# CHAPTER 1

# INTODUCTION

Customer’s opinion has changed the reality in which promotion and advertising have usually worked as a one-way message from different companies to the consumers with any mass communication channel [1]. Different brands still promote their businesses through previous media such as radio, television, and print, and some brands promote themselves through new advertising channels such as websites, social media groups, or digital marketing [2]. Companies are now using owned media to track the customer’s feedback directly, i.e. request them to provide their experiences on a review card or request them to visit the company’s website or other social media accounts. These communication channels have become more important for the companies, particularly in the context of Customers' recommendations [3].

The online reviews help other customers to make informed decisions while watching movies or purchasing any products. Now customers do not trust advertising messages only, but they learn from other possible sources of information before making any decision, particularly online reviews. Different surveys explained that maximum customers consult from online reviews [4]. The main reason is the majority of the customers prefer the consumer’s recommendations. These reviews have become the main element in marketing and have become a mandatory feature on different websites[5]. Customers have to trust the information which is provided on the website and they don’t have any option to try the product [6]. Because of their influence and prevalence, online reviews have considerably attracted the attention of both practitioners and researchers [7].

Sentiment analysis is the study of consumer's opinions and emotions expressed in written languages [8]. Sentiment analysis is a developing field in the research area. People are now using the web for business correspondence, e-commerce [9]. As the online shopping trend is growing, customers want to share their emotions and reviews on different platforms on the internet. Extraction of user's sentiment from the reviews is very important for the other users to select the right product. Sentiment analysis is also important for organizations to grow their business by tracking the customer feedback over their different products. With the development of online shopping and e-commerce, now the bulk of the users are buying their desired products from online stores. Not only for e-commerce but rather SA is also being used to predict the results of national events like elections, etc. [10].

In the comparison of physical shopping and online shopping, users are enjoying the facilities of online shopping because they can buy anything from anywhere and anytime [11]. Moreover, multiple styles and varieties of products are also available in online shopping stores and consumers have a good choice to buy variations of products without going outside [12].

While e-commerce is performing a positive role for the convenience of the customer but some problems like product originality and delivery-related issues are also associated with it. The problems with product and delivery can be like the contradiction of actual product and descriptive information available on the product, the service of product delivery and poor quality of the received product, and many more [13]. That is why it is very important to evaluate products of online shopping stores 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 reviews and detect the feelings. This is also termed as opinion mining and text analysis [14].

Sentiment analysis allows the new customer to examine previous customer's suggestions and reviews about the product. It is a basic viewpoint for the consumers when they start e-commerce.

With the advancement of the internet throughout the world, a large number of people are sharing reviews and giving feedback. The reviews 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 [9].

However, due to a large number of reviews on e-commerce websites from customers, it is not possible to analyse the opinion manually. It is a rather complex task for a customer to identify significant details from the prevalent information available on the website. Therefore, sentiment analysis is a significant approach for opinion extraction. In previous researches, sentiment analysis has been applied in many different fields i.e. Citizens' Political Preferences [15], Supply Chain Intelligence [16], Set-Based Feature Selection For Arabic Sentiment Analysis [17], Spanish Text Transformations For Twitter Sentiment Analysis [18], Odia Language Using Supervised Classifier [19], Customer Satisfaction At Aspect-Level [20], Sentiment Analysis Algorithms And Applications [21], Aspect Based Sentiment Analysis To Evaluate Arabic News Effect On Readers [22], Sentiment Analysis On Social Media For Stock Movement Prediction, Opinion Mining And Sentiment Analysis: Tasks, Approaches And Applications [23], A Review And Comparative Analysis Of Web Services [24] Sentiment Analysis Using Revised Sentiment Strength Based On SentiWordNet [25]. Sentiment analysis and its types are explained in the next section.

Multiple techniques have been utilized in sentiment analysis in past researches. One of the previously used technique is a dictionary-based technique for sentiment analysis which have been used in different researches. The efficiency of dictionary-based sentiment analysis is dependent on the accuracy and comprehensiveness of the dictionary [25]. The language, used for reviews, may be formal or maybe informal. Sentiment words are not much domain-specific and also contain short words which creates 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.

Information retrieval techniques are used to gather data from different Blogs and E-commerce websites where people share their opinion [12]. 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 [26]. 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 based on the polarity is said to be the sentiment analysis. It is also said to be the study of the attitude, emotion, and opinion of the consumers towards a particular item [11].

## 1.1 Significance of Study

E-commerce is becoming a new form of trade rapidly. Mostly, customers view the product price, quality features, and specifications on e-commerce platforms before buying a product. With reference to another analysis, retail e-commerce sales were observed 1336 billion US dollars in the world, and with the development of sales in each year. Now, the retail sales are increased from 1336 billion US dollars to 4206 in the last 6 years. It is predicted that these sales will reach 6542 billion US dollars in the next 2years [114].

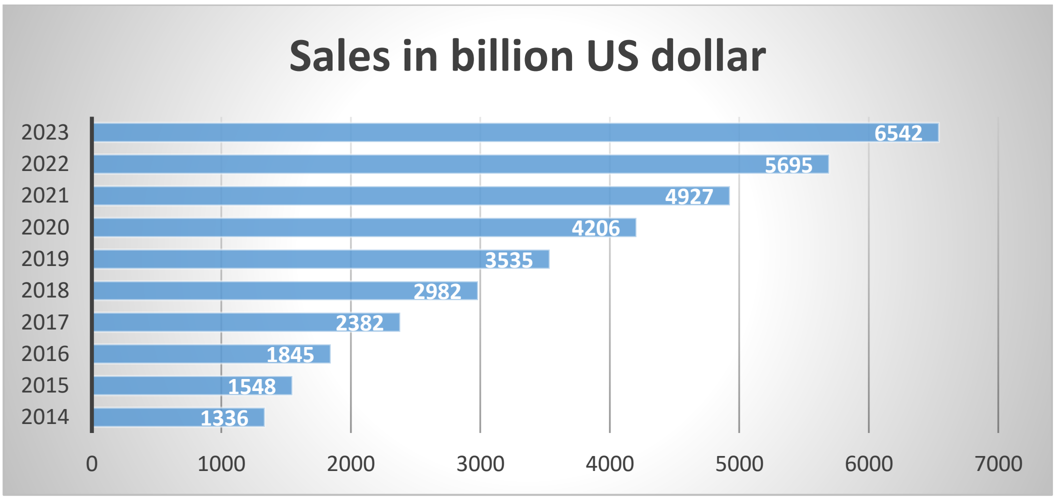

**Figure 1.1** Retail Sales of E-Commerce in worldwide (Source: www.statista.com)

During pandemic situation COVID-19, people are trying to buy their necessary things without going outside for any unnecessary reason. The sale of e-commerce is also increasing during the pandemic situation of COVID-19 [27]. Researchers present the top ten sales of e-commerce during the pandemic as the following table.

**Table 1.1:** Sales of e-commerce during COVID-19

| E-Commerce Websites | Total Business |
|---|---|
| E-bay | 1227M |
| Amazon | 4059M |
| Samsung | 648M |
| Rakuten | 804M |
| Appel | 562M |
| Walmart | 614M |
| Etsy | 395M |
| Aliexpress | 532M |
| Allegro | 272M |
| Homedepot | 292M |

## 1.2 Sentiment Analysis

Sentiment analysis is a natural language process (NLP) task in which a certain text is assessed into predefined categories (e.g., positive, negative and neutral,). Initially, lexicons based sentiment models were used for sentiment analysis that contains sentimental words with their polarities [28][13] [12]. Generally, they collect sentimental words from phrases.

Based on scattered information like strength and polarities of sentimental words, they classify the sentences in classes of sentiments with help of polarities[29] [30]. Moreover, the lexicon-based models are efficient and simple but manually sentiment lexicon creation is a time-consuming and labour-intensive job. Secondly, already static polarity is required for every sentence. For this solution, some kinds of models that automatically generate sentiment lexicons have been proposed [31][32].

Like the sentence "Sound quality of Techno mobile is not so good.". In this sense of lexicon-based approach, this sentence expresses the negative behaviour. But in the sentence "Sound quality of Samsung mobile is good." the good expresses the positive sentiment towards the sound quality of Samsung mobile. For this solution, some kind of machine learning-based models are still available.

To understand this aspect level problem, some kinds 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”. In this sentence, we can extract 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”.  In this sample sentence, both negative and positive opinions are extracted.

This kind of problem can be resolved by classifying the sentiment analysis in the following techniques mentioned in figure 1.2 [33].

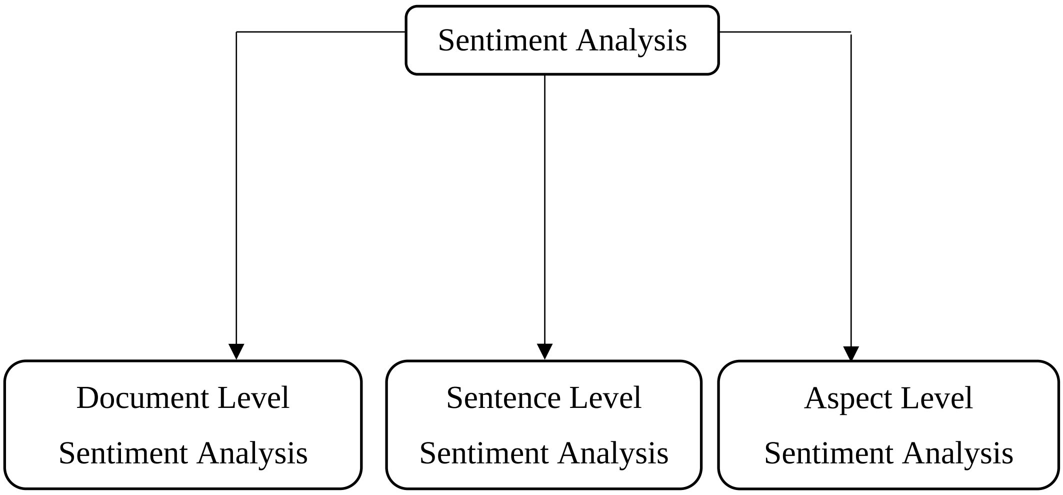
Sentiment Analysis
Document Level Sentiment Analysis
Sentence Level Sentiment Analysis
Aspect Level Sentiment Analysis

**Figure 1.2** Types of Sentiment Analysis

### 1.2.1 Document Level Sentiment Analysis

Document-level sentiment analysis is said to be the analysis of the whole document. In this approach, the complete document is considered as a single entity and it is analysed at once. The opinion about the whole document is considered as positive or negative. However, this is not a good approach because there may be e a positive specific review that has a great importance but the overall sentiment score of the document is negative and vice versa [28] [34].

### 1.2.2 Sentence Level Sentiment Analysis

Sentiment analysis at the sentence level is considered as the calculation of sentiment of each of the sentences in the document. In this approach, the document is divided into sentences, and every 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. In the above examples, we can extract the multiple meanings.

To overcome this issue the aspect level sentiment analysis has been proposed [35][36][37].

### 1.2.3 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 [38][39] researches.

## 1.3 Machine Learning

In 1997 researcher defines machine learning as the feature of computer science that aims to gain knowledge from data [25]. Machine learning is used to improve the efficiency of different analyses for example in applied Health Care and Emotion Detection etc. This is used to automate the process of flexibility and efficiency that identifies the trends from Complex data sets [40].

There are multiple steps involved to determine when ML is being used. The first step is that the machine learning technique can be used to answer the research question. In research [41], the researcher defines the three types of research problems i.e. Descriptive research, Explanatory research, and Predictive research which can be resolved through machine learning. Machine learning has been performed in many other fields and it is verified by the statistical methods which are sufficient in some cases and sometimes the questions validate the results. With respect to the working, Machine learning is divided into the following three types mentioned in figure 1.3 [33].

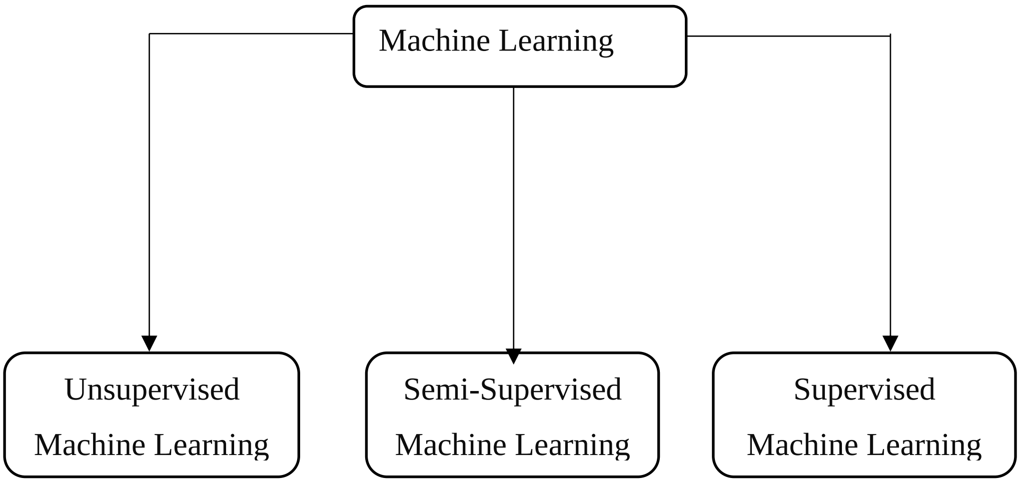
Machine Learning
Unsupervised Machine Learning
Semi-Supervised Machine Learning
Supervised Machine Learning

**Figure 1.3** Types of Machine Learning

### 1.3.1 Unsupervised Machine Learning

Unsupervised machine learning is specifically helpful for descriptive research because this research aims to find the relationship between the data structure without knowing the target outcomes. This methodology is referred to unsupervised learning because we don't have any target variable that could be happened [42].

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

Unsupervised learning is used to find trends from the dataset. There are two main types of unsupervised learning which are commonly used, principal component analysis and cluster analysis. The cluster analysis is used to achieve different qualitative groups of individuals. Principal component analysis can be used to learn the large numbers of neurons. This approach is often used as pre-processed data or to reduce the size of the forecaster from big data.

### 1.3.2 Semi-Supervised Machine Learning

Semi-supervised learning consists of both types of unsupervised and supervised learning. In this technique, the 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 [43] [44].

### 1.3.3 Supervised Machine Learning

Supervised learning is utilized by 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 [45][46][47]. Supervised machine learning can be used when prior knowledge of the predicting labels or classes is available. In this technique, the algorithm is trained with the help of a large amount of dataset first, and then the test data set is passed through the model and the efficiency of the model is measured by applying the evaluation matrices.

Supervised learning is one of the machine learning techniques in which predictive classes are known. In the case of the review detection, a review may be positive or negative or maybe neutral. So, in this technique, the predicting class of a review would be negative, positive, or neutral. The technique of supervised machine learning is worked as the data set is divided into the training dataset and the test dataset. The training of the model is performed by labelling the dataset with actual sentiment and then the 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 [48]

#### 1.3.3.1 Classification

Classification is said to be supervised machine learning [49] 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 [50] [31]. Classification is said to be the process of creating the model with the help of training data set having labels and this model can be used to predict the classes or labels of testing data. Classification in supervised machine learning is being used in several intelligence-based researches. In this study, we are going to perform analysis on the following list of classifiers:

#### 1.3.3.1.1 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 the 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 [50]. 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 [52].

#### 1.3.3.1.2 Naive Bayes classifier

Naive Bayes classifier is one of the simple statistical basin classifiers [53]. 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 [54]. 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:

$$(A|B) = P\ X|A\ (A)\ (B) \qquad \text{Eq. 1.1}$$

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

#### 1.3.3.1.3 K-Nearest Neighbours

K nearest neighbour is object-based, without a parametric learning method. It is also called lazy learners because it stores all the training samples. It does not allow to build of 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 [55][56][31].

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 finds that trend space for the K objects which are nearest to it and nominates 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 [57].

#### 1.3.3.1.4 Support vector machine

SVM has been used in many researches in the last decade and applied in multiple domain applications [53] SVM is used for regression, classification, and ranking functions. This is based on the statistical theory of learning and risk minimization principle of structure and explores the decision boundary location which is also called as a hyperplane that creates the optimum classes' separations [58] [50] [59]. SVM finds the best hyperplane to classify the data into classification. To find the best hyperplane, SVM removes outlier from data and separates to categorize with the best linear hyperplane.

SVM can be used to solve multi-dimensional problems by using different kernels [60]. Kernel changes the dimension of data space according to the nature of data.

#### 1.3.3.1.5 Random-Forest

Random forest is a type of classifier which is consisted of the collection of tree-structured classifiers h (A, $O_n$) and n=1,2,3, … where on are independently separated random vectors and every tree determines the most used class at input A. The best thing about this combination is each decision tree is made from a random vector of parameters [61].

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 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 combines the number of decision trees using the majority voting criteria in which multiple decision trees give their predictions and then the final prediction has been selected as the majority of voters on decision trees predictions. This ensemble model can give good results as compared to an individual decision tree [62]. This ensemble random forest model can also perform well on imbalanced data because of the bootstraps sampling technique [61].

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:

$$dts = dt_1\,, dt_2, dt_3, \dots, dt_n \qquad \text{Eq. 1.2}$$

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

$$rf = mode\{dt_1 , dt_2, dt_3, \dots , dt_{n\}} \qquad \text{Eq. 1.3}$$

$$rf = \sum_{i=0}^{N} dt\, i \qquad \text{Eq. 1.4}$$

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.

### 1.3.3.1.6 Logistic Regression

Logistic regression is a classification methodology that uses classes for the creation and uses a single multinomial regression model with the help of a single estimator. Logistic regression is usually used when the class boundaries are known and it is also used 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 [63]. Logistic regression uses the logistic function which can be useful when the dependent variable contains binary value[64]. Logistic regression is an advancement in linear regression. [65].

Linear regression can be defined as mathematically:

$$\hat{Y} = bX + a \qquad \text{Eq. 1.5}$$

Here, y is the prediction value; bX is the slope and $a$ 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:

$$p = \frac{e^{bX+a}}{1+e^{bX+a}} \qquad \text{Eq. 1.6}$$

Here, p is the target value between 0 and 1. $e^{bX+a}$ is the relationship between target values.

#### 1.3.3.1.7 Deep Learning

Deep learning is a part of machine learning which can be referred to as a deep neural network [59]. A neural network is influenced by 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. 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 [60].

## 1.4 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 gets trained on the imbalanced dataset they get overfitted 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 [66].

## 1.5 Evolution Measures

To determine the results of the machine learning algorithm classification techniques, following evaluation measures use to determine their results.

### 1.5.1 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:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad \text{Eq. 1.7}$$

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.

### 1.5.2 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.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Eq. 1.8}$$

### 1.5.3 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.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \text{Eq. 1.9}$$

### 1.5.4 F-Measure

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

$$F\ measure = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad \text{Eq. 1.10}$$

## 1.6 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 [67]. In research [68], researchers have achieved accuracy 94.46 from SVM. 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 at aspect-based sentiment analysis.

## 1.7 Motivation

The e-commerce business is the future of the world and gaining lots of interest from the public. People feel comfortable to buy their desired products without going outside by using online platform and also give their reviews on the product which impacts the company's worth and helps the other buyers to select the right product. Customer satisfaction is the main object of shopping websites [69]. 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. This study will contribute to find the sentiment of people on their product aspects and can improve the selection of product according to their customer's requirements.

## 1.8 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 analyses 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 analyses the reviews and predict the result accurately. So, it is required to increase the accuracy of the system.

## 1.9 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?

## 1.10 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 increase the accuracy of a classification model.

## 1.11 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.

## 1.12 Thesis Organization

Chapter I presents the introduction of the thesis-related concepts. Chapter II presents previous work and the contribution of related work. In chapter III methodology is discussed which includes pre-processing steps and classification details. In chapter IV, the discussion and the results are presented and in Chapter V, the conclusion of this study is presented.

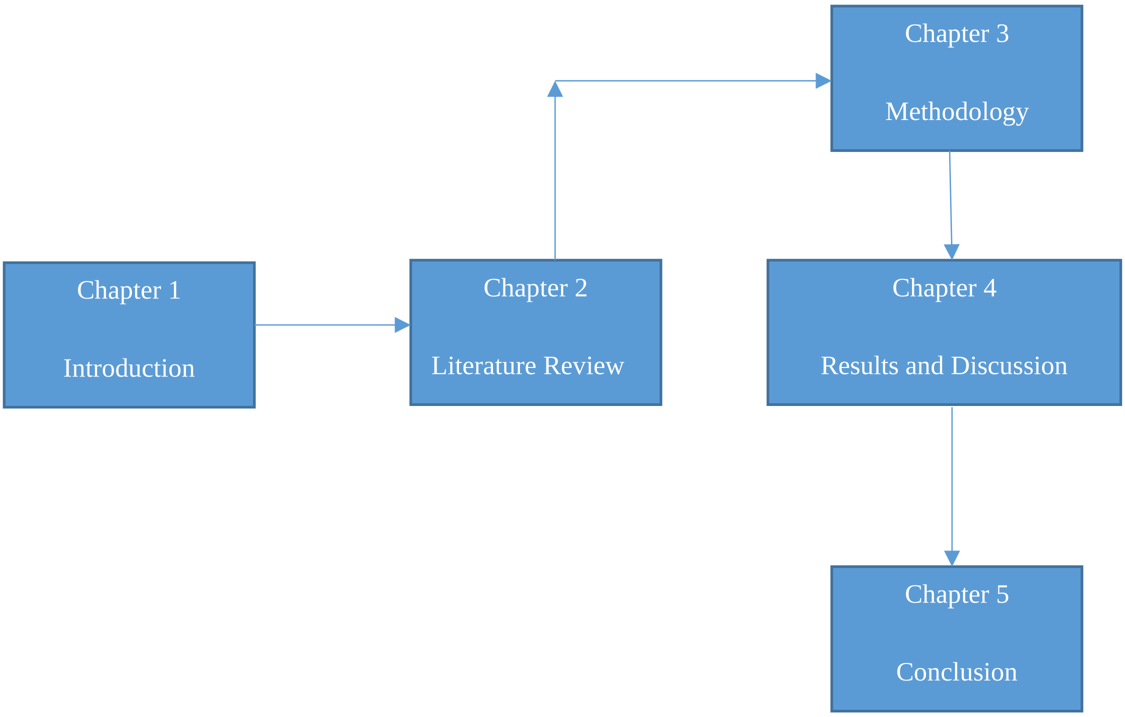

**Figure 1.4** Thesis Organization