

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



By

Furqan Durrani
Enrolment No: 03-243191-025

Supervisor: Dr. Muhammad Asim Qureshi

A thesis submitted to the Department of Computer Science, Faculty of Engineering Sciences,
Bahria University Lahore Campus in the partial fulfilment of the requirements for the award
of the Masters degree of Science in Computer Science

March 2021

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: March 12, 2021

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

DEDICATION

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

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 from 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 thank all my teachers, who contributed in my academic career.

I am grateful to my father Aurangzaib Khan and mother Sanjida Baigum who have provided me utmost support in all aspects of my life. I am also grateful to my siblings and friends who have played key role for accomplishment of this achievement.

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 classification of 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 with others in the comments section. Existing machine learning models provide a useful account of how to judge the sentiment polarity. However, the accuracy of aspect related 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 e-commerce websites eBay and Amazon. After pre-processing, subsets of data have been extracted with respect to the aspects of price, color, size, weight, and service from data, and the proposed model and different machine learning models (Naive Bayes, KNN, SVM, and Random Forest) are applied. It is observed that accuracy has been improved by using the proposed methodology which is 97%.

Table of Contents

Introduction.....	1
1.1 Sentiment Analysis.....	4
1.1.1 Document Level Sentiment Analysis.....	5
1.1.2 Sentence Level Sentiment Analysis.....	5
1.1.3 Aspect Level Sentiment Analysis	6
1.2 Machine Learning.....	6
1.2.1 Research Types	6
1.2.2 Unsupervised Machine Learning	8
1.2.3 Semi-Supervised Machine Learning.....	8
1.2.4 Supervised Machine Learning	8
1.3 Data Resampling	17
1.4 Evolution Measures	18
1.4.1 Accuracy	18
1.4.2 Precision.....	19
1.4.3 Recall	19
1.4.4 F-Measure	19
1.5 Research Gap	19
1.6 Motivation.....	20
1.7 Problem Statement.....	20
1.8 Research Questions:	21
1.9 Main Objectives	21
1.10 Main Contribution.....	21
Literature Review	22
2.1 Aspect Level Sentiment Analysis.....	32

Methodology	36
3.1 Dataset description	36
3.2 Pre-processing of data	38
3.2.1 Tokenization.....	39
3.2.2 Punctuation Removal.....	40
3.2.3 Numeric Removal.....	40
3.2.4 Lowercase Conversion	41
3.2.5 Stemming	42
3.2.6 Stop words Removal	42
3.3 Features Extraction	43
3.4 Hyperparameters for classifiers	45
3.5 Experimental Flow	46
Result and Discussion	50
4.1 LR Results	50
4.2 RF Results	53
4.3 SVM Results	56
4.4 KNN Results.....	58
4.5 GNB Results.....	61
4.6 Voting Classifier Results	63

LIST OF FIGURES

Figure 1.1: Types of Sentiment Analysis.....	5
Figure 1.2: Types of Machine Learning	7
Figure 1.3: SVM hyperplane representation	13
Figure 1.4: Random Forest	14
Figure 1.5: Linear Regression vs Logistic Regression	15
Figure 1.6: Deep learning vs machine learning	17
Figure 3.1: Methodology	47

LIST OF TABLES

Table 3.1: Sample Dataset	37
Table 3.2: Dataset attribute description	38
Table 3.3: Data before and after tokenization.....	39
Table 3.4: Data before and after punctuation removal	40
Table 3.5: Data before and after numeric removal	41
Table 3.6: Data before and after converting to lower case	41
Table 3.7: Data before and after stemming.....	42
Table 3.8: Data before and after removing stop words.....	43
Table 3.9: Data before and after pre-processing	43
Table 3.10: Hyperparameters for classifier.....	45
Table 3.11: Aspects based reviews count	48
Table 3.12: Sample of data after labeling	48
Table 3.13: Sentiment count in the dataset	48
Table 3.14: Sentiment count in the dataset after oversampling	49
Table 4.1: LR results for size aspect review dataset.....	51
Table 4.2: LR results for color aspect review dataset.....	51
Table 4.3: LR results for price aspect review dataset	52
Table 4.4: LR results for service aspect review dataset.....	52
Table 4.5: LR results for weight aspect review dataset	52
Table 4.6: LR results using 10 fold cross-validation	53
Table 4.7: RF results for size aspect review dataset	54
Table 4.8: RF results for color aspect review dataset	54
Table 4.9: RF results for price aspect review dataset	54
Table 4.10: RF results for service aspect review dataset	55
Table 4.11: RF results for weight aspect review dataset	55
Table 4.12: RF results using 10 fold cross-validation	56
Table 4.13: SVM results for size aspect review dataset	56
Table 4.14: SVM results for color aspect review dataset	57
Table 4.15: SVM results for price aspect review dataset.....	57
Table 4.16: SVM results for service aspect review dataset	57
Table 4.17: SVM results for weight aspect review dataset.....	58
Table 4.18: SVM results using 10 fold cross-validation.....	58

Table 4.19: SVM results for size aspect review dataset	59
Table 4.20: KNN results for color aspect review dataset	59
Table 4.21: KNN results for price aspect review dataset.....	59
Table 4.22: KNN results for service aspect review dataset	60
Table 4.23: KNN results for service aspect review dataset	60
Table 4.24: KNN results using 10 fold cross-validation.....	60
Table 4.25: GNB results for size aspect review dataset.....	61
Table 4.26: GNB results for color aspect review dataset.....	61
Table 4.27: GNB results for price aspect review dataset.....	62
Table 4.28: GNB results for service aspect review dataset	62
Table 4.29: GNB results for weight aspect review dataset.....	62
Table 4.30: GNB results using 10 fold cross-validation.....	63
Table 4.31: voting classifier results for size aspect review dataset	63
Table 4.32: voting classifier results for color aspect review dataset	64
Table 4.33: voting classifier results for price aspect review dataset.....	64
Table 4.34: voting classifier results for service aspect review dataset	64
Table 4.35: voting classifier results for weight aspect review dataset.....	65
Table 4.36: voting classifier results using 10 fold cross-validation.....	65

LIST OF ABBREVIATIONS

CNN	-	Convolutional Neural Network
LR	-	Logistic regression
SVM	-	Support Vector Machine
RF	-	Random Forest
VC	-	Voting Classifier
SA	-	Sentiment Analysis
DLSA	-	Document Level Sentiment Analysis
SLSA	-	Sentence Level Sentiment Analysis
ALSA	-	Aspect Level Sentiment Analysis
IEEE	-	Institute of Electrical and Electronics Engineers
ANN	-	Artificial Neural network
RNN	-	Recurrent Neural Network

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
-----------------	--------------	-------------

Chapter 1

Introduction

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 [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 [7]. Because of their influence and prevalence, online reviews have considerably attracted the attention of both practitioners and researchers [6].

Sentiment analysis is the study of people's opinions, sentiments, attitudes, 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 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 analyzing 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 analyze 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. 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 [26]. 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 analyze the reviews. Multiple Data mining and Machine Learning approaches are present for the resolution of this problem [27]. 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 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 labor-intensive job. Secondly, already static polarity is needed to be provided 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 behavior. 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 is still available. But these machine learning-based models are required a large dataset with pre-defined 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 anything which he has purchased earlier.

For understanding this problem, there are 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 following techniques mentioned in figure 1[33].

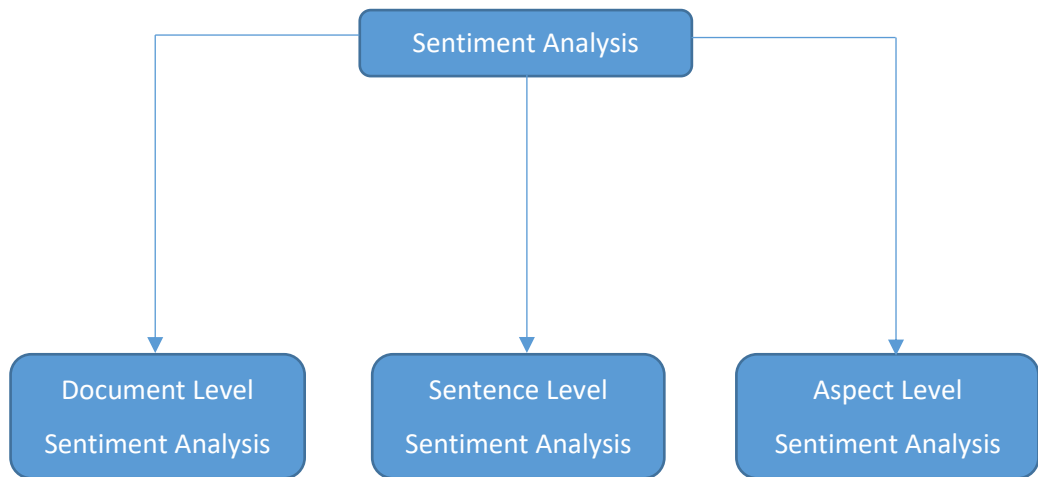


Figure 1.1: Types of Sentiment Analysis

1.1.1 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 analyzed at once. The opinion about the whole document is considered as positive or negative in this type of analysis. However, this is not a good approach because there may be a positive specific path that has great importance but the overall sentiment score of the document is negative and vice versa [28] [34].

1.1.2 Sentence Level Sentiment Analysis

Sentiment analysis at the sentence level is considered as the calculation of sentiment on 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 analyzed 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 [35][36][37].

1.1.3 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.

1.2 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. 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 questions validate the results.

1.2.1 Research Types

Research has been divided into the following three types:

1.2.1.1 Descriptive Research:

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

1.2.1.2 Predictive Research:

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

1.2.1.3 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 into three types mentioned in figure 2 [Reference].

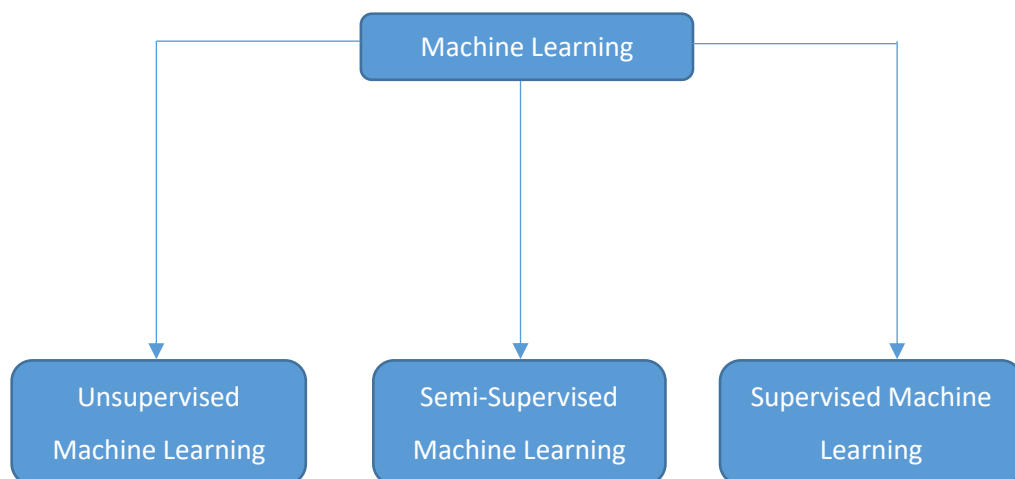


Figure 1.2: Types of Machine Learning

1.2.2 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 statistical 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 analyze the dimensions of the component's trajectories for clusters from the dataset. Multiple approaches for unsupervised learning are used i.e. Factor analysis, mixture modeling, 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.

1.2.3 Semi-Supervised Machine Learning

Semi-supervised learning consists of both types of unsupervised and supervised learning. In this technique, the dataset can be labeled or unlabelled. The labeled 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 neighbor, perceptron, neural network, convolutional neural network techniques are used [43] [44].

1.2.4 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 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 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 labeling 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]

1.2.4.1 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:

1.2.4.1.1 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 labeled data with the help of the machine learning process [50]. The decision tree algorithm works with the training samples and their labeled 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 the decision tree explains feature and every branch represents the outcome of the test and every node explains the class label [52].

Advantages

- Simple and fast.
- No requirement for prior knowledge and ability to manage high-dimensional data.
- Its representation is understandable
- Support 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 replication problem.
- Orders of the features intense are affected on the performance.

1.2.4.1.2 Naive Bayes classifier

Naive Bayes classifier is one of the simple statistical Bayesian 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 Bayes theorem which is as follows:

$$P(A|B) = \frac{P(X|A) P(A)}{P(X|B)}$$

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 Naive Bayes 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.

1.2.4.1.3 K-Nearest Neighbours

K nearest neighbor 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 lie 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].

Advantages

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

Disadvantage

- Consumes more computational cost when potential neighbors having a large labeled sample.
- Classification time is low.
- It assigns the same weight to all features but this maybe comes to the confusion where there may be many irrelevant features in the data and in this way the accuracy is affected.
- It is sensitive for local data structure
- Require more storage.

1.2.4.1.4 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. The best hyperplane selection is shown in Figure 1.

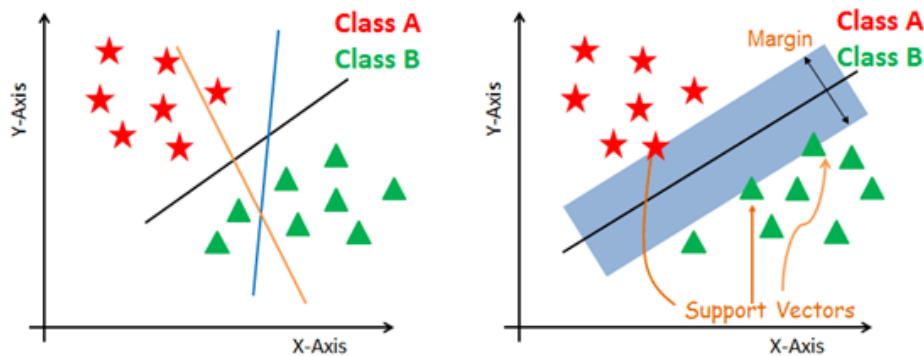


Figure 1.3: SVM hyperplane representation

Advantages

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

Disadvantages:

- Computationally expensive.
- Require a large amount of time for training data.
- Require a large amount of memory as well.

1.2.4.1.5 Random-Forest

Random forest is a type of classifier which is consisted of the collection of tree-structured classifiers $h(A, O_n)$ and $n=1,2,3, \dots$ where O_n 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].

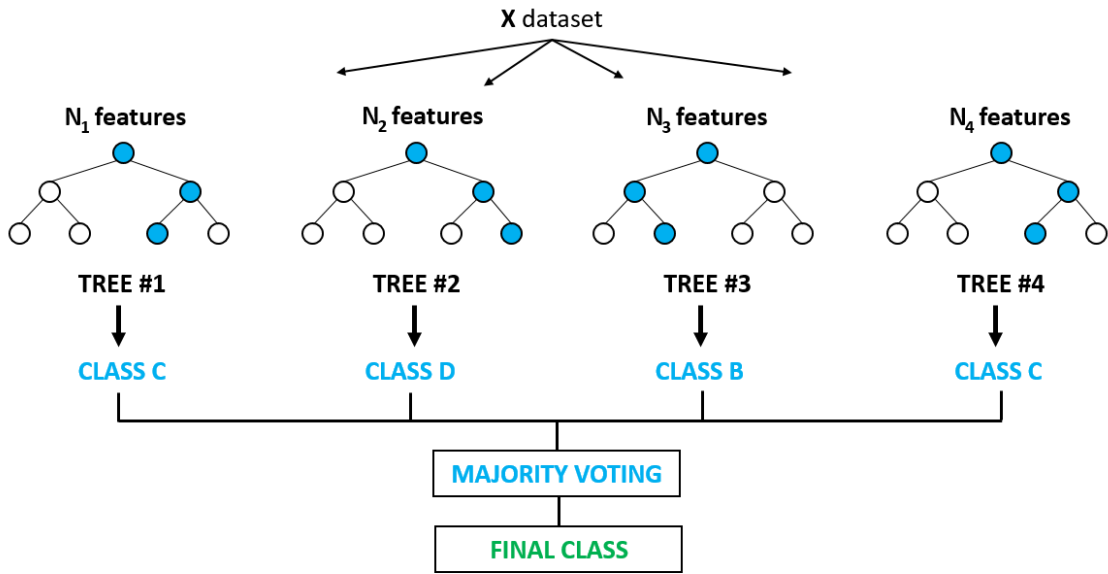


Figure 1.4: Random Forest

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:

$$dts = dt_1, dt_2, dt_3, \dots, dt_n$$

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

$$rf = mode\{dt_1, dt_2, dt_3, \dots, dt_n\}$$

$$rf = \sum_{i=0}^N dt_i$$

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.

1.2.4.1.6 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. The difference between linear regression and logistic regression is shown in Figure 2 [65].

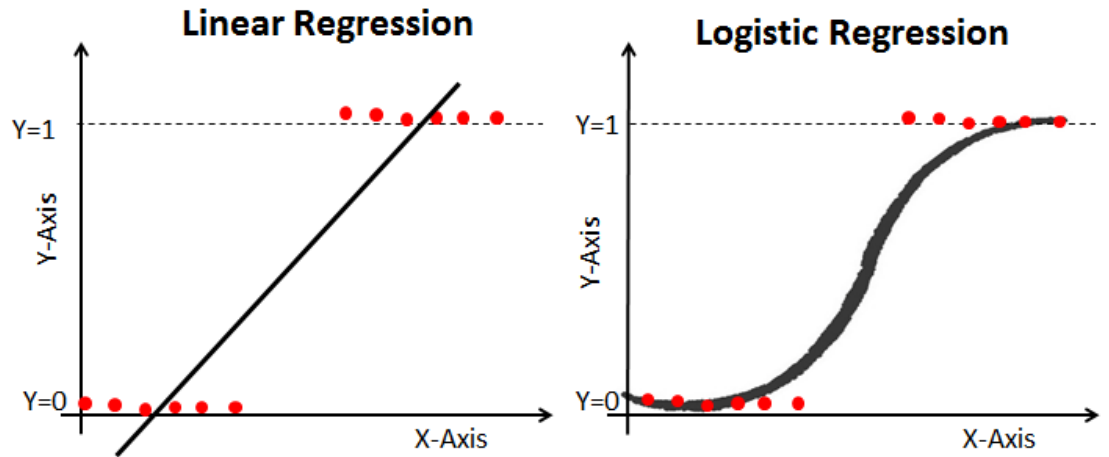


Figure 1.5: Linear Regression vs Logistic Regression

Linear regression can be defined as mathematically:

$$\hat{Y} = bX + a$$

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:

$$p = \frac{e^{bX+a}}{1 + e^{bX+a}}$$

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

1.2.4.1.7 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 modeling, 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.

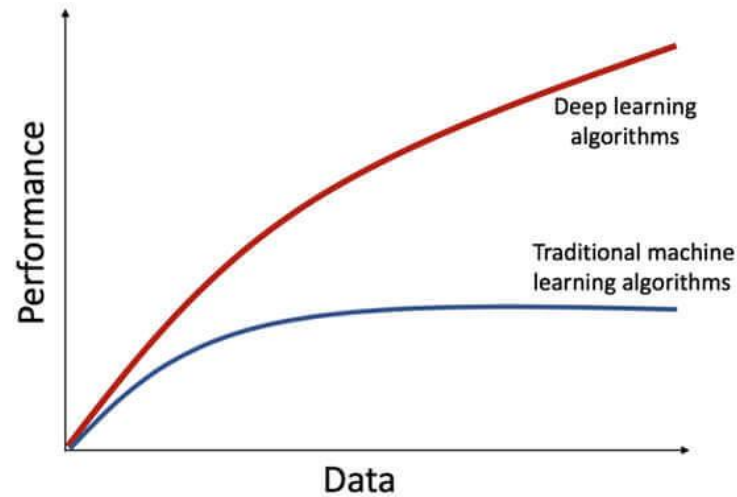


Figure 1.6: Deep learning vs machine learning

1.3 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].

1.4 Evolution Measures

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

1.4.1 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:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Or Accuracy can also be defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

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.

1.4.2 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.

$$Precision = \frac{TP}{TP + FP}$$

1.4.3 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.

$$Recall = \frac{TP}{TP + FN}$$

1.4.4 F-Measure

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

$$F\ measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

1.5 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 [31], 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 at aspect-based sentiment analysis.

1.6 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.

1.7 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 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 analyze the reviews and predict the result accurately. So, it is required to increase the accuracy of the system.

1.8 Research Questions:

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

1.9 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.

1.10 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 computer science which keeps track of all the activities in the specific area. The main reason for this analysis is to analyze the behavior of buyers, and observers of any product. With time, data is also increasing day by day on the internet and it is very difficult to analyze each review manually. To analyze the customer's sentiment, sentimental analysis comes into play. Many researches are already carried out on this topic.

In research [67] researchers extracted sentiments from the reviews and analyze the outcome to create a business model. Researchers claimed that the presented model is too much robust and has provided high accuracy. They performed analysis using Genderizer and Textblob to detect fake emotions and create a classifier to check the model's accuracy. They performed their analysis on python and R programming languages. 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 [68] the researchers used current supervised learning (Naive Bayes, Perceptron, and Multiclass SVM) algorithms to forecast ratings of a provided mathematical scale using only reviews. They collect the 1125458 reviews from the 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 analyze the results.

The researchers in [69] presented the system which classifies the customer reviews and presented the results in charts to visualization. They scrapped the data from amazon through URL and pre-processed it. In this paper, they have applied NB, SVM, and maximum entropy. The proposed methodology integrates the current sentiment analysis techniques and is claimed to increase the accuracy. They presented their result in the form of charts.

In the paper [70] authors built a model for predicting the product ratings based on text using a bag-of-words. These models utilized unigrams and bigrams techniques. They used 2,982,356 reviews of 252,331 unique products of Amazon. They performed analytical operations on the data to find the features and better understanding of features of the product. Between unigrams and bigrams, unigrams produced the most accurate result. And popular unigrams were an extremely useful predictor for ratings for their larger variance. Unigram results had a 15.89% better performance than bigrams.

In paper [71] different feature selection or extraction methods for sentiment analysis are discussed. They used the amazon dataset for feature selection and classification of sentiment. They performed pre-processing to remove special characters and stop words. They performed multiword, phrase level, and single word feature extraction and selection techniques. They used Naive Bayes as the classifier. They determined that Naive Bayes performed better at phrase level as compared to a multiword and single word. The main disadvantage of this research is, authors performed a Naive Bayes classifier only and we cannot extract a satisfactory result from which.

In this paper [72] they used easier algorithms to understand the problem easily. They collected the data of Hotel Reviews. The system gave the highest accuracy 93.50% on SVM having N-gram aspects as compared to other used algorithms logistic regression and decision trees method.

In paper [73], tf-idf is used here as an additional experiment. The main contribution of the researchers in this research is to predict rating and to separate the positive and negative words from the dataset with a linear regression model. They evaluate the model with root mean square error. They determined that Yelp reviews could be utilized for rating prediction with the help of the 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 [74], researchers described which feature of the product is good to expect the aspect-level sentiment analysis and explained the reason why this is in this way. In the 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 the 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 for each feature. Thus, this is happened by the use of a training data system.

In [75], researchers analyzed the features with the help of Information Gain, which is mostly used in conjunction with the measurement of feature selection. This approach is working in the 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 whole study concludes 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 [76], the researchers used a Convolution Neural Network for the performing of sentiment level. For aspect level sentiment classification, the authors proposed a convolution neural network. Their model first developed the CNN to extract the aspect and then applied another approach named as order labeling methodology with Conditional Random Fields (CRF) for the discovery of opinions. Then at the final point, they collect features with each word and implement the convolution neural network and define the sentiment with respect to aspect. They presented a deep learning 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 [77], they try to present the ABSA on movie review 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 labeling (M), 2: clustering(C), 3: review guided clustering (RC),

They take the 1000 movies data from the IMDB website. In this work, they have tested the efficiency of the approaches on individual sentiments. Moreover, the efficiency of cleaning the 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 [78] the researcher performed a sentiment analysis on music. In this paper, the 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. The authors used 70 % of the total data set to train the model and 30% was used for testing purposes. Researchers performed to analyze the songs on the appended subject like Baseline, Shape, and Contrast of the one thousand songs. The analysis will help the readers to analyze the consequence of emotions and observation of actuality while listening to music.

In [79] 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 with the help 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 songs. It will help to search the related songs to add them in the cluster and to analyze 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 techniques K-mean was utilized to create the cluster of the songs of jazz music and other music having different audio features of songs. Later, researchers have a plan to expand their research with the non-supervised algorithm.

In [80] researchers have worked on the integration of data on dissimilar datasets. Dataset, which is used in this research, consisted of national products of the China and United States for many years. The dataset was extracted from different central repositories. The dataset consisted of raw data and information as well. The dataset consisted of the rates of currency exchange Yuan to Dollar and vice versa. In the first iteration, currencies were exchanged with the help of math rules 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 [81] researchers performed research on data about the dataset. The dataset contained the actual standards for short-term learning. The dataset consisted of two standards, the first standard consisted of 1623 characters and the second standard

consisted of 600 samples. 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 the closest actual class. Moreover, Fine-Tune cannot do so. Researchers obtained greater accuracy with the help of K-NN of 88.42 as compared to Fine-Tune which was 73.88 %.

In [82] researchers performed different modal based classification on Hindi and Western songs. They used two types of the dataset for analysis. The first dataset consisted of audio song clips and the second dataset consisted of sentiments with eight different classes. The first dataset of AMC consisted of audio song clips while the second dataset consisted of sentiment with 8 different sentiment classes. The dataset of Hindi songs consisted of 500 songs and the second dataset consisted of 1753 audio clips. The dataset of Western songs consisted of 298 songs while 1111 clips in mp3 format. The authors utilized a supervised Machine Learning algorithm for classification purposes like SVM (Support Vector Machine) and FFNNS (Feed-forward neural networks). To determine the accuracy of classification, Ten-Fold validation was utilized. They obtained precision 58.9 % and 59.1% using SVM while on Western music they obtained 70.5 % precision. With the FFNNs algorithm, the precision value was 65.3%, 65.2% of F-measure, and 65.1 of Recall on Hindi music.

In [83] 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 purposes, the 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 Kaggle. For this purpose supervised machine learning approach was used to analyze a dataset of 25000.

In [84] Sentiment Analysis was performed on the aspect level. Mobile reviews dataset from Amazon Web Reviews used for analysis. After data pre-processing, they selected 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 comments 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 [85] Aspect level Sentiment Analysis performed on movie reviews. They classify the sentiments at the aspect level and the 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.

In [38] survey was carried on Aspect level Sentiment Analysis to aggregate the people's opinions on the entities that were mentioned within the document. At aspect level Sentiment Analysis, a single entity was analyzed 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 of Aspect Level Sentiment Analysis. Here they can discuss different ways of aspect detection like frequency-based and syntax-based aspect detection. After that in this survey, they talked about the classification methods of supervised and unsupervised. The supervised classification is used for the labeled data for training and testing both. Unsupervised classification is required to operate labeled 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 [86] solution was purposed of forms and function for Roman Urdu dataset, that was collected by surveying the local universities of Pakistan. 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 Ph.D. scholars. The messages they collected in the form of text messages, the corpus has a total of 4, 46,483 words. In their survey, they see the people prefer to write their messages using the Roman Urdu type of writing. Accordingly to their study, they analyze the female data was less romantic words than males. And the students of undergraduates have used more friendly words than a graduate students. 73 users used 20 or less friendly words those called low romantic participants. The 41 users were who used 81 or more intimate words 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 [87] emotion ontology was generated for Roma Urdu text data. Dataset was collected from different blogs by using a scraper that contains people's emotions. They classified the emotions into 5 classes (Happiness, angry, hurt, caring, and fear). After data collection, they parse the data through a syntax analyzer that recognizes the syntax structure and contracts a phrase tree, and modify the Figures into the required order. After semantic analyzer through JENA API and checking the ontology of the document classified according to their classes in happiness, angry, hurt, caring, and in fear. They experimented on four documents named (DOC1, 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%. Document 3 has 54 samples and correctly classified 46 results with recalled was 85.18% and with 93.86% with precession and document 4 carried 38 samples 'in 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 [88] 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 the continuous text. They also talked about TF & IDF. Term frequency a word present in a document and inverse document frequency talked about a word that was repeated in multiple documents. Through the paper; they tried to help the people in the field of text mining.

In [89] Sentiment Analysis was performed on multi-languages data Twitter dataset. To perform sentiment analysis, they selected multi-language tweets associated with PGE 2013. The dataset was collected from the users of five capitals of Pakistan and

tweets were belonged to legendary political parties of Pakistan. The dataset was gathered with a scrapper from 2001 to 2013. According to the results of Urdu and English tweets,

In [90] Sentiment Analysis was performed at the 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 to the extraction of data i.e. (Tokenization, Part of Speech tagging, and Lemmatization) to prepare the data and to remove the outliers as well. A spam detection model was used to avoid model 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 that has one or two star ratings classified as a low product, which a product rating was three classified as medium, and a 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 [91], a survey was performed on the supervised and unsupervised classification of the document. They classified document classification in three basic methods named rule-based classification, supervised classification, and unsupervised classification. In the rule-based classification of the documents, documents are grouped by predefined rules. This approach is good for small documents and rules are also decided by the writers. In supervised classification, the document was classified based on supervised learning. In this technique, the model is trained using training data which was known as algorithm learning. After model training, the testing data is passed through the model. The unsupervised classification was a method in which classification was achieved through clustering. It simply clusters the data according to the structure or pattern of the data. Unsupervised algorithms worked on a centered based approach, in centered based techniques each document "D" represented Documents Vector VD and centered vector of each class, and Euclidean distance between VD and centered Vector of Class was calculated. Documents having a minimum distance from centered assign a particular class.

Some of the Supervised and Unsupervised algorithms were also discussed in this paper. SVM analyses the data and classifies them after recognizing the pattern. Naive

Bayes classify the document by the calculation of posterior probability value and classify the documents according to their frequencies. Decision Tree uses a tree-based algorithm to classify the documents. In unsupervised classification, they discussed partitioned clustering. Partitional clustering algorithms develop un-nested and non-overlapping partitions of the documents. It works like first K clusters are defined and partition (P) was constructed, then clustered is redefined by moving the documents from one cluster to another iteratively. In K-mean clustering, K clusters are defined and each document move to that cluster which was near to its centered, Hierarchical clustering techniques make a cluster from top to down and 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.

In [92], research was carried on the Amazon dataset and it is organized 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 they converted the dataset into a bag of words. Pre-processing is a significant process in the field of views mining and sentiment analysis. Each sentence was analyzed and calculated by sentiment score. For calculating the sentiment score they performed a comparison between dataset and lexicon sentiments. The dataset was compared with sentiment lexicons having 2006 positive and 4783 negative review words and sentiment scores were calculated. The authors used various types of features, learning algorithms, several accuracy measurements. In this work, researchers applied different approaches as the lexicon approach, a dictionary-based approach that was usable within learning techniques. The sentiment analysis was applied to each review of products and then used the machine learning algorithms like SVM as well as NB. Naïve Bayes grew 98.17% accurateness of Camera reviews, while on the other hand Support Vector Machine got 93.

In this study [93], 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 better results for error reduction and accuracy. They proposed a method for review extraction and categorization. The purposed methodology consisted of five steps: Data Source: they

collected the most recent reviews from Rediff, Rottentomatoes, Mouthshunt, and Bollywoodhungama. Data Pre-processing: it is contained on transformation, tokenization, and removal of stop words, and mining of opinion words. Feature Selection: the system of selecting features is known as feature selection and it is called 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 weights are allocated sequentially. Sentiment Classification: it is performed by learning the model from the training dataset and classifying the data based on the trained model. The results explained that the proposed methodology showed greater accuracy than the SVM with 0.65 split ratio.

In this paper [94], the dataset was gathered by Kaggle that was consisted of food reviews collected from Amazon from October 1999 to October 2012 (29,30). Dataset consisted of a huge number of data like reviews are contained 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 the steps: took out the URLs as (www.abc.com), all tags like (#topic), removed all screen name such as (@username), took away all the punctuation marks, symbols as well as numbers, removed 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 were 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 [94], the amazon dataset 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. They explained that each sentiment sentence 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 parts of speech is the following: 'Noun, pronoun, verb, adverb, adjective, conjunction, interjection, and the preposition'. In natural language, part-of-speech (POS) producers organize the words based on POS. Some researchers collect moreover 500 sentiment of reviews on products that is related to 4 major forms: Flash drives, Computers, Mobiles, and Electronics. These were online reviews and set by almost 3.2 million people in front of 10,001 products. They proposed the process of sentiment polarity classification and POS tagging has been described along with detailed descriptions of each step.

2.1 Aspect Level Sentiment Analysis

Sentiment analysis is a fundamental task of natural language processing used to determine whether the sentiments portrayed in data are positive, negative, or neutral. Sentiment analysis is often performed to help businesses to gather customer feedback on their products and service or whether comprehend customer needs. It has been proved to be a quite efficient text mining technique while aspect-level sentiment analysis is an even more improved technique often implemented on textual data to perform a detailed analysis on customer feedback that helps businesses to produce products and services that are up to customer satisfaction and needs.

Aspect-level sentiment analysis is a text mining technique that groups data by aspect and categorizes the sentiment to each aspect of data. Aspect-level sentiment analyses are performed to evaluate customer feedback by integrating certain opinions on several aspects of any product or service. Aspects are the components or attributes of a product such as a user's review after experiencing a specific product or service, any feature of a product, or customer service provided by an organization. Aspect level sentiment analysis is specifically the positive or negative opinion on a particular aspect that can be any feature or category of a product or service.

Recently, Xiaodi Wang et al., declared that the position features can improve the performance of the model on four different datasets. Therefore, the study proposes a classification model by combining attention mechanism and position features named Multi-Level Interactive Bidirectional Gated Recurrent Unit (MI-biGRU). Word

embeddings are developed by position features of a word in a sentence so that the used approach can extract context and features of the targeted term by utilizing a well-constructed model, at last, to pay more attention to the words that are significant for sentiment classification an attention mechanism is used. The correlation between position features and the multilevel interactive attention network has been shown through experimental results [95].

A study focuses on retrieving aspect terms from each record extracted from Amazon customer review data by identifying the Parts-of-Speech and applying classification models including Naïve Bayes and Support vector machines to classify the sentiments from data. while performing the aspect level sentiment analysis feature terms of a product are the key target which depends on the product attributes. In the study, the aspect level sentiment analysis is categorized in three steps; identification, classification, and aggregation. Experiments are performed through applying preprocessing mainly Part-of-speech tagging to each term in a sentence and frequently used words are extracted. Later frequently used aspects are extracted from data through the Apriori feature extraction algorithm and opinion words such as a set of adjectives that best describe the aspect of the product are extracted after applying feature pruning. At last classification algorithms are applied to classify the sentiments into labels including positive, negative, and neutral [33].

In a study by Puspita Kencana Sari et al. Tokopedia review data is used for aspect-level sentiment analysis. The user reviews are classified based on five e-Servqual dimensions including; personalization, reliability, responsiveness, trust, web design, and sentiment analysis positive and negative. The naïve Bayes classification method is applied for the classification of the data because of its support to process large data and high-level accuracy [96].

Pratik P. Patile et al. focused on polarity prediction using a large data set extracted from Amazon customer reviews labeled into three classes such as positive, negative, or neutral. The classification task is performed by logistic regression, naïve Bayes, and support vector machines. Before performing classification, the data has been cleansed by applying some preprocessing methods including tokenization, stop word removal, and stemming. LDA and k-means algorithms are applied to extract and cluster the aspects or topics. The

experimental results show that logistic regression has outperformed the other classifiers applied [97].

Ashwath et al. performed a study on aspect-level sentiment analysis by using different notion examination techniques on client audit data. data has been preprocessed by applying stemming, stop word removal techniques and POS tagging is used to extract mainly used terms from the data, that has been further classified by two states of the art machine learning models including, naïve Bayes and support vector machines where Naïve Bayes has the high accuracy score than the other model used [36].

A study proposed a pre-trained language model based on transformer bidirectional encoder representation named SA-BERT. Semantic information of the context of data is encoded into a word vector by Bert. While the text features were extracted by using the attention mechanism on a deeper level to comprehend the semantics of the text information and to accomplish the sentiment analysis task of e-commerce data. The experiments are performed on a JD mobile review dataset that shows the efficiency of the SA-BERT model in aspect-level sentiment analysis [98].

A study focused on developing a hybrid method for aspect-level sentiment analysis with the combination of machine learning approaches and lexicons. It has been proved that the word-level analysis by word cloud visualization provides primary results regarding the customer opinion on a product or service. For review classification, two lexicons named Syuzhet and Sentiment are compared at the sentence level to classify the reviews. Syuzhet package has provided better performance and is used to train labeled text provided by it. The experimental results show that the naïve Bayes has a better accuracy score while using the Syuzhet lexicon package among other applied models [38].

Qiang Lu et al. proposed a model for aspect-level sentiment analysis named interactive rule attention network (IRAN). IRAN is proposed to designs a grammar rule encoder by standardizing the output of adjacent positions to simulates the grammatical functions in the sentence. It also learns attention information from context and target by constructing an interaction attention network. The experiments have been performed on the ACL 2014 Twitter dataset and SemEval 2014 dataset. It has been shown that the IRAN can learn effective features and has shown better performance as compared to the traditional models [99].

Another study presents a continuous learning framework based on naïve Bayes for sentiment classification of a large-scale and multi-domain e-commerce product review. The parameter estimation mechanism has been extended in naïve Bayes to support continuous learning. The experiments are performed in two different domains including; cross-domain sentiment classification and Domain-specific sentiment classification. In cross-domain classification, reviews are taken from different domains, and in domain-specific classification, reviews are taken from the same domain [100].

Yue Han et al. proposed a Pretraining and Multi-task learning model in their study based on Double BiGRU namely PM-DBiGRU. In the proposed model short text-level drug review are used to learn pre-trained weight to initialize related weight for the model for sentiment classification task model. After that two BiGRU networks are executed to produce the bidirectional semantic representations of the drug and target review, and target-specific representation is obtained attention mechanism for aspect-level sentiment classification of drug review. To transfer helpful domain knowledge from the corpus multi-task learning is further utilized [101].

A study used a mobile product review dataset for aspect-level sentiment classification. The data has been initially preprocessed by using tokenization and POS tagging. POS tagging has been used to extract the most frequent terms from the data. later the classification has been performed by using four different machine learning models including Bernoulli Naïve bayesian, multinomial naïve Bayesian, k nearest neighbor, and Support vector machines. The experimental results show that the KNN has the highest accuracy among all the other models on APPLE products and has performed comparatively better than the other classifiers [92].

Chapter 3

Methodology

This study is carried out for the classification of sentiment about the aspect of different products and we have used machine learning techniques and methods for this purpose. In this chapter, we have discussed the algorithms, datasets description, and techniques that we have used. First, we describe the used dataset in detail such as the extractions of dataset and number of sample, etc, then we explain the data preprocessing techniques and show each technique results on sample data. Second, we describe the feature engineering techniques with the example implementation next and then we provided the details about used models and provide the hyperparameters setting for each model. In the end, this chapter contains the experimental flow/proposed methodology description and evaluation criteria for the learning models.

3.1 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, unhelpful vote, 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 the 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 the dataset. The Amazon dataset also contains the text reviews of the users on different amazon products from the electronics category. We extract the 252231 reviews from Amazon and 33324 reviews from eBay.

Table 3.1: Sample Dataset

Product Name	Review Date	Review Author	Review Title	Review Text	Helpful Vote	Unhelpful Vote	Review Rating
Apple iPhone 7 - 128GB - Rose Gold (Unlocked) A1778 (GSM)	4-Sep-18	2006yushan	Volume control can be better	One thing that can 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)	5-Jul-19	chealber30	Great features, good quality for a refurbished phone	The phone was exactly as described...	1	1	5

A1778 (GSM)							
----------------	--	--	--	--	--	--	--

Table 3.2: 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.
Helpful Vote	This attribute contains the vote for the product which is helpful for the users.
Unhelpful Vote	This attribute contains the vote for the product.
Review Rating	This attribute contains the rating for the product post by the reviewer.

3.2 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 [102]. 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. Unfavorably, 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[103]. 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 [104]. Usually, initially obtained text data contains the combination of lower-case and upper-case letters, numbers, stop words, punctuations, and various forms of words that have no importance in the classification and rather take 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:

3.2.1 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 in words instead of sentences or paragraphs. In tokenization of the textual data, a sentence is transformed into an array of words or terms [105].

Table 3.3: 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 on 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'.

3.2.2 Punctuation Removal

Punctuation removal is another step of data pre-processing 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 [106].

Table 3.4: 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'

3.2.3 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 [107]. When working with reviews or textual data that is not concerned with the digits, then it is necessary to pre-process 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 3.5: 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'

3.2.4 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[108]. 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 3.6: 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'

3.2.5 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 [109]. 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 3.7: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'

3.2.6 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 pre-processing that removes the useless data for further processing of the data. Stop words are the words used to form a sentence that has no use in text classification and are meaningless to the machine learning models [9]. Stop words include words like “is, am, i, the, to, are, that, they, etc.”

Table 3.8: Data before and after removing stop words

Before Stop words Removal	After Stop words 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 3.9: Data before and after pre-processing

Before Preprocessing	After Preprocessing
I would like to introduce Mr. David as new Sales Manager, he’ll start his job on 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

3.3 Features Extraction

This study used TF-IDF for feature extraction. TF-IDF is an abbreviation of Term Frequency (TF) and Inverse Document Frequency (IDF). TF-IDF is a counting measure

technique that 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 [110]. 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 [111]. 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 [112].

Term frequency can be calculated as:

$$TF(t) = \frac{N}{T}$$

Here the N represents the number of times a term occurs in a document whereas the T represents 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 stop words in the dataset also have low IDF that indicates the feature's low value [113].

Inverse document frequency can be calculated as

$$IDF(t) = \log\left(\frac{N}{1 + D_t}\right)$$

Here D_T 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:

$$TF - IDF = tf(t) * IDF(t)$$

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 thousand 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 documents:

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 times a term occurs in a document and the total number of terms in documents whereas the IDF is a log representation between the division of the total number of documents and the number of terms 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

3.4 Hyperparameters 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 hyperparameters with the hit and trials method. These classification model achieved their best results with these hyperparameters setting.

Table 3.10 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

3.5 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.

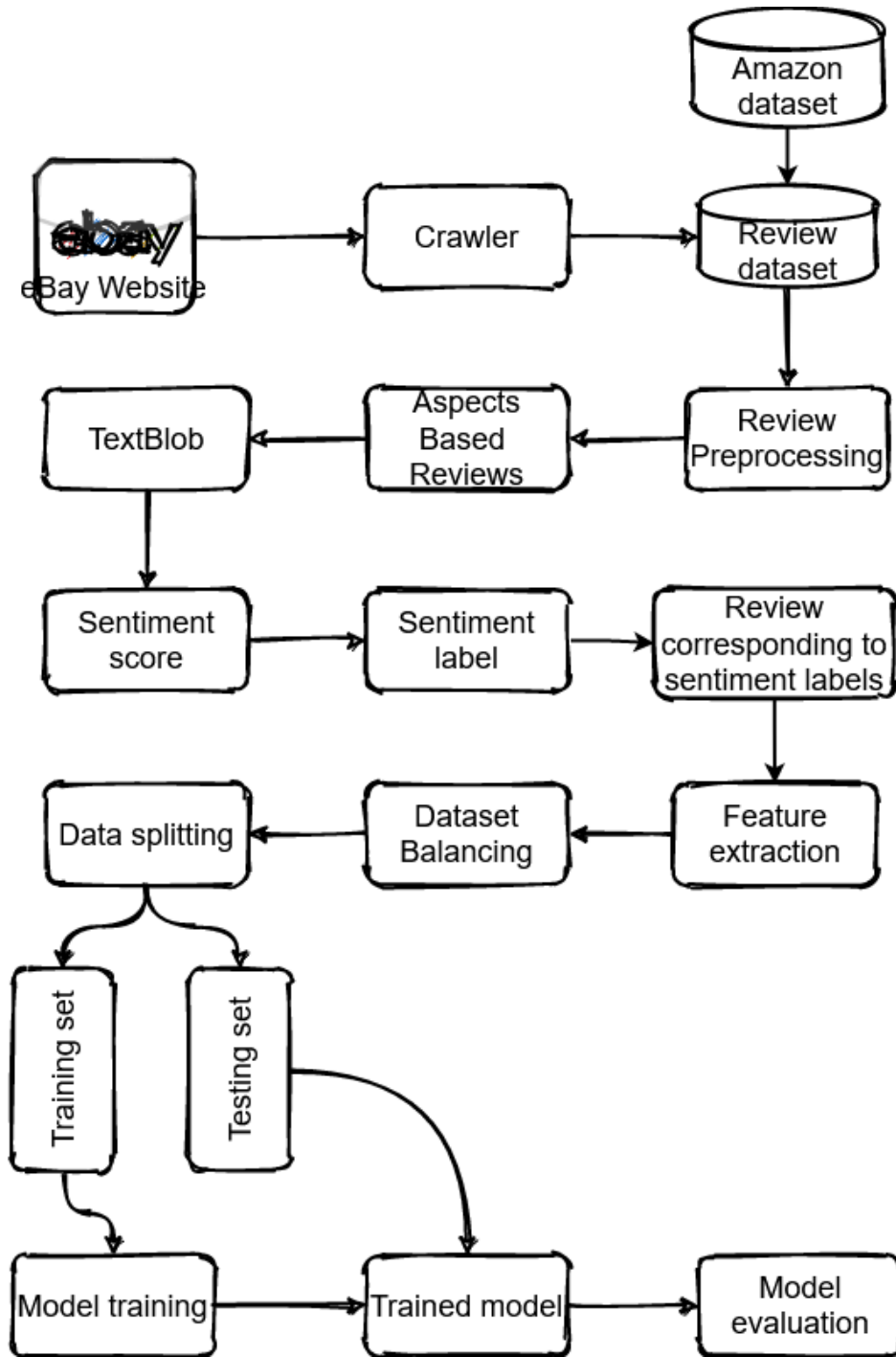


Figure 3.1: Methodology

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 the same electronics products and merge them with the eBay data

to increase the size of the data. After that, we separate the reviews based on aspects such as Color, Size, Weight, Service, and Price.

Table 3.11: 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. Pre-processing removes all raw data such as punctuation, numbers, stopwords, and then we pass this clean data to the text blob 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 3.12: 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 3.13: 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 ratios. 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 alters 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 3.14: 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

This chapter is about the results of machine learning models for aspect-based sentiment classification. We analyze the performance of machine learning models corresponding to each aspect such as price, weight, color, service, and size. The models perform differently on a different aspect dataset. We make an analysis of all performances and done a comparison in this chapter and explain the results in detail. In the end, we have done a detailed discussion about the results and also discuss the significance of the machine learning models. This chapter also contains the 10 Fold cross-validation results of each model on the whole dataset.

4.1 LR Results

This section contains the results for the LR with the undersampling and oversampling technique. Performance of LR improves with the oversampling technique because oversampling generates more features for the learning of LR while the performance of LR on imbalanced dataset show poor performance because the ratio of target classes are not equal and LR gets overfitting on majority class data and show poor performance on minority class data. Undersampling reduces the performance of LR because undersampling randomly deletes the records for data which causes the reduction of features and model get under-fitted and reduce the accuracy. Tables 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6 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 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 color aspect sentiment analysis achieve 0.94. The performance of LR on color aspect data is low as compare to the size reviews that because color 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 sampling, undersampling, and oversampling respectively. The results of color 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 4.1: 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
Undersampling	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 4.2: LR results for color 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
Undersampling	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 4.3: 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
Undersampling	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 4.4: 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
Undersampling	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 4.5: 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
Undersampling	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 4.6: LR results using 10 fold cross-validation

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

4.2 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 color aspect dataset using RF and undersampling. Tables 4.7, 4.8, 4.9, 4.10, 4.11, and 4.12 show the results of RF on all aspect datasets. Table 4.12 shows the performance of 10 fold cross-validation using the RF and results show the significance of RF.

Table 4.7: 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
Undersampling	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 4.8: RF results for color 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
Undersampling	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 4.9: 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
Undersampling	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 4.10: 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
Undersampling	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 4.11: 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
Undersampling	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 4.12: RF results using 10 fold cross-validation

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

4.3 SVM Results

This section contains the results for the SVM model using all aspect dataset with undersampling 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 4.13, 4.14, 4.15, 4.16, 4.17, and 4.18. The price aspect results are shown in Table 4.15 where it shows the highest accuracy score in all three cases such as without sampling, undersampling, and oversampling with 0.91, and 0.90 and 0.97 accuracy scores respectively. Table 4.18 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 4.13: 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
Undersampling	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 4.14: 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
Undersampling	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 4.15: 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
Undersampling	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 4.16: 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
Undersampling	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 4.17: 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
Undersampling	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 4.18: SVM results using 10 fold cross-validation

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

4.4 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 maximum 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 undersampling approach is higher as compare to oversampling. The reason is that the undersampling 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 4.19, 4.20, 4.21, 4.22, 4.23, and 4.24. The highest accuracy of KNN gets through undersampling on color aspect data because the color aspect contains less data or we can say fewer features for training.

Table 4.19: 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
Undersampling	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 4.20: 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
Undersampling	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 4.21: 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
Undersampling	0.54	Negative	0.74	0.06	0.12
		Neutral	0.42	0.97	0.59

Over-sampling	0.41	Positive	0.98	0.55	0.70
		Macro avg	0.71	0.53	0.47
		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 4.22: 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
Undersampling	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 4.23: KNN results for service 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
Undersampling	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 4.24: KNN results using 10 fold cross-validation

Aspect	Accuracy
Size	0.352752974
Color	0.609237102
Price	0.401817967
Service	0.616998901

Weight	0.653043104
--------	-------------

4.5 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 and also categorical. 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 has low fluctuation in accuracy, precision, recall, and F1 score as compared to other models. The results of GNB are presented in Tables 4.25, 4.26, 4.27, 4.28, 4.29, and 4.30. Table 4.30 show that the performance of 10 fold cross-validation to show the significance of GNB.

Table 4.25: 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
Undersampling	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 4.26: 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
Undersampling	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	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 4.27: 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
Undersampling	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 4.28: 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
Undersampling	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 4.29: 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

Undersampling	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 4.30: GNB results using 10 fold cross-validation

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

4.6 Voting Classifier Results

The voting classifier is a combination of two learning model which give the final prediction using both model output. The voting model can outperform as compared to an individual because of its ensemble architecture and can be good on both small and imbalanced datasets as shown below in results. The voting model achieved consistent performance in terms of all evaluation parameters as the voting models achieved the results of their better on size aspects 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 voting model perform all previous results and achieve a high F1 score. The performance of the models can be seen in Table 4.31, 4.32, 4.33, 4.34, 4.35, and 4.36.

Table 4.31: 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
Undersampling	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 4.32: 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
Undersampling	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 4.33: 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
Undersampling	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 4.34: 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

Undersampling	0.82	Positive	0.80	1.00	0.89
		Macro avg	0.88	0.51	0.56
		Negative	0.75	0.73	0.74
		Neutral	0.73	0.91	0.81
		Positive	0.97	0.83	0.89
Over-sampling	0.96	Macro avg	0.82	0.82	0.81
		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 4.35: 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
Undersampling	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 4.36: 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

The result of machine learning models are shown above with each aspect and the results show that all machine learning models perform better on the price aspect dataset as compare to others and reason is that the price aspects dataset is more balanced as compared to all other acts dataset and also have enough features for the good fit of learning models that are the reason each model gives its best score the price aspect dataset. All models have shown the same results with the 10 fold cross-validation which shows that the models are not overfitting. The voting classifier gives its best performance as

compared to all other models even in an imbalanced dataset case also which sho the significance of the proposed ensemble models and shows that the ensemble multiple models can generate more strong results as compare to all other individual models.

\

References

- [1] C. Campbell, L. F. Pitt, M. Parent, and P. R. Berthon, 'Understanding consumer conversations around ads in a Web 2.0 world', *J. Advert.*, vol. 40, no. 1, pp. 87–102, 2011, doi: 10.2753/JOA0091-3367400106.
- [2] M. Tichindelean and S. Vinerean, 'Disclosing the Promising Power of Social Media – an Important Digital Marketing Tool', *Stud. Bus. Econ.*, vol. 6, no. 1, pp. 37–45, 2011.
- [3] S. F. Bernritter, P. W. J. Verlegh, and E. G. Smit, 'Why Nonprofits Are Easier to Endorse on Social Media: The Roles of Warmth and Brand Symbolism', *J. Interact. Mark.*, vol. 33, pp. 27–42, 2016, doi: 10.1016/j.intmar.2015.10.002.
- [4] E. Maslowska, E. C. Malthouse, and S. F. Bernritter, 'Too good to be true: The role of online reviews' features in probability to buy', *Int. J. Advert.*, vol. 36, no. 1, pp. 142–163, 2017, doi: 10.1080/02650487.2016.1195622.
- [5] G. Cui, H. K. Lui, and X. Guo, 'The effect of online consumer reviews on new product sales', *Int. J. Electron. Commer.*, vol. 17, no. 1, pp. 39–58, 2012, doi: 10.2753/JEC1086-4415170102.
- [6] M. Information, 'Mis Technology in Electronic Markets ', vol. 26, no. 3, pp. 243–268, 2002.
- [7] J. Lee, D. H. Park, and I. Han, 'The different effects of online consumer reviews on consumers' purchase intentions depending on trust in online shopping malls: An advertising perspective', *Internet Res.*, vol. 21, no. 2, pp. 187–206, 2011, doi: 10.1108/10662241111123766.
- [8] T. Ahmad and M. N. Doja, 'Ranking System for Opinion Mining of Features from Review Documents', *Int. J. Comput. Sci. Issues*, vol. 9, no. 4, pp. 440–447, 2012.
- [9] S. Aravindan, 'Feature Extraction and Opinion Mining in Online Product Reviews', 2014, doi: 10.1109/ICIT.2014.72.
- [10] H. Xia, Y. Yang, X. Pan, Z. Zhang, and W. An, 'Sentiment analysis for online reviews using conditional random fields and support vector machines', *Electron. Commer. Res.*, vol. 20, no. 2, pp. 343–360, 2020, doi: 10.1007/s10660-019-09354-7.
- [11] G. Choi, S. Oh, and H. Kim, 'Improving document-level sentiment classification using importance of sentences', *Entropy*, vol. 22, no. 12, pp. 1–11, 2020, doi: 10.3390/e22121336.
- [12] S. Baccianella, A. Esuli, and F. Sebastiani, 'SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining', *Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010*, vol. 0, pp. 2200–2204, 2010.
- [13] M. Hu and B. Liu, 'Mining and summarizing customer reviews', *KDD-2004 - Proc. Tenth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 168–177, 2004, doi: 10.1145/1014052.1014073.
- [14] A. A. K. Author and P. P. A. D. Co-author, 'Review Analyzer Analysis of Product reviews on WEKA Classifiers', 2015.
- [15] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, 'Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France', *New Media Soc.*, vol. 16, no.

- 2, pp. 340–358, 2014, doi: 10.1177/1461444813480466.
- [16] A. K. Swain and R. Q. Cao, ‘Using sentiment analysis to improve supply chain intelligence’, *Inf. Syst. Front.*, vol. 21, no. 2, pp. 469–484, 2019, doi: 10.1007/s10796-017-9762-2.
 - [17] Q. A. Al-Radaideh and G. Y. Al-Qudah, ‘Application of Rough Set-Based Feature Selection for Arabic Sentiment Analysis’, *Cognit. Comput.*, vol. 9, no. 4, pp. 436–445, 2017, doi: 10.1007/s12559-017-9477-1.
 - [18] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, and E. A. Villaseñor, ‘A case study of Spanish text transformations for twitter sentiment analysis’, *Expert Syst. Appl.*, vol. 81, pp. 457–471, 2017, doi: 10.1016/j.eswa.2017.03.071.
 - [19] S. K. Sahu, P. Behera, D. P. Mohapatra, and R. C. Balabantaray, ‘Sentiment analysis for Odia language using supervised classifier: an information retrieval in Indian language initiative’, *CSI Trans. ICT*, vol. 4, no. 2–4, pp. 111–115, 2016, doi: 10.1007/s40012-016-0117-9.
 - [20] Z. Nasim and S. Haider, ‘ABSA Toolkit: An Open Source Tool for Aspect Based Sentiment Analysis’, *Int. J. Artif. Intell. Tools*, vol. 26, no. 6, 2017, doi: 10.1142/S0218213017500233.
 - [21] M. Al-Smadi, M. Al-Ayyoub, H. Al-Sarhan, and Y. Jararweh, ‘Using Aspect-Based Sentiment Analysis to Evaluate Arabic News Affect on Readers’, *Proc. - 2015 IEEE/ACM 8th Int. Conf. Util. Cloud Comput. UCC 2015*, pp. 436–441, 2015, doi: 10.1109/UCC.2015.78.
 - [22] T. H. Nguyen, K. Shirai, and J. Velcin, ‘Sentiment analysis on social media for stock movement prediction’, *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9603–9611, 2015, doi: 10.1016/j.eswa.2015.07.052.
 - [23] K. Ravi and V. Ravi, *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*, vol. 89, no. June. Elsevier B.V., 2015.
 - [24] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, ‘Sentiment analysis: A review and comparative analysis of web services’, *Inf. Sci. (Ny.)*, vol. 311, pp. 18–38, 2015, doi: 10.1016/j.ins.2015.03.040.
 - [25] F. H. Khan, U. Qamar, and S. Bashir, ‘A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet’, *Knowl. Inf. Syst.*, vol. 51, no. 3, pp. 851–872, 2017, doi: 10.1007/s10115-016-0993-1.
 - [26] P. Classification, ‘Patent Application Publication (10) Pub . No . : US 2019 / 0251128 A1 -’, vol. 1, 2019.
 - [27] M. Bavakhani and A. O. Mining, ‘A Deep Learning Approach for Extracting Polarity from Customers ’ Reviews’, *2019 5th Int. Conf. Web Res.*, pp. 276–280, 2019.
 - [28] P. Hoffmann, ‘Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis’, no. October, pp. 347–354, 2005.
 - [29] R. Passonneau, ‘Sentiment Analysis of Twitter Data’, no. June, pp. 30–38, 2011.
 - [30] C. Zhang, D. Zeng, and F. Wang, ‘Sentiment Analysis of Chinese Documents : From’, vol. 60, no. 12, pp. 2474–2487, 2009, doi: 10.1002/asi.
 - [31] F. Alqasemi, A. Abdelwahab, and H. Abdelkader, ‘Constructing automatic domain-specific sentiment lexicon using KNN search via terms discrimination vectors’, *Int. J. Comput. Appl.*, vol. 7074, no. December, pp. 1–11, 2017, doi: 10.1080/1206212X.2017.1409477.
 - [32] M. Song, H. Park, and K. Shin, ‘Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean’,

- Inf. Process. Manag.*, vol. 56, no. 3, pp. 637–653, 2019, doi: 10.1016/j.ipm.2018.12.005.
- [33] S. Vanaja, ‘Aspect-Level Sentiment Analysis on E-Commerce Data’, *2018 Int. Conf. Inven. Res. Comput. Appl.*, no. Icirca, pp. 1275–1279, 2018.
 - [34] D. Mohey and E. M. Hussein, ‘A survey on sentiment analysis challenges’, *J. KING SAUD Univ. - Eng. Sci.*, no. April, 2016, doi: 10.1016/j.jksues.2016.04.002.
 - [35] A. Fölling, J. Lepping, A. Fölling, and J. Lepping, ‘grids To cite this version : Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery’, 2012.
 - [36] A. Shoukry and A. Rafea, ‘Sentence-Level Arabic Sentiment Analysis’, pp. 546–550, 2012.
 - [37] Z. Bošnjak, ‘Sentiment Analysis of Customer Data’, vol. 23, no. 3, pp. 38–49, 2018.
 - [38] K. Schouten and F. Frasincar, ‘Survey on Aspect-Level Sentiment Analysis’, vol. 4347, no. c, pp. 1–20, 2015, doi: 10.1109/TKDE.2015.2485209.
 - [39] T. M. Mitchell, ‘Introduction to Machine Learning’.
 - [40] E. L. Hamaker, J. D. Mulder, and M. H. Van Ijzendoorn, ‘Developmental Cognitive Neuroscience Description , prediction and causation : Methodological challenges of studying child and adolescent development’, *Dev. Cogn. Neurosci.*, vol. 46, no. September, p. 100867, 2020, doi: 10.1016/j.dcn.2020.100867.
 - [41] *Introduction to Semi-Supervised Learning*. .
 - [42] G. James, *An Introduction to Statistical Learning*. .
 - [43] I. A. Silva, A. Boukerche, T. R. M. B. Silva, L. B. Ruiz, E. Cerqueira, and A. A. F. Loureiro, ‘Vehicular Networks : A New Challenge for Content-Delivery-Based’, vol. 49, no. 1, 2016.
 - [44] I. R. Galatzer-levy and I. R. Galatzer-levy, ‘Applications of Latent Growth Mixture Modeling and allied methods to posttraumatic stress response data Applications of Latent Growth Mixture Modeling and allied methods to posttraumatic stress response data’, vol. 8198, no. September, pp. 11–13, 2017, doi: 10.3402/ejpt.v6.27515.
 - [45] R. C. Kessler, R. M. Bossarte, A. Luedtke, A. M. Zaslavsky, and J. R. Zubizarreta, ‘Behaviour Research and Therapy Machine learning methods for developing precision treatment rules with observational data’, *Behav. Res. Ther.*, vol. 120, no. February, p. 103412, 2019, doi: 10.1016/j.brat.2019.103412.
 - [46] J. L. Gradus *et al.*, ‘Prediction of Sex-Specific Suicide Risk Using Machine Learning and Single-Payer Health Care Registry Data From Denmark’, vol. 02118, pp. 1–10, 2019, doi: 10.1001/jamapsychiatry.2019.2905.
 - [47] W. Medhat, A. Hassan, and H. Korashy, ‘Sentiment analysis algorithms and applications: A survey’, *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.
 - [48] T. Edition, *No Title*. .
 - [49] S. B. K. I. D. Z. P. E. Pintelas, ‘Machine learning : a review of classification and combining techniques’, no. 2006, pp. 159–190, 2007, doi: 10.1007/s10462-007-9052-3.
 - [50] J. R. Quinlan, ‘Induction of Decision Trees’, pp. 81–106, 2007.
 - [51] D. Grossman and P. Domingos, ‘Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood’, 2004.
 - [52] S. R. Safavian and D. Landgrebe, ‘A Survey of Decision Tree Classifier Methodology’, *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991, doi: 10.1109/21.97458.

- [53] H. Bhavsar and A. Ganatra, 'A Comparative Study of Training Algorithms for Supervised Machine Learning', no. January 2012, 2016.
- [54] B. User, 'Insight into Data Mining Theory and Practice'.
- [55] K. N. Stevens, T. M. Cover, and P. E. Hart, 'Nearest Neighbor', vol. I, 1967.
- [56] D. Mining, K. Discovery, B. Laboratories, and L. Technologies, 'A Tutorial on Support Vector Machines for Pattern Recognition', vol. 167, pp. 121–167, 1998.
- [57] M. L. Zhang and Z. H. Zhou, 'ML-KNN: A lazy learning approach to multi-label learning', *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007, doi: 10.1016/j.patcog.2006.12.019.
- [58] Q. T. Ain *et al.*, 'Sentiment Analysis Using Deep Learning Techniques : A Review', vol. 8, no. 6, 2017.
- [59] P. Vateekul and T. Koomsubha, 'A Study of Sentiment Analysis Using Deep Learning Techniques on Thai Twitter Data', 2016.
- [60] W. S. Noble and P. Street, 'What is a support vector machine ?', vol. 24, no. 12, pp. 1565–1567, 2006.
- [61] V. Y. Kulkarni, P. K. Sinha, and A. E. Classifiers, 'Pruning of Random Forest Classifiers : A Survey and Future Directions', pp. 64–68, 2012.
- [62] F. Rustam, A. Mehmood, M. Ahmad, S. Ullah, D. M. Khan, and G. Y. U. S. Choi, 'Classification of Shopify App User Reviews Using Novel Multi Text Features', vol. 8, 2020.
- [63] F. Y. Oisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, 'Supervised Machine Learning Algorithms : Classification and Comparison', no. July, 2017, doi: 10.14445/22312803/IJCTT-V48P126.
- [64] E. Ogunfiditimi, F. and Oguntade, 'Logistic Regression: A Paradigm for Dichotomous Response Data', *Int. J. Engeneeing Sci.*, vol. 3, no. 6, pp. 1–5, 2014.
- [65] D. Srilatha, D. P. Rao, C. Nikhil, and J. S. Tejumade, 'Classifying Fake News Articles Using Machine Learning Techniques', pp. 162–167, 2020.
- [66] B. Kovács, F. Tinya, C. Németh, and P. Ódor, 'Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4-yr experiment', *Ecol. Appl.*, vol. 30, no. 2, pp. 321–357, 2020, doi: 10.1002/eap.2043.
- [67] M. S. Elli and Y. Wang, 'Calendars', *Acta Archaeol.*, vol. 77, no. 1–2, pp. 62–90, 2006, doi: 10.1111/j.1600-0390.2006.00047.x.
- [68] Y. Xu, X. Wu, and Q. Wang, 'Sentiment Analysis of Yelp's Ratings Based on Text Reviews', *2015 17th Int. Symp. Symb. Numer. Algorithms Sci. Comput.*, 2015.
- [69] A. Bhatt, A. Patel, H. Chheda, and K. Gawande, 'Amazon Review Classification and Sentiment Analysis', *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 6, pp. 5107–5110, 2015.
- [70] B. Lewis, I. Smith, M. Fowler, and J. Licato, 'The robot mafia: A test environment for deceptive robots', *28th Mod. Artif. Intell. Cogn. Sci. Conf. MAICS 2017*, pp. 189–190, 2017, doi: 10.1145/1235.
- [71] T. Shaikh and D. Deshpande, 'Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews', *Int. J. Comput. Trends Technol.*, vol. 36, no. 4, pp. 225–230, 2016, doi: 10.14445/22312803/ijctt-v36p139.
- [72] E. M. S. Mona Mohamed Nasr and Ahmed Mostafa Hafez, 'Building Sentiment analysis Model using Graphlab', *Int. J. Sci. Eng. Res.*, vol. 8, no. 6, pp. 1155–1160, 2017.
- [73] M. Wang, 'Text Mining for Yelp Dataset Challenge', pp. 1–5.
- [74] Y. Liu and X. Zhu, 'Short text sentiment classification based on feature extension

- and ensemble classifier', *AIP Conf. Proc.*, vol. 1967, no. May, 2018, doi: 10.1063/1.5039023.
- [75] A. R. Nabhan and K. Shaalan, 'Text Graphlet Patterns', vol. 2, no. June, pp. 152–161, 2016, doi: 10.1007/978-3-319-41754-7.
 - [76] L. Xu, J. Liu, L. Wang, and C. Yin, 'Aspect based sentiment analysis for online reviews', *Lect. Notes Electr. Eng.*, vol. 474, pp. 475–480, 2018, doi: 10.1007/978-981-10-7605-3_78.
 - [77] D. Anand and D. Naorem, 'Semi-supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering', *Procedia Comput. Sci.*, vol. 84, pp. 86–93, 2016, doi: 10.1016/j.procs.2016.04.070.
 - [78] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Y. Sha, and Y. H. Yang, '1000 Songs for Emotional Analysis of Music', *CrowdMM 2013 - Proc. 2nd ACM Int. Work. Crowdsourcing Multimed.*, pp. 1–6, 2013, doi: 10.1145/2506364.2506365.
 - [79] G. M. M. Sarria, J. Diaz, and C. Arce-Lopera, 'Analyzing and Extending the Salsa Music Dataset', *2019 22nd Symp. Image, Signal Process. Artif. Vision, STSIVA 2019 - Conf. Proc.*, pp. 1–5, 2019, doi: 10.1109/STSIVA.2019.8730229.
 - [80] J. Hendler, 'Data integration for heterogenous datasets', *Big Data*, vol. 2, no. 4, pp. 205–215, 2014, doi: 10.1089/big.2014.0068.
 - [81] E. Triantafillou *et al.*, 'L EARNING TO L EARN FROM F EW E XAMPLES', 2020.
 - [82] B. Gopal, P. Dipankar, and S. Bandyopadhyay, 'Multimodal mood classification of Hindi and Western songs', 2018.
 - [83] M. Bilal, H. Israr, M. Shahid, and A. Khan, 'Sentiment classification of Roman-Urdu opinions using Navie Baysian, Decision Tree and KNN classification techniques', *J. KING SAUD Univ. - Comput. Inf. Sci.*, 2015, doi: 10.1016/j.jksuci.2015.11.003.
 - [84] W. P. Wong, M. Lo, and T. Ramayah, 'THE EFFECTS OF TECHNOLOGY ACCEPTANCE FACTORS ON CUSTOMER E-LOYALTY AND E-SATISFACTION IN MALAYSIA', vol. 15, no. 3, pp. 477–502, 2014.
 - [85] P. Patil and P. Yalagi, 'Sentiment Analysis using Aspect Level Classification', vol. 2, no. 4, pp. 23–27, 2016.
 - [86] A. Rextin, A. Kakakhel, and M. Nasim, 'Analyzing Emergent Users ' Text Messages Data and Exploring Its Benefits', *IEEE Access*, vol. 7, pp. 2870–2879, 2019, doi: 10.1109/ACCESS.2018.2885332.
 - [87] G. Z. Nargis and N. Jamil, 'Generating an Emotion Ontology for Roman Urdu Text', vol. 7, no. 3, pp. 83–91.
 - [88] S. Vijayarani, M. J. Ilamathi, and M. Nithya, 'Preprocessing Techniques for Text Mining - An Overview Preprocessing Techniques for Text Mining - An Overview', no. February 2015, 2020.
 - [89] I. Javed and H. Afzal, 'Opinion Analysis of Bi-Lingual Event Data from Social Networks'.
 - [90] B. R. Aditya, A. R. Sulthana, A. K. Jaithunbi, and L. Sai, 'Aspect level sentiment analysis using machine learning', 2017, doi: 10.1088/1757-899X/263/4/042009.
 - [91] D. Kalita, 'Supervised and Unsupervised Document Classification-A survey', vol. 6, no. 2, pp. 1971–1974, 2015.
 - [92] M. P. Abraham and K. R. Udaya Kumar Reddy, 'Feature based sentiment analysis of mobile product reviews using machine learning techniques', *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 2289–2296, 2020, doi: 10.30534/ijatcse/2020/210922020.

- [93] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, 'Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier', *World Wide Web*, vol. 20, no. 2, pp. 135–154, 2017, doi: 10.1007/s11280-015-0381-x.
- [94] P. Pankaj, P. Pandey, M. Muskan, and N. Soni, 'Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews', *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Perspectives Prospect. Com. 2019*, pp. 320–322, 2019, doi: 10.1109/COMITCon.2019.8862258.
- [95] X. Wang, X. Chen, M. Tang, T. Yang, and Z. Wang, 'Aspect-Level Sentiment Analysis Based on Position Features Using Multilevel Interactive Bidirectional GRU and Attention Mechanism', vol. 2020, 2020.
- [96] P. K. Sari, A. Alamsyah, and S. Wibowo, 'Measuring e-Commerce service quality from online customer review using sentiment analysis', *J. Phys. Conf. Ser.*, vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012053.
- [97] A. K. and V. J., 'Aspect Based Sentiment Analysis for E-Commerce Using Classification Techniques', *SSRN Electron. J.*, vol. 1, no. 1, pp. 35–42, 2019, doi: 10.2139/ssrn.3362346.
- [98] C. Huang and G. Liu, 'Sentiment analysis of network comments based on GCNN', *ACM Int. Conf. Proceeding Ser.*, vol. 7, no. 6, pp. 409–413, 2018, doi: 10.1145/3297156.3297189.
- [99] Q. Lu, Z. Zhu, D. Zhang, W. Wu, and Q. Guo, 'Interactive Rule Attention Network for Aspect-Level Sentiment Analysis', *IEEE Access*, vol. 8, pp. 52505–52516, 2020, doi: 10.1109/ACCESS.2020.2981139.
- [100] F. Xu, Z. Pan, and R. Xia, 'E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework', *Inf. Process. Manag.*, vol. 57, no. 5, p. 102221, 2020, doi: 10.1016/j.ipm.2020.102221.
- [101] Y. Han, M. Liu, and W. Jing, 'Aspect-Level Drug Reviews Sentiment Analysis Based on Double BiGRU and Knowledge Transfer', *IEEE Access*, vol. 8, pp. 21314–21325, 2020, doi: 10.1109/ACCESS.2020.2969473.
- [102] B. Krawczyk and F. Herrera, 'Neurocomputing A survey on data preprocessing for data stream mining: Current status and future directions', vol. 239, pp. 39–57, 2017, doi: 10.1016/j.neucom.2017.01.078.
- [103] E. Acuna, 'CO', no. January 2011, 2019, doi: 10.1007/978-3-642-04898-2.
- [104] S. Alasadi, 'Review of Data Preprocessing Techniques in Data Mining', no. September, 2017.
- [105] L. A. Mullen, K. Benoit, O. Keyes, D. Selivanov, and J. Arnold, 'Fast , Consistent Tokenization of Natural Language Text', vol. 2017, pp. 2017–2019, 2018, doi: 10.21105/joss.00655.
- [106] T. Astuti, R. Wahyudi, Z. Rifai, and R. A. Pambudi, 'An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian', *2018 3rd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng.*, pp. 230–234, 2018.
- [107] S. Xu, Y. Qian, and R. Q. Hu, 'A Data-driven Preprocessing Scheme on Anomaly Detection in Big Data Applications', pp. 814–819, 2017.
- [108] S. Yang and H. Zhang, 'Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis', no. August, 2019.
- [109] S. Al-saqqa, A. Awajan, and S. Ghoul, 'Stemming Effects on Sentiment Analysis using Large Arabic Multi-Domain Resources', *2019 Sixth Int. Conf. Soc. Networks Anal. Manag. Secur.*, pp. 211–216, 2019.

- [110] B. Yu and I. Science, ‘An evaluation of text classification methods for literary study’, vol. 23, no. 3, pp. 327–343, 2008, doi: 10.1093/lc/fqn015.
- [111] A. Companies, ‘Tweets Classification on the Base of Sentiments for US Airline Companies’, pp. 1–22, 2019, doi: 10.3390/e21111078.
- [112] W. Zhang, T. Yoshida, and X. Tang, ‘Expert Systems with Applications A comparative study of TF \tilde{A} IDF , LSI and multi-words for text classification’, *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2758–2765, 2011, doi: 10.1016/j.eswa.2010.08.066.
- [113] S. Robertson, ‘Understanding inverse document frequency: on theoretical arguments for IDF’, vol. 60, no. 5, pp. 503–520, 2004, doi: 10.1108/00220410410560582.