**Chapter 1**

**3.1 Introduction**

Sentiment analysis is the study of people`s opinions, sentiments, attitudes, and emotions expressed in written languages. [1]. Sentiment analysis is a grown upfield 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 are eagerly want to share their emotions and reviews on different platforms on the internet. Extraction of users feeling from the reviews which he provided on the internet has great importance not only for the user of the selection of right product and its variation and rather the organization needs to grow their business to check the customer feedback over their different products. It is also being used to predict the results of social and national events like elections etc.[3]. with the development of online shopping and e-commerce, now bunches of users are buying their desired products from online stores. With 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, there are multiple styles and varieties of products available on online shopping stores and consumers have a great choice to buy variations of products without going outside[5].

While e-commerce is performing a positive role inconvenience of the user but due to vitality of online shopping stores, multiple problems of products and delivery has also been arising like the contradiction between real goods and descriptive information which is provided while ordering to the product, the service of product delivery and poor quality of received product and many more [6]. That is why it is very important to evaluate purchased products from online shopping stores and the tendency of the emotion of customer towards the product with the service by performing sentiment analysis which is good for business growth and the reference for other consumers.  Sentiment analysis for customer reviews is also defined as the process of systematically analyzing the subject of emotional reviews, termed as opinion mining and text analysis [7].

Increase the usage of internet devices in business prospective sentiment analysis allows examining customer suggestions and reviews about 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. 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. Research for instance [2] of sentiment analysis on the feedback of goods. This has now become a topic of research that has been widely carried out.

However, due to the many reviews on e-commerce websites from customers, it is not possible to analyze the opinion manually. Therefore sentiment analysis is a significant approach for opinion extraction.  Sentiment analysis can be 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 types of research have presented the techniques for analyses of emotions shared in feedback depends on the dictionary. 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. These words are not much domain-specific and also short words which creates difficulty in making an accurate dictionary.  However many types of research have been performed on English text. It is observed that English words are not natural whereas 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].

There are many information retrieval techniques which 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 analyze 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 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 following types of sentiments:

There are three types of emotions:

1. Positive

Positive if the consumer is satisfied or happy after using any concerned entity.

1. Negative

Negative if the consumer is not-satisfy or unhappy after using any concerned entity.

1. Neutral

Negative if the consumer is neither sympathizes with nor disparages after any concerned entity.

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, models on sentiment analysis were using handmade sentiment lexicons 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 labor-intensive job. For this solution, some kind of models that automatically generate sentiment lexicons have been proposed.

Secondly, already fixed polarity having strength is needed to be offered to every sentiment word moreover it may be different polarity depends on the domain of application.

Like the sentence "it is very hot today". In this sense of lexicon-based approach, this sentence expresses the negative behavior 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 given by the consumer have many 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 with the help of categorizing the sentiment analysis in the following type.

1. Document Level Sentiment Analysis
2. Sentence Level Sentiment Analysis
3. Aspect Level Sentiment Analysis

**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 analyzed at once and 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 present in the document. In this approach, the document is divided into many 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 analyzed separately.  anyhow this is not the best case to find the sentiment at all because referring to the above example Samsung is a good friend 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, 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,  the 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.

There are different data mining and machine learning approaches are used for sentiment analysis purpose. Every technique has its pros and cons.  these techniques are divided into three major types:

1. **Supervised Learning**
2. **Semi-Supervised Learning**
3. **Unsupervised Learning**

In 1997 researcher [37] defines the machine learning that it is the feature of computer science which aims to gain knowledge from a bunch of data to improve the efficiency of different analysis for example in applied Health Care and Emotion Detection. 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 to solve the machine learning technique that can be used to answer the research question. In research [39], the researcher defines the three types of research problems Descriptive research, Explonatery research, and predictive research. Casual in parents' description and prediction. 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 that such questions validate the results.  let's describe the three terminologies:

1. Descriptive Research:

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

1. Predictive Research:

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

1. Explanatory Research:

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

Machine learning can be of three types:

**Supervised Learning:**

**Unsupervised Learning:**

**Semi-Supervised Learning**

**Unsupervised Learning:**

This type of 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.  This methodology is referred to as unsupervised learning because we don't have the response variable that could happen.

The main purpose of unsupervised learning is to identify or analyze the dimensions of the components trajectories for clusters from the data set. multiple approaches for unsupervised learning are used like Factor analysis mixture modeling and component analysis.

**Semi-Supervised Learning**

Semi-supervised learning consists of both types of unsupervised and supervised learning.  in this technique was the type of data is used like a label or unlabeled data.  the labeled data is utilized to train the model and the unlabeled data is utilized to purify the boundaries of classes.  In semi-supervised machine learning methodology,  K nearest neighbor, perceptron,  neural network, convolutional neural network techniques are used [40][41][42].

**Supervised Learning**

supervised learning is utilized the predictive research because the main purpose of supervised learning is to predict or cast or classify the future outcome of data. supervised machine learning is implemented on a large number of data set like reviews data clinical data to predict the uses satisfaction level for any diseases respectively [43-46], supervised machine learning can be used when prior knowledge of the predicting labels or classes are available and is based on label data.  in this technique first, the algorithm is trained with the help of the labor test data set and that test data set is Leopard on actual results. then the test data set is passed through the model and analyze the efficiency of the model by pointing accuracy.

supervised learning is one of the techniques of machine learning in which Predictive classes are known. This technique is implemented on labeled data set for example in case of the review that may be positive or negative for may be neutral in this example,  the predicting class of a review Would be a negative-positive or neutral.  by following this approach,  the data set is divided into the training dataset and the test dataset. the training of the model by labeling dataset with actual sentiment and then test data set he 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 within unsupervised learning in which there are no classes or labels inside the data. each set of data that is used in supervised or unsupervised machine learning contains a set of features properties of 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. the model 10th in result can be used to predict the classes or label of Testing data.  classification in supervised machine learning is being used in several intelligence base types of research.  a large number of Techniques are using in classification some of the widely used techniques are given below:

Logistic regression K nearest neighbor trees neural networks  SVM  NB

**Decision Tree**

decision tree classifies data set into Tu discreet by using algorithms of structure[51]  the main goal of the decision trees is to show them the information of structure present in a data set a full stop that decision tree technique is a type of supervised machine learning technique which create a decision tree from a set of class labeled data during the process of machine learning [49].  The decision tree algorithm works with the training samples and their classes of labor. then this set of training data is recursively acquitted based on feature values into a subset of data so that the data set in the subset is purer than the data set in the parent set. 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 of 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 for 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 besan classifiers [52].  is called as nave because it's 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 2a a simple framework of production which gives good result in manufacturing cases. then Airways classifier is based on  based theorem  which is as follows:

 𝑃 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

𝑃 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 a logarithms

Not require complicated recursively parameters estimation mechanism that can be applied on a large dataset

 Easy representation of   information

 May not show the best classifier results for a particular application but 808 is robust and well

**Disadvantages**

Theoretical II 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 Neighbors Classification**

 K nearest neighbor is known augmentin object-based learning method Pause it stores all the training samples and not allowed to build new classifier still a new And Labour sample data set needs to be classified.  however lazy learning algorithms demands less computational time in the training phase as compared to learning algorithms like neural networks Bayes networks and decision trees But take care of more computational time am between the classification process[54, 55,50].

 this is 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 approximate proximity[56].  when there is a labor sample is given KNN nearest neighbor points that trend space for the K objects which are closest do 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  understandable

Easy to implement

Give better for which application  that has multiple class labels

**Disadvantage**

Consumes more computational cost when potential neighbors having a large labor 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 structural minimization of lift principle and having a more explaining the decision boundary location Which is also called as a hyper train that produces the optimal classes Separations[57,49,58]. 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 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.

**Logistic Regression**

Logistic regression is a classification methodology that uses classes for creating and uses a single multinomial regression model for a single estimator.

A tree is usually used when the class boundaries at present and it also uses for or probabilities of class depending upon the distance from boundaries.  the ratio of moving towards exchange m from 0 and 1 when the director at 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 Linear Regression vs Logistic Regression

Linear regression can be defined as mathematically:

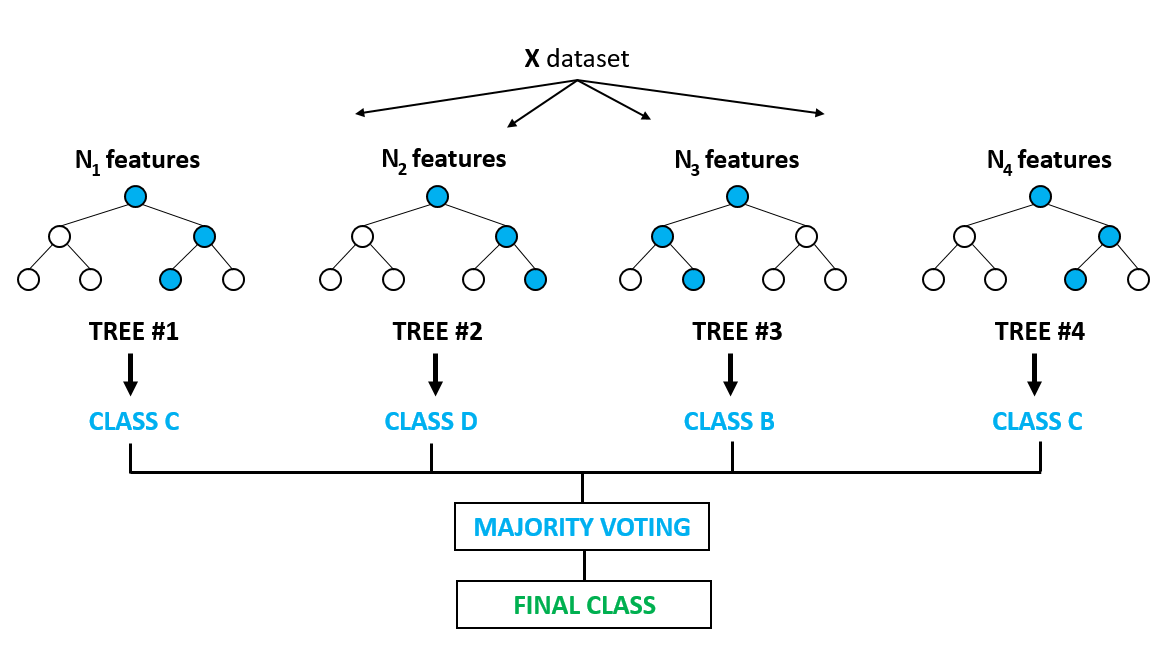
Here, y is the prediction value; bX is the slope and is the intercept. While logistic regression can be define using the linear regression function and it can be defined as using the mathematical equation as:

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

**Random-Forest**

Random forest is a type of classifier which is consisted  of the collection of  tree-structured classifiers h(A, ON) and N=1,2,3, …  where  O N are independently identical ll separated random vectors and every tree what for the most used class at input A.  the best think 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.

**Deep Learning**

Deep learning [59] Is a fart of machine learning which is referred to as a deep neural network.  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 Socher 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.

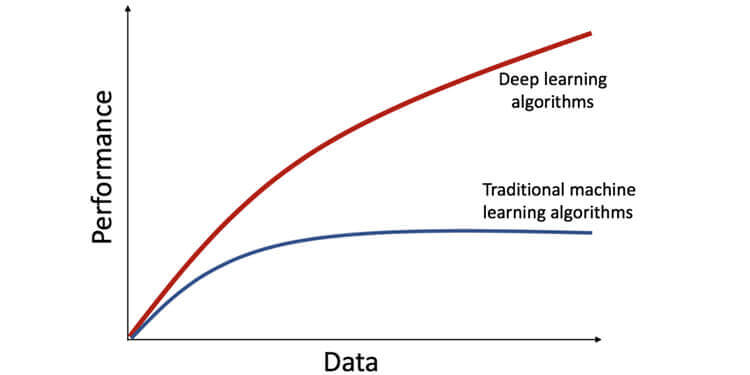


Figure Deep learning vs machine learning

**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 geet training 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 [8].

**Evolution Measures**

To determine the results of the above-mentioned classification techniques, following evaluation measures 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:

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.

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.

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.

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.

# 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] But there are some important aspects of mobiles (price, service, weight, color, and size) that still need to explore which can participate in the growth of the e-commerce business of mobiles. To the best of our knowledge, no work is performed on these aspects for mobile and smartwatches. Lots of researchers have done work in this domain but accuracy 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.

**Area of research**

Sentiment analysis is the study of people`s opinions, sentiments [1]. Sentiment analysis is now being used in multiple research areas. The public is using the web for business correspondence, e-commerce, and e-marketing now [2]. As the online shopping trends are growing, customers are eagerly want to share their emotions and reviews on different platforms on the internet. Extraction of users feeling from the reviews which he provided on the internet has great importance not only for the user of the selection of right product and its variation and rather the organization needs to grow their business to check the customer feedback over their different 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. Research for instance [2] of sentiment analysis on the feedback of goods. It has now become a topic of research that has been widely carried out. By considering this scenario, it is required to analyze the reviews providing by the user on mobile. Multiple types of research have been performed on sentiment analysis at a different level on mobile and smartwatch reviews. In earlier aspect level sentiment analysis is performed on aspects of battery and android version of mobile [6]. In this research, we have performed aspect level sentiment analysis on some important aspects of mobiles (price, service, weight, color, and size) which are still needed to explore.

**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 analyze 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 analyze such a huge amount of data so an automatic system is required which can find the people's sentiments on the product and its aspects.

Lots of researches have been carried on sentiment analysis at aspect level, sentence level, and document level on mobile and smartwatch reviews. But there are still some aspects that are required to analyze. It can help to improve e-commerce business strategies and to improve the customer’s needs according to their requirements.

Research Questions:

1. How to analyze the consumer's sentiments, opinions, and attitude?
2. How to develop a classification model to perform sentiment analysis at an aspect level?

Main Objectives

The main objective of the research:

1. To classify the eBay reviews
2. To scrape data from eBay and amazon
3. To apply different pre-processing techniques for data cleaning.
4. To check the polarity of Public opinions and sentiments.

Main Contribution

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

1. Scrape the reviews of mobile and watch from the eBay and amazon
2. Classification of mobile and watch reviews into negative, positive, and neutral sentiment
3. Data resampling to resolve the model overfitting problem.
4. Purpose the combination of Logistic Regression and Random Forest classifier with the help of Voting Classifier.

**Reference**

1. Rustam, F., Mehmood, A., Ahmad, M., Ullah, S., Khan, D.M. and Choi, G.S., 2020. Classification of shopify app user reviews using novel multi text features. *IEEE Access*, *8*, pp.30234-30244.
2. Wright, R.E., 1995. Logistic regression.
3. Zhang, M.L. and Zhou, Z.H., 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition*, *40*(7), pp.2038-2048.
4. Noble, W.S., 2006. What is a support vector machine?. *Nature biotechnology*, *24*(12), pp.1565-1567.
5. Safavian, S.R. and Landgrebe, D., 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), pp.660-674.
6. <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>
7. Louppe, G., 2014. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.
8. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.