# PURCHASING POWER PREDICTION OF CUSTOMERS USING SENTIMENT ANALYSIS

*Dissertation submitted in fulfilment of the requirements for the Degree of*

**MASTER OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

By

**VEERPAL KAUR**

**11502017**

Supervisor

**MR. VIRRAT DEVASER**


**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

MAY 2017

# PAC FORM

# ABSTRACT

In today's day, there are number of e-commerce website like Flipkart, Amazon which are working for online shopping. People are purchasing goods according to their needs from e-commerce websites. There are numerous of people who are doing online shopping. Online shopping websites can be identified the customers who having high purchasing capacity or who are their prime customers to enhance the sales. In these days, customers give the reviews about the products. These reviews show the behaviour of customer for a particular product and interest of customer about the products. Reviews are basically unstructured in nature. Opinion mining is used to get the important formation. Customer's transaction data only gives the estimation of buying capacity but reviews gives the behaviour of customers as well as giving the information of products. Reviews can also be help of customers to give information of their supporting facility. This research work proposes an outline of customers buying capacity by using their transaction data and identifying the customer's opinions about the products. In our research work we have found that customers having high purchasing power give the positive opinions.

# DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled " PURCHASING POWER PREDICTION OF CUSTOMERS USING SENTIMENT ANALYSIS" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab, is an authentic work carried out under supervision of my research supervisor Mr. Virrat Devaser. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

*Signature of Candidate*

**VEERPAL KAUR**

**Reg. No. 11502017**

# SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**PURCHASING POWER PREDICTION OF CUSTOMERS USING SENTIMENT ANALYSIS"**, submitted by **Veerpal kaur** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

*Signature of Supervisor*

**Mr. Virrat Devaser**

**Date:**

**Counter Signed by:**

1) **Concerned HOD:**
   HoD's Signature: _____

   HoD Name: _____

   Date: _____
2) **Neutral Examiners:**

   **External Examiner**

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   **Internal Examiner**

   Signature: _____

   Name: _____

   Date: _____

# ACKNOWLEDGEMENT

First and foremost praises and thanks to the God, the almighty, for his showers of blessings throughout my research work to complete the research successfully. I would like to express my deep and sincere gratitude to my research supervisor, Mr. Virrat Devaser, for giving me this opportunity to do research and providing invaluable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honour to work and study under his guidance. I am extremely grateful for what he has offered me .I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. Last but not least I would like to thank all my friends for being with me at each step when I need their support. This thesis would never be successful without your support and love.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 PURCHASING POWER:

Customer purchasing power refers to an individual customer or a market capacity to buy the quantities of goods. High buying power means a customer have good income and have the capacity to purchase costly things. Low buying power means a customer cannot afford to buy costly products but may be interested in cheaper products.

In our purposed work customer transaction data taken under considering various factors that can estimate the buying capacity of a customer. The factors that are affecting the customer buying power are Recency, Frequency and Monetary etc. We are taking transaction data it can detect the purchasing capacity of a customer but the opinion mining can predict the emotions of customer about the bought product. In our scenario, transaction data would calculated by RFM method. In RFM method, we are consider how many times recently a customer visit the shopping place taking as recurrence , how many times a customer purchasing goods in a month taking as frequency and how much money a customer spends per deal taking as a monetary.

Sentence level sentimental analysis checks whether a sentence has negative, positive or neutral opinion. Aspect level sentimental analysis do analysis on basis of sentiments present in a file and sentiment required for analysis. In our scenario, documental level sentimental analysis has been taken into account. In machine learning mostly two techniques are used for sentimental analysis which is supervised learning and unsupervised learning. In first technique, datasets are labelled and trained data is used to show reasonable output that is used for decision making. In second technique, no labelled dataset is needed and they cannot be processed easily. To solve this problem, various clustering algorithms are used.

## 1.2 BRIEF INTRODUCTION OF SENTIMENT ANALYSIS:

Sentiment analysis, also known as opinion mining, which analyzes people's opinion as well as emotions towards an entities such as products, organizations, and their associated attributes. In the present day scenario, social media play a vital role in

providing information about any product from different reviews, blogs, and comments. Scholars and researchers applied different machine learning techniques to get the meaningful information from different opinion mining resources like twitter, facebook whatsapp etc. Sentiment analysis is done in three different levels such as document level, sentence level, and aspect level. Document level determines that whether the document's opinion is positive, negative or neutral. Sentence level determines whether the sentence expresses any negative, positive or neutral opinion. Aspect level determines sentiment or expressions within given document and document to which it refers. There are two types of machine learning techniques, first one technique based on supervised learning and second is based on unsupervised learning which are used in sentiment analysis. In supervised learning technique, the dataset is labelled and thus, trained to obtain a reasonable output which help in proper decision making. Unlike supervised learning, unsupervised learning processes do not need any label data; hence they cannot be processed at ease. In order to solve the problem of processing of unlabeled data, clustering algorithms are used.

### 1.2.1 SENTIMENT ANALYSIS API'S:

Sentiment is an opinion, an attitude, thought or judgement of any by feeling. Opinion mining finds the consumers sentiments towards products or any other entities. Social networking, online shopping sites are the places where you can find sentiments of clients .This sentiments may be analysed by the researchers for their specific purpose. Application programme interfaces are provided by many sites which help to collect and analyse the data and use for decision making. For example, Twitter has three API are: REST API, Streaming API and Search API.REST API is used for collecting data and information about users and consumers where Search API helps developers to access specific Twitter content for analysis. The Streaming API helps to developer for collecting real time content. So, we can relate various applications with theses API's to solve our problems. So, Sentimental analysis can be done with Social media and online shopping data from Facebook, Twitter, whatsapp, Amazon, Flipkart and ebay etc. But there are some problems in data to be analysed by sentimental analysis because opinion may not be desired to your choice because everyone has different opinion to a single thing. There is possibility of fake opinion from different fake accounts and other problems like spam

data may also be present in sentimental analysis. Otherwise, online data is not always faithful. So, you have to use a dataset with care so that your analysis must be close to correctness and your decision are benefits to the organization for which you doing this work.

## 1.3 TEXT DATA:

Text data is basically words that are used for identifying columns of data such as headings, labels and names etc. Text data can contain numbers, special characters such as # or @ and letters. Numbers can be used for computations.

These days the majority of information is stored into database in the form of content. This content comes from enterprise data, manufacturing companies and different organizations at very high rate and in different forms. Mostly semi–organized information stored into these databases. The report may have a few to a great extent unstructured content parts for example summary of contents also some organized attributes like heading, entity name, date of birth etc.

## 1.4 TEXT MINING OR OPINION MINING:

Text mining is method of getting information out from a large content of a report, feedbacks etc. It is not easy to extract information from large content of data from different websites, content documents etc.

First of all, Data is collected by different tools and technologies. This data may be structured, unstructured, semi-structured and quasi-structured. So, this data must be pre-processed by a tool to convert a suitable format which is according to tool or technologies you are using. Next, we have to use content investigation method to getting out great values from large content data. By doing, so quality data is extracted and used as a database for further analysis.

### 1.4.1 TEXT MINING RESOURCES:

There are many data sources for text mining. We can perform text mining on papers, news articles, patents or databases. Text mining can also be performed on web sources like social media data, Wikipedia, data repositories etc. Figure 1 shows different text mining sources.

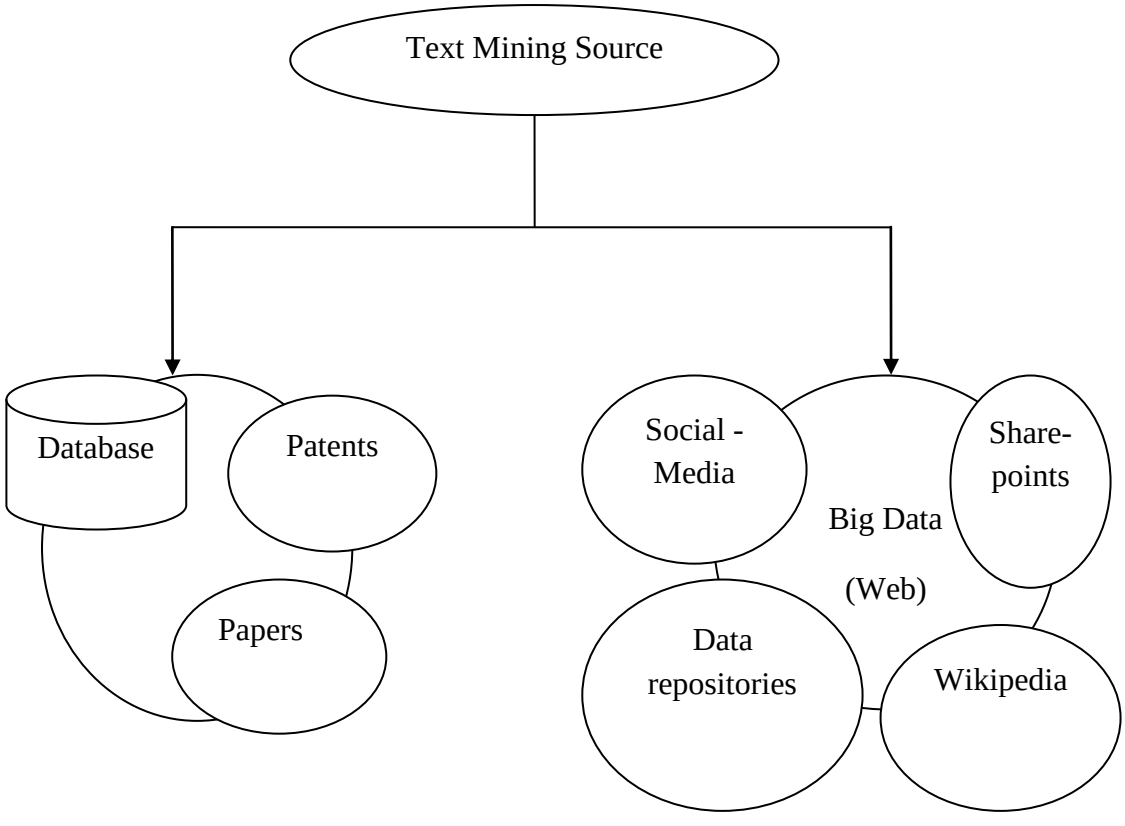


**Figure1: Text Mining Resource**

### 1.4.2 TEXTUAL ANALYTICS:

Textual analytics is a kind of information examination that is used to clarify, understand and translate a circumstance or a man's activities.

### 1.4.3 TEXTUAL ANALYSIS:

Textual analysis is the method that specialists use to clarify and interpret the qualities of a recorded or visual message. The objective of textual analysis is to explain the content, structure, and elements of the content contained in writings.

## 1.5 TEXT MINING TECHNIQUES:

There are many text mining techniques which are written below:

**1) Natural language Processing (NLP)**

It is the field of artificial intelligence and computation linguistics. It is just interactions between machines and human language called human computer interaction. This technique is used to understand the human spoken language. There are very issues in this which are following:

- Difficult to teach computers to the meaning of human language
- Human language understanding
- Human language creation

**2) Retrieval of information**

It is beginning stage in which we investigate unstructured content by discovering key expressions and connections inside content. Pattern matching is used to do this task, search for predefined sequences in content. It contains segmentation, recognizable proof of elements, sentence division, and grammatical feature task. Firstly expressions and sentences are parsed and semantically translated, afterward useful bits of data put into the databases. This technology can be extremely helpful when managing huge quantity of content.

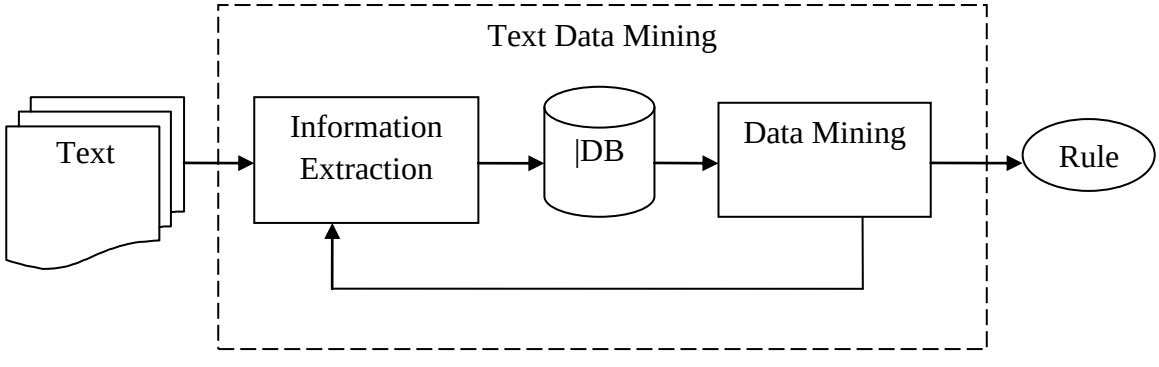


**Figure2: Information Extraction Text Mining**

**3) Classification**

Classification is a method in which the content documents are arranged according to their observed similarities. Classification techniques are k-nearest neighbour classifier, Decision Tree, Naive Bayes classifier, and Support Vector Machine.

**i. Naive Bayes**

Naive Bayes classification technique is mostly used because it is very easy to train and classify the data. In general, it is known as probabilistic classifier and can learn a set of documents to determine the pattern of document so that they have been categorized. It is comparison of contents with the list of words so that classify the documents to their right category.

When the training of any data is complete, then classification which provides the polarity of the sentiments. For example for the review comment "I am intelligent' it provide Positive polarity as result.

**ii. Maximum entropy**

Maximum entropy is based on the conditional probability distribution which maximizes the entropy. It even handles overlap feature and is same as logistic regression which finds distribution over classes. It also follows certain feature exception constraints. Where, c is the class, d is the tweet, and _ is a weight vector. The weight vectors decide the significance of a feature in classification. It follows the similar processes as naïve bayes, discussed above and provides the polarity of the sentiments.

**iii. Support vector machine**

Support vector machine is utilized to examine the information and characterize the choice limits and uses the parts for calculation which are performed in info space. The i information are two arrangements of vectors of size m each. At that point each information spoken to as a vector is grouped in a specific class. Presently the assignment is to discover an edge between two classes that is a long way from any report. The separation characterizes the edge of the classifier, amplifying the edge lessens uncertain choices. SVM additionally underpins arrangement and relapse which are valuable for factual learning hypothesis and it helps perceiving the components correctly, that should be considered, to comprehend it effectively.

**iv. Semantic Analysis**

After the training and classification we used semantic analysis. Semantic analysis is derived from the WordNet database where each term is associated with each other. This database is of English words which are linked together. If two words are close to each other, they are semantically similar. We have a capacity to find out the similar words. We map terms and examine their relationship in the ontology. The main task is to use the

stored documents that contain terms and then check the words that the user uses in their sentences are similar or not. Thus it is helpful to show the polarity of the sentiment for the users. For example in the sentence" I am happy" the word ''happy'' being an adjective gets selected and is compared with the stored feature vector for synonyms. Let us assume2 words; 'glad' and 'satisfied' tend to be very similar to the word 'happy'. Now after the semantic analysis, 'glad'

replaces 'happy' which gives a positive polarity.

**4) Clustering**

It is a method which we make groups of similar content of documents. These groups are called clusters. Number of documents is in every cluster. The qualities of cluster view better, if the one cluster contains the more similar content rather than others.

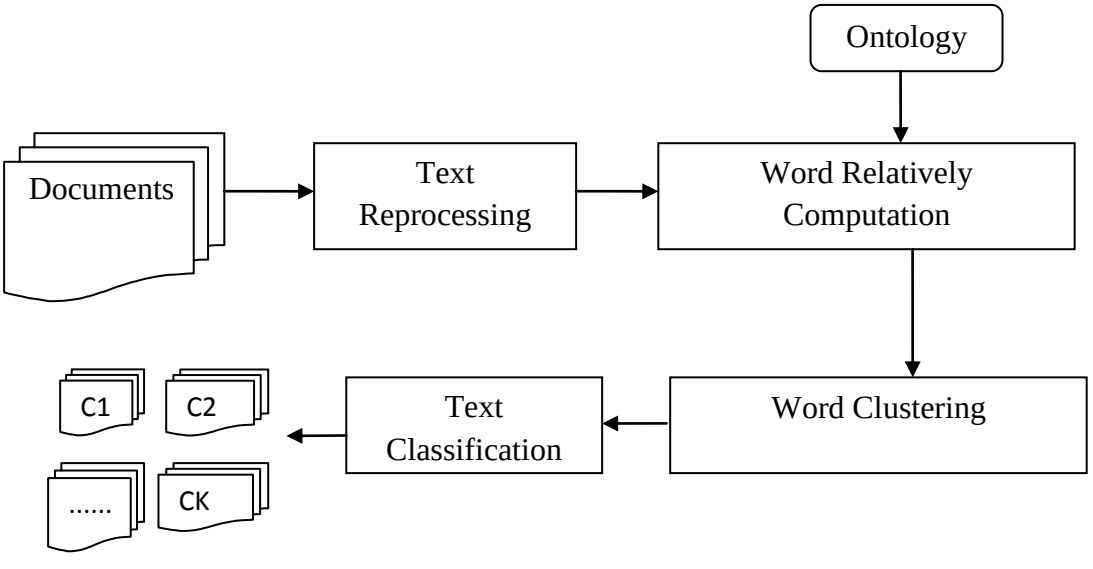


**Figure3: Text Summarization Process**

**5) Visualization**

It is a method which can enhance and make simpler the disclosure of significant data. Text flags and colours are used to identify the category of documents.

**6) Summarization**

It is a method which can be used to decrease the size and element of a document. It holds mainly significant point and common significance. It replaces the number of documents. It saves the time of clients by summarize the huge documents into small documents.

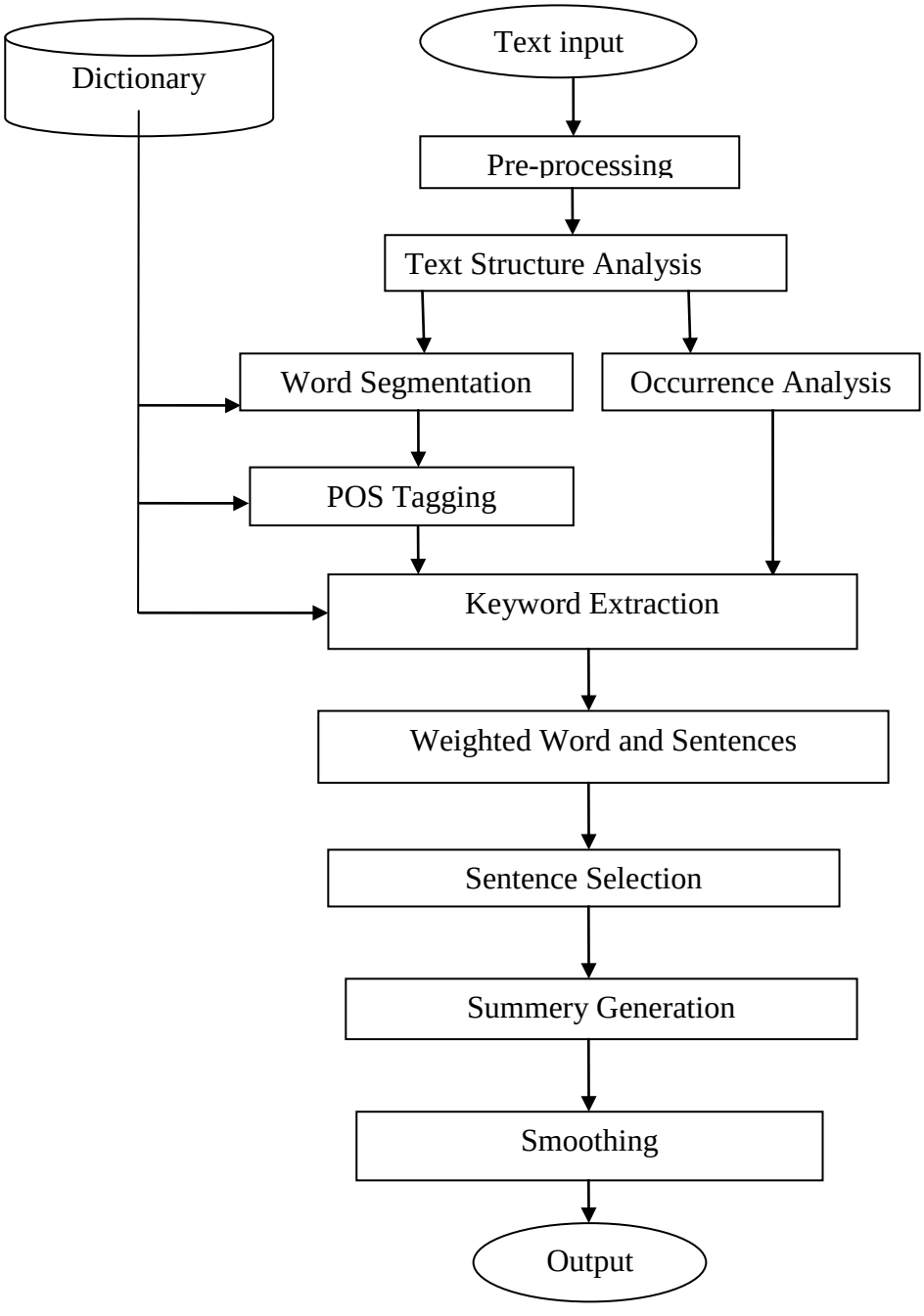


**Figure4: Clustering Process**

## 1.6 DATA MINING

It is the way of extraction of information and patterns from the very large number (or quantity) of data. In the present day, the use of data mining is growing day by day. It is used for decision making activities. Sometimes data mining are called data or Knowledge discovery because it is the process of analysing the data from different perspectives and summarizing it into useful information. Microsoft Academic Research provides ranking for data mining key words which have been grown recently (used in publications, organizations etc.). Few of them are listed below

- Association rules: These are if/then statements that are used to display the relationships between apparently independent data in a relational database or any other information store.
- Machine Learning: Both data mining and machine learning used same methods. But there is difference, machine learning focused on prediction, based on known properties, whereas data mining focuses on identification of unknown properties.
- Support Vector Machine: It is supervised learning algorithm which analyzes data used for classification.
- Cluster Algorithms: Clustering is one of the emerging research fields in data mining due to its numerous applications. Example : K-means
- Information Retrieval
- Search Engine
- Web Search
- Indexation
- Social Network

Data Mining is the process to find the hidden information as well as pattern from a bulk amount of data i.e. the data should be coming from different sources such as data ware house, Data mart etc.

### 1.6.1 DATA MINING TECHNIQUES:

**1.6.1.1 Classification:** The classification is done because of exactly guess the aimed class for all case in the data. One of the example of this model is it is help to predict the student performance.

- In inclusion, there are two stages in classification. The initial part is the learning process. In this part, the training data or facts are examined by classification algorithm and rules and design are created which are based on learned model or classifier.
- In the second part the model is used for classification and testing data are used for gaining the accuracy of classification design. Then, establish on the sufficient accuracy, the rules can be used for the classification of new or recently developed data or for unseen data.

**1.6.2.2 Decision Tree**: Decision trees are broadly used in the classification procedure. With the help of this, the model can be predicted and classified. Decision trees shows rules, which may followed by individual and used in knowledge structure like a database.

Example: if the attendance is not matching the giving criteria then the chances of giving the exam is less.

- Decision Tree can be represented as:

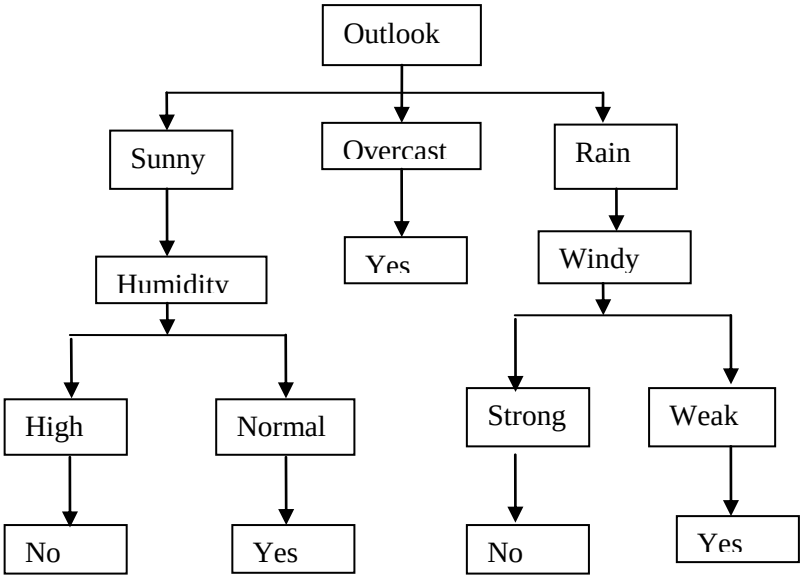


**Figure 5: Decision Tree**

It's like a flowchart. In this rectangular shapes of boxes are called node internal nodes are those nodes that have a child and the leaf node are those who don't have children. The top nodes are root node. In the given figure outlook is a root node. Humidity and Windy is an internal node

**1.6.2.3 Naïve Bayes:** It is a conditional probability approach. In which a mention problem case is to be classified, and it can be elected by a vector x=(x1, x2……xn) representing a few *n* features. Using Bayes theorem it is written as

Posterior=prior X likelihood/evidence.

**1.6.2.4 Fuzzy logic:** It is a method to determine the "degree of facts" instead of the general "true or false" (1 or 0).Data mining uses different methods (approaches) and assumption from a broad areas or fields for the knowledge extraction from huge amount of data. But uncertainty is a general phenomenon in data mining problems. Therefore, it is applied to manage with the uncertainty in actual world.

#### 1.6.2.5 Clustering:

Clustering is a procedure of dividing a group of data (or objects) into a set of significant family, called clusters. Clustering can be used as stand–alone tool to get inside into data distribution or it can be used as pre-processing step for other algorithms.

#### 1.6.2.6 Association rules:

Association rules are if/then statements that are used to display the relationships between apparently independent data in a relational database or any other information store. Example of this would be "If a customer buys a dozen of eggs, he is 80% possibility also to purchase the milk."

# CHAPTER 2

# REVIEW OF LITERATURE

Anshul Mittal et.al has worked on sentiment analysis to discover relationship among community emotion and bazaar emotion. Author proposed Self Organizing Fuzzy Neural Networks (SOFNN) algorithm for classification of community emotion into 4 categories that is cool, happy, attentive and mercy. Author developed his own methodology which is word list generation based on questionnaire. In this paper, author asked 65 different questions to find out the current mood of user. The answers of 65 questions are feed into the pre processor. They made a standard of 6 words is called Profile Of Mood State (POMS). These 65 words were mapped with POMS. Author used a scale rate from 1 to 5. Author proposed SentWordnet approach to extend their list by considers all relevant synonyms. Author used only some words to filter the tweets.
Total no. of matches of all words in tweets varies every day. This algorithm worked for maximum only 140 characters. Author proposed static correlation mapping techniques to map the score of each word and also cross validated their results by comparing the values returned by algorithm and sentiment analysis technique. He compared the events like Michael Jackson's death which was on 25th June of 2009 and thanks giving day which was on 26th of November 2009 and found that the value of line in graph suddenly decreased after 25th June 2009. A naive greedy strategy to make the decision about buy and sell decision was created. Buy decision have to be taken if predicted stock value is greater than standard deviation for next day. Sell decision are required to be taken if predicted stock value and standard deviation are more than the actual value at that time. [1]

Hiroshi Sugimura et.al has proposes a framework in which feature patterns were extracted. This data is extracted from fund, medical research, modern sensors, and so on. The framework extracted the features that describe similar information in database. They concentrate on two parts of the characteristic design: universal and limited occurrence. The framework cut out sub successions from this data. Through utilizing grouping, numerous representatives' sequence was separated from these sub successions. They made a strategy that applies TF*IDF weight method to time series data, which was used

as a part of content mining. The time series information was grouped by using the gained characteristic design. The characteristic designs get to be characteristics in the machine learning and were enhanced with use of hereditary calculation. They settled on a choice tree that decided future practices. They clarified how these two instruments were combinatory connected in the whole information disclosure process. [2]

Xing Huang et.al has discovered a number of micro-blog business-related mining algorithms which consider the connection among the expressions and the how many times it appears in the text, and ignores the expression which is divided in a specific category, that decreased correct levels of micro-blog business-related word mining. This Issue is solved by using TF-IDF (Term Frequency- Inverse Document Frequency) algorithm which calculates the weight of each word. To check the possibility of the better algorithm, they made classes the enormous micro-blog information into definite types and after that they used better TF-IDF algorithm to evaluate word weight between the categories. They computation was worked in the Hadoop circulated system. The test marks showed that the better TF-IDF algorithm was valuable. When they compared with conventional algorithms, this enhanced algorithm has very much better accuracy. The speed of data processing has also really enhanced by using Hadoop. The micro-blog called mini blog. The micro-blog database size was exploding by an exponential rate when the rate of use of micro-blog users increased. Micro-blog data compared with conventional text data had its own distinctive features: less word, better arbitrary of the content, less effective features, and vague concept description. These features have creating trouble in processing data. They faced a hard problem in micro-blog data processing that how to discover the precious information exactly and fast among such huge text information in micro-blog. An expression plays a vital role on community view investigation, but also it has a very high commercial value. So in the field of enormous social network data study, it was very necessary to pull out the expensive commercial phrase fast and precisely in distributed computing environment., The objective of author to get better conventional mining algorithm and after that transplant it into the Hadoop distributed framework, so that words with commercial value more fast and precisely extracted among the huge micro-blog information. [3]

Xiuzhen Zhang et.al has discussed Internet business applications used reputation based trust models to figure vendors reputation trust scores. The reputation scores were all around high for merchants and it was troublesome for purchasers to choose reliable vendors. Purchasers could express their sentiment straightforwardly in free content

criticism remarks. Author proposed Comment-based Multi-dimensional trust display for purchasers to assess reputation score from client input remarks. They proposed an algorithm for mining input remarks for measurement ratings and weights, joining strategies of natural language processing, conclusion mining and subject displaying. eBay and Amazon information demonstrated that CommTrust could efficiently deal with "all great reputation" matter and level vendors adequately.

eBay and Amazon reputation framework figure reputation scores for merchants were processed by totalling feedback ratings. The fundamental issue with the eBay reputation administration framework was "all great reputation" issue, where input feedback is more than 99% positive by and large. eBay detailed seller rating for sellers (DSRs) on four parts of exchanges, to be specific thing, communication, delivery time, delivery and paying were likewise examined. The clients who leave the negative criticism their remarks could draw in the negative ratings, in this way they harm their own reputation. They used vocabulary opinion mining strategy to separate the feeling from criticism remarks. Author proposed Latent Dirichlet Allocation (LDA) cluster strategy to process the collected measurement ratings and weights. The entire algorithm was called Lexical-LDA. They worked fundamental in three zones was trust assessment, examination of criticism remarks and the motion picture review and item survey. The EigenTrust algorithm was used to process the trust evaluations. [4]

Thin Nguyen et.al has examined that online people group were used by expansive number of individuals to talk about emotional wellness issues. The principle thought process to concentrate the attributes of online sadness groups (CLINICAL) in correlation with those joining other online groups (CONTROL). They used machine learning and statistical techniques to segregate online mail amongst depressions and control groups via state of mind, psycholinguistic procedures and substance subjects pulled out from posts created via individuals from groups. Each and every one viewpoints counting frame of mind, composed substance and composing manner are observed to be essentially unique among two groups. Opinion examination demonstrated the clinical gathering have been small valence than individuals in the group. They demonstrated great analytical strength in depression order utilizing points and psycholinguistic pieces of information like components. Unbiased understanding between composing styles and substance, great validity influence was a vital stride in comprehension web-based community networking and their utilization in emotional well-being. They have explored online depression groups and concentrated their differential essentials to other online groups. Three aspects

were inspected: influence, psycholinguistic procedures and points inside substance. Device education and numerical strategies were utilized to depression online post amongst depression and control groups. Every one of viewpoints: influence, composed substance and composing style were observed to be altogether extraordinary among these two gatherings. Underlying themes were bringing into to have more noteworthy extrapolative influence than language characteristics for expectation of despair groups. [5]

Renata L. Rosa et.al has determined that sentiment analysis an order a verdict with positive, neutral or negative power. This paper displayed a music proposal framework in light of a sentiment power metric, called enhanced Sentiment Metric (eSM) which was the relationship of a vocabulary opinion metric between client's profile views. The clients' opinions were taken out from reviews which were given on informal organizations and composition proposal framework was worked throughout a structure of near to the ground many-sided quality designed for cell phones that recommends tunes in light of the present client's conclusion power. The subjective tests comes about highlighted the significance of taking into account client's report into a feeling metric. In light of the outcomes, a correction component which was relies on become old, instructive point and femininity. The correction factor was used to acquire all the more genuine sentiment force esteem. The eSM, enhanced the music suggestion framework, demonstrating that feelings could change which relies upon client's profile. [6]

Yuefeng Li et.al has clarified creative model for significance include in content mining. It separated the both positive and negative examples from the content record. The target of significance attribute detection was to easily detect both important and unimportant content from document. There were two fundamental issues in pattern matching procedures. The one was the low-support issue. Long patterns were generally more particular for the subject; however they normally show low support or in documents. While base support was diminished, then a lot of noise patterns could be found. The other issue was distortion issue in pattern mining. There were many earlier strategies in content mining to understand these issues. Pattern taxonomy mining (PTM) models have been used shut consecutive patterns in content sections. Content based model (CBM) used natural language processing (NLP) procedures to discover the ideas. To discover ideas from sentences they proposed verb-contention structure. Throughout the years, individuals have created numerous common natural language processing term-based strategies for positioning records, data separating and message arrangement. There

several hybrid approaches were proposed to learn term features within only relevant documents for text classification. It used a Rocchio classifier to separate arrangement significant reports from the unlabeled set in the first stage. They fabricated a SVM classifier to arrange content records in the second stage. A two-organize model was demonstrated that the coordination of the rough investigation (a term-based model) and pattern taxonomy mining is the most ideal approach to outline a two-arrange display for data separating framework. [7]

Lorenzo Gatti et.al has examined that inferring polarity lexica for opinion investigation in which polarity rate were connected by way of words outside the realm of relevance was a difficult issue. For the most part, an exchange off amongst accuracy and scope was hard, and it relies on upon the system used to make the vocabulary. Physically commented on lexica give a high exactness however need in scope, though programmed determination from previous learning ensures high scope at the charge of a minor accuracy. Programmed deduction of earlier polarities was less tedious than manual explanation. There has been an awesome blossom of these methodologies, specifically in view of the SentiWordNet asset. They looked at the most every now and again used procedures in light of SentiWordNet by fresher and mixed them in a knowledge formation. By exploiting physically made earlier extremity lexica, ensemble technique was better ready to predict the earlier estimation of hidden words and well performed the various SentiWordNet techniques. Using this system they made SentiWords, an earlier extremity dictionary of numerous words that had a high exactness and a high scope. They at last demonstrated that in assessment examination assignments, using their dictionary permitted us to perform well both the on its own measurements got from SentiWordNet and well known physically clarified sentiment lexica. [8]

Mondher Bouazizi et.al has discussed that most part analysts are work on the sentiment analysis which restricts the interruption of human and gathered sentiment from interpersonal organizations using content mining strategies. Long range informal communication sites are adjusted towards the grouping of writings into positive and negative. Author proposed an approach which was based on pattern that works on data gathered from Twitter. They characterize the tweets into 7 distinct classes specifically, “happiness”, “sadness” , “anger”, “love”, “hate”, “sarcasm” and “neutral”. The approach ends up being exceptionally exact in double grouping ("positive" and "negative") than multi-class arrangement ( "positive", "negative" and "neutral"). Twitter turned into an

extremely mainstream stage for individuals to express their considerations about items or films, share their every day encounter and their conclusion about forthcoming occasions, for example, sports or political decisions. There are some pattern- based elements: Sentiment-based elements are those which in light of the opinion extremity (i.e., "positive"/"negative") of the distinctive parts of tweets. They first calculate emotional scores using SentiStrength. [9]

Pang et.al has been used sentiment classification which was based on categorization such as negative and positive sentiments. In this research, these three machine learning algorithms, such as Naive Bayes, Support Vector Machine, and Maximum Entropy were under-taken for experiments. The classification process was attempted using the n-gram technique like unigram, bigram, and combination of both unigram and bigram. They have implement the machine learning algorithm using framework of bag-of- word features. According to their result analysis among the three algorithms SVM algorithm shows the better result and Naive Bayes yields poor result. [10]

Salvetti et.al has discussed on Overall Opinion Polarity (OvOp) concept using machine learning algorithms such as NB and Markov model for classification. In this paper, Part Of Speech tag work like a lexical filter and wordnet provided hypernym for classification. As their result analysis shows that the result obtained POS filter is more accurate in comparison with that of wordnet. In the field of OvOp, recall was given less importance in comparison with that of accuracy. The authors presented a system in which they given rank reviews based upon function of probability. According to them, their approach shows better result in case of web data. [11]

Beineke et.al has used Naive Bayes model for sentiment classification. They have extracted pair of derived features which are linearly combinable to predict the sentiment. In order to improve the accuracy result, they have added additional derived features to the model and used labelled data to estimate relative influence. They have followed the approach of Turney which effectively generates a new corpus of label document from the existing document. This idea allows the system to act as a probability model which is linear in logistics scale. The authors have chosen five positive and negative words as anchor words which produce 25 possible pairs and they used them for the coefficient estimation. [12]

Mullen and Collier has applied SVM algorithm for sentiment analysis where values are assigned to few selected words and then combined to form a model for classification Mullen and Collier (2004). Along with this, different classes of features having closeness to the topic are assigned with the favourable values which help in classification. The authors have presented a comparison of their proposed approach with data, having topic annotation and hand annotation. The proposed approach has shown better result in comparison with that of topic annotation where as the results need further improvement, while comparing with hand annotated data. [13]

Dave et.al has used a tool for collecting reviews, then move them and using aggregation sites for sorting. These organized reviews are used for testing and training. After that features of reviews were identified and using scoring methods to determine whether the re- views were positive or negative. In this paper, classifier was used to classify the sentences obtained from web-search through search query using product name as search condition. [14]

Matsumoto et.al has used the syntactic relationship among words as a basis of document level sentiment analysis. In this paper, frequent word sub- sequence and dependency sub-trees are extracted from sentences, which act as features for SVM algorithm. They extract unigram, bi- gram, word subsequence and dependency sub-tree from each sentence in the dataset. They used two different datasets for conducting the classification, IMDb dataset and Polarity dataset. In case of IMDb dataset, the training and testing data are provided separately but in Polarity dataset 10-fold cross validation technique is considered for classification as there is no separate data designated for testing or training. [15]

Zhang et.al has proposed the classification of Chinese comments based on word2vec and SVM performance. Author's approach was based on two parts. In first part, they have used word2vec tool to cluster similar features in order to capture the semantic features in selected domain. Then in second part, the lexicon based and POS based feature selection approach was adopted to generate the training data. Word2vec tool adopts Continuous Bag-of-Words (CBOW) model and continuous skip-gram model to learn the vector representation of words. SVM performance is an implementation of SVM for multivariate performance measures, which follows an alternative structural formulation of SVM optimization problem for binary classification. [16]

Liu and Chen have proposed various multi-label classifications on sentiment classification. They have used eleven multilevel classification methods compared on two micro- blog dataset and also eight different evaluation matrices for analysis. Apart from that, they have also used three different sentiment dictionaries for multi-level classification. According to the authors, the multi-label classification process perform the task mainly in two phases i.e., problem transformation and algorithm adaptation. In problem transformation phase, the problem is transformed into multiple single-label problems. During training phase, the system learns from these transformed single label data, and in the testing phase, the learned classifier makes pre- diction at a single label and then translates it to multiple labels. In algorithm adaption, the data is transformed as per the requirement of the algorithm. [17]

Luo et.al has proposed an approach to convert the text data into low dimension emotional space (ESM). They have annotated small size words, which have definite and clear meaning. They have also used Ekman Paul's research to classify the words into six basic categories such as anger, fear, disgust, sadness, happiness and surprise .They again have considered two different approaches for assigning weight to words by emotional tags. The total weight of all emotional tags were calculated and based on these values, the messages are classified into different groups. Although their approach yields reasonably a good result for stock message board, the authors claim that it can be applied in any dataset or domain. [18]

Niu et.al has proposed a Multi-View Sentiment Analysis (MVSA) dataset, including a set of image-text pair with manual annotation collected from Twitter. Their approach of sentiment analysis can be categorized into two parts, i.e., lexicon based and statistic learning. In case of lexicon based analysis, a set of opinion words or phrases are considered which have pre-defined sentiment score. While in statistic learning, various machine learning techniques are used with dedicated textual features. [19]

Gayatree Ganu et.al said that online reviews are an important asset for users deciding to buy a product, see a movie, or go to a restaurant, as well as for businesses tracking user feedback. However, most reviews are written in a free-text format, and are therefore difficult for computer systems to understand, analyze, and aggregate. One consequence of this lack of structure is that searching text reviews is often frustrating for users. User experience would be greatly improved if the structure and sentiment conveyed

in the content of the reviews were taken into account. Our work focuses on identifying this information from free-form text reviews, and using the knowledge to improve user experience in accessing reviews. Specifically, we focused on improving recommendation accuracy in a restaurant review scenario. In this paper, they report on our classification effort, and on the insight on user-reviewing behaviour that they gained in the process. We propose new ad-hoc and regression-based recommendation measures, that both take into account the textual component of user reviews. Their results show that using textual information results in better general or personalized review score predictions than those derived from the numerical star ratings given by the users. [20]

Geetika Gautam et.al has discussed that wide spread of World Wide Web has brought a new way of expressing the sentiments of individuals. It is also a medium with a huge amount of information where users can view the opinion of other users that are classified into different sentiment classes and are increasingly growing as a key factor in decision making. This paper contributes to the sentiment analysis for customers' review classification which is helpful to analyze the information in the form of the number of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of these two. For this we first pre-processed the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms namely: Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature. Finally we measured the performance of classifier in terms of recall, precision and accuracy. In this paper, author proposed a set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on twitter data. The key aim is to analyze a large amount of reviews by using twitter dataset which are already labeled. The naïve byes technique which gives us a better result than the maximum entropy and SVM is being subjected to unigram model which gives a better result than using it alone. Further the accuracy is again improved when the semantic analysis WordNet is followed up by the above procedure taking it to 89.9% from 88.2%. The training data set can be increased to improve the feature vector related sentence identification process and can also extend WordNet for the summarization of the reviews. It may give better visualization of the content in better manner that will be helpful for the users. [21]

Philip Beineke et.al discussed that sentiment classification is the task of labeling a review document according to the polarity of its prevailing opinion (favorable or unfavorable). In approaching this problem, a model builder often has three sources of information available: a small collection of labelled documents, a large collection of unlabeled documents, and human understanding of language. Ideally, a learning method will utilize all three sources. To accomplish this goal, we generalize an existing procedure that uses the latter two. They extend this procedure by re-interpreting it as a Naive Bayes model for document sentiment. Viewed as such, it can also be seen to extract a pair of derived features that are linearly combined to predict sentiment. This perspective allows them to improve upon previous methods, primarily through two strategies: incorporating additional derived features into the model and, where possible, using labelled data to estimate their relative influence. In business settings, there is growing interest in learning product reputations from the Internet. For such problems, it is often difficult or expensive to obtain labeled data. As a result, a change in modelling strategies is needed, towards approaches that require less supervision. In this paper they provide a framework for allowing human-provided information to be combined with unlabeled documents and labeled documents. They have found that this framework enables improvements over existing techniques, both in terms of the speed of model estimation and in classification accuracy. As a result, we believe that this is a promising new approach to problems of practical importance. [22]

Kunpeng Zhang et.al have discussed that large numbers of customers are choosing online shopping because of its convenience, reliability, and cost. As the number of products being sold online increases, it was becoming increasingly difficult for customers to make purchasing decisions based on only pictures and short product descriptions. On the other hand, customer reviews, particularly the text describing the features, comparisons and experiences of using a particular product provide a rich source of information to compare products and make purchasing decisions. Online retailers like Amazon allow customers to add reviews of products they have purchased. These reviews have become a diverse and reliable source to aid other customers. Traditionally, many customers have used expert rankings which rate limited a number of products. Existing automated ranking mechanisms typically rank products based on their overall quality. However, a product usually has multiple product features, each of which plays a different role. Different customers may be interested in different features of a product, and their preferences may vary accordingly. In this paper, they present a feature-based product

ranking technique that mines thousands of customer reviews. They first identify product features within a product category and analyze their frequencies and relative usage. For each feature, we identify subjective and comparative sentences in reviews. They assign sentiment orientations to these sentences. By using the information obtained from customer reviews, they model the relationships among products by constructing a weighted and directed graph. After that they mine this graph to determine the relative quality of products. Experiments on Digital Camera and Television reviews from real-world data on Amazon.com are presented to demonstrate the results of the proposed techniques. Recent trends have indicated that large numbers of customers are switching to online shopping. Online customer reviews are an unbiased indicator of the quality of a product. However, it is difficult for users to read all reviews and perform a fair comparison. We describe Recent trends have indicated that large numbers of customers are switching to online shopping. Online customer reviews are an unbiased indicator of the quality of a product. However, it is difficult for users to read all reviews and perform a fair comparison. They describe a methodology and algorithm to rank products based on their features using customer reviews. First, they manually define a set of product features that are of interest to the customers. They then identify subjective and comparative sentences in reviews using text mining techniques. Using these, they construct a feature-specific product graph that reflects the relative quality of products. By mining this graph using a page-rank like algorithm (pRank), they are able to rank products. They implement ranking methodology on two popular product categories (Digital Camera and Television) using customer reviews from Amazon.com. They believed that our ranking methodology is useful for customers who are interested in specific product features, since it summarizes the opinions and experiences of thousands of customers. [23]

Kushal Dave et.al said that the web contains a wealth of product reviews, but sifting through them is a daunting task. Ideally, an opinion mining tool would process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good). They begin by identifying the unique properties of this problem and develop a method for automatically distinguishing between positive and negative reviews. Their classifier draws on information retrieval techniques for feature extraction and scoring, and the results for various metrics and heuristics vary depending on the testing situation. The methods work best as well as or better than traditional machine learning. When operating on individual sentences collected from web searches, performance was limited due to

noise and ambiguity. But in the context of a complete web based tool and aided by a simple method for grouping sentences into attributes, the results are qualitatively quite useful. [24]

Nikolay Archak et.al has posited that the information embedded in product reviews cannot be captured by a single scalar value. Rather, they argue that product reviews are multifaceted, and hence the textual content of product reviews is an important determinant of consumers' choices, over and above the valence and volume of reviews. To demonstrate this, they use text mining to incorporate review text in a consumer choice model by decomposing textual reviews into segments describing different product features. They estimate model based on a unique data set from Amazon containing sales data and consumer review data for two different groups of products (digital cameras and camcorders) over a 15-month period. They alleviate the problems of data sparsity and of omitted variables by providing two experimental techniques: clustering rare textual opinions based on point wise mutual information and using externally imposed review semantics. Author demonstrates how textual data can be used to learn consumers' relative preferences for different product features and also how text can be used for predictive modeling of future changes in sales. [25]

# CHAPTER 3
# PRESENT WORK

## 3.1 PROBLEM FORMULATION

In our research work we have designed a RFM technique as well as naive bayes classification that compute the purchasing power of customers using sentiment analysis. In our research work we are taking customer's transaction data in addition to sentiment data. This study will help to the business to improve the sales for a business entity. This study will help to maintain the prime customers as well as finding the customers having high purchasing power.

## 3.2 OBJECTIVES OF THE STUDY

Main objectives of this research work are:

1. The research work is predicting customer's purchasing power depending upon various factors like Receny, Frequency, and Monetary and sentiment analysis.
2. Identifying customer behaviour for a particular product.

## 3.3 RESEARCH METHODOLOGY

The proposed system uses the RFM method for the prediction of customer purchasing power. The inputs to the system are customer ID, start date, end date, amount, recurrence, monetary, frequency and output is total purchasing power score. In RFM method, recurrence means how a customer recently visited, frequency means how many times the customer visited and monetary means how much a customer spends the money per deal in RFM method some ranges are set to calculate the total purchasing power of each customers.

RFM functions:

For recency ranges:

Recency ranges set as 0-120 days, 120-240 days, 240-450 days, 450-500days, and more than 500days.

For frequency ranges:

Frequency ranges set as 0-2times, 2-5 times, 5-8 times, 8-10 times, and more than 10 times.

For monetary ranges:

Monetary ranges set as 0-10 dollars, 10-20 dollars, and so on.

In this system 250 total score set as a threshold value. If the total score lies above 250, it means that customer has the higher purchasing power and they are prime customers.

Range: 0-250

| No. | Amount | Classification label |
|---|---|---|
| 1. | 250-500 | High |
| 2. | 0-249 | Low |

**Table1: Classification of Customer Purchasing Power**

The purposed system also used the naïve bayes classification method to calculate the positive and negative reviews of customers.

**Naive Bayes (NB) Method:**

This method is used for both classifications as well as training purposes. This is a probabilistic classifier method based on Bayes' theorem. In this paper, multinomial Naive Bayes classification technique is used. Multinomial model considers word frequency information in reviews for analysis, where a document is considered to be an ordered sequence of words obtained from vocabulary 'V'. Thus, each document d i obtained from multinomial distribution of word is independent of the length of d i. *N* it is the count of occurrence of w t in document d i. The probability of a document belonging to a class can be obtained using the following equation:

$$P(d_i|c_j;0) = P(|d_i|)|d_i|!\prod_{t=1}^{V}\frac{P(w\ t|c_j;0)^{N_{it}})}{N_{it}}$$

Where P (d i | c j; θ) refers to the probability of document‘d’ be- longing to class ‘c’. *P* (| d i |) is the probability of document‘d’ and *P* (w t | c j; θ) is the probability of occurrence of a word ‘w’ in a class ‘c’. After estimating the parameters calculated from training document, classification process is carried out on text document by calculating posterior probability of each class and selecting the highest probable class.

Research methodology is organized into:

Step1: Collect the data for customers as well as reviews of customers.

Step2: After insert the data set into Rstudio, it is pre-processed for better decisions. Pre-processing remove the duplicate entries from data as well as fill the missing value in customer purchasing data. In the review data set it is pre-processed by removing the stop word, numerical data and special characters etc.

Step3: RFM technique is applied to calculate the purchasing power of customers. Vectorization is applied on the pre-processed review data to get the required data.

Step4: In the parallel process the naive bayes classification technique is also applied for classification of reviews.

Step5: Classification gives the result of positive and negative polarity according to the purchasing power of customers.

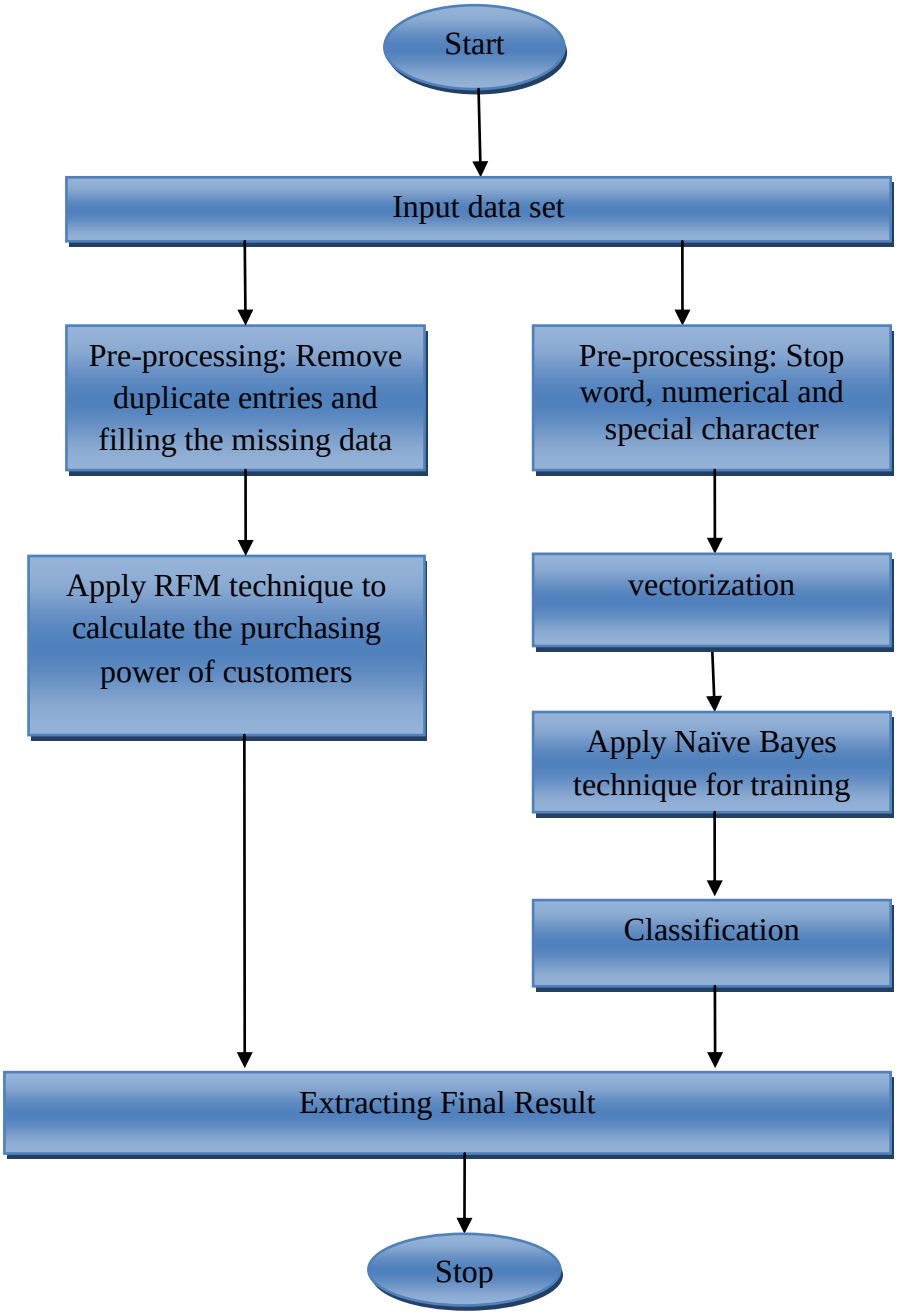


**Figure 6: Research Methodology of Purposed System**

# CHAPTER 4

# SOFTWARE USED

## 4.1 R

R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R is a programming language and software environment for statistical analysis, graphics representation and reporting.

This programming language was named **R**, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka).

**Features of R**

1. R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
2. R has an effective data handling and storage facility.
3. R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
4. R provides a large, coherent and integrated collection of tools for data analysis.
5. R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

**Application of R**

1. Political analysis: it will help to analyse the election estimates by collecting public opinion about the different parties.
2. Data science analysis: it will help to analyse the important information that stored in the dataware house

3. Statistical analysis: it will help to analyse the numerical data such as student record of marks, employee record of salaries etc.

4. Data visualization: it helps to show the data analysis in the different graphs, trees etc.

5. Weather forecasting: it will also help to predict the weather by using the historical data.

6. Credit risk analysis: it will help to predict the fraud customers by doing the analysis of customer's data.

7. Graphical representation: this tool is used to represent the data in pictorial manner.

### 4.2 RStudio

RStudio is an open source integrated development environment (IDE) for the R programming language.

**Features of RStudio**

1. All required things like console, source, plots, workspace, help, history, exists at one place

2. Syntax highlighting editor with code completion.

3. Execute code directly from the source editor.

4. Full support for authoring numeric data and text data.

5. Runs on all major platforms (Windows, Mac, and Linux) and can also be run as a server, enabling multiple users to access the RStudio IDE using a web browser.

### 4.3 Operating Environment

Windows, macOS and Linux can be used as the operating system. For RStudtio system Windows 7 is used.

# CHAPTER 5

# RESULTS AND DISCUSSION

## 5.1 EXPERIMENTAL RESULTS

This chapter presents the work that has been carried out in this thesis. The following results were obtained. The below snapshot shows the result of attributes which required for analysis of customer purchasing power.

| | ID | Date | Amount | Recency | Frequency | Monetary |
|---|---|---|---|---|---|---|
| 4 | 4 | 1997-12-12 | 26.48 | 201 | 4 | 25.125 |
| 6 | 21 | 1997-01-13 | 11.77 | 534 | 2 | 37.555 |
| 7 | 50 | 1997-01-01 | 6.79 | 546 | 1 | 6.790 |
| 8 | 71 | 1997-01-01 | 13.97 | 546 | 1 | 13.970 |
| 9 | 86 | 1997-01-01 | 23.94 | 546 | 1 | 23.940 |
| 25 | 111 | 1998-06-20 | 55.47 | 11 | 16 | 69.190 |

**Figure 7: Data Frame of Customer Purchasing Power**

The below snapshot shows our prime customers who have high purchasing and having total scores lies above 250. There are 17 prime customers who‘s total score lies above 250.

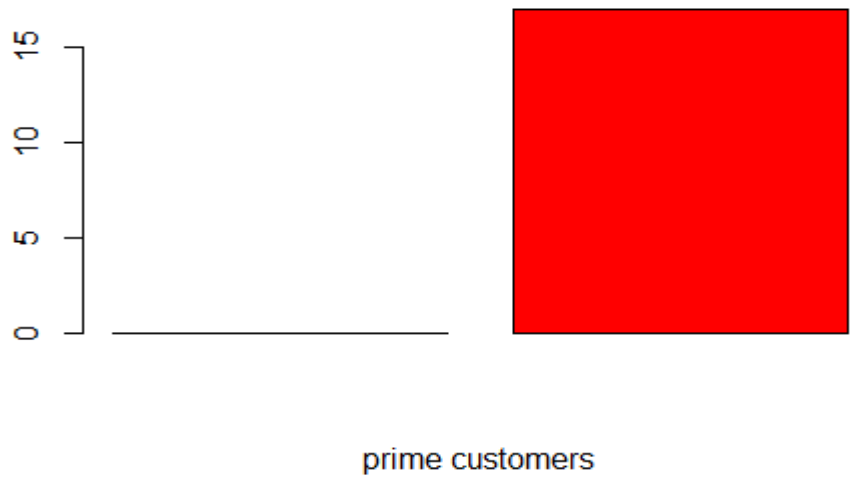


**Figure 8: Prime Customers**

The below snapshot shows total observations of reviews with there positive and negative dimenations.

```
Observations: 2,000
Variables: 2
$ class <chr> "Pos", "Pos", "Pos", "Pos", "Pos", "Pos", "Pos", "Pos", "Pos", "Pos", "Pos", "...
$ text  <chr> " films adapted from comic books have had plenty of success   whether they re ...
```

**Figure 9: Data Frame of Review Data**

The below snapshot shows train data set of reviews.

```
1500 12144
```

**Figure 10: Train Data of Reviews**

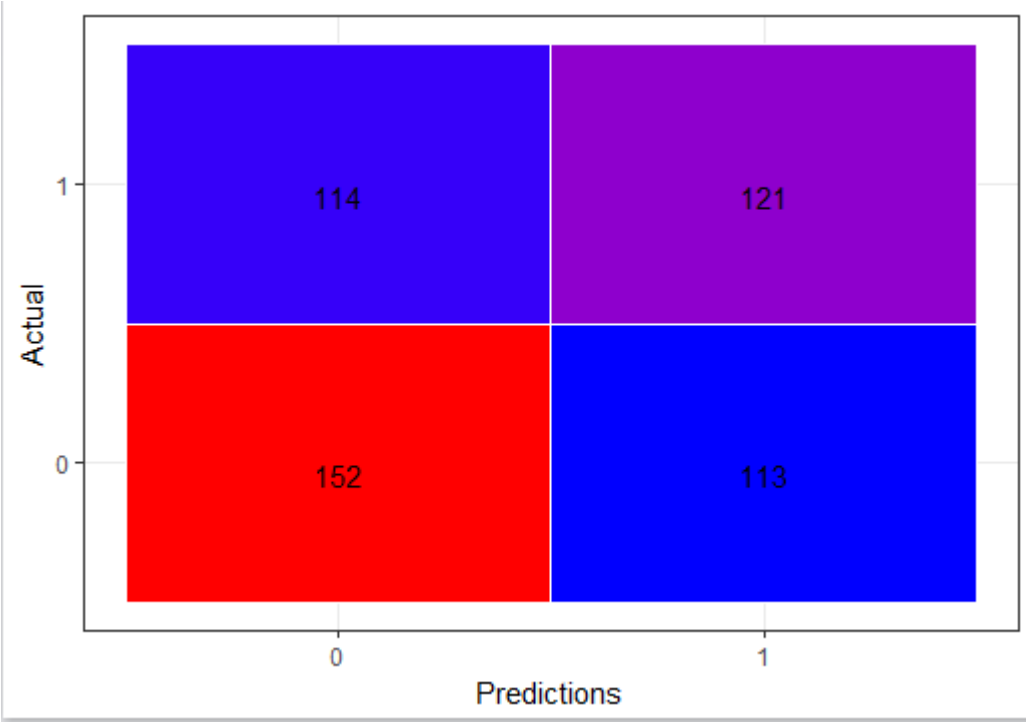


**Figure 11: Prediction of Reviews by Confusion Matrix**

The accuracy of our research work is 94.117. The below graph represents result of accuracy of our research work.

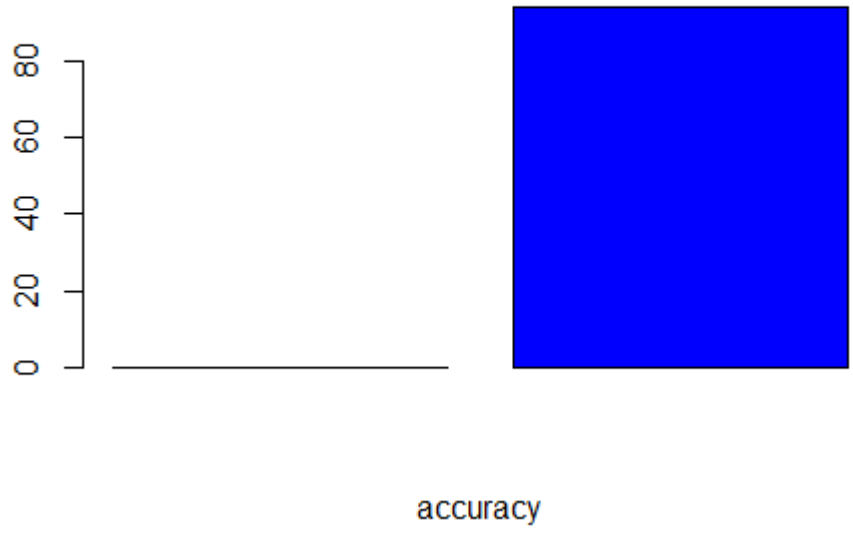


**Figure 12: Accuracy Result**

| **PURPOSED APPROACH** | |
|---|---|
| **1.Target** | To predict the purchasing power of customers |
| **2.Parameters** | Recency, frequency ,monetary and reviews |
| **3.Technique Used** | RFM and addition with naive bayes classification |
| **4.Result** | 94.117 |

**Table 2: Purposed Work Details**

# CHAPTER 6

# CONCLUSION AND FUTURE SCOPE

## 6.1 CONCLUSION

In this research, a system is purposed to predict the purchasing power of the customers by considering different attributes .These factors can affect the purchasing power of the customers. This prediction system tries to estimate the purchasing power of customers by considering a threshold value which denotes the purchasing power of customers. If the customer purchasing power value is above than threshold value, then the customer has higher purchasing power otherwise the customer has lower purchasing power.

Opinion mining is also used to find the customers opinion about various bought products. Sentiment classification is used to classify various customers who have positive and negative opinion about a bought product. Sentiment classification is under taken by gathering the reviews of customers. Reviews are unstructured in nature that cannot provide the important information. Opinion mining is used to extract the important information from reviews that can help the customer management relationship in future.

The purposed system using Naive Bayes and RFM technique for data mining and the accuracy of the system by using these techniques is 94.11.

## 6.2 FUTURE SCOPE

The purposed work is detecting the customer's purchasing power by considering various factors like recency, frequency, monetary etc. These factors are helpful for predicting a customer's purchasing power on monthly basis. This work will be helpful with business perspective to improve one's sale. This system can be used in various e-commerce websites to check their customer's purchasing power and to compare their product reviews with other e-commerce website to increase their sales. We can predict the interest of customer's by using the click stream or using clicking time retain on website by using web mining techniques in future.

# REFERENCES


[1] Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." 15 (2012).

[2] Sugimura, Hiroshi, and Kazunori Matsumoto. "Classification system for time series data based on feature pattern extraction." Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on. IEEE, 2011.

[3] Huang, Xing, and Qing Wu. "Micro-blog commercial word extraction based on improved tf-idf algorithm." TENCON 2013-2013 IEEE Region 10 Conference (31194). IEEE, 2013.

[4] Zhang, Xiuzhen, Lishan Cui, and Yan Wang. "Commtrust: Computing multi-dimensional trust by mining e-commerce feedback comments." IEEE Transactions on Knowledge and Data Engineering 26.7 (2014): 1631-1643.

[5] Nguyen, Thin, et al. "Affective and content analysis of online depression communities." IEEE Transactions on Affective Computing 5.3 (2014): 217-226.

[6] Rosa, Renata L., Demsteneso Z. Rodriguez, and Graça Bressan. "Music recommendation system based on user's sentiments extracted from social networks." IEEE Transactions on Consumer Electronics 61.3 (2015): 359-367.

[7] Li, Yuefeng, et al. "Relevance feature discovery for text mining." IEEE Transactions on Knowledge and Data Engineering 27.6 (2015): 1656-1669.

[8] Gatti, Lorenzo, Marco Guerini, and Marco Turchi. "SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis." IEEE Transactions on Affective Computing 7.4 (2016): 409-421.

[9] Bouazizi, Mondher, and Tomoaki Ohtsuki. "Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter." Communications (ICC), 2016 IEEE International Conference on. IEEE, 2016.

[10] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

[11] Salvetti, Franco, Stephen Lewis, and Christoph Reichenbach. "Automatic opinion polarity classification of movie." Colorado research in linguistics 17.1 (2004): 2.

[12] Beineke, Philip, Trevor Hastie, and Shivakumar Vaithyanathan. "The sentimental factor: Improving review classification via human-provided information." Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2004.

[13] Mullen, Tony, and Nigel Collier. "Sentiment Analysis using Support Vector Machines with Diverse Information Sources." EMNLP. Vol. 4. 2004.

[14] Dave, K. , Lawrence, S. , & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519–528). ACM .

[15] Matsumoto, Shotaro, Hiroya Takamura, and Manabu Okumura. "Sentiment classification using word sub-sequences and dependency sub-trees." Advances in Knowledge Discovery and Data Mining (2005): 21-32.

[16] Zhang, Dongwen, et al. "Chinese comments sentiment classification based on word2vec and SVM perf." Expert Systems with Applications 42.4 (2015): 1857-1863.

[17] Liu, Shuhua Monica, and Jiun-Hung Chen. "A multi-label classification based approach for sentiment classification." Expert Systems with Applications 42.3 (2015): 1083-1093.

[18] Luo, B. , Zeng, J. , & Duan, J. (2016). Emotion space model for classifying opinions in stock message board. Expert Systems with Applications, 44 , 138–146 .

[19] Niu, Teng, et al. "Sentiment analysis on multi-view social data." International Conference on Multimedia Modeling. Springer International Publishing, 2016.

[20] Ganu, Gayatree, Noemie Elhadad, and Amélie Marian. "Beyond the Stars: Improving Rating Predictions using Review Text Content." WebDB. Vol. 9. 2009.

[21] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.

[22] Beineke, Philip, Trevor Hastie, and Shivakumar Vaithyanathan. "The sentimental factor:Improving review classification via human-provided information."Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2004.

[23] Zhang, Kunpeng, Ramanathan Narayanan, and Alok N. Choudhary. "Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking." WOSN 10 (2010): 11-11.

[24] Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." Proceedings of the 12th international conference on World Wide Web. ACM, 2003.

[25] Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis. "Deriving the pricing power of product features by mining consumer reviews." Management Science 57.8 (2011): 1485-1509.

[26] Beineke, Philip, Treor Hastie, and Shivakumar Vaithyanathan. “The sentimental factor: Improving review classification via human-provided information. “Proceedings of the 42nd annual meeting on association for computational linguistics. Assocition for Computational Linguistics, 2004.

[27] Luo, Banghui, Jianping Zeng, and Jiangjiao Duan. "Emotion space model for classifying opinions in stock message board." Expert Systems with Applications 44 (2016): 138-146.

[28] Mouthami, K., K. Nirmala Devi, and V. Murali Bhaskaran. "Sentiment analysis and classification based on textual reviews." Information communication and

embedded systems (ICICES), 2013 international conference on. IEEE, 2013. Computational Linguistics, 2009.

[29] Arora, Shilpa, Mahesh Joshi, and Carolyn P. Rosé. "Identifying types of claims in online customer reviews." Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for

[30] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

**APPENDIX**

**LDA**- Latent Dirichlet Allocation

**TF-IDF**- Term Frequency- Inverse Document Frequency

**SVM**- Support Vector Machine

**RFM-** Recency Frequency Monetary