Aspect Level Sentiment Analysis of E-Commerce: A case study of eBay and Amazon



Furqan Durrani

03-243191-025

A thesis submitted in the fulfilment of the

requirements for the award of the degree of

Master of Science in (Computer Science)

Department of Computer Science

BAHRIA UNIVERSITY LAHORE

February

i

**Author’s Declaration**

I, Furqan Durrani hereby state that my MS thesis titled“ Aspect Level Sentiment Analysis of E-Commerce: A Case Study of eBay and Amazon” is my own work and has not been submitted previously by me for taking any degree from this university Bahria University Lahore Campus or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Name of Student: Furqan Durrani

Date: February 10, 2021

ii

**Plagiarism Undertaking**

I, solemnly declare that research work presented in the thesis titled “Aspect Level Sentiment Analysis of E-Commerce: A Case Study of eBay and Amazon” is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the university reserves the right to withdraw / revoke my MS degree and that HEC and the University has the right to publish my name on the HEC

* University website on which names of students are placed who submitted plagiarized thesis.

Student / Author’s Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of the Student: Furqan Durrani

iii

**DEDICATION**

*I would like to dedicate my thesis to ALLAH Almighty, who constantly supported me and motivates me. My parents who teaches me to have faith in ALLAH and to remain determined and confident.*

iv

**ACKNOWLEDGMENTS**

I would like to thank ALLAH Almighty, who gave me the strength and courage to accomplish this research and make me able to contribute my little efforts in the field of data science.

This thesis is the first-fruit of the training and efforts, which I have been receiving form my respected professor, and as a gratitude, I primary dedicate my thesis to my professor Dr. M. Aasim Qureshi. Without his constant guidance, it would not be possible for me to achieve my goal. I would like to acknowledge my brothers, who waited hours and hours outside university for me. I would like to thank all my teachers, who contributed in my academic career.

v

ABSTRACT

Sentiment analysis refers to the study of people's sentiments, opinions, emotions, and attitudes expressed in the form of written language or the process of classifies the opinions expressed in the form of text. Specifically, it is used to determine whether the person’s attitude towards a particular subject or product is positive, negative, or neutral. The e-commerce websites provide the facility to users to share their feedback, opinions, or reviews about any particular product of the market. With the advancement of technology, every person has easy access to the internet and they feel comfortable with online purchasing and share their experience for others in the comments section. Existing machine learning models provide a useful account of how to judge the polarity. However, the accuracy of context relative position information for the target terms is still required to be done. Hence, this study proposed the model which contains the combinations of two machine learning models Logistic Regression and Random Forest under the architecture of voting classifiers. For this study, data is collected from ecommerce websites eBay and Amazon. After pre-processing, 5 sub data set with respect to the aspects of price, colour, size, weight and service have been extracted and the proposed model and different machine learning models (Naive Bayes, KNN, SVM, Random Forest) are applied. We have observed that accuracy has been improved by using proposed methodology which is 97%.

**Table of Contents**

[Introduction 19](#_Toc63587347)

[1.1. Sentiment Analysis 21](#_Toc63587348)

[1.1.1 Document Level Sentiment Analysis 22](#_Toc63587349)

[1.1.2 Sentence Level Sentiment Analysis 22](#_Toc63587350)

[1.1.3 Aspect Level Sentiment Analysis 22](#_Toc63587351)

[1.2 Machine Learning 23](#_Toc63587352)

[1.2.1 Research Types 23](#_Toc63587353)

[1.2.2 Unsupervised Machine Learning 24](#_Toc63587354)

[1.2.3 Semi-Supervised Learning 24](#_Toc63587355)

[1.2.4 Supervised Learning 25](#_Toc63587356)

[1.3 Data Resampling 33](#_Toc63587357)

[1.4 Evolution Measures 34](#_Toc63587358)

[1.4.1 Accuracy 34](#_Toc63587359)

[1.4.2 Precision 35](#_Toc63587360)

[1.4.3 Recall 35](#_Toc63587361)

[1.4.4 F-Measure 35](#_Toc63587362)

[1.5 Research Gap 35](#_Toc63587363)

[1.6 Motivation 36](#_Toc63587364)

[1.7 Problem Statement 37](#_Toc63587365)

[1.8 Research Questions: 37](#_Toc63587366)

[1.9 Main Objectives 37](#_Toc63587367)

[1.10 Main Contribution 37](#_Toc63587368)

[Literature Review 39](#_Toc63587369)

[Methodology 50](#_Toc63587370)

[3.1 Dataset description 50](#_Toc63587371)

[3.2 Pre-processing of data 52](#_Toc63587372)

[3.2.1 Tokenization 52](#_Toc63587373)

[3.2.2 Punctuation Removal 53](#_Toc63587374)

[3.2.3 Numeric Removal 54](#_Toc63587375)

[3.2.4 Lowercase Conversion 54](#_Toc63587376)

[3.2.5 Stemming 55](#_Toc63587377)

[3.2.6 Stop words Removal 56](#_Toc63587378)

[3.3 Features Extraction 57](#_Toc63587379)

[3.4 Hyper parameters for classifiers 58](#_Toc63587380)

[3.5 Experimental Flow 59](#_Toc63587381)

[Result and Discussion 63](#_Toc63587382)

[4.1 LR Results 63](#_Toc63587383)

[4.2 RF Results 66](#_Toc63587384)

[4.3 SVM Results 68](#_Toc63587385)

[4.4 KNN Results 71](#_Toc63587386)

[4.5 GNB Results 73](#_Toc63587387)

[4.6 Voting Classifier Results 76](#_Toc63587388)

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **TITLE** | **PAGE** |

xiii

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **TABLE NO.** | **TITLE** | **PAGE** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**LIST OF ABBREVIATIONS**

**LIST OF APPENDICES**

|  |  |  |
| --- | --- | --- |
| **APPENDIX** | **TITLE** | **PAGE** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Chapter 1

# Introduction

Sentiment analysis is the study of people's opinions, sentiments, attitudes, and emotions expressed in written languages. [1]. Sentiment analysis is a grown-up field in the research area. People are now using the web for business correspondence, e-commerce, and e-marketing [2]. As the online shopping trends are growing, customers eagerly want to share their emotions and reviews on different platforms on the internet. Extraction of users feeling from the reviews is very important for the user to select the right product and its variations. Sentiment analysis is also important for organizations to grow their business by tracking the customer feedback over their different products. It is also being used to predict the results of national events like elections, etc. [3]. With the development of online shopping and e-commerce, now the bulk of the users are buying their desired products from online stores.

In the comparison of local shopping and online shopping, users are enjoying the facilities of online shopping because they can buy anything from anywhere and anytime [4]. Moreover, multiple styles and varieties of products are available in online shopping stores and consumers have a good choice to buy variations of products without going outside [5].

While e-commerce is performing a positive role for the convenience of the user but some problems like products and delivery are also associated with it. The problems with product and delivery can be like the contradiction between real things and descriptive information available on the product, the service of product delivery and poor quality of the received product, and many more [6]. That is why it is very important to evaluate products from online shopping store and to check the tendency of the sentiment of the customer towards the product. It also helps the organization for business growth and the reference for other consumers. Sentiment analysis for customer reviews is also defined as the process of systematically analysing the subject of emotional reviews, termed as opinion mining and text analysis [7].

Sentiment analysis allows the new customer to examine previous customer's suggestions and reviews about the product. Sentiment analysis is a basic view point when people start e-commerce. With the advancement of the internet throughout the world, a large number of people are engaged in writing reviews and giving feedback. The reviews, which are written by the consumers, help other buyers to make informed decisions about the product(s) they want to purchase. These reviews are also beneficial for the manufacturers of products. They need to go through the reviews of buyers. It is a rather complex task for a customer to identify significant details from the prevalent information available on the website due to a large number of reviews. Sentiment analysis has now become a topic of research that has been widely carried out [2].

However, due to a large number of reviews on e-commerce websites from customers, it is not possible to analyse the opinion manually. Therefore, sentiment analysis is a significant approach for opinion extraction. In previous researches, sentiment analysis has been applied in many different fields [8-19]. One of the important prerequisites of sentiment analysis is the selection of words that is used as an aspect in the analysis. Many techniques depend on the dictionary for analysis of sentiment has been performed in different researches. The efficiency of dictionary-based sentiment analysis is dependent on the accuracy and comprehensiveness of the dictionary [20]. The language used for reviews is between formal and informal language. Sentiment words are not much domain-specific and also contains short words which create difficulties for making an accurate dictionary. However, many types of research have been performed on English text. It is observed that English words are not natural. The Chinese words in sentences are required to splits into segments for sentiment analysis. The accuracy of sentiment analysis in the Chinese language is dependent on the segmentation of the sentences [21].

Information retrieval techniques are used to gather data from different Blogs and E-commerce websites where the people share their opinion. [22]. Once the reviews are collected, then the next problem is to analyse the reviews. Multiple Data mining and Machine Learning approaches are present for the resolution of this problem [23]. From the bulk of reviews, some opinions are positive and some are negative. The negative and positive opinions represent the polarity of review, and the analysis of a large number of opinions on the basis of the polarity is said to be the sentiment analysis. It is also said to be the study of attitude, emotion, and opinion of the consumers towards a particular item [24]. Sentiment analysis is also said to be the process of classification of the reviews on the three types of sentiments like positive, negative and neutral.

## Sentiment Analysis

Sentiment analysis could be a natural language process (NLP) task within which a certain text is assessed into predefined categories (e.g., positive, negative and neutral,). Initially, handmade sentiment lexicons models were used for sentiment analysis were that contain emotional words explained with polarities [25] [26] [27]. Generally, they collect sentimental words from phrases. Based on scattered information like strength and polarities of emotional words, they categorize sentences in classes of sentiments with help of polarities [28] [29].

Moreover, the lexicon-based models are efficient and simple. Sentiment lexicon construction manually is a time-consuming and labour-intensive job. Secondly, already static polarity is needed to be provided for every sentiment. For this solution, some kind of models that automatically generate sentiment lexicons have been proposed.

Like the sentence "it is very hot today". In this sense of lexicon-based approach, this sentence expresses the negative behaviour because it is using "very hot” means extreme hot weather. But in the sentence "the boy is very hot" the very hot expresses that the feature of a boy that boy is looking so smart. For this solution, some kind of machine learning-based models has been proposed. But these machine learning-based models have required a large set of data with their polarity for the training of the model. And this is not a critical problem nowadays because several blogs and e-commerce websites available which is being used to share their opinion about any good or anything which he has purchased earlier by the user.

For understanding this problem, there is some kind of reviews have multiple useful meanings. Like this sentence "Samsung is a good brand of mobile" in this specific sentence, clear positive opinion can be extracted and for example “techno brand of mobile is not a good brand”. From this sentence, we can understand the negative review of the consumer. But what in the case when the user shares the opinion like “Samsung is a good brand but techno is not a good brand in the same sentence”. This kind of problem can be resolved by categorizing the sentiment analysis in the the following techniques mentioned in figure 1.

Sentiment Analysis

Document Level Sentiment Analysis

Sentence Level Sentiment Analysis

Aspect Level Sentiment Analysis

(Caption)

### **Document Level Sentiment Analysis**

Document-level sentiment analysis is said to be the sentiment analysis of a whole document. In this approach, the complete document is considered as a single entity and it is analysed at once. The review of the document is considered as positive or negative. Anyway, this is not a good approach because there may be e a positive specific path that has great importance but the overall sentiment score of the document is negative and vice versa [30] [31].

### **Sentence Level Sentiment Analysis**

Sentiment analysis at the sentence level is considered as is the calculation of sentiment on the sentences in the document. In this approach, the document is divided into sentences, and his sentence is considered as an entity. This is a better way to find the sentiment clarity as compared to the whole document because in this technique every sentence is analysed separately. Anyhow this is also not the best case to find the sentiment because referring to the above example Samsung is a good brand but that techno is not a good brand. Another example sentence for the reference is "the functionality of Samsung mobile is too much smooth but very short battery life". In the above examples. We can extract the multiple meanings. To overcome this issue the spirit level sentiment analysis has been proposed [32, 33, and 34]

### **Aspect Level Sentiment Analysis**

Aspect level sentiment analysis is said to be the analysis in which every feature or aspect is considered as an entity like price, size, and weight of mobile. a feature is said to be the instance or attribute of anything. In this approach, the main focus is to find out the feature of an entity and to find out the sentiment according to the feature. Aspect level sentiment analysis has been performed in many fields so far like explain in [35,36] researches.

## Machine Learning

In 1997 researcher [37] defines the machine learning as it is the feature of computer science which aims to gain knowledge from data. Machine learning is used to improve the efficiency of different analysis for example in applied Health Care and Emotion Detection etc. This is used to automate the process, for flexibility and efficiency that identify the trends from Complex data sets [38].

When to use Machine Learning?

There are multiple steps involved to determine when it is being used. The first step is that the machine learning technique can be used to answer the research question. In research [39], the researcher defines the three types of research problems i.e. Descriptive research, Explanatory research, and Predictive research can be resolved through machine learning. For the mentioned task, machine learning has been performed and it is verified by the statistical methods which are sufficient in some cases and sometimes the such questions validate the results.

### Research Types

Research has been divided into the following three types:

#### Descriptive Research:

The main purpose of descriptive research is to provide a summary of the properties of the data.

#### Predictive Research:

The main purpose of predictive research is to forecast the future outcomes that would utilize for money think screening and selection:

#### Explanatory Research:

The main purpose of this research is to understand the informal mechanism that would be used to create future interventions

To solve the above research categories, machine learning has been divided in the following three types [Reference].

Machine Learning

Unsupervised Machine Learning

Supervised Machine Learning

Semi-Supervised Machine Learning

(Caption)

### Unsupervised Machine Learning

Unsupervised machine learning methodology is specifically helpful for descriptive research because this research aims to find the relationship between the data structure without knowing the statistical outcomes [Reference]. This methodology is referred to as unsupervised learning because we don't have any target variable that could be happened.

The main purpose of unsupervised learning is to identify or analyse the dimensions of the component’s trajectories for clusters from the data set. Multiple approaches for unsupervised learning are used like Factor analysis, mixture modelling and component analysis.

### Semi-Supervised Learning

Semi-supervised learning consists of both types of unsupervised and supervised learning. In this technique, dataset can be labelled or unlabelled. The labelled data is utilized to train the model and the unlabelled data is utilized to purify the boundaries of classes. In semi-supervised machine learning methodology, K nearest neighbour, perceptron, neural network, convolutional neural network techniques are used [40] [41] [42].

### Supervised Learning

Supervised learning is utilized by the predictive research because the main purpose of supervised learning is to predict or classify the future outcome of data. Supervised machine learning is implemented on a large number of datasets like reviews dataset to predict the user satisfaction level for any product [43-46]. Supervised machine learning can be used when prior knowledge of the predicting labels or classes are available. In this technique, the algorithm is trained with the help of the large amount of dataset first to train the model and then the test data set is passed through the model and the analysis of the efficiency of the model is measured by calculating the accuracy.

Supervised learning is one of the machine learning technique in which predictive classes are known. This technique is implemented on labelled data sets. In case of the review detection, review may be positive or negative or may be neutral. So, in this example, the predicting class of a review would be a negative, positive or neutral. The technique of supervised machine learning is as the data set is divided into the training dataset and the test dataset. The training of the model is performed by labelling dataset with actual sentiment and then test data set is passed over the model and results are observed.

There are mainly two techniques are used in supervised machine learning regression and classification [47]

#### Classification

Classification is said to be supervised machine learning [48] because the labels are already given with the data in contradiction with unsupervised learning in which there are no predefined classes or labels inside the data. Each set of data that is used in supervised machine learning contains a set of features or attributes that may be continuous or categorical [49], [50].  Classification is said to be the process of creating the model with the help of training data set having labels. Then, the model can be used to predict the classes or label of testing data.  Classification in supervised machine learning is being used in several intelligence base researches.  A number of techniques are being used in classification some of the widely used techniques are given below:

##### Decision Tree

Decision tree classifies data set into trees by using algorithms of the data structure [51]. The main goal of the decision trees is to show the information of structure present inside the dataset. The decision tree technique is a type of supervised machine learning technique that creates a tree from a set of class labelled data with the help of the machine learning process [49].  The decision tree algorithm works with the training samples and their labelled classes. Then this training dataset is recursively divided, based on features, into a subset of data so that the data set in the subset is purer than the data set in the parent set. In the subset of data, each internal node present in e decision tree explains feature and every branch represent the outcome of the test and every node explains the class label [5].

**Advantages**

* Simple and fast.
* No requirement for prior knowledge and ability to manage high dimensional data.
* Its representation is understandable
* Sport incremental learning

**Disadvantages**

* It takes a long time to train the data.
* Require a large number of available memory when dealing with a large data set.
* Does not perform well while using the diagonal partitioning data set.
* More complex for some replication problem.
* Orders of the features intense are affected on the performance.

##### Naive Bayes classifier

Naive Bayes classifier is one of the simple statistical baysen classifiers [52].  It is called naive because it is supposed that all the variables are mutually correlated and participate in classification.  This is also called conditional independence [53]. This supposition is unrealistic for the maximum data set and it may lead to a simple framework of production which gives good results in manufacturing cases. Naive Bayes classifier is based on based theorem which is as follows:

(A|B) = 𝑃 𝑋|A (A) (B)

A - Hypothesis, (such that tuple B belongs to class X)

B - Evidence, explained by measure onset of attributes

P (A|B) - Posterior probability which hypothesis A holds the evidence B

P (A) - Prior probability of A, independent on B

(𝑋|A) - Posterior probability which B conditioned on A

**Advantages**

* Needs a small computational time frame for training data.
* Easy to construct.
* Model created from Navy Base is a type of product that can be z-transform into logarithms.
* Not require complicated recursively parameters estimation mechanism that can be applied on a large dataset.
* Easy representation of information.
* It may not show the best classifier results for a particular application but it is robust.

**Disadvantages**

* Theoretically, the error rate of Naive Bayes with other classifiers is minimum but in practice, it is not true always.
* Accuracy is not well as compared to the other classifiers.

##### K-Nearest Neighbours

K nearest neighbour is the object-based, without parametric learning method. It is also called lazy learners because it stores all the training samples. It does not allow to build a new classifier until a new unlabelled data sample requires to be classified.  However lazy learning algorithms demands less computational time in the training phase as compared to other machine learning algorithms like neural networks, Bayes networks and decision trees but take more memory for the classification process [54, 55,50].

This is the simplest algorithm among all machine learning algorithms. This is based on the rule that the sample data which are similar to one another will lies in near proximity [56].  When the unlabelled sample is given, KNN find that trend space for the K objects which are nearest to it and nominate the class by finding the very most frequent class label. When the value of k is 1 then nominate the class from the training sample which is the closest with the unknown sample inside the pattern space [3].

**Advantages**

* Easy to understand.
* Easy to implement.
* Give better for which application that has multiple class labels.

**Disadvantage**

* Consumes more computational cost when potential neighbours having a large labelled sample.
* Classification time am is low.
* It assigns the same wait squalor tribute but this maybe comes to the confusion where there may be many e irrelevant features in the data can and in this way the accuracy is affected.
* It is sensitive for local data structure
* Require more storage.

##### Support vector machine

SVM was used to integrate numbers in the last decade and applied in multiple domain applications. SVM is used for regression learning classification.  This is based on the statistical theory of learning and risk structural minimization principle and explore the decision boundary location which is also called as a hyperplane that creates the optimum classes’ separations [57, 49, and 58]. SVM finds the best hyperplane to classify the data into categories. To find the best hyperplane SVM removes outlier from data and can separate to categories with the best linear hyperplane. SVM can be used to solve multi-dimensional problems by using different kernels [4]. Kernel changes the dimension of data space according to the nature of data. Best hyperplane selection is shown in Figure 1.



Figure 1 SVM hyperplane representation

**Advantages**

* The most accurate and robust method among all the well-known algorithms of classification.
* SVM has a strengthened theoretical bass and requires just dozens of examples for training data.
* Search the best classification method to analyze between 2 Class bus from the training data set.
* It requires less planning overfitting as compared to other methods.

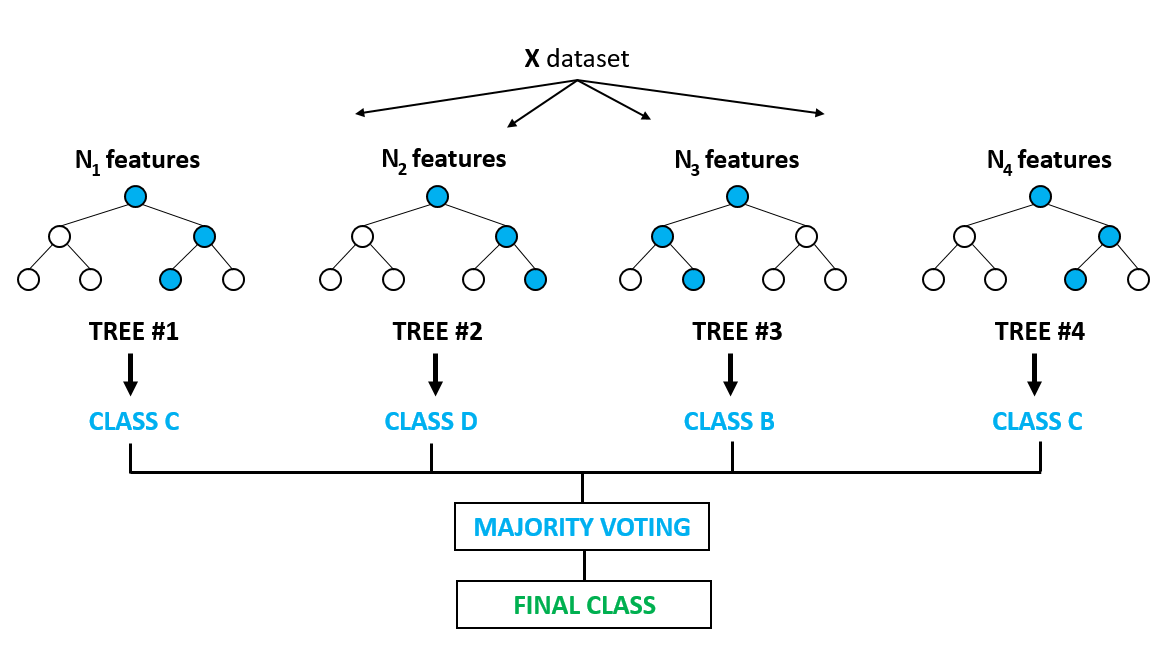
**Disadvantages:**

* Computationally expensive.
* Require a large amount of time for training data.
* And require a large amount of memory as well.

##### Random-Forest

Random forest is a type of classifier which is consisted on the collection of  tree-structured classifiers h(A, ON) and N=1,2,3, …  where  O(N) are independently identically separated random vectors and every tree determine the most used class at input A.  The best thing of this combination is each decision tree is made of from random vector of parameters [1].

Random forest develops a group of decision trees. The randomization to create different decision trees has proved apart equally efficient by using the method of random subspace or bagging as compared to 2 different approaches that produce a group of different classifiers.   The base classifier of random forest is a decision tree. Random forest is an ensemble model which combine the number of classifiers using the majority voting criteria in which multiple decision trees give their predictions and then final prediction make using the majority voting on decision trees predictions. This ensemble model can give good results as compared to an individual decision tree [1]. This ensemble random forest model can also perform well on imbalanced data because of the bootstraps sampling technique [7].



Random forest is an ensemble model that combine the number of decision trees in the prediction procedure as we mention above so we can define RF as:

Here, dts are the decision tree in a random forest, and n is the number of trees.

Here rf is the prediction by the random forest using the majority voting criteria. And N is the number of decision trees in the prediction procedure.

##### Logistic Regression

Logistic regression is a classification methodology that uses classes for creating and uses a single multinomial regression model for a single estimator. Logistic regression is usually used when the class boundaries are present and it also uses for probabilities of class depending upon the distance from boundaries.  The ratio of moving towards extreme from 0 and 1 when the data is large.  Logistic regression uses the logistic function which can be useful when the dependent variable contains binary value [2]. Logistic regression is an advancement in linear regression the difference between linear regression and logistic regression is shown in Figure 2 [6].



Figure 2 Linear Regression vs Logistic Regression

Linear regression can be defined as mathematically:

Here, y is the prediction value; bX is the slope and is the intercept. While logistic regression can be define using the linear regression function and it can be defined as using the mathematical equation as:

Here, p is the target value between 0 and 1. is the relationship between target values.

##### Deep Learning

Deep learning is a part of machine learning which can be referred to as a deep neural network [59].  Neural network is affected with the human brain and it holds many neurons which create a magnificent network. The deep learning networks can provide training to both unsupervised and supervised categories of machine learning [60]. Deep learning involves several networks such as RNN (Recurrent Neural Networks), CNN (Convolutional Neural Networks), DBN (Deep Belief Networks), Recursive Neural Networks, and many more. The neural networks are very helpful in vector representation, text generation, vector representation, sentence modelling, feature present word, sentence classification, and representation estimation.

Deep learning is very important in both supervised and unsupervised learning; several researchers are performing sentiment analysis with the help of deep learning. It contains numerous effectual and famous models and the concerned models are utilized to resolve the diverse problems successfully [61]. The most popular example Soccer has utilized the Recursive Neural Network (RNN) for the depiction of reviews of movies from the rottentomatoes.com website.

Deep learning is more effective when we have a large dataset for training. If we will increase the size of training data the performance of the deep learning model will be increasing while the machine learning model performance will not be increase after a certain limit of data as shown in Figure 3. Deep learning model uses the neural network in learning procedure and it didn’t need any feature to extract technique it can automatically find important feature from the data while machine learning models need handcraft features so that the reason deep learning approach have lots of benefit on machine learning models but it can be only useful when we have a large dataset for the training.

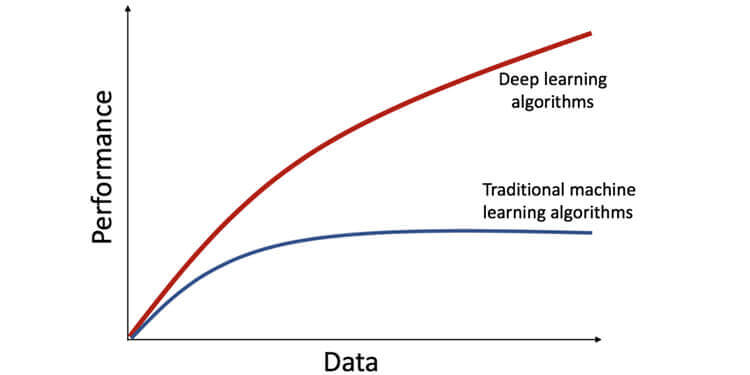


Figure 3 Deep learning vs machine learning

## Data Resampling

Data resampling is one of the important techniques in machine learning when a dataset is imbalanced. In the classification task, the imbalanced dataset is considered a major issue. The imbalanced dataset contains more records for one class known as majority class and other classes contain fewer data and known as the minority class.

In the machine learning model when the model get trained on the imbalanced dataset they get over fitted on the majority class data and show poor performance on the minority class data. To solve this problem there is multiple data resampling technique which can reduce the problem of the imbalanced dataset problem by generating the data for artificially for the minority class.

In this study, SMOTE over-sampling technique is used to make the dataset balanced. SMOTE stands for synthetic minority oversampling technique which increases the number of samples for minority class data according to majority class data [8].

## Evolution Measures

To determine the results of the above-mentioned classification techniques, following evaluation measures to determine their results.

### Accuracy

Accuracy is a factor for the assessment of the classification model. It is a metric for the evaluation of the classification model. It can define the correct prediction of the classifier, how many variables are classified correctly. Formally accuracy is defined as follows:

Or Accuracy can also be defined as:

Here,

TP is true positive which shows that when the model predicts the instance as True and the actual label is the instance was also True.

TN is true negative which shows that when the model predicts the instance as False and the actual label is the instance was also False.

FP is a false positive which shows that when the model predicts the instance as True and the actual label is the instance was also False.

FN is a false negative which shows that when the model predicts the instance as False and the actual label is the instance was also True.

### Precision

Precision is the ratio of correctly predicted positive variables by the total predicted positive variables. It can also be called a percentage of the relevant results. Precision is a factor that defines "how useful the results" of the classifier.

### Recall

A recall is the ratio of correctly predicted positive variables by the total variables of an actual class. It can also be called, the rate of true positive in the total number of positive samples. In binary classification, recall is the sensitivity of the classifier. Recall refers to the percentage of the total relevant results that were correctly classified by the algorithm.

### F-Measure

F-score is the weighted harmonic mean of precision and recall. It reaches the best value which means perfect precision and recall.

## Research Gap

Many researches have been performed on document-level sentiment analysis, sentence-level sentiment analysis, and aspect level sentiment analysis on mobile and smartwatch reviews. Previously, Aspect Level Sentiment Analysis is performed on aspects of battery and android version of mobile [6]. In research [68], researchers have achieved accuracy 94.46 from SVM and 87 from NB. Lots of researchers have done work in this domain but accuracy is still a gap for other researchers to work in this domain so we all contribute in this domain by achieving high accuracy for aspect-based sentiment analysis.

## Motivation

As the usage of internet devices increases from a business perspective, sentiment analysis allows the customer to suggest and reviews the product. Sentiment analysis is a basic perspective when people start e-commerce. With the advancement of the internet throughout the world, a large number of people engage in writing reviews and giving feedback. The reviews are written by users to help other buyers make informed decisions about the product(s) they want to purchase. These reviews are also beneficial for the manufacturers of products. They need to go through the reviews of buyers. For this, it is necessary to analyse the reviews providing by the user on mobile and smartwatches at the aspect level.

The e-commerce business is the future of the world and gaining lots of interest from the public. People buy products using an online platform and give their reviews on the product which impacts the company's worth. So companies try to find the sentiment of their customer on the products so they can make better policies in the future to increase the sale. So we propose an approach for these companies in which they can find the sentiment of people on their product aspects and can improve the product according to their customer's requirements.

## Problem Statement

Customer reviews have a greater impact on e-commerce business so companies collect their customer reviews using different platforms such as social media pages and their website. In this way, they collect lots of data which is very difficult for a human being to analyse such a huge amount of data so an automatic model is developed which can find the people's sentiments on the product and its aspects. The accuracy of the model is much important to analyse the reviews and predict the result accurately. So, it is required to increase the accuracy of the system.

## Research Questions:

1. How to analyse the consumer's sentiments?
2. How to develop a classification model to perform sentiment analysis at an aspect level with higher accuracy?

## Main Objectives

The main objective of the research:

1. To classify the eBay and Amazon reviews
2. To scrape data from eBay and amazon
3. To apply different pre-processing techniques for data cleaning.
4. To increase the accuracy of classification model.

## Main Contribution

The main contributions of the thesis can be summarized as the following:

1. Data resampling to resolve the model overfitting problem.
2. Increased accuracy.
3. Purpose the combination of Logistic Regression and Random Forest classifier with the help of Voting Classifier.

Chapter 2

# Literature Review

Sentiment Analysis has become the fastest-growing research area in the computer science which keeps track of all the activities in the specific area [pp1]. The specific reason behind this analysis is the analysis of the behaviour of buyers, and observers of any product or anything. With time, data is also increasing day by day on the internet and that’s why it’s very necessary to know the attitude of people for any product. For analysing observer or customer behaviour, sentimental analysis comes into play. Several kinds of research have already been done on this topic. Sentimental analysis has three levels of classification i.e. Aspect Level, Sentence Level and Document Level.

Sentiment Inquiry is said to be the research field in natural language processing (NLP), having process of detecting and extracting sentiment from the textual data and Categorizing their emotion. Sentiment analysis follows the people’s emotions, judgement, attitude and opinion to single, movies, issues, products, events, organization, etc.

In research [1] researchers mined sentiments from the reviews and analyse the outcome to create a business model. Researchers claimed that presented tools is too much robust and have provided high accuracy. They used the business analytics which made their results more suitable. They also performed analysis on detection of emotions from gender base review, and fake reviews. They performed their analysis on python and R programing language. They used Support Vector Machine and Multinomial Naïve Bayesian as their leading classifiers. They achieved 80% accuracy with SVM and 72.95 with MNB.

In research [2] the researchers used current supervised learning (Naive Bayes, Perceptron, and Multiclass SVM) algorithms to forecast reviews rating on a provided mathematical scale using only reviews. They collect the 1125458 reviews from Yelp dataset challenge. They used hold out cross validation by using 70% of data as training dataset and 30% of data as testing dataset. In this research the author used precision and recall evaluation matrices to analyse the results.

The researchers in [3] targeted to create a system which visualizes the textual reviews sentiments in form of charts. They scrapped the data from amazon through URL to collect the data and pre-processed it. In this paper they have applied NB, SVM and maximum entropy. The proposed the methodology which integrates the current sentiment analysis techniques and claimed to increase the accuracy. They presented their result in chart.

In the paper [4] authors built a model for predicting the product ratings based on rating text using a bag-of-words. These models tested utilized unigrams and bigrams. They used 2,982,356 reviews of 252,331 unique products of Amazon video game user 2,982,356 reviews from UCSD Time-based models didn’t work well as the variance in average rating between each year month, or day was relatively small. Between unigrams and bigrams, unigrams produced the most accurate result. And popular unigrams were extremely useful predictor for ratings for their larger variance. Unigram results had a 15.89% better performance than bigrams.

In paper [5] different feature selection or extraction methods for sentiment analysis are discussed. They gathered Amazon data at first and performed pre-processing to remove special characters and stop words. They performed multiword, phrase level and single word feature extraction and selection technique. They used Naive Bayes as the classifier. They determined that Naive Bayes performed better at phrase level as compared to multiword and single word. The main disadvantage of this research is, authors performed naive Bayes classifier only and we cannot extract a satisfactory result from which.

In this paper [6] they used easier algorithms to understand the problem easily. They collect data of Hotel Reviews. The system gives highest accuracy 93.50% on SVM having N-gram aspects as compared to other used algorithms logistic regression and decision trees method.

In paper [7], tf-idf is used here as an additional experiment. It can predict rating by using bag of words. They used root mean square error and linear regression model. They determined that Yelp reviews could be utilized for rating prediction with the help of bag of words model. They claimed that TF-IDF searches the most relevant terms from the reviews but unfortunately these terms do not participate in improving the prediction. However they used unigrams which shows positive effects on their results.

In [9], researchers described that which feature of product is good to expect the aspect-level sentiment analysis and explained the reason why this is in this way. In first iteration, all the textual data is categorized and it is considered an important aspect for the sentiment analysis. For this purpose natural language was used on Stanford's CoreNLP package. They collected the restaurant reviews for the analysis. In start all textual reviews data was pre-processed by using tokenization, lemmatization, removing parts of speech. The next stage is regarding SVM in which authors extract the features through SVM. After the acknowledgment of all features, obtain score can be added in for each feature. Thus, this is happened by the use of training data system.

Researchers analysed the features by the help of Information Gain, which is mostly used in conjunction for the measurement of feature selection. This approach is working in following steps: First of all, Information Gain is used for every feature, second IG scores related to all properties are arranged high to low and in SVM used k % features. The conclusion of the whole study is that when we used 1% best features for information gain, then the accuracy decreased with the value of 2.9% when we utilized complete features.

In this paper [10], the researchers used a Convolution Neural Network for the performing of sentiment level. For feature extraction, authors proposed a convolution neural network first. And then they applied another approach for the discovery of opinions this approach is known as order labelling methodology with Restricted Random Fields (RRF). Then at the final point, they collect characteristics with in the pushing of each word and finally, apply the convolution neural network and define the sentiment in the direction of aspect. They presented a deep learning-level approach to aspect-level sentiment analysis, which employs a convolutional neural network for aspect extraction and sentiment analysis and CRF for opinion target expression.

In the current study [11], they try to present the ABSA regarding movie reviews data. They collected the movie reviews. They proposed the methodology to extract the aspect and the sentiment related to the aspect by using the appended crafted rule. They defined the three patterns for the selection of required words:1: Manual labelling (M), 2:clustering(C), 3: review guided clustering (RC),

They take the 1000 movies' plots from the IMDB website. In this work they have tested the efficiency of the approaches on individual sentiments. Moreover the efficiency of cleaning the plot sentences have a good effect on the general aspect sentiment mining from the emotions as a whole but it needs to be fixed with effective approaches for combining opinions through different sentiment.

In [12] the researcher performed a sentiment analysis on music. In this paper, dataset of one thousand songs was collected from the web for analysis. They collected around 20,000 emotions from one thousand songs for sentiment analysis towards music. Linear Regression was used to calculate the polarity of sentiments in this research. Authors used 70 % of total data set to train the model and 30% was used for testing purpose. Researchers performed analyse the songs on the appended subject like, Baseline, Shape, and Contrast of the one thousand song. The analysis will help the readers to analyse the consequence of expressive and observation of actuality while listing music.

In [13] researchers performed the analysis and expand the Jazz music dataset. They used the dataset of having 21651 songs of jazz music. The dataset was collected by the support of techniques of information retrieval. The data about songs and the audio features of the songs were also present in the dataset. Researchers worked dedicatedly on searching the features of the song. It will help to search the related songs to add them in cluster and to analyse the song either it was the type of jazz music or it is not the jazz music. It will also clarify the purposed of the used dataset. One of the unsupervised Machine Learning technique K-mean was utilized to create the cluster of the songs of jazz music and other music having different audio feature of songs. Later, researchers have a planned to expand their research with non-supervised algorithm.

In [14] researchers have worked on integration of data on dissimilar datasets. Dataset used in this research was consisted on national products of the China and United States of many years. The dataset was extracted from different central repositories. The dataset was consisted on raw data and information as well. Dataset consisted on the rates of currency exchange Yuan to Dollar and vice versa. In first iteration currencies were exchanged with the help of math rule and produced values. Then, they divided the information by eliminating other values or data formats and tags. Researchers gathered data from multiple repositories so that integration of data could be performed. After the integration of data, validation of data was performed to confirm the quality of data.

In [15] researchers performed research on data about dataset. The dataset contained the actual standards for short term learning. The dataset consisted on two standards, first standard consisted on 1623 characters and the second standard consisted on 600 sample. Classification, multiple classification algorithms were used like K-NN and Fine-Tune. One of them K-NN is used to classify each considered item to closest actual class. Moreover Fine-Tune cannot do so. Researchers obtained grearter accuracy with the help of K-NN of 88.42 as compare to Fine-Tune which was 73.88 %.

In [16] researchers performed different modal based classification on Hindi and Western songs. They used two types of dataset for analysis. The first dataset was consisted on the audio song clips and the second dataset was consisted on sentiments with eight different classes. The first dataset of AMC was consisted on audio song clips while the second dataset was consisted on sentiment with 8 different sentiment classes. The dataset of Hindi songs was consisted on 500 songs and the second dataset consisted on 1753 audio clips. The dataset of Western songs was consisted on 298 songs while 1111 clips in mp3 format. Authors utilized a supervised Machine Learning algorithm for classification purpose like SVM (Support Vector Machine) and FFNNS (Feed-forward neural networks). To determine the accuracy of classification, Ten-Fole validation was utilized. They obtained precision 58.9 % and 59.1% using SVM while on Western music they obtained 70.5 % precision. With FFNNs algorithm, the precision value was 65.3%, 65.2% of F-measure and 65.1 of Recall on Hindi music.

In [17] researchers performed Sentiment Analysts on Roman Urdu with the help of different machine learning algorithm methods on mobile reviews. The dataset was gathered from sites and blogs. For classification purpose, Rapid Miner tool was used.

A study [18] was performed on the Dimensionality Reduction on Bag of “Pop Corns, Bag of Words” a set of data which was collected from on Kagygle. For this purpose supervised machine learning approach was used to analysys dataset of 25000.

In [19] Sentiment Analysis performed on Aspect level. Mobile reviews dataset from Amazon Web Reviews used for analysis. After data pre-processing, select features and determine the rating of selected features based on Sentiment Analysis. Data mining association rule was used for the segmentation of the sentences using NL Processing. After opinion orientation of the words simply count the total numbers of positive and negative comment for each feature and finally rate the aspect top to down which has maximum numbers of positive reviews, and at last which aspect that gets minimum positive reviews and maximum the negative.

In [20] Aspect level Sentiment Analysis performed on movie reviews. To classify the sentiments at aspect level and at document level by exploring a sentiment-based scheme. Classification of document level has some linguistic feature, it can range from adverb + adjective to adverb + adjective + adverb combination. While in this paper they advised a domain specific-heuristic approach for aspect level classification. Dataset for analysis was collected from different sites. To classify these reviews in positive, negative and neutral by using the publically available library. To indicating the Aspect, 5- Gram technique was used. Sentiment Analysis performed by using feature-based heuristic scheme at Aspect level.

In [21] survey was carried on Aspect level Sentiment Analysis to aggregate the people opinion on the entities that were mentioned within the document. At Aspect level Sentiment Analysis, a single entity was analysed at a single time. Aspect level Sentiment Analysis, a review would generally refer to the entity from the document, so the aspect detection was the major/important part in Aspect Level Sentiment Analysis. Here they can discuss different ways of aspect detection like frequency-based and syntax-based aspect detection focused on frequencies of the aspect and in syntax-based methods find the means of the syntactical relaxation in which aspect are present. After that in this survey, they talked about the classification methods supervised and unsupervised. The supervised classification used for the labelled data for training and testing both. Unsupervised classification required to operate labelled data only for training the algorithm at testing it can classify the unlabelled data. State of the art Aspect Level Sentiment Analysis proposed in this survey.

In [22] solution was purposed of forms and function for Roman Urdu dataset, that was collected by making a survey of the local universities of Pakistan. In survey total 116 participants (58 male and 52 female). The average age of both male and female was 21.01 years in which 103 (88.8%) members from undergraduates 10 (8.6%) were from graduates while 3 (2.6%) from them were PhD scholars. The messages they collected in the form of text messages, the corpus has total of 4, 46,483 words. In their survey, they see the people prefer to write their messages using Roman Urddu type of writing. Accordingly to their study, they analyse the female data was less romantic words than males. And the students of undergraduates has used more intimate words than graduate student. 73 users, used 20 or less intimate words those called low romantic participants. 31 such users that used 21 to 80 intimate words they were classified as medium Romantic Peoples. The 41 users were those who used 81 or more intimate words they were classified as high romantic participants, The research was carried to understand the population way of messaging and classify the users were they adopted low, medium or high Romantic way to communicate with others.

In [23] emotion ontology was generated for Roma Urdu text data. Dataset was collected from different blogs by using a scraper that contains people emotions. They classified the emotions into 5 classes (Happiness, angry, hurt, caring, and fear). After data collection, they parse the data through syntax analyser that recognize the syntax structure and contract a phrase tree and modify the Figures into the required order. After semantic analyser through JENA API and checking the ontology of the document classified according to their classes in happiness, angry, hurt, caring, and in fear. They performed an experiment on four documents named (DOCI, DOC2, DOC3 and DOC4). In document 1, 30 sample data was taken ‘n which from 27 were classified correctly and their results precession 93.10% and with recall 90%. In doc2, carried 33 sample data and their algorithm correctly classified them in number 28. Results with recall a DOC2 was 84.84% and with recession 93.33%. The document has 54 samples and correctly classified 46 results with recalled was 85.18% and with 93.86% with precession and document carried 38 samples ‘n which from 31 correctly classified and recall gives 81.57% and precessions give 91.17% results. Their main goal to design an algorithm to identify the emotions from Roman Urdu text and their algorithm gave better results in the form of precessions and recall.

In [24] research was carried to find the hidden pattern in raw text and different pre- processing techniques in text mining were discussed. They extracted useful text data from the raw data with the help of a scraper. Unstructured text data may be extracted from files, spreadsheets or from rational database which contained noisy data as well as HTML tags or Stops Words. Remove these outliers and noise present in data with the help of pre- processing techniques. They discussed a few of them pre-existing pre-processing techniques. Firstly they talked about data then stemming a process in which, identify the stem/root of the word. Then they discussed N-gram techniques in which N-gram a string, where character extracted from continues text. They also talked about TF & IDF. Term frequency a word present in a document and inverse document frequency talked about a word which was repeated in multiple documents. Through the paper; they tried to help the people in the field of text mining.

In [26] Sentiment Analysis was Performed on multi-languages data Twitter dataset. To perform sentiment analysis, they selected multi-language tweets associated with PGE 2013. The users in dataset was form the five capitals of Pakistan and tweets were belonged to legendary political parties of Pakistan. The dataset was gathered with scrapper from 2001 to 2013. According to the results of urdu and english tweets,

In [25] Sentiment Analysis was performed at Aspect level by using different Machine Learning Techniques and purposed a system for the Aspect level Detection of reviews. Different pre-processing techniques were applied on extracted data i.e. (Tokenization, Part of Speech tagging and Lemmatization) to prepare the data and to remove the outliers as well. Spam detection model was used to avoid the model the spam or noisy data. After pre-processing and classify the reviews (Positive, Negative, and Neutral) Machine Learning algorithm i.e. (Support Vector Machine and Naive Bayes) were applied to check the rating of the product. According to their reviews products were classified into three categories (Low, Medium, & High). A product which has one or two star ratings classified as the low product, which product rating was three classified as medium and product having 4 and 5 ratings classified as a high product. Their proposed system classified the comment into two classes as Negative and Positive very fast and correctly.

In [27] carried a survey on Supervised and Unsupervised classification of document. There were three basic methods to classify the document. Rule-based supervised and unsupervised document classification. In the rule-based classification of the documents according to the defined rules, actually these rules were query phrase. In supervised classification, the document was classified on the basis of supervised learning. Actually first trained the algorithm using training data which was known as algorithm learning. After learning the algorithm classified the testing data. The unsupervised classification was a method in which classification was achieved through clustering. It can simply cluster the data according to the structure or pattern of the data. It can simply cluster the data according to the structure or pattern of the data. Unsupervised algorithms worked on centred based approach, in centred based techniques each document "D" represented Documents Vector VD and centred vector of each class and Euclidean distance between VD and centred Vector of Class was calculated. Documents having a minimum distance from centred assign a particular class.

Some of the Supervised and Unsupervised algorithms were discussed. SVM analyse the data and classify them after recognizing the pattern. Naive Bayes classify the document by the calculation of posterior probability value and classify the documents into classes according to their frequencies. Decision Tree uses a tree-based algorithm to classify the documents. In unsupervised classification, they discussed partitioned clustering. In partition clustering used a un-nested partition, first K clusters are defined and partition (P) was constructed, then redefined clustering solution was by moving the documents from one cluster to other iteratively. In K-mean clustering, K clusters are defined and each document move to that cluster which was near to its centred, Hierarchical clustering techniques make a cluster from top to down oF bottom to up and documents are divide into the cluster. Research in the field of Sentiment Analysis was conducted at different times In different domains.

The dataset [28] was consisted on Amazon and it is organised in Jason format. All the Jason files are consist of many reviews. The dataset contains on review of different sources such as TV, Mobile Phone, Camera, Laptops, tablets etc. In pre-processing term researcher passes through the reduction process of stemming, punctuation, stop word, repeated word etc. and then it has converted in bag of words. Pre-processing is a significant process in the field of views mining and sentiment analysis. Each sentence was analysed and calculated by sentiment score. For calculating the sentiment score there has been a comparison done among dataset and lexicon opinion. The dataset was compared with sentiment lexicons having 2006 positive and 4783 negative review words and sentiment scores were calculated. Authors used various types of features, learning algorithms, several accuracy measurements calculated. In this work, researchers applied different approaches as lexicon approach, a dictionary-based approach that was usable within learning techniques. This famous analysis (sentiment analysis) has applied on each review of products and then used the machine learning algorithms like SVM as well as NB. Figure 2 presents the measurements and dimensions related to NB and SVM. Naïve Bayes grew 98.17% accurateness of Camera reviews, while on the other hand Support Vector Machine got 93.

In this study [29], the researcher used a Gini Index for feature selection with one of the classifier SVM was used for the classification of sentiments on large movie review. The results of this study presented that this Gini Index method is helpful for the better results for error reduction and accuracy. They proposed a method for the reviews extraction and categorization. The perposed methodology was consisted on five steps: Data Source: they shared their views and experiments on underneath corpora. Data Preprocessing: it is contained on transformation, tokenization, removal of stop words and mining of opinion words. Feature Selection: the system of selecting feature is known as feature selection and it is called the attribute selection, and it is mostly used and helpful for the model construction. Representation: the consequence of attributes on the Gini Index is planned and weightiness are allocated in order ways. Sentiment Classification: it is the type of classification and its main focus is on the opinions, it is contained on input as attribute and selection of top attributes on the weight-based and the term frequency is removed from the last data by the use of SVM.

They collected the dataset from the reviews present on the sites like Bollywood movies like Rottentomatoes, Hungama, Times of India, Rediff and Mouthshut. For preprocessing, tokenization, normalization, stemming was performed on extracted dataset. The results explained that the proposed methodology showed the greater accuracy than the SVM with 0.65 split ratio.

In this paper [30], the dataset was gatherd by Kaggle that was consisted on food reviews collected from Amazon from October 1999 to October 2012 (29,30). Dataset was consist of a huge number of data like as reviews are contain on 568,454, users consisted of 256,059, products was 74,258, and 260 users that had more than 50 evaluations. In this pre-processing stages, scholars followed that steps: took out the URLs as ([www.abc.com](http://www.abc.com/)), all tags like (#topic), remove all screen name such as (@username), took away all the punctuation marks, symbols as well as numbers, remove all stop words, replacement of emotions within sentiments, the transformation of text to lowercase, exchange the words with roots, reduction of repeated words and retweets. This type of analysis provides help in judging the customers' sentiment. In the same way, in a product-centric approach, the researcher gets success towards the best-reviewed product that was done by several customers. This research explained the consumers' views and emotions towards the products. And the results of this research was additional evidence about the significance of customers' reviews for digital as well as online marketing research. Researchers' most of the work contained on the investigation of customers' opinions and reviews that they took from various E-commerce sites.

In this paper [31], the amazon dataset of ranging from august 2018 to December 2018 was used in this study. Despite reduced objectives from this research, all the subjective contents were detached for the upcoming examinations of sentiment sentences. A sentiment sentence is must contain one negative and one positive word. All the sentences first of all organized into English words. Each word exists on its semantic role that elaborates on the meaning of words and how a word is used. The semantic role is also known as parts of speech. The English language commonly consists of 8 famous parts of speech. The name of these part of speech are the following: 'Noun, pronoun, verb, adverb, adjective, conjunction, interjection and preposition'. In natural language, part-of-speech (POS) producer organize the words based on POS. Some researchers collect moreover 500 sentiment of reviews on products that is related to 4 major forms: such as Flash drives, Computers, Mobiles and Electronics. These were online reviews and set by almost 3.2 million people in front of 10,001 products. Every review gives the following background knowledge of the reviewer: 1) Reviewer ID, 2) The exact date and time of review, 3) Mention the model of Product, 4) The text that customer write as review.

Chapter 3

# Methodology

## Dataset description

This study performs experiments on user reviews using a supervised machine learning approach. For this, we extract the eBay reviews using the Crawler. The dataset sample is shown in Table 1. The dataset contains the product name, review date, review author, review title, reviews text, helpful vote, unhelpfulvote, and review rating as all attribute descriptions are shown in Table. We used only the review text attribute for the experiments and then find the sentiment using the lexicon technique from text reviews. We also collect data from Amazon site related to user sentiment on electronic devices . <https://jmcauley.ucsd.edu/data/amazon/>. We merge this dataset with the eBay data to increase the size of dataset. The Amazon dataset also contain the the text reviews of the users on different amazon products from electronics category. We extract the 252231 reviews from Amazon and 33324 reviews from eBay.

Table 1 Dataset Sample

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Product Name | Review Date | Review Author | Review Title | Review Text | HelpfulVote | UnhelpfulVote | Review Rating |
| Apple iPhone 7 - 128GB - Rose Gold (Unlocked) A1778 (GSM) | 4-Sep 18 | 2006yushan | Volume control can be better | One thing that can definitely be better is the … | 0 | 2 | 5 |
| Apple iPhone 7 - 128GB - Rose Gold (Unlocked) A1778 (GSM) | 9-Oct-18 | tcpeterson62758 | Nice Phone - Not Cleared | The pro is the phone looks brand new…. | 11 | 1 | 3 |
| Apple iPhone 7 - 128GB - Rose Gold (Unlocked) A1778 (GSM) | 7-Apr-20 | radiodoc0 | Great value. Good as new | Like brand new, installed sim card from old phone … | 0 | 0 | 5 |
| Apple iPhone 7 - 128GB - Rose Gold (Unlocked) A1778 (GSM) | 5-Jul-19 | chealber30 | Great features, good quality for refurbished phone | Phone was exactly as described… | 1 | 1 | 5 |

Table 2 Dataset attribute description

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Product Name | This attribute contains the product name. |
| Review Date | This attribute contains the date for the review when the user posts it. |
| Review Author | This attribute contains the author name who posted the review. |
| Review Title | This attribute contains the short title for reviews. |
| Review Text | This attribute contains the text for reviews. |
| HelpfulVote | This attribute contains the vote for the product which is helpful for the users. |
| UnhelpfulVote | This attribute contains the vote for the product. |
| Review Rating | This attribute contains the rating for the product post by the reviewer. |

## Pre-processing of data

Pre-processing of the obtained data is one of the significant tasks that need to be performed. Through data pre-processing we generally manage to transform the unstructured data into an organized and structural format [1]. The basic purpose of pre-processing is to enhance the quality of the input data by reducing its quantity so that the machine can understand the patterns from data that further helps to extract more useful and relevant features from the pre-processed data. It aids the machine to learn more accurate patterns from data which also improves the performance of the machine learning classifier in terms of accuracy. Input data should be delivered in the required format, amount, and structure that is appropriate to the required task. Unfavourably, real-world data is vastly inclined by the inappropriate factors, the performance of the analysis being performed depends on the quality of data ultimately the low-quality data will provide low quality performance [2]. Several existing data pre-processing techniques are being used to attain structured data from unorganized data. These techniques are used to remove the least useful and unnecessary data that devours machine process time and power.

As it helps to attain only the data that is useful and important to make further analysis, pre-processed data acts a vital role in the decision-making of the machine learning model [3]. Usually, initially obtained text data contains the combination of lower-case and upper-case letters, numbers, stopwords, punctuations, and various forms of words that has no importance in the classification and rather takes a lot of processing time which further leads to misclassification of data. Such types of data have no meaning to the text and have no role in the decision-making process of the machine. As it is conclusive that the initial format of obtained data is inappropriate to the classifier, so that removing such patterns from data will not cause any harm and no important information will be lost, it will become more valuable for classification tasks instead.

It is the initial processing of data to prepare it for further major processing or analysis. There are several steps required to prepare data for further processing. Some of the core applied steps are discussed below:

### Tokenization

The input data is unstructured thus tokenization is a way to convert the raw data into small tokens or chunks by splitting the long strings of text. Larger strings of data can be tokenized into sentences and sentences can be tokenized into words. It is a dissection process that exclusively results into words instead of sentences or paragraphs. In tokenization of the textual data, a sentence is transformed into an array of words or terms [4].

Table 3 Data before and after tokenization

|  |  |
| --- | --- |
| Before Tokenization | After Tokenization |
| I would like to introduce Mr. David as new Sales Manager, he’ll start his job from Oct 01, 2021. | ‘I’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘Mr.’, ‘David’, ‘as’, ‘new’, ‘Sales’, ‘Manager’, ‘,’, ‘he’ll’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’, ‘01’ ‘,’, ‘2021’. |
| I would love to help you with your homework tomorrow be ready at 6:00. | ‘I’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’, ‘6:00’. |
| He is going to be the Employee of The Month again, who else could get a chance as long as he is THERE. | ‘He’, ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘Employee’, ‘of’, ‘The’, ‘Month’, ‘again’, ‘,’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘THERE’. |

### Punctuation Removal

Punctuation removal is another step of data preprocessing that aims to remove the punctuations including “[!’;’.&? #\*^%/(){}\|-@\_,]” from the data. Punctuations are removed from data because they do not have any impact on the data as they are meaningless to the machine. It also reduces the capability of a machine to discriminate between other characters and punctuation [5].

Table 4 Data before and after punctuation removal

|  |  |
| --- | --- |
| Before Punctuation Removal | After Punctuation Removal |
| ‘I’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘Mr.’, ‘David’, ‘as’, ‘new’, ‘Sales’, ‘Manager’, ‘,’, ‘he’ll’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’, ‘01’ ‘,’, ‘2021’. | ‘I’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘Mr’, ‘David’ ‘as’, ‘new’, ‘Sales’, ‘Manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’, ‘01’, ‘2021’ |
| ‘I’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’, ‘6:00’. | ‘I’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’, ‘600’, |
| ‘He’, ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘Employee’, ‘of’, ‘The’, ‘Month’, ‘again’, ‘,’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘THERE’. | ‘He’, ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘Employee’, ‘of’, ‘The’, ‘Month’, ‘again’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘THERE’ |

### Numeric Removal

In this step, numeric values are removed from the data to improve its quality. Since in the text data where the numeric values like digits are not of any use in the decision-making process, which rather trouble the machine in the feature extraction process. Usually, the values containing numbers do not contribute to the classification of data [6]. When working with reviews or textual data that is not concerned with the digits, then it is necessary to preprocess data to remove the numeric values. The same applies to the null values since the null values do not add to the performance of the model.

Table 5 Data before and after numeric removal

|  |  |
| --- | --- |
| Before Numeric Removal | After Numeric Removal |
| ‘I’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘Mr’, ‘David’ ‘as’, ‘new’, ‘Sales’, ‘Manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’, ‘01’, ‘2021’ | ‘I’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘Mr’, ‘David’, ‘as’, ‘new’, ‘Sales’, ‘Manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’ |
| ‘I’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’, ‘600’, | ‘I’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’ |
| ‘He’, ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘Employee’, ‘of’, ‘The’, ‘Month’, ‘again’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘THERE’ | ‘He’, ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘Employee’, ‘of’, ‘The’, ‘Month’, ‘again’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘THERE’ |

### Lowercase Conversion

In another step, all the letters are converted to lowercase. Machine learning models are case sensitive so that this conversion has major importance [7]. For example, if the conversion is not applied to data the model will count the existence of “Sales” and “sales” as two different words. Sometimes in an informal or formal record, a blend of upper-case and lower-case letters is used to give significant attention to the specific words.

Table 6 Data before and after converting to lower case

|  |  |
| --- | --- |
| Before Lowercase Conversion | After Lowercase Conversion |
| ‘I’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘Mr’, ‘David’, ‘as’, ‘new’, ‘Sales’, ‘Manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’ | ‘i’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘mr’, ‘david’, ‘as’, ‘new’, ‘sales’, ‘manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’ |
| ‘I’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’ | ‘i’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’ |
| ‘He’, ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘Employee’, ‘of’, ‘The’, ‘Month’, ‘again’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘THERE’ | ‘he’ ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘employee’, ‘of’, ‘the’, ‘month’, ‘again’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘there’ |

### Stemming

Stemming is performed on the input data to convert all the words or terms used in the data into their first form. In other words, it is a preprocessing technique that has been used to transform words into their root form to improve machine learning model performance [8]. For instance, the word “records”, “recording”, “recorded” are different forms of the same word that might confuse the classification model. Therefore, after stemming techniques these forms of words will be converted to their root form “record”.

Table 7 Data before and after stemming

|  |  |
| --- | --- |
| Before Stemming | After Stemming |
| ‘i’, ‘would’, ‘like’, ‘to’, ‘introduce’, ‘mr’, ‘david’, ‘as’, ‘new’, ‘sales’, ‘manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’ | ‘i’, ‘will’, ‘like’, ‘to’, ‘introduce’, ‘mr’, ‘david’, ‘as’, ‘new’, ‘sale’, ‘manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’ |
| ‘i’, ‘would’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’ | ‘i’, ‘will’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’ |
| ‘he’ ‘is’, ‘going’, ‘to’, ‘be’, ‘the’, ‘employee’, ‘of’, ‘the’, ‘month’, ‘again’, ‘who’, ‘else’, ‘could’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘there’ | ‘he’, ‘is’, ‘go’, ‘to’, ‘be’, ‘the’, ‘employee’, ‘of’, ‘the’, ‘month’, ‘again’, ‘who’, ‘else’, ‘can’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘there’ |

### Stop words Removal

One of the major tasks in preprocessing is to remove the data that has no use in the classification. Stop words are the words that are worthless for the model to make the decision. In this step of pre-processing the stop words are removed from the dataset. It is the most vital task in preprocessing that removes the useless data for further processing of the data. Stopwords are the words used to form a sentence that has no use in text classification and are meaningless to the machine learning models [9]. Stopwords include words like “is, am, i, the, to, are, that, they, etc.”

Table 8 Data before and after removing stopwords

|  |  |
| --- | --- |
| Before Stopwords Removal | After Stopwords Removal |
| ‘i’, ‘will’, ‘like’, ‘to’, ‘introduce’, ‘mr’, ‘david’, ‘as’, ‘new’, ‘sale’, ‘manager’, ‘hell’, ‘start’, ‘his’, ‘job’, ‘from’, ‘Oct’ | ‘like’, ‘introduce’, ‘david’, ‘new’, ‘sale’, ‘manager’, ‘hell’, ‘start’, ‘job’, ‘Oct’ |
| ‘i’, ‘will’, ‘love’, ‘to’, ‘help’, ‘you’, ‘with’, ‘your’, ‘homework’, ‘tomorrow’, ‘be’, ‘ready’, ‘at’ | ‘love’, ‘help’, ‘homework’, ‘tomorrow’, ‘ready’ |
| ‘he’, ‘is’, ‘go’, ‘to’, ‘be’, ‘the’, ‘employee’, ‘of’, ‘the’, ‘month’, ‘again’, ‘who’, ‘else’, ‘can’, ‘get’, ‘a’, ‘chance’, ‘as’, ‘long’, ‘as’, ‘he’, ‘is’, ‘there’ | ‘go’, ‘employee’, ‘month’, ‘again’, ‘who’, ‘else’, ‘get’, ‘chance’, ‘long’ |

Table 9 Data before and after preprocessing

|  |  |
| --- | --- |
| Before Preprocessing | After Preprocessing |
| I would like to introduce Mr. David as new Sales Manager, he’ll start his job from Oct 01, 2021. | like introduce david new sale manager hell start job Oct |
| I would love to help you with your homework tomorrow be ready at 6:00. | Love help homework’ tomorrow ready’ |
| He is going to be the Employee of The Month again, who else could get a chance as long as he is THERE. | go employee month again who else get chance long |

## Features Extraction

This study used TF-IDF for feature extraction.TF-IDF is an abbreviation Term Frequency (TF) and Inverse Document Frequency (IDF). TF-IDF is a counting measure technique which is typically reproduced in information retrieval (IR) and clarification. It is assumed that TF-IDF will show how a term is represented in the analysis. TF and IDF are used in feature extraction techniques [10]. This is a very common algorithm for converting text into a meaningful number representation that is used to suit the prediction machine algorithm. The words are calculated to be more important with higher frequency ratings. TF-IDF is different from the BoW method because the BoW is the basic word count in a script, but the TF-IDF finds weighted text data features such that machine learning models can train themselves to increase their accuracy on important aspects [11]. The frequency of words shows us how often a word is used in a script. The huge volume document may contain many term frequency for a word so there is a better possibility that a term will be available for more time in the large documents [12].

Term frequency can be calculated as:

Here the N represents the number of time a term occurs in a document whereas the T represent the total terms in a document.

While the frequency of the inverse document (IDF) shows us how continually a word appears in a corpus document. If a more frequent word in a document having a low score of inverse document frequency. Then the stopwords in the dataset also have low IDF that indicates the feature's low value [13].

Inverse document frequency can be calculated as

:

Here Dt is the number of documents where term t appears when the term frequency function satisfies TF ~~+~~ 0 then 1 will be added into the formula to avoid zero-division.

So the complete TF-IDF can be defined as:

**Example:**

Consider a document containing 100 words wherein the word cat appears 3 times. The term frequency (i.e., TF) for cat is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousands of these. Then, the inverse document frequency (i.e., IDF) is calculated as log (10,000,000 / 1,000) = 4. Thus, the TF-IDF weight is the product of these quantities: 0.03 \* 4 = 0.12.

Let’s have another example, there are two document:

|  |  |
| --- | --- |
| **Doc 1** | The sky is blue. |
| **Doc 2** | The sun is bright. |

The frequency of the terms are shown in below table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **blue** | **bright** | **sky** | **sun** |
| **Doc 1** | 1 | 0 | 1 | 0 |
| **Doc 2** | 0 | 1 | 0 | 1 |

So the TF is the division between the number of time a term occurs in a document and total number of terms in a documents whereas the IDF is a log representation between the division of total number document and the number of term represented in these. So TF-IDF is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **blue** | **bright** | **sky** | **sun** |
| **Doc 1** | 0.707107 | 0.000000 | 0.707107 | 0.000000 |
| **Doc 2** | 0.000000 | 0.707107 | 0.000000 | 0.707107 |

## Hyper parameters for classifiers

The classification of the sentiment using the text is done in this study using the supervised machine learning approach. All used classifiers are discussed in the Introduction chapter, we train all these models with different hyperparameters setting and we find these hyper parameters with the hit and trials method. These classification model achieved their best results with these hyper parameters setting.

Table 10 Hyperparamters for classifier

|  |  |
| --- | --- |
| Classification Model | Hyperparameters |
| Random Forest | n\_estimators=300  max\_depth= 300 |
| Decision Tree | Max\_depth=300 |
| Logistic Regression | solver=’liblinear’  multi\_class=’multinomial’  C=3.0 |
| Support Vector Machine | Kernel=’linear’  C=3.0 |
| K Nearest Neighbour | **n\_neighbors=3**  **leaf\_size=30** |
| Gaussian Naïve Bayes | **Default Setting** |

## Experimental Flow

In this study, experiments are performed on a Corei7 7th generation machine with 8 GB RAM and 500 GB ROM with a windows 10 operating system. For experiments, python language and Jupyter Notebook we used with sci-kit learn library. The experiment flow diagram is shown in Figure 1.

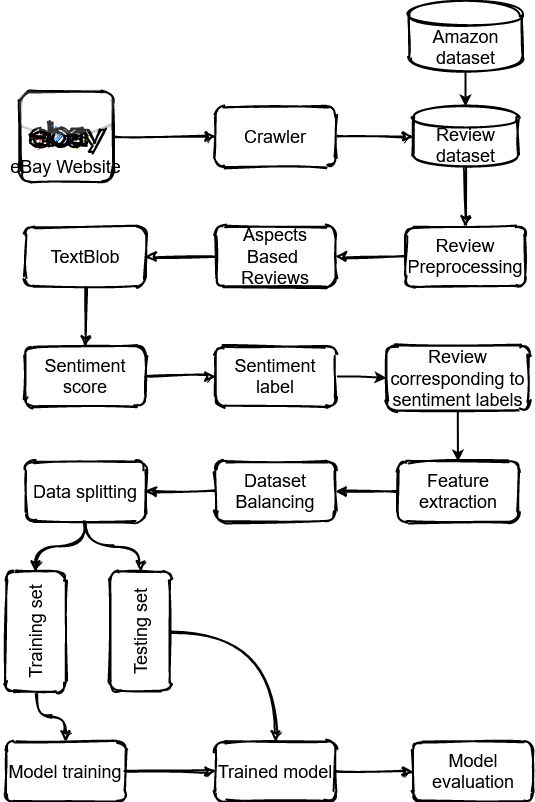


Figure 1 Experimental Flow Diagram

In the experiment approach first, we collect the data using a crawler from the eBay site. In data collection, we collect data contain the reviews of eBay users. These reviews are on different services provided by eBay as shown in Table 1. We also extract the reviews from Amazon related to same electronics products and merge it with the eBay data to increase the size of data. After that we separate the reviews on the base of aspects such as Colour, Size, Weight, Service, and Price.

Table 11 Aspects based reviews count

|  |  |
| --- | --- |
| Aspect | Count |
| Size | 6969 |
| Price | 32139 |
| Service | 6901 |
| Weight | 2068 |
| Color | 3074 |
| Total |  |

For the classification purpose, we extract the target label for reviews as positive, negative, and neutral. For that, we used textblob library. Before passing the data to textblob we have done preprocessing of text reviews to clean text reviews. Preprocessing removes all raw data such as punctuation, numbers, stopwords, and then we pass this clean data to textblob to extract the sentiment from the data as shown in Table 9. The dataset sample after finding the sentiment is shown in Table 11 and the ratio of sentiment is shown in Table 12.

Table 12 Sample of data after labeling

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Sentences | Polarity Score | Sentiment |
| 1 | one thing definitely better volume control nev... | 0.257143 | Positive |
| 2 | pro phone looks brand new con phone gift daugh... | 0.087273 | Positive |
| 3 | like brand new installed sim card old phone pr... | 0.245455 | Positive |

Table 13 Sentiment count in the dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Aspect | Count | Positive | Negative | Neutral |
| Size | 6969 | 5262 | 1358 | 345 |
| Price | 32139 | 24105 | 6592 | 1417 |
| Service | 6901 | 5023 | 1515 | 360 |
| Weight | 2068 | 1524 | 467 | 75 |
| Color | 3074 | 2239 | 680 | 153 |

The feature extraction technique was used to extract the feature after extracting the sentiment and for that, we used the TF-IDF features extract technique which gives the weighted features which are more suitable for the learning of the model (see Section Feature Extraction). The used dataset is imbalanced because the ratio of the data for each sentiment is not equal which can be cause for the overfitting of the model for majority class data. To solve this problem, SMOTE technique is used to make the dataset balanced. The data count after balancing the dataset is shown in Table 5. SMOTE technique generates mock data to create a balance between the target class ratio. An argument is passed to the SMOTE algorithm to set a threshold value for mock data to balance minority and majority classes. In this technique, SMOTE chooses comparative records and alter those records one column at a time by a random value within the difference to the adjacent records. We get a 1:1 ratio of each class negative, positive and negative examples by using SMOTE techniques in this experiment as shown 13.

Table 13 Sentiment count in the dataset after oversampling

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Aspect | Total Count | Positive | Negative | Neutral |
| Size | 15786 | 5262 | 5262 | 5262 |
| Price | 72315 | 24105 | 24105 | 24105 |
| Service | 15069 | 5023 | 5023 | 5023 |
| Weight | 4572 | 1524 | 1524 | 1524 |
| Color | 6717 | 2239 | 2239 | 2239 |

After that, we split the dataset into training and testing sets with the ratio of 80 and 20. The 80 percent of data used for the training of machine learning models and 20 percent of data used for the testing of models. Random forest, logistic regression, support vector machine, k nearest neighbor, and decision tree are gets fitted on the training set and after that evaluation is done using the test data. For the evaluation, accuracy, precision, recall, and F1 score are used and on an imbalanced dataset F1 score can be preferred.

Chapter 4

# Result and Discussion

## LR Results

This section contains the results for the LR with the under sampling and oversampling technique. Performance of LR improves the oversampling technique because oversampling generate more features for the learning of LR while the performance of LR on imbalanced dataset show performance because the ratio of target classes is not equal and LR get overfitting on majority class data and show poor performance on minority class data. Under-sampling reduces the performance of LR because under-sampling randomly deletes the records for data which causes the reduction of features and model get under fitted and reduce the accuracy. Tables 1, 2, 3, 4, 5, and 6 contain the sentiment analysis results for the different aspects of the product using the reviews dataset.

LR achieve a 0.95 accuracy score with the oversampling technique for size aspect sentiment analysis and also achieve high precision, recall, and F1 score 0.95, 0.95, 0.95, and 0.95 respectively. The accuracy score of LR without any resampling is also given as 0.83 but there is lots of fluctuation in precision, recall, and F1 score. The accuracy score is 0.83 while the F1 score is 0.59 which is not acceptable for the well-fitted model. The accuracy score with the oversampling technique and LR for colour aspect sentiment analysis achieve 0.94. The performance of LR on colour aspect data is low as compare to the size reviews that because colour aspect data contain very few records and highly imbalanced as ‘Negative’: 680, 'Neutral': 153, 'Positive': 2239. So it shows poor performance. The performance of LR is 0.79, 0.73, and 0.94 in terms of without any sampling, under-sampling, and oversampling. The results of colour aspect reviews show that the size of data highly impacts the accuracy of machine learning models. LR performs best on the price aspect using the oversampling technique with 0.96 accuracies because the price aspect have the large dataset as compare to other all other aspect and LR have more records to learn sentiment. Table 6 shows the results using the 10 fold cross-validation after oversampling which shows the significance of LR.

Table 1 LR results for size aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.83 | Negative | 0.83 | 0.43 | 0.57 |
| Neutral | 1.00 | 0.17 | 0.29 |
| Positive | 0.83 | 0.99 | 0.90 |
| Macro avg | 0.89 | 0.53 | 0.59 |
| Under-sampling | 0.83 | Negative | 0.78 | 0.80 | 0.79 |
| Neutral | 0.78 | 0.74 | 0.76 |
| Positive | 0.90 | 0.91 | 0.91 |
| Macro avg | 0.82 | 0.82 | 0.82 |
| Over-sampling | 0.95 | Negative | 0.94 | 0.97 | 0.95 |
| Neutral | 0.95 | 1.00 | 0.97 |
| Positive | 0.97 | 0.89 | 0.93 |
| Macro avg | 0.95 | 0.95 | 0.95 |

Table 2 LR results for colour aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.79 | Negative | 0.87 | 0.32 | 0.47 |
| Neutral | 1.00 | 0.10 | 0.18 |
| Positive | 0.78 | 0.99 | 0.87 |
| Macro avg | 0.88 | 0.47 | 0.51 |
| Under-sampling | 0.73 | Negative | 0.74 | 0.66 | 0.69 |
| Neutral | 0.62 | 0.81 | 0.70 |
| Positive | 0.88 | 0.75 | 0.81 |
| Macro avg | 0.74 | 0.74 | 0.73 |
| Over-sampling | 0.94 | Negative | 0.90 | 0.98 | 0.94 |
| Neutral | 0.95 | 1.00 | 0.97 |
| Positive | 0.98 | 0.86 | 0.91 |
| Macro avg | 0.94 | 0.94 | 0.94 |

Table 3 LR results for price aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.90 | Negative | 0.87 | 0.74 | 0.80 |
| Neutral | 0.91 | 0.34 | 0.50 |
| Positive | 0.91 | 0.98 | 0.94 |
| Macro avg | 0.90 | 0.69 | 0.75 |
| Under-sampling | 0.88 | Negative | 0.87 | 0.86 | 0.87 |
| Neutral | 0.84 | 0.88 | 0.86 |
| Positive | 0.95 | 0.91 | 0.93 |
| Macro avg | 0.89 | 0.88 | 0.88 |
| Over-sampling | 0.96 | Negative | 0.93 | 0.97 | 0.95 |
| Neutral | 0.96 | 1.00 | 0.98 |
| Positive | 0.98 | 0.90 | 0.94 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 4 LR results for service aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.83 | Negative | 0.81 | 0.50 | 0.62 |
| Neutral | 1.00 | 0.12 | 0.21 |
| Positive | 0.84 | 0.98 | 0.90 |
| Macro avg | 0.88 | 0.53 | 0.58 |
| Under-sampling | 0.74 | Negative | 0.76 | 0.66 | 0.70 |
| Neutral | 0.63 | 0.85 | 0.72 |
| Positive | 0.88 | 0.75 | 0.81 |
| Macro avg | 0.75 | 0.75 | 0.74 |
| Over-sampling | 0.95 | Negative | 0.94 | 0.97 | 0.95 |
| Neutral | 0.94 | 1.00 | 0.97 |
| Positive | 0.98 | 0.88 | 0.93 |
| Macro avg | 0.95 | 0.95 | 0.95 |

Table 5 LR results for weight aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.79 | Negative | 0.76 | 0.19 | 0.30 |
| Neutral | 0.00 | 0.00 | 0.00 |
| Positive | 0.79 | 0.99 | 0.88 |
| Macro avg | 0.52 | 0.39 | 0.39 |
| Under-sampling | 0.60 | Negative | 0.50 | 0.60 | 0.55 |
| Neutral | 0.62 | 0.56 | 0.59 |
| Positive | 0.73 | 0.67 | 0.70 |
| Macro avg | 0.62 | 0.61 | 0.61 |
| Over-sampling | 0.95 | Negative | 0.87 | 0.98 | 0.93 |
| Neutral | 0.99 | 1.00 | 0.99 |
| Positive | 0.98 | 0.86 | 0.92 |
| Macro avg | 0.95 | 0.95 | 0.95 |

Table 6 LR results using 10 fold cross-validation

|  |  |
| --- | --- |
| Aspect | Accuracy |
| Size | 0.95502495 |
| Color | 0.94447475 |
| Price | 0.95740908 |
| Service | 0.94883654 |
| Weight | 0.95320248 |

## RF Results

RF is an ensemble model that uses different decision trees to make a final prediction and uses the bootstrap method to train each decision tree in a random forest. Bootstrap method randomly selects the records from dataset to make a subsample dataset for decision tree so in this way random forest can be good somehow on the imbalanced and small size of datasets.

RF achieved the highest accuracy of 0.97 with oversampling on all aspects except size and service. RF shows a better result as compared to the LR. RF outperforms in terms of all evaluation parameters. The lowest accuracy is achieved on the colour aspect dataset using RF and under-sampling. Tables 7, 8, 9, 10, 11, and 12 show the results of RF on all aspect datasets. Table 12 shows the performance of 10 fold cross-validation using the RF and results show the significance of RF.

Table 7 RF results for size aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.83 | Negative | 0.83 | 0.43 | 0.57 |
| Neutral | 1.00 | 0.17 | 0.29 |
| Positive | 0.83 | 0.99 | 0.90 |
| Macro avg | 0.89 | 0.53 | 0.59 |
| Under-sampling | 0.83 | Negative | 0.78 | 0.80 | 0.79 |
| Neutral | 0.78 | 0.74 | 0.76 |
| Positive | 0.90 | 0.91 | 0.91 |
| Macro avg | 0.82 | 0.82 | 0.82 |
| Over-sampling | 0.96 | Negative | 0.94 | 0.98 | 0.96 |
| Neutral | 0.96 | 1.00 | 0.98 |
| Positive | 0.98 | 0.90 | 0.94 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 8 RF results for colour aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.75 | Negative | 0.95 | 0.14 | 0.25 |
| Neutral | 0.71 | 0.17 | 0.27 |
| Positive | 0.74 | 0.99 | 0.85 |
| Macro avg | 0.80 | 0.43 | 0.46 |
| Under-sampling | 0.65 | Negative | 0.86 | 0.32 | 0.46 |
| Neutral | 0.47 | 0.92 | 0.62 |
| Positive | 0.89 | 0.86 | 0.87 |
| Macro avg | 0.74 | 0.70 | 0.65 |
| Over-sampling | 0.97 | Negative | 0.97 | 0.95 | 0.96 |
| Neutral | 0.97 | 1.00 | 0.99 |
| Positive | 0.97 | 0.96 | 0.96 |
| Macro avg | 0.97 | 0.97 | 0.97 |

Table 9 RF results for price aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.80 | Negative | 0.98 | 0.20 | 0.33 |
| Neutral | 0.97 | 0.13 | 0.22 |
| Positive | 0.79 | 1.00 | 0.88 |
| Macro avg | 0.92 | 0.44 | 0.48 |
| Under-sampling | 0.89 | Negative | 0.92 | 0.86 | 0.89 |
| Neutral | 0.86 | 0.89 | 0.87 |
| Positive | 0.90 | 0.92 | 0.91 |
| Macro avg | 0.89 | 0.89 | 0.89 |
| Over-sampling | 0.97 | Negative | 0.97 | 0.95 | 0.96 |
| Neutral | 0.98 | 1.00 | 0.99 |
| Positive | 0.96 | 0.96 | 0.96 |
| Macro avg | 0.97 | 0.97 | 0.97 |

Table 10 RF results for service aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.77 | Negative | 0.93 | 0.18 | 0.30 |
| Neutral | 0.80 | 0.12 | 0.21 |
| Positive | 0.77 | 1.00 | 0.87 |
| Macro avg | 0.83 | 0.43 | 0.46 |
| Under-sampling | 0.81 | Negative | 0.76 | 0.64 | 0.69 |
| Neutral | 0.72 | 0.92 | 0.81 |
| Positive | 0.94 | 0.86 | 0.90 |
| Macro avg | 0.81 | 0.81 | 0.80 |
| Over-sampling | 0.96 | Negative | 0.96 | 0.93 | 0.94 |
| Neutral | 0.97 | 1.00 | 0.99 |
| Positive | 0.95 | 0.94 | 0.94 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 11 RF results for weight aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.77 | Negative | 0.78 | 0.08 | 0.15 |
| Neutral | 0.00 | 0.00 | 0.00 |
| Positive | 0.77 | 0.99 | 0.87 |
| Macro avg | 0.52 | 0.36 | 0.34 |
| Under-sampling | 0.69 | Negative | 1.00 | 0.27 | 0.42 |
| Neutral | 0.65 | 0.94 | 0.77 |
| Positive | 0.67 | 0.83 | 0.74 |
| Macro avg | 0.77 | 0.68 | 0.64 |
| Over-sampling | 0.97 | Negative | 0.95 | 0.95 | 0.95 |
| Neutral | 1.00 | 1.00 | 1.00 |
| Positive | 0.96 | 0.96 | 0.96 |
| Macro avg | 0.97 | 0.97 | 0.97 |

Table 12 RF results using 10 fold cross-validation

|  |  |
| --- | --- |
| Aspect | Accuracy |
| Size | 0.96275297 |
| Color | 0.960707854 |
| Price | 0.96934366 |
| Service | 0.96463119 |
| Weight | 0.96568230 |

## SVM Results

This section contains the results for the SVM model using all aspect dataset with under sampling and oversampling. SVM is a linear model that can perform better with a linear kernel. Similarly, SVM outperforms using oversampling like RF and LR. SVM achieved the highest accuracy of 0.97 only on the price aspect dataset because the price aspect has large data and SVM performs when the number of features is more. The results of the SVM model shown in Tables 13, 14, 15, 16, 17, and 18. The price aspect results are shown in Table 15 where it shows the highest accuracy score in all three cases such as without any sampling, under-sampling, and oversampling with 0.91, and 0.90 and 0.97 accuracy scores. Table 18 shows the performance of SVM with 10 fold cross-validation approach using after oversampling of data and results show the SVM significance on all aspects of data.

Table 13 SVM results for size aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.85 | Negative | 0.80 | 0.55 | 0.65 |
| Neutral | 0.85 | 0.24 | 0.37 |
| Positive | 0.86 | 0.97 | 0.91 |
| Macro avg | 0.84 | 0.59 | 0.64 |
| Under-sampling | 0.82 | Negative | 0.78 | 0.79 | 0.78 |
| Neutral | 0.77 | 0.74 | 0.75 |
| Positive | 0.90 | 0.91 | 0.91 |
| Macro avg | 0.81 | 0.81 | 0.81 |
| Over-sampling | 0.96 | Negative | 0.94 | 0.98 | 0.96 |
| Neutral | 0.96 | 1.00 | 0.98 |
| Positive | 0.98 | 0.90 | 0.94 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 14 SVM results for color aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.80 | Negative | 0.81 | 0.42 | 0.56 |
| Neutral | 0.60 | 0.10 | 0.17 |
| Positive | 0.80 | 0.97 | 0.87 |
| Macro avg | 0.73 | 0.50 | 0.53 |
| Under-sampling | 0.72 | Negative | 0.73 | 0.63 | 0.68 |
| Neutral | 0.60 | 0.81 | 0.69 |
| Positive | 0.88 | 0.75 | 0.81 |
| Macro avg | 0.73 | 0.73 | 0.72 |
| Over-sampling | 0.94 | Negative | 0.91 | 0.98 | 0.94 |
| Neutral | 0.95 | 1.00 | 0.97 |
| Positive | 0.98 | 0.86 | 0.91 |
| Macro avg | 0.95 | 0.95 | 0.94 |

Table 15 SVM results for price aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.91 | Negative | 0.85 | 0.79 | 0.82 |
| Neutral | 0.86 | 0.51 | 0.64 |
| Positive | 0.93 | 0.97 | 0.95 |
| Macro avg | 0.88 | 0.76 | 0.80 |
| Under-sampling | 0.90 | Negative | 0.88 | 0.89 | 0.89 |
| Neutral | 0.85 | 0.89 | 0.87 |
| Positive | 0.96 | 0.92 | 0.94 |
| Macro avg | 0.90 | 0.90 | 0.90 |
| Over-sampling | 0.97 | Negative | 0.95 | 0.99 | 0.97 |
| Neutral | 0.98 | 1.00 | 0.99 |
| Positive | 0.99 | 0.93 | 0.96 |
| Macro avg | 0.97 | 0.97 | 0.97 |

Table 16 SVM results for service aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.84 | Negative | 0.76 | 0.60 | 0.67 |
| Neutral | 0.75 | 0.22 | 0.34 |
| Positive | 0.86 | 0.96 | 0.91 |
| Macro avg | 0.79 | 0.59 | 0.64 |
| Under-sampling | 0.82 | Negative | 0.76 | 0.80 | 0.78 |
| Neutral | 0.75 | 0.83 | 0.79 |
| Positive | 0.96 | 0.84 | 0.89 |
| Macro avg | 0.82 | 0.82 | 0.82 |
| Over-sampling | 0.96 | Negative | 0.94 | 0.98 | 0.96 |
| Neutral | 0.95 | 1.00 | 0.97 |
| Positive | 0.98 | 0.90 | 0.94 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 17 SVM results for weight aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.79 | Negative | 0.67 | 0.26 | 0.37 |
| Neutral | 0.00 | 0.00 | 0.00 |
| Positive | 0.80 | 0.97 | 0.87 |
| Macro avg | 0.49 | 0.41 | 0.42 |
| Under-sampling | 0.60 | Negative | 0.47 | 0.53 | 0.50 |
| Neutral | 0.65 | 0.61 | 0.63 |
| Positive | 0.73 | 0.67 | 0.70 |
| Macro avg | 0.61 | 0.60 | 0.61 |
| Over-sampling | 0.96 | Negative | 0.89 | 0.99 | 0.93 |
| Neutral | 0.99 | 1.00 | 1.00 |
| Positive | 0.99 | 0.88 | 0.93 |
| Macro avg | 0.96 | 0.96 | 0.95 |

Table 18 SVM results using 10 fold cross-validation

|  |  |
| --- | --- |
| Aspect | Accuracy |
| Size | 0.962752974 |
| Color | 0.949237102 |
| Price | 0.971817967 |
| Service | 0.956998901 |
| Weight | 0.963043104 |

## KNN Results

The performance of the KNN model on aspect reviews data shown in this section. KNN is not a good performer on the used dataset and achieved the highest accuracy of 0.75. The reason behind the bad performance of KNN is the large feature set. As the text data have a large feature set which will increase more after oversampling that the reason the accuracy of KNN in the oversampling case is very low. The accuracy of the KNN model using the under-sampling approach is higher as compare to oversampling. The reason is that the under-sampling approach reduces the size of data which causes a reduction in the size of the feature set. The results of KNN are shown in Tables 19, 20, 21, 22, 23, and 24. The highest accuracy of KNN gets through under-sampling on color aspect data because the color aspect contains less data or we can say fewer features for training.

Table 19 SVM results for size aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.75 | Negative | 1.00 | 0.01 | 0.01 |
| Neutral | 0.00 | 0.00 | 0.00 |
| Positive | 0.75 | 1.00 | 0.86 |
| Macro avg | 0.58 | 0.34 | 0.29 |
| Under-sampling | 0.67 | Negative | 0.57 | 0.55 | 0.56 |
| Neutral | 0.64 | 0.44 | 0.52 |
| Positive | 0.74 | 0.95 | 0.83 |
| Macro avg | 0.65 | 0.64 | 0.64 |
| Over-sampling | 0.35 | Negative | 1.00 | 0.01 | 0.03 |
| Neutral | 0.35 | 1.00 | 0.52 |
| Positive | 0.00 | 0.00 | 0.00 |
| Macro avg | 0.45 | 0.34 | 0.18 |

Table 20 KNN results for color aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.72 | Negative | 1.00 | 0.01 | 0.01 |
| Neutral | 0.00 | 0.00 | 0.00 |
| Positive | 0.71 | 1.00 | 0.83 |
| Macro avg | 0.57 | 0.34 | 0.28 |
| Under-sampling | 0.72 | Negative | 0.67 | 0.68 | 0.68 |
| Neutral | 0.74 | 0.54 | 0.62 |
| Positive | 0.65 | 0.79 | 0.71 |
| Macro avg | 0.68 | 0.67 | 0.67 |
| Over-sampling | 0.61 | Negative | 0.99 | 0.88 | 0.94 |
| Neutral | 0.46 | 1.00 | 0.63 |
| Positive | 1.00 | 0.00 | 0.00 |
| Macro avg | 0.82 | 0.63 | 0.52 |

Table 21 KNN results for price aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.65 | Negative | 0.64 | 0.04 | 0.07 |
| Neutral | 0.10 | 0.53 | 0.17 |
| Positive | 0.81 | 0.82 | 0.82 |
| Macro avg | 0.52 | 0.46 | 0.35 |
| Under-sampling | 0.54 | Negative | 0.74 | 0.06 | 0.12 |
| Neutral | 0.42 | 0.97 | 0.59 |
| Positive | 0.98 | 0.55 | 0.70 |
| Macro avg | 0.71 | 0.53 | 0.47 |
| Over-sampling | 0.41 | Negative | 0.95 | 0.20 | 0.33 |
| Neutral | 0.37 | 1.00 | 0.54 |
| Positive | 1.00 | 0.01 | 0.02 |
| Macro avg | 0.77 | 0.40 | 0.30 |

Table 22 KNN results for service aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.74 | Negative | 0.64 | 0.09 | 0.16 |
| Neutral | 0.33 | 0.40 | 0.36 |
| Positive | 0.77 | 0.96 | 0.86 |
| Macro avg | 0.58 | 0.48 | 0.46 |
| Under-sampling | 0.68 | Negative | 0.60 | 0.45 | 0.52 |
| Neutral | 0.54 | 0.69 | 0.61 |
| Positive | 0.85 | 0.84 | 0.84 |
| Macro avg | 0.66 | 0.68 | 0.66 |
| Over-sampling | 0.62 | Negative | 0.99 | 0.87 | 0.93 |
| Neutral | 0.47 | 1.00 | 0.64 |
| Positive | 1.00 | 0.00 | 0.01 |
| Macro avg | 0.82 | 0.62 | 0.52 |

Table 23 KNN results for weight aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.72 | Negative | 0.34 | 0.22 | 0.27 |
| Neutral | 0.50 | 0.07 | 0.12 |
| Positive | 0.78 | 0.89 | 0.83 |
| Macro avg | 0.54 | 0.39 | 0.41 |
| Under-sampling | 0.60 | Negative | 0.35 | 0.53 | 0.42 |
| Neutral | 0.40 | 0.22 | 0.29 |
| Positive | 0.58 | 0.58 | 0.58 |
| Macro avg | 0.44 | 0.45 | 0.43 |
| Over-sampling | 0.65 | Negative | 0.59 | 0.98 | 0.74 |
| Neutral | 0.71 | 1.00 | 0.83 |
| Positive | 1.00 | 0.02 | 0.04 |
| Macro avg | 0.77 | 0.67 | 0.53 |

Table 24 KNN results using 10 fold cross-validation

|  |  |
| --- | --- |
| Aspect | Accuracy |
| Size | 0.352752974 |
| Color | 0.609237102 |
| Price | 0.401817967 |
| Service | 0.616998901 |
| Weight | 0.653043104 |

## GNB Results

Naïve Bayes family models are probability-based models that can perform better on data where features are highly correlated with the target class. the results of the GNB on the used dataset are somehow good after oversampling of data. The highest accuracy of 0.92 achieved using the weight aspect data. Overall the performance of GNB is good after oversampling of data and performs very poorly on the imbalanced data in terms of the accuracy score. GNB is somehow better on imbalanced as it have low fluctuation in accuracy, precision, recall, and F1 score as compare to other models. The results of GNB show in Table 25, 26, 27, 28, 29, and 30. Table 30 show that the performance of 10 fold cross-validation to show the significance of GNB.

Table 25 GNB results for size aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.61 | Negative | 0.23 | 0.35 | 0.27 |
| Neutral | 0.92 | 0.31 | 0.46 |
| Positive | 0.78 | 0.71 | 0.74 |
| Macro avg | 0.64 | 0.45 | 0.49 |
| Under-sampling | 0.57 | Negative | 0.49 | 0.85 | 0.62 |
| Neutral | 0.84 | 0.26 | 0.40 |
| Positive | 0.64 | 0.59 | 0.61 |
| Macro avg | 0.66 | 0.57 | 0.54 |
| Over-sampling | 0.88 | Negative | 0.73 | 0.98 | 0.84 |
| Neutral | 0.99 | 1.00 | 0.99 |
| Positive | 0.99 | 0.65 | 0.78 |
| Macro avg | 0.90 | 0.88 | 0.87 |

Table 26 GNB results for color aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.65 | Negative | 0.33 | 0.36 | 0.34 |
| Neutral | 0.83 | 0.17 | 0.28 |
| Positive | 0.76 | 0.78 | 0.77 |
| Macro avg | 0.64 | 0.43 | 0.46 |
| Under-sampling | 0.62 | Negative | 0.57 | 0.71 | 0.64 |
| Neutral | 0.69 | 0.42 | 0.52 |
| Positive | 0.66 | 0.68 | 0.67 |
| Macro avg | 0.64 | 0.60 | 0.61 |
| Over-sampling | 0.91 | Negative | 0.79 | 1.00 | 0.88 |
| Neutral | 1.00 | 1.00 | 1.00 |
| Positive | 0.99 | 0.74 | 0.85 |
| Macro avg | 0.93 | 0.91 | 0.91 |

Table 27 GNB results for price aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.62 | Negative | 0.33 | 0.36 | 0.34 |
| Neutral | 0.83 | 0.17 | 0.28 |
| Positive | 0.76 | 0.78 | 0.77 |
| Macro avg | 0.62 | 0.43 | 0.55 |
| Under-sampling | 0.61 | Negative | 0.57 | 0.71 | 0.64 |
| Neutral | 0.69 | 0.42 | 0.52 |
| Positive | 0.66 | 0.68 | 0.67 |
| Macro avg | 0.64 | 0.60 | 0.61 |
| Over-sampling | 0.89 | Negative | 0.79 | 1.00 | 0.88 |
| Neutral | 1.00 | 1.00 | 1.00 |
| Positive | 0.99 | 0.74 | 0.85 |
| Macro avg | 0.88 | 0.89 | 0.88 |

Table 28 GNB results for service aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.64 | Negative | 0.31 | 0.43 | 0.36 |
| Neutral | 0.52 | 0.19 | 0.28 |
| Positive | 0.80 | 0.74 | 0.77 |
| Macro avg | 0.54 | 0.45 | 0.47 |
| Under-sampling | 0.65 | Negative | 0.50 | 0.77 | 0.61 |
| Neutral | 0.83 | 0.45 | 0.59 |
| Positive | 0.76 | 0.71 | 0.73 |
| Macro avg | 0.70 | 0.65 | 0.64 |
| Over-sampling | 0.88 | Negative | 0.75 | 0.97 | 0.85 |
| Neutral | 0.99 | 1.00 | 0.99 |
| Positive | 0.98 | 0.69 | 0.81 |
| Macro avg | 0.90 | 0.89 | 0.88 |

Table 29 GNB results for weight aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.72 | Negative | 0.22 | 0.18 | 0.20 |
| Neutral | 0.00 | 0.00 | 0.00 |
| Positive | 0.77 | 0.85 | 0.81 |
| Macro avg | 0.33 | 0.34 | 0.33 |
| Under-sampling | 0.44 | Negative | 0.40 | 0.53 | 0.46 |
| Neutral | 0.75 | 0.17 | 0.27 |
| Positive | 0.43 | 0.75 | 0.55 |
| Macro avg | 0.53 | 0.48 | 0.43 |
| Over-sampling | 0.92 | Negative | 0.80 | 0.99 | 0.89 |
| Neutral | 1.00 | 1.00 | 1.00 |
| Positive | 0.99 | 0.78 | 0.88 |
| Macro avg | 0.93 | 0.93 | 0.92 |

Table 30 GNB results using 10 fold cross-validation

|  |  |
| --- | --- |
| Aspect | Accuracy |
| Size | 0.872752974 |
| Color | 0.919237102 |
| Price | 0.881817967 |
| Service | 0.886998901 |
| Weight | 0.903043104 |

## Voting Classifier Results

The voting classifier is an combination of two learning model which give the final prediction using the both model output. The voting model can outperform as compare to and individual because of its ensemble architecture and can be good on both small and imbalanced dataset as shown below in results. The voting model achecived consistent perform in terms of all evaluation parameters as the voting models achecived its betters results on size aspect 0.80, 0.91 and 0.96 which are highest as compare to all previous results on size aspects in their terms of F1 score. On the service aspect data voring model perform all previous results and achieve high F1 score. The performance of the models can be seen in Table 31, 32, 33, 34, 35 and 36.

Table 31 voting classifier results for size aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.80 | Negative | 0.97 | 0.21 | 0.35 |
| Neutral | 1.00 | 0.22 | 0.36 |
| Positive | 0.79 | 1.00 | 0.88 |
| Macro avg | 0.92 | 0.48 | 0.53 |
| Under-sampling | 0.81 | Negative | 0.84 | 0.62 | 0.71 |
| Neutral | 0.70 | 0.87 | 0.78 |
| Positive | 0.89 | 0.91 | 0.90 |
| Macro avg | 0.81 | 0.80 | 0.80 |
| Over-sampling | 0.96 | Negative | 0.96 | 0.96 | 0.96 |
| Neutral | 0.96 | 1.00 | 0.98 |
| Positive | 0.97 | 0.93 | 0.95 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 32 voting classifier results for color aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.76 | Negative | 0.96 | 0.16 | 0.27 |
| Neutral | 1.00 | 0.13 | 0.24 |
| Positive | 0.75 | 1.00 | 0.85 |
| Macro avg | 0.90 | 0.43 | 0.45 |
| Under-sampling | 0.67 | Negative | 0.73 | 0.42 | 0.53 |
| Neutral | 0.52 | 0.88 | 0.66 |
| Positive | 0.88 | 0.82 | 0.85 |
| Macro avg | 0.71 | 0.71 | 0.68 |
| Over-sampling | 0.96 | Negative | 0.93 | 0.97 | 0.95 |
| Neutral | 0.96 | 1.00 | 0.98 |
| Positive | 0.98 | 0.90 | 0.94 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 33 voting classifier results for price aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.87 | Negative | 0.95 | 0.53 | 0.68 |
| Neutral | 0.96 | 0.20 | 0.33 |
| Positive | 0.86 | 0.99 | 0.92 |
| Macro avg | 0.92 | 0.58 | 0.65 |
| Under-sampling | 0.90 | Negative | 0.92 | 0.86 | 0.89 |
| Neutral | 0.85 | 0.90 | 0.88 |
| Positive | 0.93 | 0.93 | 0.93 |
| Macro avg | 0.90 | 0.90 | 0.90 |
| Over-sampling | 0.97 | Negative | 0.95 | 0.98 | 0.96 |
| Neutral | 0.97 | 1.00 | 0.99 |
| Positive | 0.99 | 0.94 | 0.96 |
| Macro avg | 0.97 | 0.97 | 0.97 |

Table 34 voting classifier results for service aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.81 | Negative | 0.91 | 0.31 | 0.47 |
| Neutral | 0.93 | 0.21 | 0.34 |
| Positive | 0.80 | 1.00 | 0.89 |
| Macro avg | 0.88 | 0.51 | 0.56 |
| Under-sampling | 0.82 | Negative | 0.75 | 0.73 | 0.74 |
| Neutral | 0.73 | 0.91 | 0.81 |
| Positive | 0.97 | 0.83 | 0.89 |
| Macro avg | 0.82 | 0.82 | 0.81 |
| Over-sampling | 0.96 | Negative | 0.95 | 0.96 | 0.96 |
| Neutral | 0.95 | 1.00 | 0.98 |
| Positive | 0.96 | 0.92 | 0.94 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 35 voting classifier results for weight aspect review dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sampling | Accuracy | Class | Precision | Recall | F1-score |
| Without any sampling | 0.77 | Negative | 0.80 | 0.09 | 0.17 |
| Neutral | 0.00 | 0.00 | 0.00 |
| Positive | 0.77 | 0.99 | 0.87 |
| Macro avg | 0.52 | 0.36 | 0.35 |
| Under-sampling | 0.64 | Negative | 0.55 | 0.40 | 0.46 |
| Neutral | 0.64 | 0.78 | 0.70 |
| Positive | 0.75 | 0.75 | 0.75 |
| Macro avg | 0.64 | 0.64 | 0.64 |
| Over-sampling | 0.96 | Negative | 0.90 | 0.98 | 0.94 |
| Neutral | 0.99 | 1.00 | 1.00 |
| Positive | 0.98 | 0.89 | 0.93 |
| Macro avg | 0.96 | 0.96 | 0.96 |

Table 36 voting classifier results using 10 fold cross-validation

|  |  |
| --- | --- |
| Aspect | Accuracy |
| Size | 0.972752974 |
| Color | 0.969237102 |
| Price | 0.971817967 |
| Service | 0.966998901 |
| Weight | 0.963043104 |