Analysis for Text MessagesAbhishekBhagat,AkashSharma,Sarat Kr. Chettri (June-2020)	a)The IMDB dataset: a large movie review dataset which consists of 50,000 polarized movie review posts. b) The Sentiment 140 dataset:Consists of 1.6 million Twitter messages. c) The SemEval-2013 dataset:Consists of comments taken from a variety of topics discussed on Twitter. d) The SemEval-2014 dataset:Deal	Naïve Bayes, Decision Tree and Support Vector Machine (SVM) for sentiment analysis.	We find that the results obtained from the Decision Tree and SVM have a lower mean square error or higher accuracy with most of the datasets and are considered to be good classifiers.	Drawback of using a machine learning approach to opinion mining is that each opinion is treated as a single uniform statement and assigns a sentiment score to the post as a whole.Also since the social networking data can be accessible in various languages,it becomes an obstacle to sentiment analysis.

	s with product reviews made by consumers.			
7)Sentimental Analysis of Tweets using Naive Bayes Algorithm M. Vadivukarassi, N. Puviarasan, P. Aruna (2017)	2 kinds of APIs to extract the tweets:Search API used for dumping old tweets and Streaming API used for dumping live Tweets.	In this paper, the keywords are collected from Twitter using Twitter API.The extracted raw data is preprocessed using Natural Language Toolkit techniques. Algorithms used are Naïve Bayes & Chi-Square test.	The proposed system would be easy for users to obtain the summarized report about the opinion from Twitter. It also supports them in the decision making process in their daily life activities.	Application settings that should always be kept private.
8)Natural Language Processing for Sentiment Analysis- Wei Yen Chong, BhawaniSelvaretnam, Lay-Ki Soon (2014)	A total of 1513 tweets were extracted from Twitter and manually labelled.	SVM,Decision Tree ,Naïve bayes. The processed tweets will proceed to sentiment classification, in order to predict the sentiment of tweets. a)Subjectivity Classification b)Semantic Association c)Polarity classification.	The results are tabulated in a confusion matrix. It records the predicted result and the actual result.	It is also noticeable that due to the highly appeared misspelled words and slangs in tweets, it is not easy to extract the sentiment lexicons if it is not preprocessed to formal language.
9) Analysis of Feature Selection Methods for Text Classification using Multiple Datasets Archit Aggarwal, BhavyaGola,Tush	a)20Newgroups Dataset b)Polarity Dataset c)Reuters21578 Dataset	Naive Bayes, Bagging, Random Forest and Naive Bayes Multinomial classifiers for text	It usually gives more often than not, better or equally good results without using any feature selection as	…….

ar Sankla (June 2020)		classification on various datasets.	compared to using feature selectors taking into account all the evaluation measures.	
10) Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier - Arif Abdurrahman Farisi, Yuliant Sibaroni ,Said Al Faraby. (2019)	In this research the dataset is derived from Data finitis's Business Database which contains hotel reviews of as many as 5000 English sentences in CSV File.	Multinomial Naïve Bayes Model K-Fold Cross Validation	The use of preprocessing in figure 1 greatly affects the performance of the system so it can be seen if using preprocessing F1- Score average results can improve performance optimally.	By doing a test scenario will result in different performance because each scenario can affect the model built. Each scenario is validated using 10 fold cross validation each time

CHAPTER 3

PROBLEM DEFINITION

The aim of this project is to make an application in the field of natural language processing in order to find and implement a novel algorithm to solve the problem of measuring real-time comments made by users on products. For this we have collected the reviews from amazon that have been written by different users about a particular product and then the polarity of each feature of the mentioned product is determined that either it is negative, positive and neutral. For determining the polarity of the feature of product, the sentiment of the feature word (noun) of the text is calculated and corresponding scores are given to public opinions of the product's feature which could being used to compare between the product based on a particular feature and a statistical output was produced to show the results. The methods used in this project can be used for any specific product with public opinions. The customer's reviews reflect the customer's sentiments and have a substantial significance for the products being sold online including electronic gadgets, movies, household appliances and books. Hence, extracting the exact features of the products by analyzing the text of reviews requires a lot of effort and human intelligence.

CHAPTER 4

SYSTEM REQUIREMENT SPECIFICATION

4.1 System Requirements

All computer software needs certain hardware or software resources to be present to be used efficiently.These prerequisites are called system requirements.

4.2 Functional Requirements

A functional requirement defines a function of a system or its component, where a function is described as a specification of behavior between outputs and inputs.These can be of two types, software and hardware.

4.2.1 Software Requirements

- Operating system: Windows 7 and above
- Anaconda Navigator(anaconda3.x)
- Tool used : Jupyter Notebook
- .Python Libraries such as Numpy,Matplotlib,pandas,condas,nltk.

4.2.2 Hardware Requirements:

- Processor – i3 and above
- Hard Disk – 500 GB
- Memory – 4 GB RAM

CHAPTER 5
PROPOSED ARCHITECTURE

5.1 System Architecture

The architecture diagram provides an overview of the entire system, identifying the main components.The idea is to mention every work area briefly.

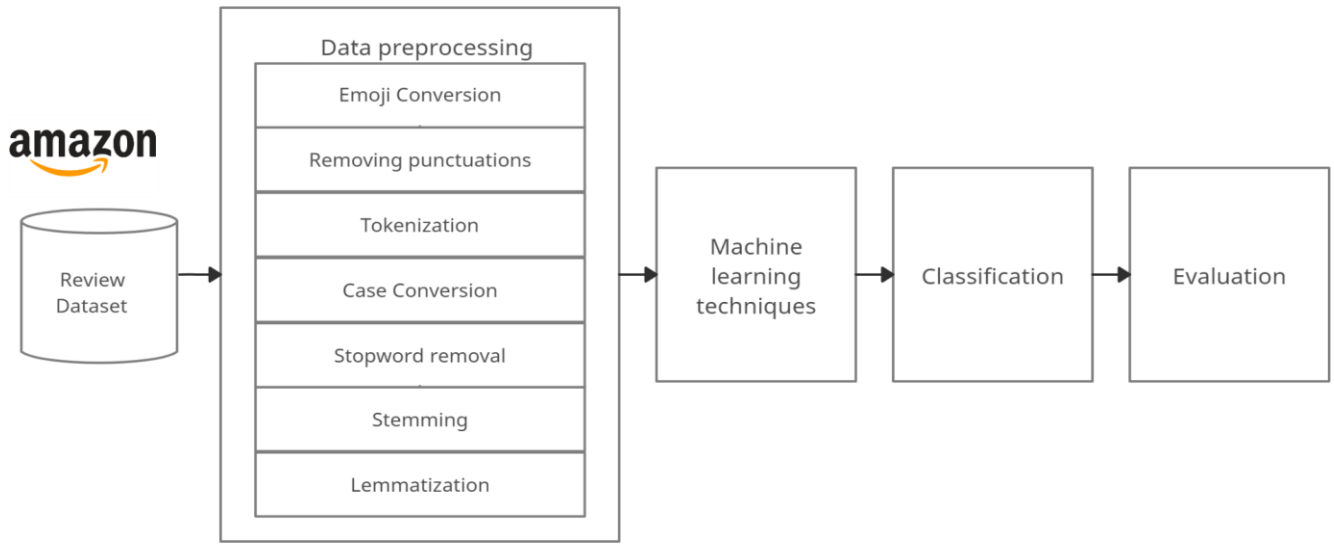


Fig 5.1 System architecture

- The proposed system is basically composed of 5 main modules namely, Data acquisition, Data preprocessing, Machine learning techniques, Classification and Evaluation.
- The first process is Data acquisition which is the process of gathering Amazon product reviews to perform sentiment analysis.
- In the second module, this dataset undergoes various steps of preprocessing to transform and refine the reviews into a data set that can be easily used for subsequent analysis.
- The third module concerns applying different machine learning techniques for building a classification model.
- The resulting classification models are then compared using various performance metrics to determine the best algorithm.

CHAPTER 6

SYSTEM DESIGN

Sentiment analysis is a type of natural language processing for tracking the sentiments of the public about a particular product or topic. Sentiment analysis, which is also called opinion mining, involves building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets.

6.1 Data acquisition

The dataset used for training consists of a big dataset (4 million reviews) available on Kaggle. This Kaggle dataset consists of Amazon customer reviews (input text)

Categories	Number of reviews
Data science book	20648
Electronic products	35633
Musical Instruments	10259
Fine foods	10000

Data specifications

6.2 Data pre-processing

Raw reviews are full of noise, misspellings, and contain numerous abbreviations and slang words. Such noisy characteristics often influence the performance of sentiment analysis approaches. Thus, some preprocessing approaches are applied prior to feature extraction.The preprocessing of reviews includes the following steps:

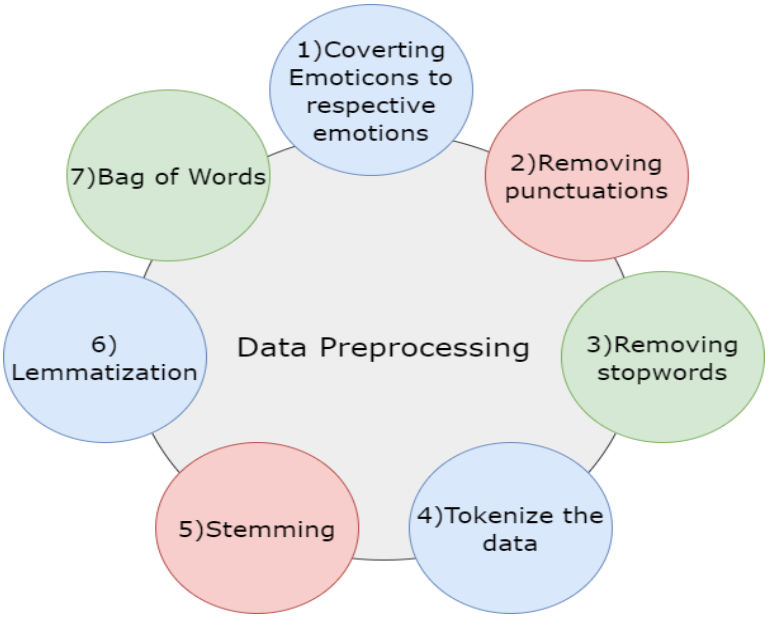


Fig 6.2 Steps in data preprocessing

a)Converting emojis to their respective emotions

The use of emoji on the internet rapidly increased in recent years.People often use them when it is difficult to describe their expressions only with words.A single Emoji character may enhance the expressivity of a text message. A name of a city has no sentiment value when it is posted alone. However, if the user used an Emoji along with this name, the text may have a sentiment value. For example, a smiling Emoji character can express someone's positive feeling towards the city. In contrast, using the angry face Emoji along with some brand name may reveal negative feelings towards the brand.In this model we convert Emojis to their respective emotions.

b)Removing the punctuations in the reviews

Often used to divide a text into sentences, phrases, punctuation marks must be removed.It is a standardization process that treats "excited" and "excited!!!" the same way.

c)Tokenize the data

Tokenization is the process of breaking up a given text into units called tokens. Tokens can be individual words, phrases or even whole sentences.These tokens are very useful for finding such patterns as well as is considered as a base step for stemming and lemmatization.Stemming and Lemmatization both generate the root form of the inflected words obtained from tokenization.

d)Case conversion

Converting all the words to lowercase.Words like “sad” and “Sad” are the same, but when not converted to lowercase they will be considered as two different words.

e)Removing Stopwords

Stopwords are often added to sentences to make them grammatically correct, for example, words such as a, is, an, the, and etc.These should be removed so machine learning algorithms can better focus on words which define the meaning/idea of the text.

f)Stemming

Stemming is a normalization technique used in Natural language processing that reduces words to their common root by removing plurals, genders, and conjugation We can do stemming in NLP using libraries such as PorterStemming, Snowball Stemmer, etc.For example, we have the word ‘Eating', we remove the suffix ‘ing’ and bring that word to a base word ‘eat’.Stemming is mainly used to reduce the dimensionality of data. In simple words, if there we have words like walk, walks, waited, waiting that are different but similar contextually. We bring these words to the base word ‘walk’ by removing suffixes from all the words.As there is no accuracy in stemming,we will not get a meaningful word, it only cuts off the suffix.So we are doing Lemmatization to overcome this issue.

f)Lemmatization

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization the root word is called Lemma. A lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words.For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words. Because lemmatization returns an actual word of the language, it is used where it is necessary to get valid words.

g)Bag of Words

Bag of Words (BOW) is a method to extract features from text documents. These features can be used for training machine learning algorithms. It creates a vocabulary of all the unique words occurring in all the documents in the training set.

CHAPTER 7
SYSTEM IMPLEMENTATION

We have implemented 5 Algorithms and tested the accuracy for each by splitting the dataset into test and train. The 'TextBlob' Library is used to get the subjectivity and polarity of each review.It gives some positive values and some negative values for each review.To achieve this we should load and normalize the column 'Stop reviews' The 'CleanedText' is the column containing normalized data and it is in the text format.

```
[ ] from textblob import TextBlob
    #load the descriptions into textblob
    desc_blob = [TextBlob(desc) for desc in df['CleanedText']]
    #add the sentiment metrics to the dataframe
    df['tb_Pol'] = [b.sentiment.polarity for b in desc_blob]
    df['tb_Subj'] = [b.sentiment.subjectivity for b in desc_blob]
    #show dataframe
    df.head(3)
```

	AMAZON_TEXT_REVIEWS	CATAGORIES	StopWReviews	CleanedText	tb_Pol	tb_Subj
0	I order 3 of them and one of the item is bad q...	Amazon Electronics Products	order one item bad quality miss backup spring ...	order one item bad quality miss backup spring ...	-0.700000	0.666667
1	Bulk is always the less expensive way to go fo...	Amazon Electronics Products	bulk always less expensive way go product like	bulk always less expensive way go product like	-0.333333	0.383333
2	Well they are not Duracell but for the price i...	Amazon Electronics Products	well duracell price happy	well duracell price happy	0.800000	1.000000

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.It is available in the NLTK package and can be applied directly to unlabeled text data.VADER sentiment analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores.The sentiment score of a text can be obtained by summing up the intensity of each word in the text.We check only if the text expresses a positive, negative or neutral opinion and not the subjective fact or opinion.VADER's `SentimentIntensityAnalyzer()` takes in a string as input and returns a dictionary of scores in each of four categories:

1)negative

2)neutral

3)positive

4)compound

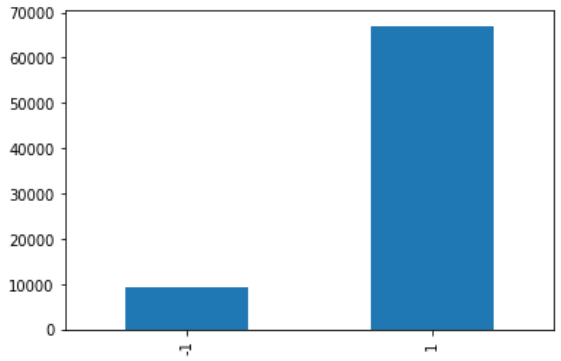
```
pip install vaderSentiment
```

```
Requirement already satisfied: vaderSentiment in c:\users\avinash\anaconda3\lib\site-packages (3.3.2)Note: you may need
Requirement already satisfied: requests in c:\users\avinash\anaconda3\lib\site-packages (from vaderSentiment) (2.24.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\avinash\anaconda3\lib\site-packages (from requests->vaderS
Requirement already satisfied: idna<3,>=2.5 in c:\users\avinash\anaconda3\lib\site-packages (from requests->vaderSentimer
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\avinash\anaconda3\lib\site-packages (from requests->vaderSe
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in c:\users\avinash\anaconda3\lib\site-packages (
```

compound	neg	neu	pos
-0.6249	0.298	0.702	0.000
0.3612	0.000	0.737	0.263
0.7003	0.000	0.256	0.744

```
df["sentiment"].value_counts().sort_values().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c8c91aca48>
```




```
[ ]  newdf = df[["CleanedText", "sentiment"]]
     newdf.columns = ["reviews", "score"]
     newdf.tail()
```

	reviews	score
76534	disabled retire rn always wish librarian!happy...	1
76535	one point consider library work oppose work li...	1
76536	overall think excellent resource anyone pursue...	1
76537	great	1
76538	execellent info	-1

We have applied VaderSentimentAnalyzer() to get labels like neg,pos,neu since all the algorithm needs a labeled dataset for further implementation.Our dataset contains 2 columns i,e AMAZON_TEXT_REVIEWS(text reviews) and CATEGORIES(Types of Reviews combined).We check if the values for 'VaderSentimentAnalyzer' after loading

the normalized column is the same as 'TextBlob' or not.It is successful that we got the same positive and negative values for each row on implementing the libraries .Hence comparison done successfully.

We have defined another 2 columns for better understanding.

1)Sentiment -1 means negative review and 1 means positive review.

2)Label 'pos' means positive review and 'neg' means negative review.

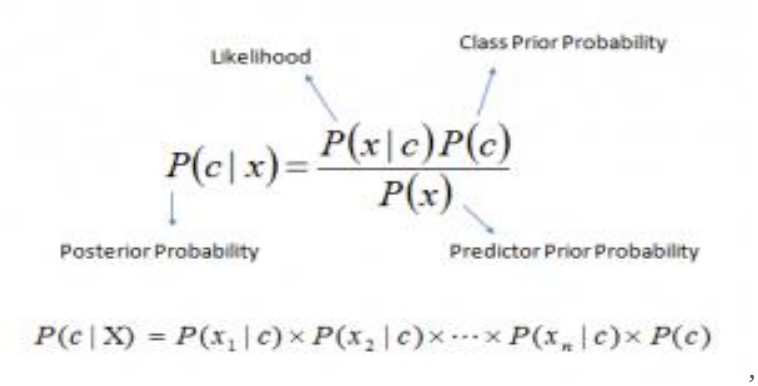
We have appended the above 2 columns named sentiment and Label,which gives a clear picture of the analysis for checking the customer reviews 'positive or negative' .

In this step we have achieved our task of checking the 'sentiment analysis of amazon reviews' i.e positive or negative.Below is the screenshot for the same.We need only 2 columns to work with all the algorithms,to get accuracy of each.For this we have taken 'CleanedText' column data.'Sentiment' column contains -1 and 1 scores that represent positive and negative reviews.For better understanding we have re-named those 2 columns into 'reviews' and 'score'.

7.1 Naive Bayes Implementation

Naive Bayes is a supervised learning algorithm that's typically used for classification problems. Naive Bayes is simple, intuitive, and yet performs well

- It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.
- They are probabilistic, which means that they calculate the probability of each tag for a given text, and then output the tag with the highest one.
- The way they get these probabilities is by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature.
- In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.It predicts the tag of text.
- Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



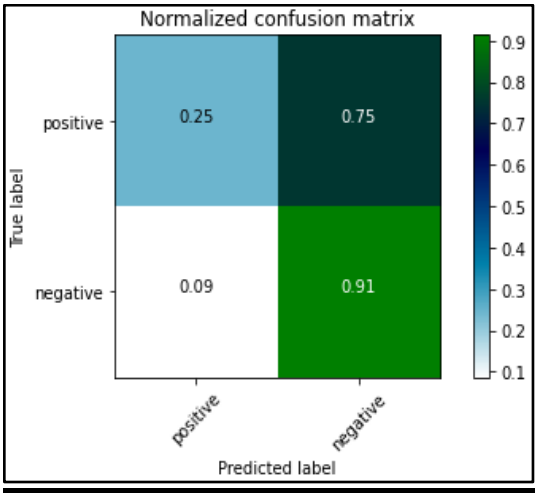
1)P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

2)P(c) is the prior probability of class.

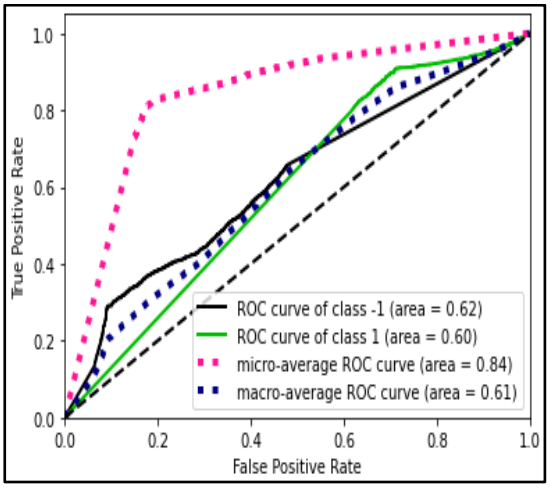
3)P(x|c) is the likelihood which is the probability of the predictor given class.

4)P(x) is the prior probability of the predictor.

Accuracy	Precision	Recall	F1-score	Specificity
0.82	0.88	0.91	0.89	0.25



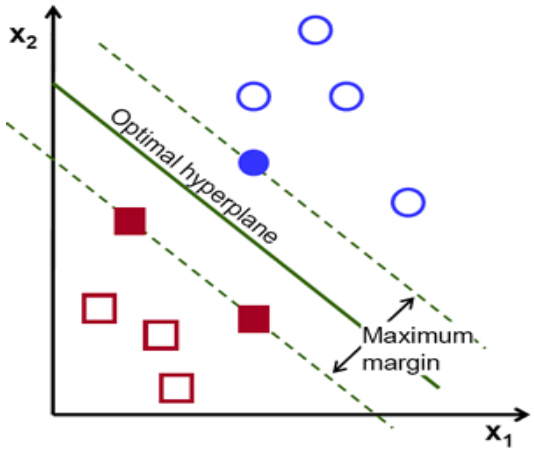
a)Confusion Matrix



b)ROC Curve

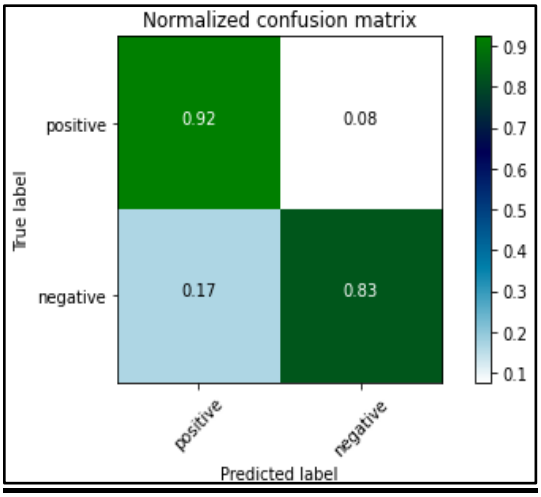
7.2 Support Vector Machine(SVM)

A support vector machine is a supervised machine learning model that uses classification algorithms for two-group classification problems.After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. So we're working on a text classification problem.Objective is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

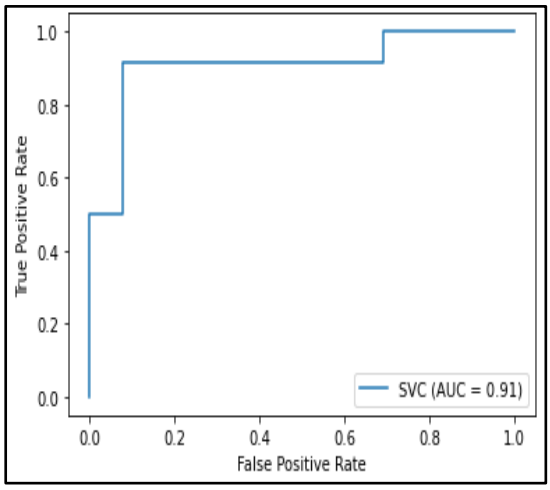


There are many possible hyperplanes that could be chosen.Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes.Here we use LinearSVC method to implement SVM

Accuracy	Precision	Recall	F1-score	Specificity
0.88	0.91	0.83	0.87	0.92



a)Confusion Matrix



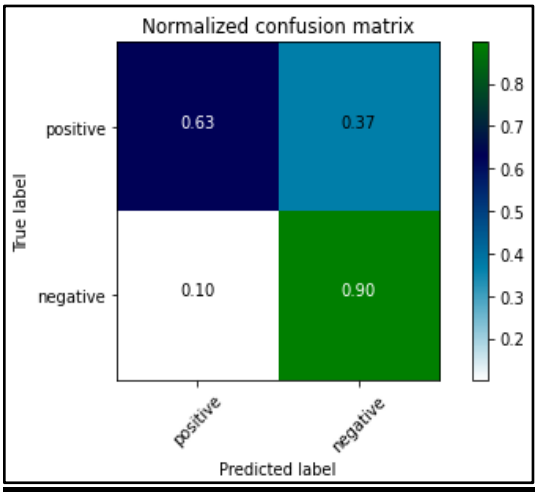
b)ROC Curve

7.3 Logistic Regression(BOW method)

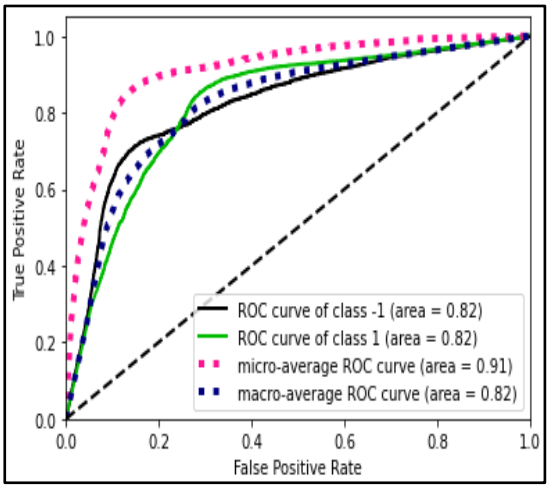
Logistic Regression is a 'Statistical Learning' technique categorized in 'Supervised' Machine Learning methods dedicated to 'Classification' tasks.Logistic regression is a classification algorithm, used when the value of the target variable is categorical in

nature.Logistic regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.

Accuracy	Precision	Recall	F1-score	Specificity
0.86	0.93	0.90	0.91	0.63



a)Confusion Matrix



b)ROC Curve

7.4 Decision Tree Algorithm

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter.Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes.The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.To get corresponding output we take training data samples.

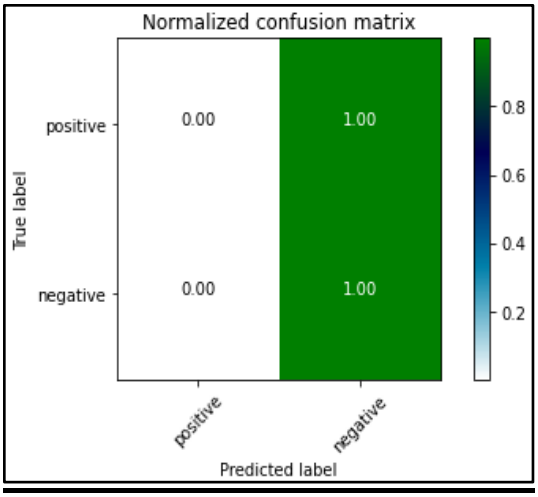
Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.The logic behind the decision tree can be easily understood because it shows a tree-like structure.But here we are just implementing a decision to get accuracy.Decision trees can be divided into two types; categorical variable and continuous variable decision trees.Decision tree contains categorical data(Yes/No)and numerical data.As a standard practice, you may follow 70:30 to 80:20 as needed.But we

should estimate how accurately the classifier predicts the outcome. The accuracy is computed by comparing actual test set values and predicted values.

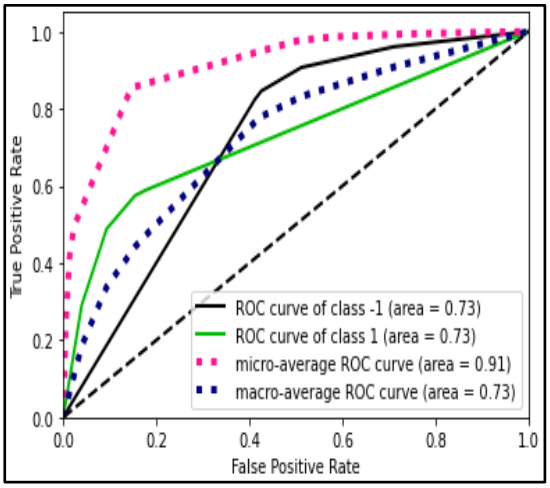
There are several advantages of using decision tree for predictive analysis:

- Decision trees can be used to predict both continuous and discrete values i.e. they work well for both regression and classification tasks.
- They require relatively less effort for training the algorithm.
- They can be used to classify non-linearly separable data

Accuracy	Precision	Recall	F1-score	Specificity
0.85	0.85	1.00	0.92	0.00



a)Confusion Matrix

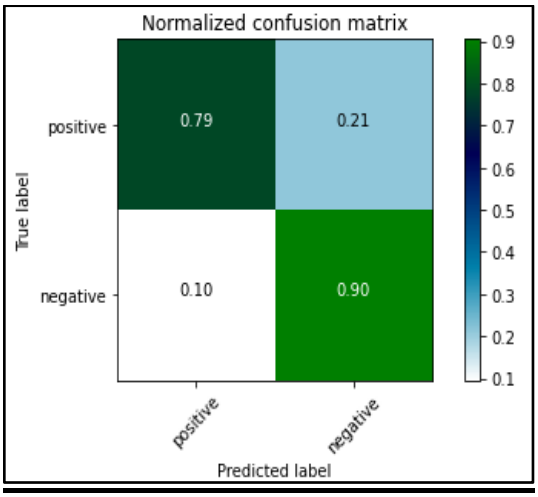


b)ROC Curve

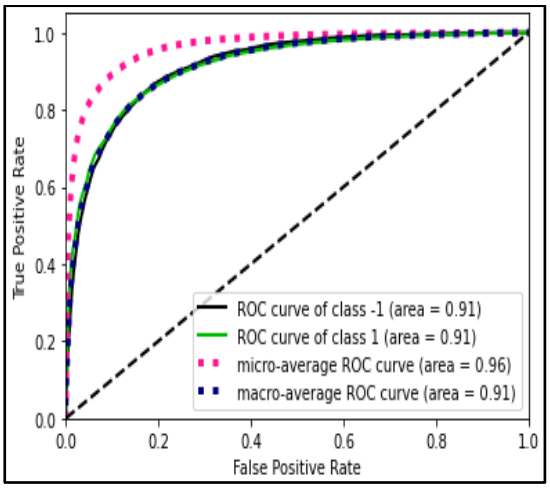
7.5 Random Forest Algorithm

Random forest is a supervised learning algorithm which is used for both classification as well as regression.It creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier.Random forest adds additional randomness to the model, while growing the trees.But as stated, a random forest is a collection of decision trees.With that said, random forests are a strong modeling technique and much more robust than a single decision tree.

Accuracy	Precision	Recall	F1-score	Specificity
0.90	0.98	0.90	0.94	0.79



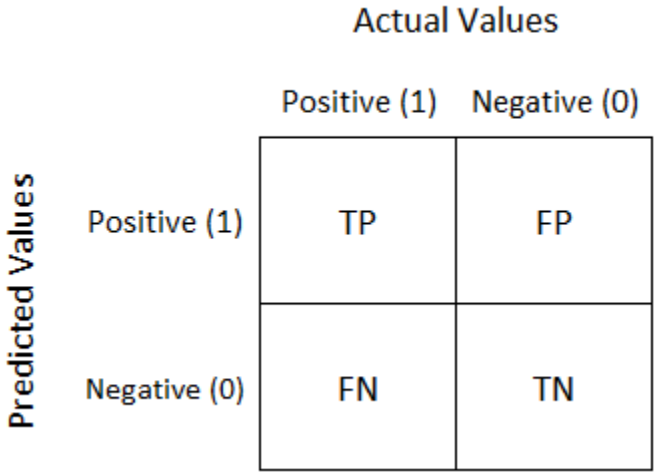
a)Confusion tree



b)ROC Curve

CHAPTER 8
EVALUATION PROCESS

For the evaluation of the proposed work we have used various metrics such as accuracy, precision, recall,f1-score,specificity, confusion matrix and roc curves.



A confusion matrix, also known as error matrix is a NxN matrix that gives a summary for evaluating the performance of a classification model.Since we are taking only positive and negative values the number of target classes N is 2.Hence it is a binary classification problem.Hence we would have a 2x2 matrix as shown above with 4 values.

1. True Positive(TP) : The actual value was positive and the model predicted a positive value.
2. True Negative(TN) : The actual value was negative and the model predicted a negative value.
3. False Positive(FP) : The actual value was negative but the model predicted a positive value.
4. False Negative(FN) : The actual value was positive but the model predicted a negative value.

We have used 6 performance metrics to determine which algorithm is the best:

1)Accuracy : It is defined as the percentage of correct predictions for the test data i.e it compares the predicted sentiment with the real sentiment based on the text.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

2)Precision : Precision is the ratio between the True Positives and all the Positives i.e It is the quantity of the right predictions that the model made.

$$precision = \frac{TP}{TP + FP}$$

3)Recall(Sensitivity) : The recall is the measure of our model correctly identifying True Positives.It tells us how many of the actual positive sentiments we were able to predict correctly with our model.

$$recall = \frac{TP}{TP + FN}$$

4)F1-score : Sometimes when we try to increase the precision of the model, recall goes down, and vice versa.F1-score captures both the trends in a single value.It is the harmonic mean of precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

5)Specificity : It is the proportion of truly negative cases that were classified as negative; thus, it is a measure of how well your classifier identifies negative cases. It is also known as the true negative rate.

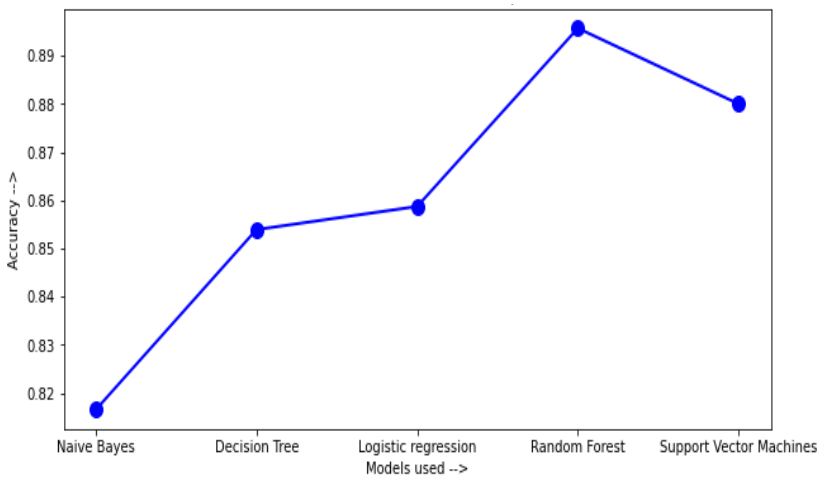
$$specificity = \frac{TN}{TN + FP}$$

6)ROC : ROC stands for Receiver Operating Characteristics.It is a performance measure for classification problems which shows the trade-off between Sensitivity(True Positive Rate) and Specificity(1-False Positive Rate).Classification algorithms that give curves closer to the top-left corner indicates a better performance.The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the algorithm.

CHAPTER 9

RESULTS AND DISCUSSION

In our study, we have used five machine learning techniques to perform sentiment analysis of Amazon reviews.For the evaluation of the proposed work we have used various metrics such as accuracy, precision, recall,f1-score,specificity, confusion matrix and roc curves.Table below summarizes the values obtained for each of the performance metrics



As shown we get the highest accuracy with the Random Forest algorithm. The other performance metrics, such as accuracy or F1-score, are also higher with the Random Forest algorithm when compared with other algorithms.Therefore, we consider this algorithm as the most suitable for the sentiment analysis of Amazon reviews.

Our approach aims to improve quality of sentiment analysis on textual product reviews and simple visual representation of obtained results which will be useful for nontechnical users.Individual customers can take its benefit for decision making and service providers can take advantage to improve quality of service as well as for new product design.Sentiment Analysis can be used to determine attitude of people towards particular product or service.

CHAPTER 10

CONCLUSION

The Internet is a rich source of reviews on ecommerce products or online services. Customers always prefer to read reviews before paying money to the service provider. But it is hardly possible to read all reviews in today's fast life. Also every review may provide new information about the product or feature of the product. So there is a probability of missing any important review given by the consumer.

We are going to identify the polarity of review i.e. whether it is positive, negative or neutral. Sentiment analysis will assist us to find out the polarity of reviews. Due to the large number of reviews we are going to convert textual reviews into visual format by using NLTK it will enhance reliability in decision making.

CHAPTER 11

FUTURE SCOPE

The future looks bright for the use of machine learning in the e-commerce sector.Sentiment analysis is getting better because people on social media are more emotive and expressive than ever.Brands will continue to use this tool but so will the public, government, education centers and many other organisations.Sentiment analysis is on the verge of breaking into new areas of applications. With more and more businesses turning to sentiment analysis to measure and predict results, as well as to understand consumer behaviors, these tools are quickly gaining a reputation that is going to help push it forward into the future and towards deeper and more accurate conclusions and insights.The increasing efficiency and efficacy of the machine learning algorithms with an ever-increasing amount of customer data are paving the path for a machine learning dominated business environment. It is predicted that the price variation, product listing etc will become more and more tailored to the customer and

will be instrumental in customer retention. Also, the sellers will receive deep market analysis helping them in bettering their offering per demographic area.

CHAPTER 12
REFERENCES

[1] Xing Fang* and Justin Zhan "Sentiment analysis using product review data" Fang and Zhan Journal of Big Data (2015) 2:5

[2] Muhammad Taimoor Khan, Mehr Durrani2, Armughan Ali, IrumInayat, Shehzad Khalid and Kamran Habib Khan "Sentiment analysis and the complex natural language" Khan et al. Complex Adapt Syst Model (2016) 4:2

[3] Ji fang, Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification", Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pp. 94–100.

[4] Turney, Peter D., "Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of Association for Computational Linguistics, Philadelphia, PA. July 2002, pp. 417-424.

[5] Turney, Peter D., "Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of Association for Computational Linguistics, Philadelphia, PA. July 2002, pp. 417-424.

[6] AshishShukla* Rahul MisraM.tech Scholar, CSE Department Assistant Professor, CSE Department , Pranveer Singh Institute of Technology, Kanpur Pranveer Singh Institute of Technology, Kanpur U.P.T.U., Luck now, Uttar Pradesh, India U.P.T.U., Luck now, Uttar Pradesh, India "Sentiment Classification and Analysis Using Modified K-Means and Naïve Bayes Algorithm"

[7] Negar Hariri, Carlos Castro-Herrera, Member, IEEE, Mehdi Mirakhorli, Student Member, IEEE, Jane Cleland-Huang, Member, IEEE, and BamshadMobasher, Member, IEEE "Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings" IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 39, NO. 12, DECEMBER 2013.

[8] H. Cherfi · A. Napoli · Y. Toussaint "Towards a text mining methodology using association rule extraction".

[9] Opinion mining and sentiment analysis, Bo Pang1 and Lillian Lee2. 1 Yahoo! Research, 701 First Ave. Sunnyvale, CA 94089, U.S.A., bopang@yahooinc.com, 2 Computer Science Department, Cornell University, Ithaca, NY 14853, U.S.A., llee@cs.cornell.edu

[10] Sentiment Analysis and Opinion Mining: A Survey G.Vinodhini* Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002. RM.Chandrasekaran, Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002. India.

[11] Customers behaviour prediction using artificial neural network BichenZheng, Keith Thompson, Sarah S.Lam, Sang Won Yoon, Nathan Gnanasambandam.

[12] An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques PranaliBorele, Dilipkumar A. Borikar

[13] Review on classification based on artificial neural networks, Saravanan K1 and S. Sasithra2.