***Performing sentiment analysis on online customer reviews using classification algorithms to enhance the business opportunities***

# Abstract

It is a known fact that e-commerce industry is now interested in ‘opinion mining’, to classify the reviews for their products and services (Santhosh Kumar K L, 2016). If a person is purchasing a product from an online shop, then, he or she will often comment about the product online. These kind of reviews can be isolated as positive, negative or neutral in order to create ‘opinion’ by the industry people. This activity is referred to as opinion mining. It is about analyzing the emotions of the customers on the purchased product with the help of user reviews. Since, classification of these reviews manually is a very tedious process; the automated function will help in mitigating the issue. This project aims to classify the reviews of amazon customer on electronic devices using algorithms such as logistic regression, naïve Bayes and support vector machine. The Weka tool can also be used for classifying the reviews by applying the above-mentioned algorithms. According to the performance metrics of these algorithms, performance comparison for these algorithms can be made so that effective and efficient algorithm can be selected for classifying the reviews to get more accurate results (Hui Song, 2013).

***Keywords: Opinion mining, logistic regression, naïve Bayes, support vector machine***

**Table of Contents**

[1. Abstract 2](#_Toc7122181)

[2. Chapter A 5](#_Toc7122182)

[2.1 Introduction 5](#_Toc7122183)

[2.2 Hypothesis 5](#_Toc7122184)

[2.3 The problem 5](#_Toc7122185)

[2.4 The project aim and objective 6](#_Toc7122186)

[2.5 Mind map 6](#_Toc7122187)

[2.6 Project research schedule 7](#_Toc7122188)

[2.7 Merits of project 7](#_Toc7122189)

[3. Chapter B 7](#_Toc7122190)

[3.1 Background information 7](#_Toc7122191)

[3.2 Literature review 9](#_Toc7122192)

[3.3 Collection of Dataset 14](#_Toc7122193)

[3.4 Ethical issues 15](#_Toc7122194)

[3.5 Dataset description 15](#_Toc7122195)

[4. Chapter C 18](#_Toc7122196)

[4.1 Material and methodology 18](#_Toc7122197)

[*Data classify* 19](#_Toc7122198)

[*Classifier models* 19](#_Toc7122199)

[4.2 Research Methodologies 20](#_Toc7122200)

[4.2.1 SVM – Support vector machine 20](#_Toc7122201)

[4.2.2 Naïve Bayes algorithm 20](#_Toc7122202)

[4.2.3 Logistic algorithm 21](#_Toc7122203)

[5. Chapter D 23](#_Toc7122204)

[5.1 Evaluation 23](#_Toc7122205)

[5.1.1 Implementing sentimental analysis 23](#_Toc7122206)

[5.1.2 Working of Python in Weka 23](#_Toc7122207)

[5.1.3 Working concepts of python code in opinion mining 23](#_Toc7122208)

[5.1.4 Working Flow of python in opinion mining 24](#_Toc7122209)

[5.2 Source Code work-flow explanation 25](#_Toc7122210)

[5.3 Practical work- Evaluation 29](#_Toc7122211)

[6. Chapter E 34](#_Toc7122212)

[6.1 Final results 34](#_Toc7122213)

[6.2 Comparison of classification algorithms-Results 34](#_Toc7122214)

[6.3 Graphical results 35](#_Toc7122215)

[7. Chapter E 39](#_Toc7122216)

[7.1 Conclusion 39](#_Toc7122217)

[7.2 Recommendation 40](#_Toc7122218)

[References 41](#_Toc7122219)

# Chapter A

## 2.1 Introduction

E-commerce technology plays an important role in everybody’s day to day life. E-commerce websites allow their users to convey what they feel about the particular product that they have purchased from the website (Raheesa Safrin, 2017). This is called a review or an opinion. It clearly depicts the attitude or mood of the reviewer and it may end as a negative or positive review towards the product. The emotions which are expressed in the e-commerce websites by the users are known as sentiments. Sentiment analysis is the process of text analysis to detect and extract subjective data in the source material.

Opinion mining is essential for e-commerce websites and it provides maximum benefits to the individuals (Alexandra Cernian, 2015). Every day people post their experience in e-commerce website as their opinion. Mining the opinion of the users regarding the product will help the organization as well as individuals to make a decision on buying the product. Different e-commerce service providers follow different review structures to allow users to express their opinion regarding the product. However, users need to read the opinion of others in e-commerce from the start to the end to make a decision on purchasing the product. This issue will be addressed by applying sentiment analysis to the dataset. The analysis will cluster the dataset and group data as positive and negative through which the people can easily make a decision on purchasing products from e-commerce sites.

## 2.2 Hypothesis

There is a significant relationship between the online customer's review and consumer’s decision making in purchasing products. Technology advancement obtains a huge volume of information by monitoring the existing e-commerce user’s decision in purchasing products. Conducting sentiment analysis on the customer review dataset will help to identify the relationship between the review and decision making in an online product purchase. There are the number of classification algorithms available to process the dataset to perform the sentiment analysis. This research work completely focused on testing the hypothesis by processing the dataset.

## 2.3 The problem

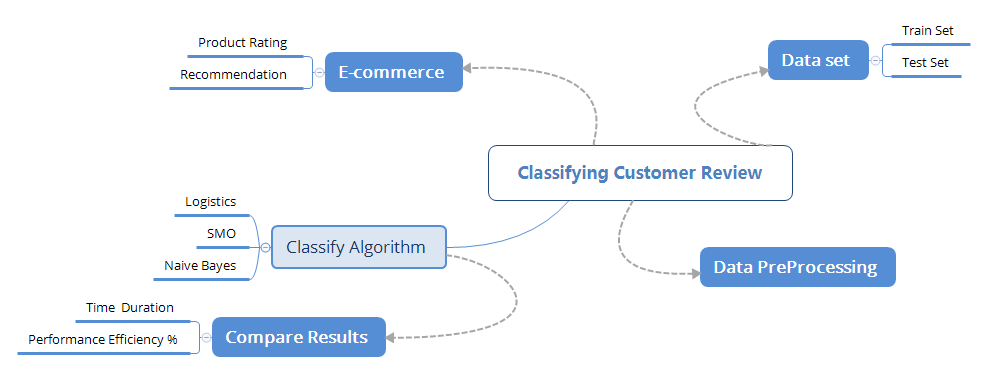
E-commerce is one of the emerging technologies which plays an important role in promoting internet related business strategies. Analysing the e-commerce customers review dataset will help to understand the pros and cons of the product. As well as an organization can easily encounter the problems that are associated with the product to meet its customer expectation (Kanellopoulos, 2006). At the same time, individuals can easily make a decision on purchasing the products from the internet by viewing the others review. However, individuals or organization need to read all the reviews or opinion of the customers from the start to end which consumes a lot of time and creates issues in understanding their opinion. So accessing the huge volume of data related to the e-commerce products will help to obtain the most useful information. Moreover, people discuss both the positive and negative of the products so grouping the opinion will help to know that the product meets their expectation or not. Applying classification algorithms will help to perform the sentiment analysis on the collected dataset. At the same time, it helps to group the positive and negative opinion of the products to help the organization or individuals to make the decision quickly (IraklisVarlamis, 2017).

## 2.4 The project aim and objective

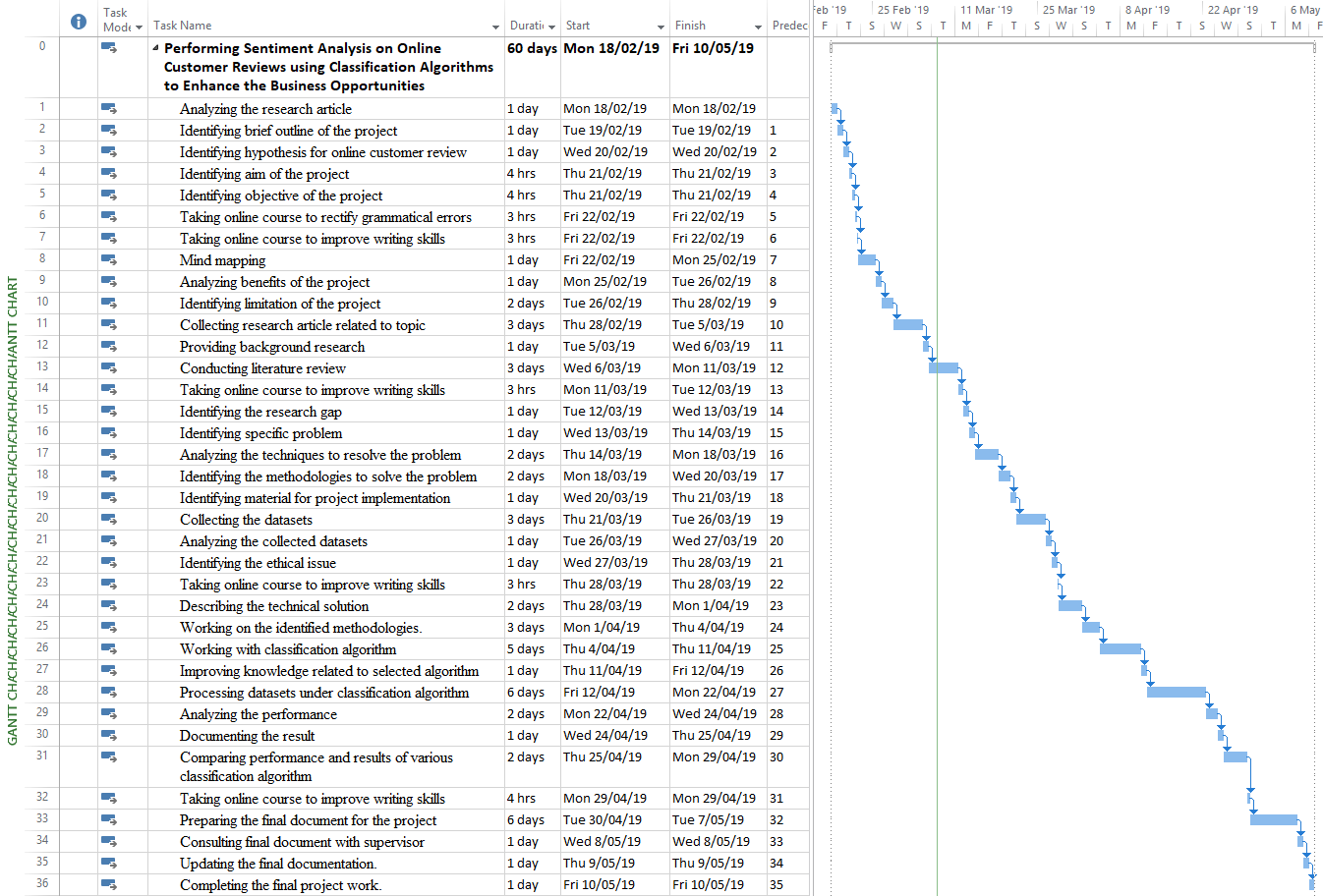
To classify the positive and negative reviews of the clients from their comments posted online with the help of classified algorithms such as logistic regression, naïve Bayes and Support vector machine. The performances of these algorithms will be compared in terms of accuracy with one another. The main objective of this algorithm is to classify the e-commerce customer data to predict the customer’s behavior in purchasing products and to help the individuals to obtain more knowledge about the product quality.

## 2.5 Mind map

The mind map section explicates the workflow of the project concept i.e. analysing the customer review about their purchased product. One of the essential processes to obtain best and accurate results is uploading the valid datasets into the tool.

***Figure 1: Mind Map for classifying the customer review***

## 2.6 Project research schedule

  
***Figure 2: Project schedule***

## 2.7 Merits of project

This project helps to showcase the recommendation and review rating of the product made by the customer on their purchased product. This will highly helpful for another customer to purchase a particular product by knowing the rating and recommendation of a product made by the previous customer. Thus, purchasing the damaged product over the online shopping can be avoided to a large extent.

# Chapter B

## 3.1 Background information

Sentiment analysis is an important aspect in opinion mining research (Hani Nurrahmi, 2016). A number of research works are conducted to automate the sentiment analysis from text data sources. It will be helpful for the organization to make a prediction on their product or service to implement effective future strategies. Sentiment analysis is gaining more power in providing valuable data in various (M Hu, 2006). It also has the ability to predict the user’s behaviour pattern and preference to recommend a particular product to the user in a particular time interval. A number of researches are conducted to obtain useful content from the e-commerce site and to categorize them into positive and negative.

Analysing customer review plays a vital role in enhancing the quality of the product to meet the expectation of customers. However, it helps the organization to increase sales and also helps the customer to make a decision in purchasing the product. The number of researches has been done on sentiment analysis to enhance the performance and accuracy of the opinion mining. It is essential to classify the customers in the e-commerce industry according to their behaviours in buying products. Support Vector Machine learning method is used for predicting the future behaviours of the customers. It is possible to predict the time difference in purchasing the product and other distinct factors among every individual client. While comparing with other models such as linear regression and another existing method, SVM performs well in classifying the data.

Sentimental analysis is used for classifying the reviews of customers in e-commerce and this is the most widely used analysis in this current domain. Comments posted by the consumer of the e-commerce will be used for classifying the opinions. These kinds of activities help in understanding the emotions of the customers. They have mentioned their view in terms of text, video or any other media format and these reviews are collected as an input for classification and phrase identification algorithm has been used for categorizing the data (Raheesa Safrin, 2017). This article helps in determining the opinions of the customer and the comments are differentiated as positive and negative groups. This type of analysis will help in enhancing the nature of products and the betterment of sales. The current system for manufacturing, marketing and delivering terms of the company can be modified to attract the customers. Classification of reviews will act as feedback for the product manufacturers and another line of suppliers.

Website contents are used for performing opinion mining and people are ready to express their views on a daily basis in e-commerce to help others. Obtaining the opinion from the e-commerce user will provide many benefits to the business. The current market trend will be observed from the opinions. Machine learning algorithms such as support vector machines, logistic and naïve Bayesian algorithms are used to classify the collected data.

## 3.2 Literature review

Tracking and tracing customer behaviour and their choice of purchasing the product are monitored by introducing the RFID (Radio-frequency identification) technology. This technology helps in improving the collection of moving data gathered from various individual review around various resources in an efficient way (Kwon Y S, 2011). The vast amount of customer behaviour is collected and gathered effectively with the help of RFID tag. Using the RFID tag technology, customer behaviour characteristics and shopping path of the products in the market are analysed and determined. Over this methodology of RFID, the time consumption and data distorting error are lessened at large extent. RFID technology offers an effective communication system which is highly integrated with each category in the store or market. The system works by separating and clustering paths and fields of the product and customer behaviour. RFID was used in the retails store to predict the need for a particular product and to maintain the stock level but in the e-commerce it’s possible. An organization needs a new way to find the most liked product by the customers so they have added an additional feature to obtain review from the customers about their product.

Customer behaviours while shopping is collected and recorded by applying the spatial pattern clustering over the LCS (Longest common subsequence) technique where the choice of customer opinion is addressed in an effective manner. The operation happened with the technique is been capable of identifying the hotspots where most of the customer visit has happened. The statistical analyses are made to establish the shopping path of the customer. The effectiveness of RFID technology is been carried out by the author for establishing the reviews of the purchased product. Three different possible moves made over the RFID based recording of wayfinding behaviour are the movement of the shopping cart, basket, and shopper. The communications are made with the help of transponder and server. The antenna receives response from various signals.

But, the system is not effective in the privacy and security aspects of customer credential details (Kwon Y S, 2011). In the previous system, retail business providers obtained data by using the RFID technology and analysed the data to enhance their service to meet the customer’s expectation. The advanced of technology has led the retailer to migrate to the online business and to enhance their business as well as to meet the customer expectation they introduced product review on the e-commerce service. The data which obtained using the product review functionality will help in conducting the sentiment analysis.

The author (Liu T, 2018) uses the smart tag technology of RFID for deep analysis of data acquisition and shopping system. Over the research work, the deep shopping data and information are getting recorded to find out how the goods are getting received and the shoppers intention in purchasing a certain product. As the project works effectively by detecting the motion of the product thereby it is been named as tag booth which offers the best results with high accuracy. The intended goal of this research work is to collect fine-grained deep shopping data in an automatic manner. With the features of RF signal purchased goods and signals are carried out in certain order to recognize the individual purchased product details in the specified retail stores. These effective working models are carrying out with the help of COTS with RFID devices. The location of the purchased product or object and customer movement is tracked with deep shopping system as high accuracy. The accuracy of the results is monitored with two important metrics of tag booth such as TPR (true positive rate) and FPR (false positive rate). The errors in the system are also controlled by using the time error indicator.

With the system effective performance and accuracy, it is been found that the operating system is considered to face certain limitation in processing all the collected information. The technology used within the business faces limitation to identify the exact relationship processed among the customer and the business staff. The technology with tag booth system is not highly recommended to identify the physical movement of the person in the shopping market or store.

The author (Elwalda, 2014) addresses influences of online customer review and their purchasing intention which provides greater problem over the complex nature. The intended goal of the research work is to relate two sources together by minimizing the complex problem and analyse consumer expectation while shopping. The sales and services of the OCR (online customer reviews) over the retail management system and shopping processes are determined in a well-versed manner. With the help of integrated work of OCR framework, non-numerical attributes and purchase intention level are traced out to better manner. Over this work, business commodities and development are gained as per consumer feedback. The marketing strategies and perceived risk of the products are managed.

As the author (Shafi S I, 2014) considered to analyse the behaviour of each customer and their behaviour in purchasing the product. With this operation, the impact and primary goal of the research work are attained in a well-versed manner. The demographic descriptive is processed by the author by undergoing various methods such as questionnaires, data analysis and interpretation, and data samples. The simple methodology for processing customer behaviour has some of the common limitations to establish effectiveness to the next level (Shafi S I, 2014). The data sampling and acquisition processes have been failed to explore and been carried out with different peculiarities.

The author (A.Halawaa, 2018) discusses the classification accuracy level for Arabic tweets. By applying the ensemble methods include bagging boosting and SMO on the same dataset and verify the classification results in order to identify the effectiveness of the accuracy. The intended goal of this research work by the author is to enhance the strong learning algorithms functions using ensemble approaches. The success of approaches relies on the misclassified instance in the dataset. The effectiveness of the approach is to reduce the false positive levels in order to minimize the error rate. The accuracy of each algorithm is measured with the help of accuracy, precision, and recall /f measure value. The text classifier’s performance accuracy for each algorithm is measured with the help of F-measure. After successful completion of the evaluation and test process, the author recommended the SMO classifier model as the best one in order to attain the best performance accuracy in this research work.

The author (Shafi S I, 2014) discusses customer intention towards the apparel products in the supermarket or store. The research work is made by collecting a variety of details about the purchased product in the retail stores such as consumer characteristics, store attributes, reference group, promotion, and product attributes. This effective information is linked with the purchase intention to gather the consumer buying attributes a purchasing behaviour. The intended goal of this research work is to analyse and explore the relationship between the variables and behaviour level of apparel products. The author uses one of the best and effective methodologies for his research work as examine and questionnaires.

The author (Andersen P, 2017) examined the negative reviews over the purchased product by the customer. The intended goal of the research is to establish and examine the reviews of the purchased product by analysing the effectiveness of price satisfaction, purchase intention, purchase goal and received values. The research model proposed by the author to enhance the customer intention over the purchased product is linking the shopping orientation, intention shop using the internet, product types and demographics together. Over this model flow, the relationship between the shopper intension with product shopping is made and traced out easily and effectively. With high internet connectivity, information is transformed from one place to another.

The system has limitation by focusing the two different types such as a low proportion of the negative impacts or reviews with a high proportion of positive reviews commands and vice versa. Thereby, neutral and mixed reviews are no easily gathered and collected by the shoppers in order to explore customer intention over the purchased products and goods. By sharing the attractive and retaining customer details over the internet connectivity, the misuse level of credential data is increased to a large extent. Thus the security and privacy level of data are leaking day by day thereby online shopper intension towards smart shopping are also decreasing periodically. As the limitation faced by the author over this research work is poor sampling covering area and self-selection of the products for purchasing which fails to measure the constraints.

Nowadays online shopping intentions are becoming increases in order to collect the trendy and variety of products from anywhere and at any time with comfortably. The reason behind the dramatic growth of online shopping is a customer can purchase their products over the internet connectivity, the online shopping business contains all kinds of product to sell. Over online shopping, customer can post their comments about their purchased product. By reviewing the comments of each product, another customer will purchase the product. Here, the author worked with online shopping website where the purchased customer only can post the reviews about the product whereas the other customer is not allowed to post the comments over the product. Hence, fake reviews are totally avoided. These effective processes are made with the help of opinion mining methodology.

The sentimental analysis processes are carried out with simple plain text or language processing. The intended function of sentimental analysis is to identify and extract subjective information in all sources. Most of the sentimental analysis processes are achieved over review about their product, customer product suggestion, and expressing the review over the social media. With the help of SVM and opinion mining, customer reviews can separate or categories into positive, negative and neutral review with high accuracy and efficiency (Asghar, 2018).

Shopping online through e-commerce using different application has turned to be important for most of the customer which is considered to save a lot of time. Customers review on e-commerce products creates a strong impact on the purchasing decisions of others. The research of the author (Ruba Obiedat, 2013) is completely focused on finding e-commerce consumer's review impact on buying intention of the others in the context of the UK. Quantitative data is collected in this work from 120 consumers in the UK to understand various customer decision in purchasing products. Finally, the author has concluded that buying intention of the online customer is directly affected by existing e-commerce customer review. Most of the e-commerce customers spend a lot of time reading online product review prior to purchasing decision.

The author (Dhokrat, 2015) discusses the efficiency of two different algorithms such as J48 and SMO on different datasets. Here, J48 uses the C4.5 version decision algorithm and SMO uses the SVM classification algorithm; by comparing two algorithms the SMO delivers the best performance. Both algorithms are applied over all dataset in order to confirm the outcomes of each algorithm. From the outcome, it is clearly noted that SMO provides classification accuracy. Here the author undergoes his research work on iris and Yuta selection, although the iris and Yuta offer equal accuracy, Sequential minimal optimization offers quite higher than that along with clear and better dataset accuracy. In this research work, the author suggested the Sequential minimal optimization algorithm for analysing the classification algorithm in order to gain a better outcome compared with all other classification algorithms.

Product review described by every buyer for the products will help another customer to find whether the product they choose is good or bad and also will help the seller to identify the feedback based on their sold product. As the number of product increases day by day every e-commerce customer is allowed to write the desired review which creates a positive and negative impact based on their experience they gain in using the product. The performed operation is defined as a free format. The author (Hani Nurrahmi, 2016) proposed concept in this paper is based on developing a system which would extract the product review specified by the different customer and classify their opinion automatically. Proposed concept by the author uses the class sequential method to extract a feature of every product and the other concept based on using Opinion Lexicon is been described for feature opinion classification.

Product review processed and analysed for every product by online e-commerce customer is been continuously increasing. The author (Li Lin, 2016) proposed a new concept in this paper is to filter unwanted sentence, noise, and words that are considered to be meaningless in the provided review. For this case, a new informative review identification method is proposed along with the usage of parsing and sentiment analysis. To provide the specified operation two linguistic concepts such as feature extraction and classifier learning are being got used which will help in effective extraction of a certain feature in the provided review. The author experimented his result in Amazon product review which is considered to define the best performance in identifying the right customer review. With the proposed concept there arise certain factor for the concept to be getting used for pre-processing in the review analysis and extraction of information more effectively.

Online shopping processed in these days is considered to be getting increased which are more common in buying various goods from different resources. In this case, each product that is being got purchased online is been provided with a different review from a different customer to increase the sale of specified product and to introduce quality in the refined product. With the great volume of review processed from multiple customer experience, it becomes a certain difficulty to analyse all the review and to choose the best product. As a solution to this factor, the author (Neena Devasia, 2016) in this research work has proposed the concept of semantic-based approach which will help in extracting the best product features and will help the product to be get analysed at feature level from the customer review. Along with this approach, using recursive deep model and algorithm will help in providing sentiment analysis for the product review sentences. The experiment conducted by the author is considered to be effective which obtain the best result in analysing all the product review.

Various e-commerce application like Amazon and Flip kart plays a major role in selling the various product to the customer through online. This increases the decision-making operation for every manufacturer and the customer to identify the best product and to improve the quality of the product. Most of the purchase happened over this website is by gathering the positive and negative review which are being got provided by the customer based on their gained experience. This makes the customer find out the best product and the product manufacturer to deliver the best product with the improvement in quality, performance, and recommendation. The author (Zeenia Singh, 2017) proposed concept of performing sentiment analysis help in extracting the subjective information from the text of online reviews processed over mobile phones and helps in identifying all the positive and negative review based on the product along with the identification anger, anticipation, disgust, fear, joy, sadness and trust processed over the purchased product.

## 3.3 Collection of Dataset

E-commerce offers several datasets to the data scientists to perform more research work on it, and helps to develop and promote the business to peak. For this project work, “customer’s opinion about their online purchased product” related dataset is collected from the open source dataset which is available on website “data.world”.

## 3.4 Ethical issues

The customer opinion and reviews about their online purchased products are evaluated easily and effectively by collecting the customer’s shopping behaviour record. For business development, e-commerce offers online shopping customer intention /opinion details to the data scientists. In simple words, for doing more research work over an online datasets, e-commerce offers accurate and valid results with free of cost. The open datasets are highly helpful to predict an accurate result. Moreover, datasets are obtained from the e-commerce for doing the research work by data scientists thereby the privacy and security level of customer details won’t get affected. Thus, ethical issues are avoided to a larger extent.

The datasets are obtained from the data.world website and the website contains plenty of datasets, among that “customer opinion over online product” datasets is chosen. As customer behaviour datasets are offered and revealed by the company, therefore no unethical activities involved in collecting and analysing customer opinion about their online shopping. Moreover, the collected dataset has ratings and recommendation details which are made by the customer about their purchased product, thereby the customer personal and credential details are not included in it. Therefore, the privacy and security level of personal details are not affected by processing this online dataset.

## 3.5 Dataset description

From e-commerce website, the online dataset is download to analyse the customer opinion about their purchased product. The dataset contains nine different attributes with 7299 customer’s reviews about their product. The 9 attributes are,

* brand
* colors
* name
* primaryCategories
* reviews.doRecommend
* reviews.numHelpful
* reviews.rating
* reviews.title
* reviews.username

***Brand***

As the dataset is related to the electronics product, this attribute shows the brand details of each purchased product by the customer. Some of the brands listed in the dataset are, Microsoft, boytone, sanus, ultimate Ears, lowepro, corsair, sdi technologies, inc., Verizon wireless, JVC, JBL, Lenovo, siriusxm, Pny, sling media, Sony, midland, Toshiba, Yamaha, DreamWave, glengery, dell, MEE audio, Samsung, bose, Logitech, Motorola, definitive technology, alpine, Belkin, bowers & wilkins, JBL, CLARITY-TELECOM, house of marley, Toshiba, kicker, SVS, WD, Netgear, and peerless-AV.

***Colours***

This attribute shows the colour variation of each purchased product by the customer. From the dataset, it is clearly noted that most of the customer purchased their product in black.

***Name***

This attribute shows the name of each product purchased by the customer.

***Primary categories***

This attribute shows the category details (electronics) for each purchased product.

***Reviews.dorecommend***

This attribute collects the reviews of the customer about their purchased product in terms of true and false.

***Reviews.numhelpful***

This attribute shows the reviews details which are highly helpful by other customer to purchase the product.

***Review.rating***

This attribute shows the review rates of product purchased by the customer. The review ratings are made out of 5 for each product.

***Review.title***

This attribute shows review statement and comments made for each product purchased by the customer.

***Reviews.username***

This attribute shows the registered name of each customer.

***Tools for data mining***

The various data mining tools are available in open online sources to perform and analyze the activity in a clear manner. Some of the open source tools available in the online market are Weka tool, Tanagra, mat lab, orange, .net framework, and rapid miner. These tools help to classifies and categories the attributes in the desired format to meet accurate results.

***Selected data mining tool***

From these open source data mining tools, I have selected the Weka tool as best to compare and processes the results. As the Weka tool works with any data format, the classification of attributes become simple and easy. It is highly helpful to address how a customer feels about the brand, product, and services. This also helps to identify the spikes in sentiment by using e-commerce influences. The following classification algorithms are used to classify the customer review through SVM (support vector machine), naïve Bayes and logistic algorithms. The intent reason for choosing the classification algorithms for the customer opinion analysis is to obtain an accurate result.

***Reason for selected tool***

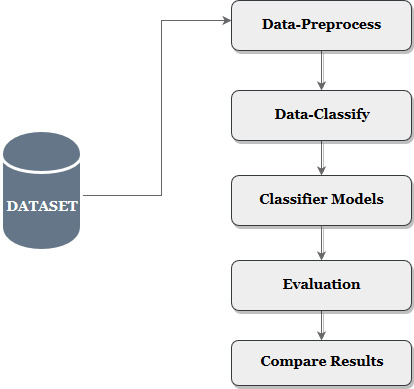
***Table 1: Data mining tools***

|  |  |  |
| --- | --- | --- |
| *Rapid miner* | *Orange* | *Weka* |
| It does not supports java and python interface modules. | It supports visual programming such as python and machine learning | Works effectively on python and java modules |
| Takes more memory capacity thereby the system performance are minimize | Takes long time to build the dataset and works effectively for smaller dataset | No limitations on size of the dataset |
| Commercial- expensive licenses | Freely available | Freely available under General Public Licenses |

# Chapter C

## 4.1 Material and methodology

Nowadays, the social application and microblogging website accessibility levels are increases to express their feeling or intention about the product. As the online marketing /stores work with high flexibility, whereby people can purchase their product over the online stores from anywhere and at any time. Thus, online shopping customer levels are increases day by day in the modern technology. Before purchasing the product from the online store, the review and rating of the product which are made by the previous customer will helps the new customer to purchase that particular product. To analyse the customer opinion about their purchased product, the following machine learning algorithms are highly helpful (Dhokrat, 2015).

******

***Figure 3: Flow chart of data processing***

***Data pre-process***

The intended function of the data preprocesses stages in the SVM is transforming the raw data format into understandable functions. One of the data mining techniques i.e. support vector machine helps to identify the lacking or missing of attributes in the dataset. Hereby, the discrepancies of data processing are controlled to a large extent without any errors. The dataset is collected from the open source which is highly available for the data scientists in order to carry out their research work.

Uploading the collected dataset into the data mining tool (Weka), under the stage of data preprocessing. Here, customer review details about their purchased products are uploaded into the data mining tool called Weka. The dataset contains the details about the customer’s purchased product review. There are nearly seven thousand ninety-nine customer reviews are available with eight different attributes. To process these review details within the short span of time, the multithreading techniques are widely used. Thereby, the processing time of the customer review details is reduced. Thus the pre-processes stage is also called as the filtering stage and finally the resultant output is made as normalization function (Asghar, 2018).

## *Data classify*

Under the classifier models, different classifier functions are applied over the collected dataset by choosing the effective classifier model under functions and Bayes. The train set and test set functionalities are used for the uploaded dataset. Further, efficiency and performance levels obtained over the dataset are addressed in well versed manner. The intent difference of train set and test set is functions. The train set is used to train the dataset according to the classifier model by taking the small amount of dataset whereas test dataset is used to test the corresponding selected train dataset.

## *Classifier models*

The support vector machine is one of the classifier models which is used to classify the hyperplane by performing the train set for the selected dataset. The SVM works with two different purposes i.e. classifies the dataset and regressing the dataset. In the classification process, it divides the hyperplane into two different vectors. For the smaller redefined datasets, it offers the high accuracy and efficacy by taking the short span of time to train the dataset.

The SVM classifier is readily working to analysis the customer review about their purchased product. An attribute in the given dataset is classified accordingly and analyzed the results “as a recommended product for other customers to purchase”. Here, the dataset contains different attributes such as brand, color, categories, review title, rating, helpful, name, and comment. By working with these different attributes the corrected and accurate output will be obtained based on the customer review about their purchased product.

## 4.2 Research Methodologies

The section covers various methodologies to analyse the customer opinion about their purchased product. The reviews, comments, and rating of purchased product of one customer will help other customers to purchase that particular product. To perform this operation, data mining and classifier models are works in a parallel manner. Some of the methodologies used to predict an accurate result are, SVM, Naive Bayes and logistics algorithms. By analysing the performance levels of these classification algorithms, the best algorithm can predict with high accuracy (Jebaseeli, 2016).

### 4.2.1 SVM – Support vector machine

As SVM works with the linear way, thereby it replaces all missing value and transforms the data of each attribute into binary format. The memory type required to processes or handle a large number of data, to classify and to test is linear type. Here, the performance accuracy of the SVM classifier algorithm depends on the correctly classified (A.Younessb, 2018).

By uploading customer review datasets into Weka tool and applying the classifying algorithms SMO, the final results obtained will be displayed on the screen. This technology is in high demand and the quality of service that is provided by the E-commerce websites are highly commendable. Support Vector Machine (SVM) is considered as one of the fastest and better algorithms that can classify even big data and produce more accuracy on results. This method is also considered to be one of the powerful prediction models and it works on complex pattern analysis effectively in the form of high dimensional data sets (D. Ben Ayed Mezghani, 2010).

Support vector machine works with a supervised learning algorithm which has been mostly used for solving problem-related classification and also used for the challenges related to regression. The main challenge faced over the SVM is constructing the classifier model. The concept of SVM mainly depends on the decision plane which helps to defining the decision boundaries (Belbag, 2016).

### 4.2.2 Naïve Bayes algorithm

Naive Bayes algorithm is deliberated as one of the best machine learning algorithms which is mainly used for text classification problem in high dimensional training datasets. It allows building the model in a faster way and also permits for quicker prediction. It also performs the calculation to predict the conditional probability, where the probability mentioned here is been based on the previous knowledge available on the events. Naïve Bayes algorithm resembles as (Zhang, 2011).

Based on the formula the components used here are described as,

**P(A|B)** – Conditional probability of the event A with the given event B is true

**P(A) and P(B)** – P(A) and P(B) are considered to be the probability of occurrence of event A and event B.

**P(B|A) –** Probability of the event B with the given event A is true.

In the given equation,

A - Proposition

B - Evidence.

P(A) – Prior probability of proposition

P(B) – Prior probability of Evidence

P(A|B) – Posterior

P(B|A) – Likelihood

With the defined information the given equation will resemble as,

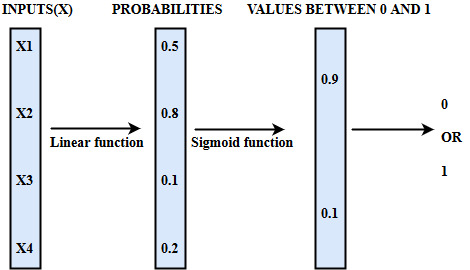
Some of the uses that are getting defined with the usage of the Naïve Bayes algorithm are,

* The algorithm is defined as one of the most useful algorithms to classify the text document.
* It helps in filtering and eliminating spams from the user system.

### 4.2.3 Logistic algorithm

The main functionality of logistic algorithm is very simple and it can be used as a performance baseline for tasks. The technique used to define the algorithm is been borrowed from statistics which helps in predicting the continuous outcome. This method is mainly used for binary classification and its outcome will be defined between 0 and 1.

Working operation with the logistic regression algorithm is used to measure the relationship between the dependent variable and the independent variable by estimating the probability of the logistic function. In Dependent variable, the prediction happens from the e-commerce and in Independent variable; the features are based on the operation. The prediction of any operation needs to be transformed to binary value 0 and 1. The operation performed with logical function is otherwise defined as a sigmoid function which is defined in the S-Shaped curve and it represents any of the real-valued numbers. The numbers can be mapped between the values 0 and 1. The value represented between the value 0 and 1 can be transformed using threshold classifier. The working operation of the logistic regression has been described in the below image to derive the expected output.

******

***Figure 4: Working model of Logistic regression***

***Summary***

***Table 2: Comparison of SVM vs. logistics vs. naïve Bayes***

|  |  |  |
| --- | --- | --- |
| *Support vector machine* | *Logistics regression* | *Naïve Bayes* |
| Effective in high dimension spaces (Hani Nurrahmi, 2016). | Used to determine the output when two or more than one independent variable. | Categorical variable has the sub-category in the train data set then model will considered as 0. |
| Works effectively where, number of dimension > number of samples | Predicts only continuous numeric variables | To solve the zero frequency, the Laplace estimation are highly required. |
| Memory efficient | Poor performance when there are non-linear relationship | Bad estimator which fails to offer the probability of good outcomes. |
| Versatile with different kernel functions which helpful to take best decision function | Poor flexibility to work on more complex dataset | It assumes the independent predictors which lead to fails of accurate outcome. |

# Chapter D

## 5.1 Evaluation

### 5.1.1 Implementing sentimental analysis

Sentimental analysis is the process of identifying and extracting the subjective information such as brand, product name, colour and review about the product. This process is termed as contextual mining of text. This information are obtained by monitoring the online conversation of customer review about their purchased product (Himani Bhavsar, 2012).

According to the project, it is foremost important to analysis the key aspects of review and title made by the customer about their purchased product. Here, the sentiment analysis is nothing but text classifier; which classifies the incoming message or comment of the customer about their purchased products in terms of positive/negative /neutral review.

### 5.1.2 Working of Python in Weka

The general working process of python in weka tool is addressed in the following steps which will help to analyse the opinion mining (sentimental analysis) of customer.

* First, start JVM from python.
* Load the data in the appropriate format.
* Build the decision tree by choosing any of the classifier models.
* Evaluate the classifier using cross-validation.
* Visualize the classifier, based on filtered data.
* Visualize the class probabilities.
* Plot the grids for more accuracy in predictions.
* Cluster the data and visualize the cluster.
* Select the Attribute.
* Build a classifier with attribute selection.
* Manage and obtain the results.

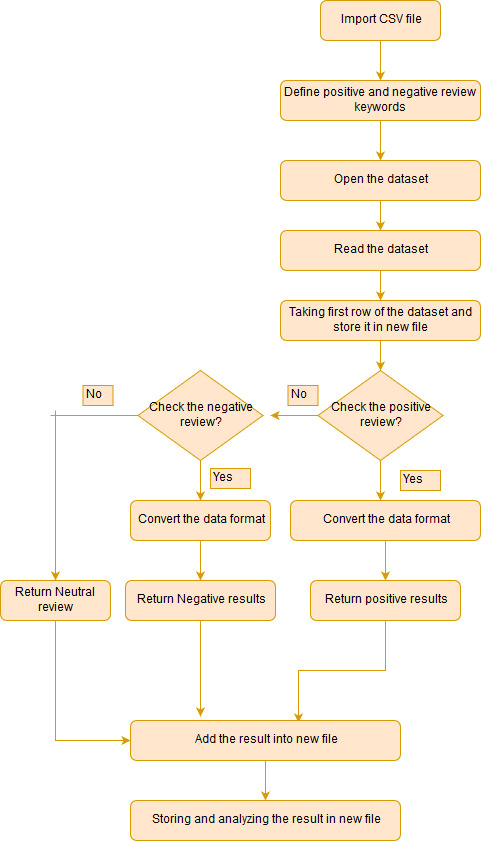
### 5.1.3 Working concepts of python code in opinion mining

Importing CSV file (dataset) and defining the keywords for the positive and negative reviews effectively. The attributes in the dataset are started analysing in row-wise. Here, the basic code “row [:7]” is used to read the data of attributes in each row. This [:7] indicates particular attributes starts from 1to 7 in the row are only considered into the sentimental analysis part whereas other attributes in the row 8 is not taking into the analysis part.

As the first row of the dataset contains the name of the attributes which are not included in the sentimental analysis process, therefore taking the particular row (row 1) as separately and saving into the new file i.e. new.row. Further, sentimental analysis processes are starts from row 2 along with row [:7]. Finally, writing the obtained results into the results.csv file.

### 5.1.4 Working Flow of python in opinion mining

The below image shows the simple and understandable flow chart of python code for this project to analyse the customer opinion about their purchased product.

******

***Figure 5: Working Flow of python in opinion mining***

## 5.2 Source Code work-flow explanation

**Step 1: Importing the csv file**

Import the csv file to analyse the dataset, using the code *“import csv”.*

**Step 2: Defining the keywords**

* Define the positive and negative keywords to analysis the positive/negative/neutral reviews of the customer purchased product.

*positive\_reviews = ["Love", "Nice", "New", "Improvement", "Perfect", "Good", "Is worth", "Best", "Amazing", "Great", "Flawless", "Cool", "Eases", "Better", "Excellent", "Fantastic", "Perfected", "Awesome", "improved"]*

*negative\_reviews = ["Not worth", "Pricey", "Must", "Flaw", "Not", "Useless"]*

* These different keywords for positive and negative reviews are analysed and obtained from various research work.

**Step 3: Reading the data file**

* The dataset in the csv file is opened, by using the code as

“*with open('DatafinitiElectronicsProductData.csv', 'r') as file:” //\* open the dataset\*//*

* The basic code for reading the data in the file is given below,

*first\_row = True //\*defining first row as true to take the row separately\*//*

*with open('DatafinitiElectronicsProductData.csv', 'r') as file:*

*reader = csv.reader(file) //\* read the dataset\*//*

*for row in reader:*

*if(first\_row is True):*

*# creating first row*

*new\_row = row[:7]*

*new\_row.append("Sentiment Analysis")*

*csv\_data.append(new\_row)*

*first\_row = False*

* Here, as the first row is defined as true, the following process are carried out.
* Creating the first row separately by taking the attributes of row 1, from 1 to 7 (i.e. until rating attribute).
* These attribute details from 1 to 7 of row 1 are storing into a separate file as new.row.
* For storing the results of the sentimental analysis, the sentimental analysis attribute is added as an additional attribute (row 8) in the new.row file.
* Again, defining the first row as false to continue the further processes of sentimental analysis.

**Step 4: Checking the positive review of the data**

*if(check\_positive\_review(row[8])): //\* function call \*//*

* This checks the customer comment is positive or not by calling the sub function.
* To check the positive data, it calls the function “def check\_positive\_review(data)”

*def check\_positive\_review(data):*

*data = data.lower(); //\*convert the data format\*//*

*for positive in positive\_reviews: //\*checking with positive review keywords\*//*

*positive = positive.lower();*

*if(positive in data):*

*return True; //\*customer review is positive\*//*

*return False; //\*customer review is negative\*//*

* In order to avoid the upper case and lower case error or bugs in the customer’s review keywords/title/comment,
* *data = data.lower();*  is used to convert the review in the form of lower case.
* Checks the data (i.e. review comment) is positive or not, if the review keyword is positive then it returns the positive results.
* These results are stored into the separate file as new.row.

*# creating row with positive review*

*new\_row = row[:7]*

*new\_row.append("Positive")*

*csv\_data.append(new\_row)*

***For example,***

* ***“The product is prefect and happy to use”*** is the review comment made by the customer about their purchased product.
* Here, the key word “perfect” is matches with the positive review keywords. Thus it returns the result as Positive and add the result in the sentimental analysis column in the new.row.

**Step 5: Checking the Negative review of the data**

* If the review comment made by the customer was not satisfied with the condition of positive review then it checks the negative review condition by calling the negative review function.

*if(check\_positive\_review(row[8]))://\*function call\*//*

* If the above condition gets failed to satisfy, then it checks next condition i.e. negative review.

*elif(check\_negative\_review(row[8])):*

* This checks the customer comment is negative or not by calling the sub function as,

*def check\_negative\_review(data):*

*data = data.lower(); //\*converts data format\*//*

*for negative in negative\_reviews*

*negative = negative.lower();*

*if(negative in data):*

*return True; //\*customer review is negative\*//*

*return False //\*customer review is positive\*//*

* Here, it converts the negative reviews into lower case to avoid error in processing the data.
* Checks the negative keyword in the comment made by the customer about their product.
* If the comment is matches with the negative keyword, then it displays the results as negative and get stored in the sentimental analysis column in the new.row.

*# creating row with positive review*

*new\_row = row[:7]*

*new\_row.append("Positive")*

*csv\_data.append(new\_row)*

***For example,***

* If the review comment was in negative aspects as, ***“I’m not satisfied with this product”.***
* Here, the keyword “not” is matches with the negative keyword thereby it displays the sentimental analysis results as negative.

**Step 6: Checking the neutral review of the data**

* If both positive and negative conditions of the customer review or comment gets failed, then checks with the neutral condition which is given below.

*# creating row with neutral review*

*new\_row = row[:7]*

*new\_row.append("Neutral")*

*csv\_data.append(new\_row)*

* Simply, adding the neutral review in the sentimental analysis column in the new.row.

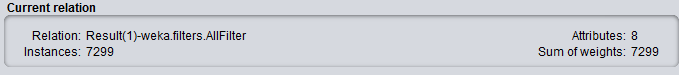
**Step 7: Writing the new dataset to result.csv file**

* Using the code, *writer = csv.writer(open('Result.csv', 'w+', newline=''))* the data in the new.row is writing into the result.csv file.
* Here, w+ is used to write the data by creating a new file called result.csv.
* Thus, writing process of obtained results from the csv file to result.csv is completed.

## 5.3 Practical work- Evaluation

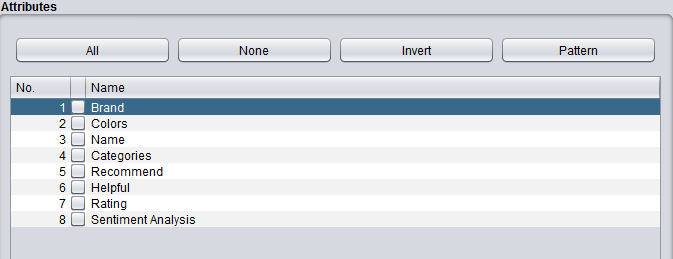
***Stage 1: Pre-process***

In the pre-process, the results.csv file is opened in the Weka tool by using the option as “open file”.

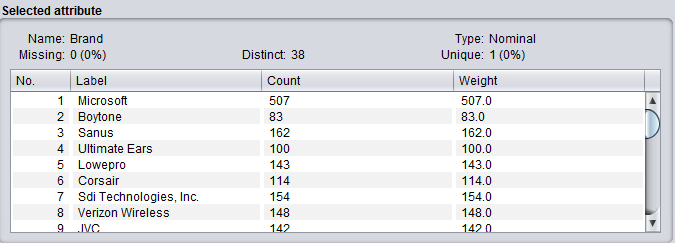


***Figure 6: Pre-process stage***

Choosing all filters in the filter field, results are analysed. The total number of instance in the dataset is 7299 with eight different attributes which are available in the result file.

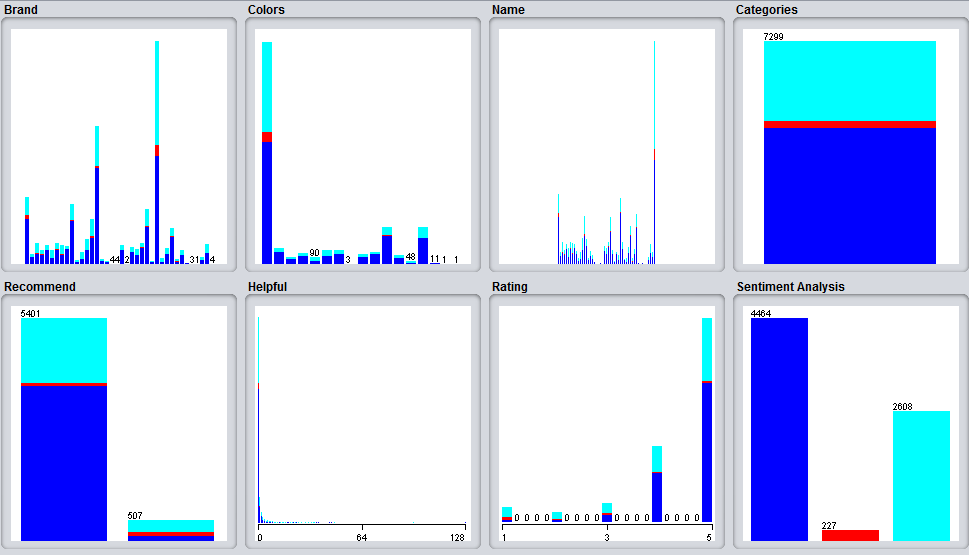


***Figure 7: Attributes list***



***Figure 8: Detailed description of selected attribute***

The attribute field contains list of attributes used in the dataset such as brand, colours, name, categories, sentimental analysis, rating, helpful, recommend. The detailed information about the selected brand are shown in the selected attribute fields.



***Figure 9: Graphical representation of different attributes***

The graphical representation of sentimental analysis results for all attributes is shown in the class field. This graphical view shows the clear sentimental results for each attribute in the dataset.

***Stage 2: Classify***

***For Naïve Bayes,***

In classify stage, applying the different classifier model and analysing the classifier results by using the test option as “use training set”. First, the classifier model is chosen as naive Bayes in the Bayes and choosing the test option as “use of a training set”. Start analysing the results for sentimental analysis attributes.

***Classifier Results for naïve Bayes***

Time taken to build model: 0.08 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.84 seconds

=== Summary ===

Correctly Classified Instances 4769 65.3377 %

Incorrectly Classified Instances 2530 34.6623 %

Kappa statistic 0.2554

Mean absolute error 0.2703

Root mean squared error 0.4031

Relative absolute error 81.5294 %

Root relative squared error 99.0124 %

Total Number of Instances 7299

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.843 0.564 0.702 0.843 0.766 0.309 0.704 0.758 Positive

0.229 0.023 0.241 0.229 0.235 0.211 0.834 0.169 Negative

0.365 0.164 0.554 0.365 0.440 0.228 0.663 0.510 Neutral

Weighted Avg. 0.653 0.404 0.635 0.653 0.633 0.277 0.693 0.651

=== Confusion Matrix ===

a b c <-- classified as

3764 39 661 | a = Positive

68 52 107 | b = Negative

1530 125 953 | c = Neutral

***For Logistics,***

The logistic function results are analysed by choosing the classifier model as logistics in the functions. Use training set option is chosen in the test option field and start analysing the results for logistics in the classifier output field.

***Classifier results for logistics***

Time taken to build model: 7.88 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.53 seconds

=== Summary ===

Correctly Classified Instances 4784 65.5432 %

Incorrectly Classified Instances 2515 34.4568 %

Kappa statistic 0.2286

Mean absolute error 0.2948

Root mean squared error 0.3843

Relative absolute error 88.902 %

Root relative squared error 94.3881 %

Total Number of Instances 7299

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.870 0.626 0.686 0.870 0.767 0.285 0.703 0.759 Positive

0.018 0.003 0.167 0.018 0.032 0.045 0.855 0.175 Negative

0.344 0.154 0.554 0.344 0.424 0.219 0.672 0.530 Neutral

Weighted Avg. 0.655 0.438 0.623 0.655 0.622 0.254 0.697 0.659

=== Confusion Matrix ===

a b c <-- classified as

3884 5 575 | a = Positive

77 4 146 | b = Negative

1697 15 896 | c = Neutral

***For SMO,***

In classify, the classifier model is chosen as SMO under functions and use training set option is chosen in the test field option. Finally, SMO’s classification output is obtained in the classifier output screen.

***Classifier results***

Time taken to build model: 42.28 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.42 seconds

=== Summary ===

Correctly Classified Instances 4830 66.1734 %

Incorrectly Classified Instances 2469 33.8266 %

Kappa statistic 0.2046

Mean absolute error 0.3043

Root mean squared error 0.3952

Relative absolute error 91.7702 %

Root relative squared error 97.0557 %

Total Number of Instances 7299

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.937 0.726 0.670 0.937 0.781 0.292 0.605 0.666 Positive

0.000 0.000 ? 0.000 ? ? 0.500 0.031 Negative

0.248 0.088 0.612 0.248 0.353 0.219 0.580 0.421 Neutral

Weighted Avg. 0.662 0.475 ? 0.662 ? ? 0.593 0.559

=== Confusion Matrix ===

a b c <-- classified as

4182 0 282 | a = Positive

98 0 129 | b = Negative

1960 0 648 | c = Neutral

# Chapter E

## 6.1 Final results

This section covers the detailed analysis of the final outcome of the project by undergoing different classification algorithms. The customer’s opinion about their purchased product is analysed with the help of dataset. The dataset is obtained from the online source which has eight different attributes to find out the customer opinion about their purchased product. Here, the name of the product, review title and recommendation or rating attributes are highly helpful to categories the customer reviews in terms of positive, negative and neutral review. The following graphs of different algorithms will help to find the best algorithms in terms of performance and accuracy. To obtain best and accurate results, the different attributes in the dataset is correctly classified and analysed.

## 6.2 Comparison of classification algorithms-Results

An e-commerce result for “an online shopping customer opinion about their purchased product” is analysed and evaluated in a well-versed manner. The customer's review and opinion about their purchased product are separated in terms of positive, negative and neutral review which is shown clearly in the above discussion. The valid input attributes are processed and evaluated as per the e-commerce dataset. Compare with logistics algorithm, SMO/ support vector machine algorithm offers the high throughput and minimum processing speed and time. The prediction analysis is made simpler and easier with the help of confusion matrix where, the positive, negative and neutral review results or customer opinion are obtained.

This section covers, the comparison analysis of three different classifier model such as SMO/ support vector machine, logistics, and naïve Bayes. The performance /efficiency of three different models is shown in tabular format. Here, the results of classification algorithms are obtained with the help of data mining tool i.e. over the Weka. Uploading the large datasets into the data mining tool, the prediction accuracy is obtained for each classifier models. The prediction accuracy depends on the percentage of correctly classifier instances by each classifier model (Zadgaonkar, 2013).

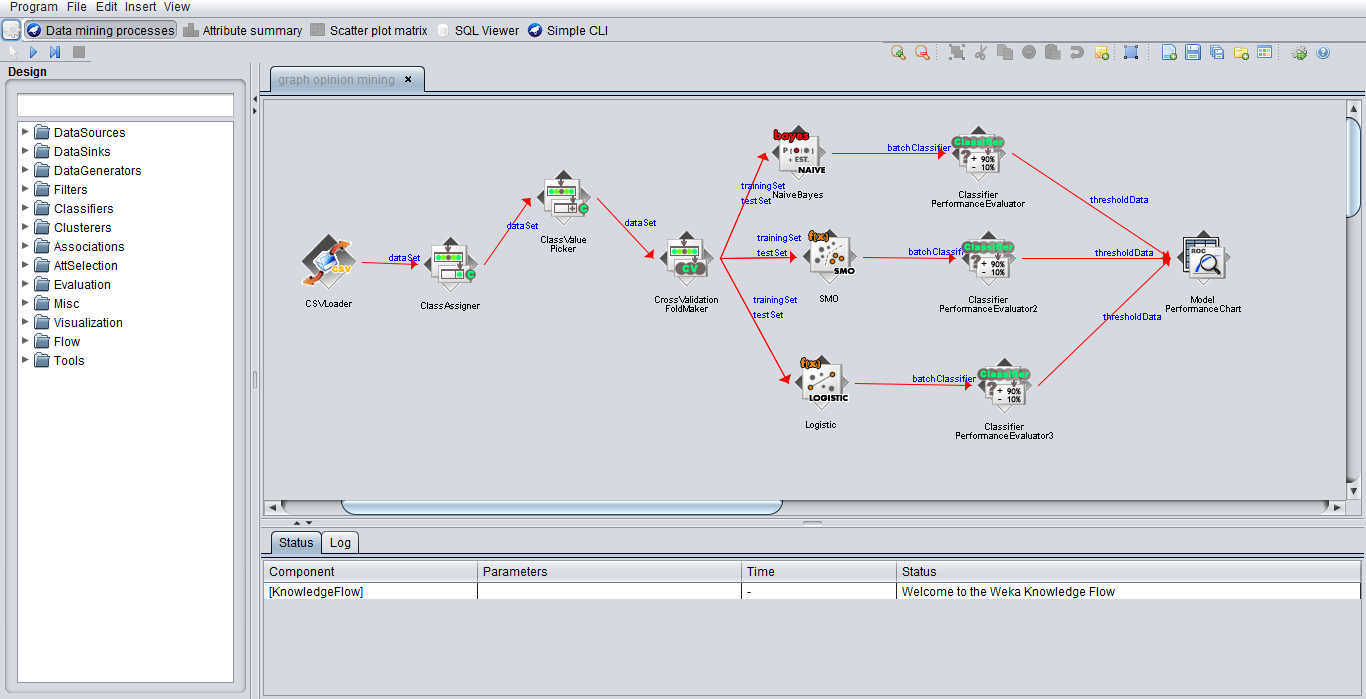
***Table 2: Classifier results comparison table***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Classifier model* | *Test model time*  *(in second)* | *Correctly classifier instance* | *Incorrectly classifier instance* | *Efficiency in*  *%* |
|
| Naïve Bayes | 0.84 | 4769 | 2530 | 65.3377 % |
| Logistics | 0.53 | 4784 | 2525 | 65.5432 % |
| SMO | 0.42 | 4830 | 2469 | **66.1734 %** |

The above tabulation shows the comparative evaluation of three different classifier models. These classifier results are obtained over the Weka tool. Among all three classifier models, SMO performance is high and best in all aspects. Here, the SVM takes 0.4 seconds to test the dataset whereas the logistics and naïve Bayes takes 0.5 and 0.84 seconds respectively. Hence, it is clearly noted that the time taken to test the model by SMO is very less compared with other algorithms. The efficiency obtained by the SVM over the correct instance is 66.1734 % which is highest among all. The naive Bayes and logistics classifier models have quite closer efficiency such as 65.3377 % and 65.5432 % respectively but still, ***SMO has the highest efficiency as 66.1734%.***

## 6.3 Graphical results

The graphical analysis of the “customer opinion about their purchased product” is made with the help of weka tool. In the Weka tool, under the knowledge flow environment function, the following processes are carried out effectively and efficiently.

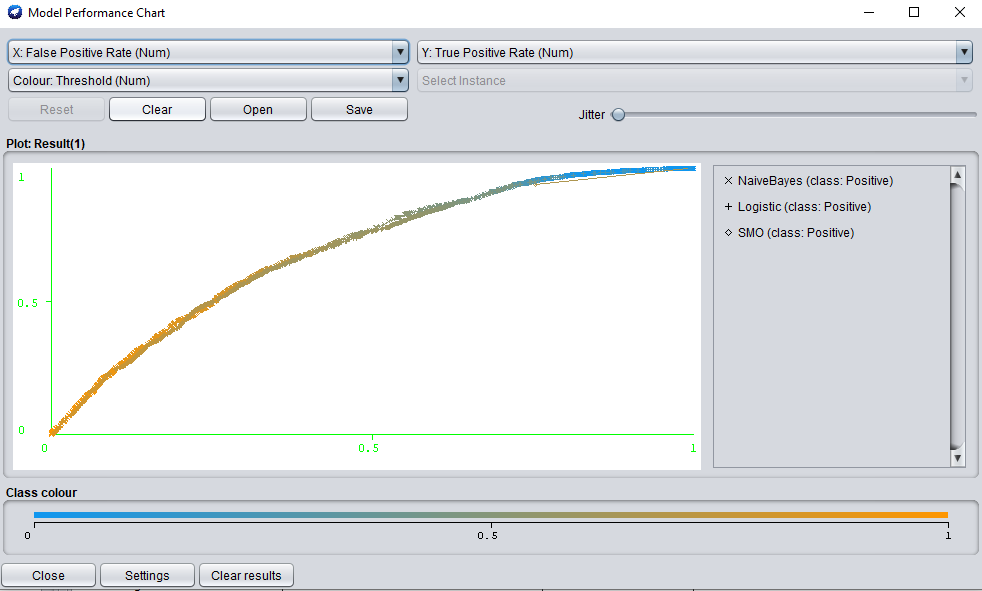
***Figure 10: Knowledge flow design***

This image represents the performance comparison of three different classifier models in order to analyse the customer opinion about their purchased product. Here, the SMO, naive Bayes and logistics classifier models are chosen for the graphical-comparison evaluation. First, selective modules are chosen and link each module with their respective function. After the successful processes of linking one module with another along with their functions, configure process is made in order to upload the dataset.

By right-clicking the class value picker, the configuration processes are done by choosing any of classes as a positive, negative and neutral class. After successful execution of the configuration process, right-clicking the model performance chart in order to visualize the changes of the results in a graphical manner. The below graphs shows the performance results for three different algorithms with respect to true positive and false positive. Here, the customer opinion about their purchased product is classified into three different classes such as positive review, negative review and neutral about the product. The following graphs shows the performance comparison of each class over three different algorithms/ techniques in weka tool.

***Positive review analysis***

The positive reviews from all different algorithms are collected together in a well-versed manner. For the positive review, the performance accuracy of each algorithm is detected easily and effectively with the help of weka tool. The below graph shows the positive review achieved by each algorithm.



***Figure 11: Analysing positive review of customer opinion***

Compare with naïve Bayes and logistic algorithms, SMO detects the positive reviews made by each customer about their purchased product, clearly and effectively. In graph, SMO detects the positive reviews of customer at the top of the graph with high accuracy.

Here, the X-axis represents the false positive rate of a positive review

Y-axis represents the true positive rate of a positive review

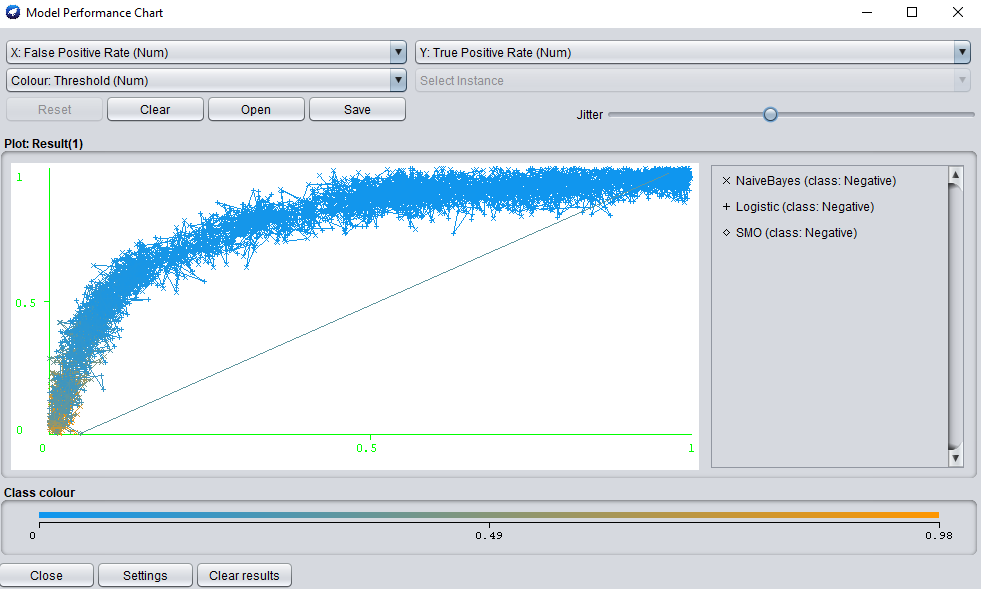
***Negative review analysis***

The below graph shows a comparison of negative review detected by each algorithm in weka tool. The graph contains the negative class (i.e. negative review made by the customer about their purchased product) detection made by three different algorithms effectively. From the graphical results, it is clearly noted that SMO detects the negative review of the customer about their product with accuracy. The obtained results of logistic and naive Bayes algorithms have fluctuation and high noisy in the results and fails to meet an accurate results.

Here, the X-axis represents the false positive rate of a negative review

Y-axis represents the true positive rate of a negative review

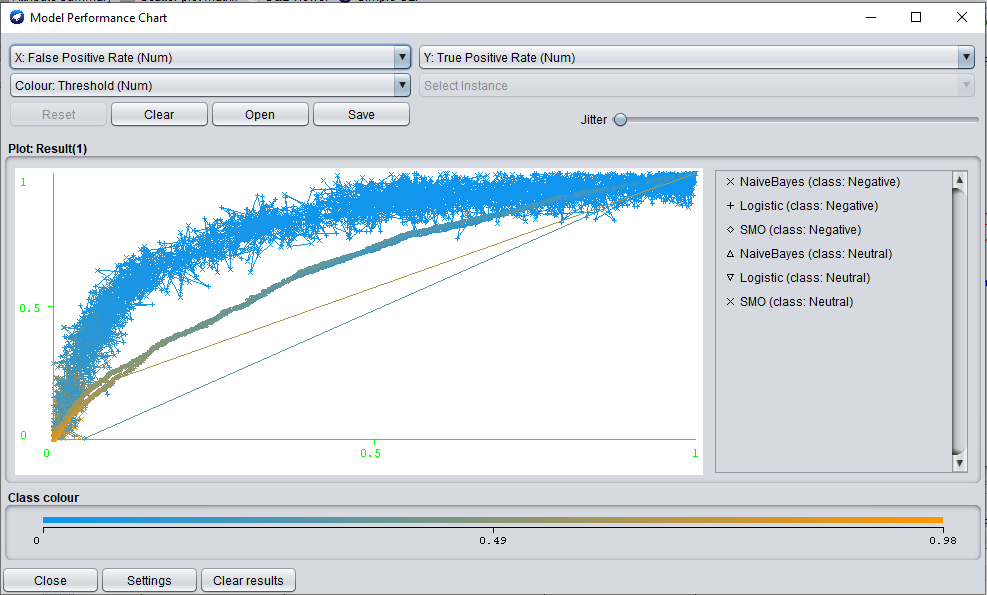
Thus, over the SMO algorithm, when the negative review made by the customer about their product are also be known clearly and easily.



***Figure 12: Analysing negative review of customer opinion***

***Neutral review analysis***

The below image shows the neutral review made by the customer about their purchased product and it will also shows the accuracy of results obtained by each algorithm. From the three different classifier algorithms, the neutral reviews are collected in a well-versed manner. After collecting all neutral reviews made by the customer from three different algorithms, the best classifier model is analysed over the graphical representation. Here compare with naïve Bayes and logistic, SMO detects the neutral results with the best accuracy.



***Figure 13: Analysing neutral review of customer opinion***

Here, X-axis represents the false positive rate of neutral review

Y-axis represents the true positive rate of neutral review

From the obtained graphical results, neutral review of customer opinion is analysed in a well-versed manner. SMO delivers better accuracy (offers consistency results in all stages without any fluctuation or noisy) and performance level of data over detecting the neutral reviews compared with other algorithms. As it processes the data in a linear way, it works better than two different algorithms.

# Chapter E

## 7.1 Conclusion

The online shopping customer’s review about their purchased product is analysed in advance by using different machine learning algorithm easily and effectively. Here, the support vector machine algorithm is chosen to processes the customer review and categorizing the review into three different states such as positive, negative and neutral. The obtained dataset from the online resource i.e. from e-commerce website for doing the research purpose. The e-commerce dataset is upload into the data mining processes and conducting the support vector machine algorithm over the dataset. In order to obtain the best and effective results in all aspects especially in terms of efficiency and high performance, various research works are reviewed (Yong Shi, 2012).

From the research work analysis, the support vector machine algorithm is chosen to proceed. The classification of SVM offers the best performance and it also works better in the textual format of the dataset. Over this project work, the intended purpose of the project “analysing the customer opinion about their purchased product and suggesting the other new customer purchase the particular product” is achieved. Hence, the new customer will be highly benefited to purchase a high-quality product without any damage by knowing the previous opinion of customer’s about their purchased products. Meanwhile, product demand is also increasing among the customer.

The effectiveness of support vector machine (SVM) algorithm over the opinion mining is analysed with the help of different research work of various authors. Thus, both theoretical and practical work is made to obtain the best results with high efficiency. The support vector machine algorithm works effectively and efficiently with online an e-commerce dataset. Over the practical work, the performance and an efficiency level of SVM are obtained as higher than the other two algorithms such as logistics and naïve Bayes. Over weka, the correct and incorrect results for each classifier models are obtained which are not the same for all classifiers. Among all classifier model results, SMO or SVM is considered as the best because which offers high efficiency by consuming minimum time to processes the dataset. The graphical comparison is also shown in this project report. The intended function of SVM in the data mining project is to gather and maintain the customer relationship effectively and efficiently.

Thus, the customer opinion about their purchased product has been analysed and predicted easily and effectively by using the SVM / SMO technique. Performance of SMO is evaluated and compared with the other two different algorithms (naïve Bayes and logistics). From the comparison of results, it is obtained that SMO reduces the speed complexity and offers the best technique to improve accuracy.

## 7.2 Recommendation

The current results of SVM offer efficiency as 67 percent to “analyse the customer review about their purchased product”. An efficiency range of SVM in the current results may higher than all other algorithms such as logistics and naïve Bayes but it is quite low. As the SVM works effectively only with the smaller dataset, hereby the efficiency of the SVM for large data will be automatically reduced i.e. when larger the dataset range, then an efficacy level of SVM will be minimized (Yong Shi, 2012).

According to this project work, the accuracy level of SMO is obtained as 66% which is almost better than other algorithms but still, it is quite low. In order to increase or improvise the classification accuracy (support vector machine/SMO) on the test data, cross-validation test will high recommended (Tong, 2011). Thus, more number of cross-fold (10 cross fold) validation will improvise the classification accuracy on the test data. By undergoing cross-validation tests, the SMO classifier’s prediction and detection accuracy for a large number of training data will be achieved in a well-versed manner.

In the future, to overcome the problem of poor efficacy range of SVM for a large dataset, cross-validation test processes is highly recommended. The smart text classifier will also be helpful to collect the review made by the customer from any language and to processes the operation accordingly, these effective text classifier functions will achieve by integrating the function of SVM with fuzzy logic.

Using the SVM with fuzzy logic, the text classification processes will carry out as the best accuracy when compared with SVM. The fuzzy logic is one of the techniques used to extract the customer intention about the purchased product with the help of text information. The effectiveness of fuzzy logic over the opinion mining is simple and easy to understand. Meanwhile, the SVM is also offered effective and efficient training to analyse the customer intention about their purchased products. As the processes involves with fuzzy logic techniques which will delivers the better accuracy than simple SVM algorithm. Hence, fuzzy logic based SVM will offers 97% accuracy (Kadyan, 2016).

Over the fuzzy logic based SVM algorithms, the complex text classification can also be processed effectively and efficiently without any confusion. The SVM may get fails when the text information of customer review is from other languages. But, fuzzy based SVM will highly work with any type of languages. Hence, it is considered as the best option to analyse the customer review about their purchased product i.e. sentimental analysis (Raheesa Safrin, 2017). Finally, to analyse the customer intention about their purchased product and to classify the products review in terms of a positive, negative and neutral basis, the fuzzy logic-based SVM is highly recommended. By implementing the fuzzy logic based SVM algorithms, the customer review and prediction about their product can be processed with high accuracy as 97 cents.

# References

A.Halawaa, A. (2018). Improve the automatic classification accuracy for Arabic tweets using ensemble methods. *Journal of Electrical Systems and Information Technology*, 8.

A.Younessb, H. M. (2018). Improve the automatic classification accuracy for Arabic tweets using ensemble methods. *Journal of Electrical Systems and Information Technology*, 8.

Abdulaziz Elwalda. (2014). The Influence of Online Customer Reviews on Purchase Intention: the Role of Non-numerical Factors. *European Marketing Conference LCBR, At Munich, Germany*, 1 - 13.

Alexandra Cernian, V. S. (2015). Sentiment analysis from product reviews using SentiWordNet as lexical resource. *ECAI 2015 - International Conference – 7th Edition*, 15 - 18.

Andersen P. (2017). Examining impacts of negative reviews and purchase goals on consumer purchase decision. *Journal of Retailing and Consumer Services,* , 1-8.

Asghar, N. (2018). Sentiment Analysis of Customer Reviews based on Integration of Structuredand Unstructured Texts. *https://cs.uwaterloo.ca/~nasghar/dmproj.pdf*, 7.

Belbag, S. (2016). Prediction of Consumer Behavior Regarding Purchasing Remanufactured Products: A Logistics Regression Model. *International Journal of Business and Social Research*, 1-10.

D. Ben Ayed Mezghani, S. Z. (2010). Evaluation of SVM Kernels and Conventional Machine Learning Algorithms for Speaker Identification . *International Journal of International Journal of Hybrid Information Technology* .

Dhokrat, A. (2015). Review on Techniques and Tools used for Opinion Mining. *International Journal of Computer Applications Technology andResearch*, 6.

Grazyna Suchacka, A. P. (2015). Classification of E-Customer sessions based on support vector machine. *Proceedings 29th European Conference on Modelling and Simulation*, 1 - 7.

Hani Nurrahmi, W. M. (2016). Feature Extraction and Opinion Classification Using Class Sequential Rule on Customer Product Review. *Fourth International Conference on Information and Communication Technologies (ICoICT)*, 1 -5.

Himani Bhavsar, M. H. (2012). A Review on Support Vector Machine for Data Classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*.

Hui Song, J. C. (2013). Semantic Analysis and Implicit Target Extraction of Comments from E-commerce Websites. *Fourth World Congress on Software Engineering*.

IraklisVarlamis. (2017). Large scale opinion mining for social, news and blog data. *Journal of Systems and Software*, 13.

Jebaseeli, N. (2016). An Opinion Mining and Sentiment Analysis Techniques: A Survey. *International Research Journal of Engineering and Technology*, 3.

Kadyan, D. (2016). Sentiment Analysis using Fuzzy based SVM Classifier . *International Journal of Innovations in Engineering and Technology*, 5.

Kanellopoulos. (2006). Data Preprocessing for Supervised Leaning. *International journal of computer science*, 1-8.

Kwon Y S. (2011). Grocery Customer Behavior Analysis using RFID-based Shopping Paths Data. *World Academy of Science, Engineering and Technology*, 1-5.

Liu T. (2018). TagBooth: Deep Shopping Data Acquisition Powered by RFID Tags. *IEEE*, 1-9.

Lopamudra Dey, S. C. (2014). Sentiment Analysis of Review Datasets using Naïve Bayes’ and K-NN Classifier . *https://arxiv.org/ftp/arxiv/papers/1610/1610.09982.pdf*, 7.

M Hu, B. L. (2006). Opinion feature extraction using class sequential rule. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 61 - 66.

Raheesa Safrin, K. T. (2017). Sentiment analysis on online product review. *International Research Journal of Engineering and Technology (IRJET), 4*(4), 8.

Santhosh Kumar K L, J. D. (2016). Opinion Mining and Sentiment Analysis on Online Customer Review. *IEEE International Conference on Computational Intelligence and Computing Research*, 1 - 4.

Shafi S I. (2014). An Investigation on Shoppers' Buying Behaviour Towards Apparel Products in Bangalore City. . *Pacific Business Review International*, 1 - 6.

Sharmista. (2018 ). SVM and Fuzzy SVMBased Opinion Mining In Tamil Using R. *American Journal of Engineering Research*, 9.

Tong, S. (2011). Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, 1-22.

Yong Shi. (2012). Recent advances on support vector machines research. *Journal Technological and Economic Development of Economy*, 1-6.

Zadgaonkar, A. (2013). Opinion Mining and analysis:A survey. *International Journal on Natural Language Computing*, 10.

Zhang, W. (2011). An Improvement to Naive Bayes for Text Classification. *Procedia Engineering*, 1-7.