**CHAPTER ONE**

**INTRODUCTION**

* 1. **BACKGROUND OF THE STUDY**

With the recent fast growth of e-commerce, large number of products is sold online, and a lot more people are purchasing products online. In most cases, people while buying also give feedback of product purchased in form of reviews or ratings. The user generated reviews for products and services are largely available on internet. Since information available on internet is so widespread, there is need to extract the needful information for which is needed to make use of sentimental analysis. Sentiment analysis extracts abstract and to the point information required for source materials by applying concept of Natural Language Processing (NLP). It is used to deal with identification and aggregation of the opinions given by the customers. These reviews play vital role in determining potential customer for the products as well as market trend for product.

In past days, purchasing of products was more based on getting product review from nearby neighbors, relatives etc. as products were purchased directly from merchants. People believed relatives, and friends review about product helpful. But with the rapid change in technology, the World observed development of e-commerce industry with sites flooded by products from different brands made available to customers at the touch of one click. The availability of product-based sites with doorstep delivery has made it convenient for customers to shop online. It provides one stop shop for all needs of customers. With so much change and dynamics in shopping pattern, we see merchants providing customers with feedback option about the product. Customers write reviews from all parts of the world. There are thousands, millions of reviews being written. So, a question arises on how to get fundamental judgment about product without going through each of them separately. A lot of reviews are very long, making it difficult for a potential customer to review them to make an informed decision on whether the customer should purchase the product or not. A vast number of reviews also make it difficult for product manufacturers to keep log of customer opinions and sentiments expressed on their products and services. It thus becomes necessity to produce a summary of reviews. Summarization of reviews is done using sentiment analysis.

Sentiment analysis tends to extract subjective information required for source materials by applying natural concept of natural language processing (Pang *et al.* 2012). The main task lies in identifying whether the opinion stated is positive, negative or neutral. Since customers usually do not express opinions in simple manner, sometimes it becomes tedious task to judge an opinion stated. Some opinions are comparative ones while others are direct. Sentimental analysis helps customer visualize satisfaction while purchasing by simple summarization of these reviews into positive or negative; two broader classified classes. Feedbacks are mainly used for helping customers purchase online and for knowing current market trends about products which is helpful for developing market strategies by merchants.

This research examines the effectiveness of applying machine learning techniques to the sentiment classification problem. Machine learning is divided into: supervised and unsupervised approaches (Liu *et al. 2012*). Supervised learning tends to be more accurate because each of the classifiers is trained on a collection of representative data known as corpus in contrast to unsupervised learning which does not require prior training. In order to mine the data instead; it measures how far a word is inclined towards positive and negative.

With an ever increasing demand of smart phones, the mobile phone market is expanding at an exponential pace. With such a boom in the smart-phone industry, there is a need to realize the holistic review of the brand and the model of phone. There are numerous brands present in the market, out of which some are dominant and occupy quite a big part of the industry. For instance, Samsung, Apple, etc. are names associated with brands which are famous throughout the world. Electronic commerce plays a vital role in increasing the sales of the mobile phones and influencing consumer buying patterns. Reviews available on such e-commerce platforms act as a guiding tool for the consumers to make informed decisions. Retail websites like Amazon.com offer different options to the reviewers for writing their reviews. For instance, the consumer can provide numerical rating from 1 to 5 or write comments about the product. As there are innumerable products manufactured by many different brands, so providing relevant reviews to the consumers is the need of hour. Number of reviews associated with a product or a brand is increasing at an alarming rate, which is no less than handling the big data. Classifying the reviews on the basis of sentiment of customers into positive and negative sentiment provides sentiment orientation of the review, hence results in better judgment. Segregation of reviews on the basis of their sentiment can help future buyers to evaluate positive and negative feedback constructively and reach at better decisions as per their requirements. This evaluation often acts as a testimony to the users who are looking to know the details and specifications of the smartphones; thereby increasing user credibility.

Furthermore, as online marketplaces have been popular during the past decades, the online sellers and merchants ask their purchasers to share their opinions about the products they have bought. Everyday millions of reviews are generated all over the Internet about different products, services and places. This has eventually made the Internet the most important source of getting ideas and opinions about a product or a service.

However, as the number of reviews available for a product grows, it is becoming more difficult for a potential consumer to make a good decision on whether to buy the product. Different opinions about the same product on one hand and ambiguous reviews on the other hand makes customers more confused to get the right decision. Thus, the need for analyzing these contents seems crucial for all e-commerce businesses.

Sentiment analysis and classiﬁcation is a computational study which attempts to address this problem by extracting subjective information from the given texts in natural language, such as opinions and sentiments. Different approaches have used to tackle this problem from natural language processing, text analysis, computational linguistics, and biometrics. In recent years, Machine learning methods have got popular in the semantic and review analysis for their simplicity and accuracy (Sepideh Paknejad, 2018).

Amazon is one of the e-commerce giants that people are using every day for online purchases where they can read thousands of reviews dropped by other customers about their desired products. These reviews provide valuable opinions about a product such as its property, quality and recommendations which helps the purchasers to understand almost every detail of a product. This is not only beneﬁcial for consumers but also helps sellers who are manufacturing their own products to understand the consumers and their needs better.

* 1. **PROBLEM STATEMENT**

The rapid growth in online purchase via e-commerce and other channels with feedback mechanism has increased the volume of customer review data. Potential customers are fond of reading product’s review before buying the product. Unfortunately, due to the unstructured and large volume of this data, potential customers are not able to make use of it to make inform decision prior to their purchase.

Since human sentiment is multidimensional in nature, thus, sentiment classiﬁcation aims to determine the overall intention of a written text which can be of admiration or criticism type. This can be achieved by using Machine Learning algorithms such as Na¨ıve Bayes and Support Vector Machine.

* 1. **AIM AND OBJECTIVES OF THE STUDY**

The general objective of this research is to perform sentiment analysis on product reviews using Machine Learning techniques. The specific objectives are:

1. Analyze and categorize review data.
2. Analyze sentiment on dataset from document level (review level).
3. Categorization or classification of opinion sentiment into Positive and Negative using Na¨ıve Bayes and Decision Tree.
   1. **RESEARCH METHODOLOGY**

When conducting a research, researchers often make use of several fact finding techniques such as interviews, reviews of journals/articles, questionnaire, observations etc. The particular method to be used strictly depends on the nature or purpose of the study. The source of data to be used in this researchis product reviews collected from amazon website from 1996 - 2014.

Each review includes the following information:

1. Review text
2. Satar rating
3. Date of review
4. Variation
5. Feedback

Every rating is based on a 5-star scale, resulting all the ratings to be ranged from 1-star to 5-star with no existence of a half-star or a quarter-star.

* 1. **EXPECTED CONTRIBUTION TO KNOWLEDGE**

The benefits of sentiment analysis and its use by business owners help them gain an advantage over their competitors. Terms like ‘opinion mining’ and ‘text identification’ often describe the meaning of sentiment analysis as a suitable method used by marketers to recognize customers’ preferences. The data gathered from customers’ responses like tweets, comments, feedback and any writing that’s related to products or services are studied and this process is called sentiment analysis. Marketers and organizations are pursuing this process to stay relevant in the competitive field and to find a suitable way to advance their business

**CHAPTER TWO**

**LITERATURE REVIEW**

**2.1 INTRODUCTION**

Sentiment analysis aims to categorize text as positive or negative on the basis of the positive or negative sentiment (opinion) expressed in the document towards a topic. A document with positive or negative sentiment is also said to be of positive or negative polarity respectively. The granularity of the polarity can be up to the level of words. That is, there can be polar (subjective) and non-polar (objective/neutral) sentences and words. Also, sentiment analysis is a technique to classify people’s opinions in product reviews, blogs or social networks. It has different usages and has received much attention from researchers and practitioners lately.

Many researchers have worked in the field of sentiment analysis, each one proposing new way of getting better efficiency from machine learning approaches.

Many researchers have worked on sentiment analysis of product review, each one proposing new way of getting better efficiency from machine learning approaches using different algorithms.

* 1. **SENTIMENT CLASSIFICATION AND ANALYSIS**

Electronic commerce is becoming increasingly popular due to the fact that e-commerce websites allow purchasers to leave reviews on different products. Millions of reviews are being generated everyday by costumers which makes it difficult for product manufacturers to keep track of customer opinions of their products. Thus, it is important to classify such large and complex data in order to derive useful information from a large set of data. Classiﬁcation methods are the way to tackle such problems. Classiﬁcation is the process of categorizing data into groups or classes based on common traits (Pandey *et al.* 2016). A common concern for organizations is the ability to automate the classiﬁcation process when big datasets are being used (Liu *et al.* 2014).

Sentiment analysis, also known as opinion mining, is a Natural Language Processing (NLP) problem which means identifying and extracting subjective information of text sources. The purpose of sentiment classiﬁcation is to analyze the written reviews of users and classify them into positive or negative opinions, so the system does not need to completely understand the semantics of each phrase or document (Liu *et al.* 2014). This however is not done by just labeling words as positive or negative. There are some challenges involved. Classifying words and phrases with prior positive or negative polarity will not always work. For example, the word “amazing” has a prior positive polarity, but if it comes with a negation word like “not”, the context can completely change (Singla *et al.* 2017).

As Ye *et al.,* (2015) state the word “unpredictable” camera has a negative meaning to that camera while “unpredictable” experience is considered as positive for tourists. Sentiment classiﬁcation has been attempted in different ﬁelds such as movie reviews, travel destination reviews and product reviews. Lexicon based methods and machine learning methods are two main approaches that are usually used for sentiment classiﬁcation.

**2.3 SENTIMENT CLASSIFICATION USING MACHINE LEARNING METHODS**

There are a large number of papers that have been published in the ﬁeld of Machine Learning (ML). One of the most used approaches for sentiment classiﬁcation is Machine Learning Algorithms.

One of the ﬁrst deﬁnitions of Machine Learning that has been provided by Tom Mitchell (1997) in his book Machine Learning is as follow: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”

Machine learning aims to develop an algorithm in order to optimize the performance of the system by using example data. The solution that machine learning provides for sentiment analysis involves two main steps. The ﬁrst step is to “learn” the model from the training data and the second step is to classify the unseen data with the help of the trained model (Khairnar an Kinikar 2013). Machine learning algorithms can be classiﬁed in different categories:

1. Supervised learning
2. Semi-supervised learning
3. Unsupervised learning

i. In **Supervised Learning,** the process where the algorithm is learning from the training data can be seen as a teacher supervising the learning process of its students (Brownlee 2016). The supervisor is somehow teaching the algorithm what conclusions it should come up with as an output. So, both input and the desired output data are provided. It is also required that the training data is already labeled. If the classiﬁer gets more labeled data, the output will be more precise. The goal of this approach is that the algorithm can correctly predict the output for new input data. If the output were widely deferent from the expected result, the supervisor can guide the algorithm back to the right path.

There is however some challenges involved when working with supervised. The supervised learning works ﬁne as long as the labeled data is provided. This means that if the machine faces unseen data, it will either give wrong class label after classiﬁcation or remove it because it has not “learnt” how to label it (Cunningham *et al*., 2018).

ii. The **Semi-supervised** learning which has the beneﬁt of both supervised and unsupervised learning refers to problems in which a smaller amount of data is labeled, and the rest of the training data set is unlabeled. This is useful when collecting data that can be cheap but labeling it can be time consuming and expensive. This approach is highly favorable both in theory and practice because of the fact that having lots of unlabeled data during the training process tends to improve the accuracy of the ﬁnal model while building it requires much less time and cost (Zhu 2015).

iii. The **Unsupervised learning** is difference from supervised learning; it is trained on unlabeled data with no corresponding output. The algorithm should ﬁnd out the underlying structure of the data set on its own. This means that it has to discover similar patterns in the data to determine the output without having the right answers. One of the most important methods in unsupervised learning problems is clustering. Clustering is simply the method of identifying similar groups of data in the data set (Kaushik 2016).

For sentiment classiﬁcation in an unsupervised manner it is usually the sentiment words and phrases that are used. This means that the classiﬁcation of a review is predicted based on the average semantic orientation of the phrases in that review (Turney 2012). This is obvious since the dominating factor for sentiment classiﬁcation is often the sentiment words (Berk 2016).

**2.3.1 SUPPORT VECTOR MACHINES**

Support vector machines (SVM) are supervised learning method that can be used for solving sentiment classiﬁcation problems (Cristianini and ShaweTaylor 2016). This technique is based on a decision plane where labeled training data is placed and then algorithm gives an optimal hyper plane which splits the data into deferent groups or classes. As seen in Figure 2.1. The best hyper plane is the one that separates the classes with the largest margin. This is achieved by choosing a hyper plane so that its distance from the nearest data on each class is maximized (Berk 2016).

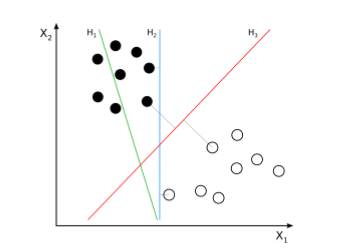


Figure 2.1: Sample Support Vector Machines Plot

“H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin.” Source: Nello *et al,* 2012.

**2.3.2 NA¨IVE BAYES**

Na¨ıve Bayes is another machine learning technique that is known for being powerful despite its simplicity. This classiﬁer is based on Bayes theorem and relies on the assumption that the features (which are usually words in text classiﬁcation) are mutually independent. In spite of the fact that this assumption is not true (because in some cases the order of the words is important), Na¨ıve Bayes classiﬁers have proved to perform surprisingly well (Rish 2017). The ﬁrst step that should be carried out before applying the Na¨ıve Bayes model on text classiﬁcation problems is feature extraction.

**2.3.3 FEATURE EXTRACTION**

Since Machine Learning algorithms work only with ﬁxed-length vector of numbers rather than raw text, the input (in this case text data) need to be parsed. The method for transforming the texts into features is called the Bag of words model of text, which is a commonly used method of feature extraction. The approach works by creating deferent bags of words that occur in the training data set where each word is associated with a unique number. This number shows the occurrence of each word in the document. A simple illustration of the Bag of words model can be seen in Figure 1.2. The model is called a bag of words because the position of the words is in the document discarded (Rish, 2017).

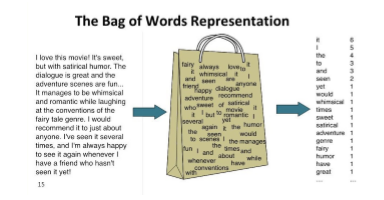


Figure 2.2: A simple illustration of the Bag of words model. Source: Peter et al, 2012.

**2.3.4 DECISION TREE**

A tree has many analogies in real life, and turns out that it has influenced a wide area of **machine learning**, covering both **classification and regression**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, it’s also widely used in machine learning.

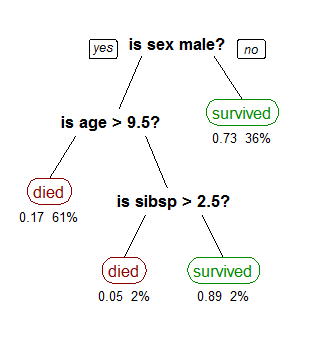


Figure 2.3 Sample Decision Tree. Source: Mudinas *et al.,* 2014.

A decision tree is drawn upside down with its root at the top. In the Figure 2.3, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn’t split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

* 1. **CLASSIFICATION OF SENTIMENT ANALYSIS APPROACHES**

Machine learning based approach uses classification technique to classify text; it consists of two sets of documents: training and a test set. The training set is used for learning the differentiating characteristics of a document, while the test set is used for checking how well the classifier performs. The features of machine learning based approach for sentiment classification are:

1. **Term presence and their frequency:** that includes uni-grams or n-grams and their presence or frequency.
2. **Part of speech information:** used for disambiguating sense which is used to guide feature selection
3. **Negations:** has the potential of reversing sentiment’s opinion words/phrases: that expresses positive or negative sentiments.

The lexicon-based approach uses sentiment dictionary with opinion words and match them with the data for determining polarity. There are three techniques to construct a sentiment lexicon: manual construction, corpus-based methods and dictionary-based methods. The manual construction is a difficult and time-consuming task. Corpus-based methods can produce opinion words with relatively high accuracy. Finally, in the dictionary-based techniques, the idea is to first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in the WordNet dictionary for their synonyms and antonyms.

Finally, in the hybrid approach, the combination of both the machine learning and the lexicon-based approaches has the potential to improve the sentiment classification performance.

There are some advantages and limitations in using these different approaches depending on the purpose of the analysis as depicted in Table 2.1.

The main advantage of machine learning approaches is the ability to adapt and create trained models for specific purposes and contexts, while the limitation is that it is difficult integrating into a classifier, general knowledge which may not be acquired from training data. Furthermore, learnt models often have poor adaptability between domains or different text genres because they often rely on domain specific features from their training data. Lexicon-based approaches have the advantage that general knowledge sentiment lexicons have wider term coverage, however these approaches have two main limitations. Firstly, the number of words in the lexicons is finite, which may constitute a problem when extracting sentiment from very dynamic environments. Secondly, sentiment lexicons tend to assign a fixed sentiment orientation and score to words, irrespective of how these words are used in a text. The main advantages of hybrid approaches are the lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level and the lesser sensitivity to changes in topic domain. While the main limitation is that reviews are with a lot of noise (irrelevant words for the subject of the review) are often assigned a neutral score because the method fails to detect any sentiment. The process of sentiment analysis is depicted in Figure 2.3. The process starts form collection of review data, follows by identifying the sentiment, after which the feature is being extracted from the data and lastly, the polarity of the sentiment is concluded.

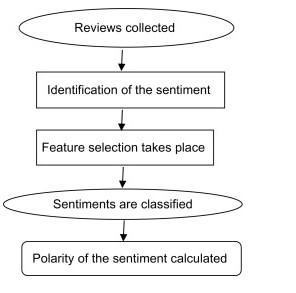


Figure 2.4 Process of analyzing sentiments of reviews. Source: Mudinas *et al.,* 2014.

**Table 2.1: Sentiment classification approaches**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/No** | **Sentiment Classification Approaches** | | **Features/Techniques** | **Advantages and Limitations** |
| 1 | Machine learning | Bayesian Networks. Naive Bayes Classification. Maximum Entropy Neural Networks. Support Vector Machine | Term presence and frequency. Part of speech information Negations Opinion. Words and phrases | **Advantages**: The ability to adapt and create trained models for specific purposes and contexts  **Limitations:** The low applicability to new data because it is necessary the availability of labeled data that could be costly or even prohibitive |
| 2 | Lexicon based | Dictionary based approach. Novel Machine Learning Approach. Corpus based approach. Ensemble Approaches. | Manual construction,  Corpus-based Dictionary based. | **Advantages**: Wider term coverage.  **Limitations:** Finite number of words in the lexicons and the assignation of a fixed sentiment orientation and score to words. |
| 3 | Hybrid | Machine learning. Lexicon based. | Sentiment lexicon constructed using public resources for initial sentiment detection.  Sentiment words as features in machine learning method. | **Advantages:** Lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level and the lesser sensitivity to changes in topic domain.  **Limitations:** Noisy reviews. |

## 

## 2.5 REVIEWS ON RELATED WORKS

**2.5.1 Techniques and Applications for Sentiment Analysis (Feldman Ronen, 2013).** The author talks about the specific problems within sentiment analysis field which includes; document level, sentence level, feature level, comparative opinion and sentiment lexicon problem.

**2.5.2 Thumbs up? Sentiment Classification Using Machine Learning Techniques (Bo *et al*. 2014),** considers classifying documents not by topic, but by overall sentiment, concluding whether a review is positive or negative. Reviews are converted to simple decision by making use of approaches such as Naïve Bayes, Support Vector Machine by initially counting the number of positive and negative words in a document. Since opinions are not always direct e.g. “the Nokia phone is good” but also it can be a comparative opinion like “Nokia phone has better battery life than Samsung”. There exist three levels at which opinions are classified: sentence level, document level, and feature level. At sentence level, subjective and objective opinions exist, at document level, a document is classified based on overall sentiment expressed by opinion holder. However, at feature level, attributes of products are taken into consideration, which provides classification in depth.

**2.5.3 A holistic lexicon-based approach to opinion mining (Ding *et al.* 2016).**  The authors proposed a holistic lexicon-based approach that allows the system to handle opinion words that are context dependent. It takes into account the counting of the number of positive and negative opinion words near the product feature in each sentence. If count of positive opinion words is more than that of negative opinion words, the final prediction on the feature is positive else negative.

**2.5.4 Comparative experiments on sentiment classification for online product reviews (Cui *et al*. 2017).**  The authors make use of higher n-gram model using three classifiers. The first one being language model which is a generative method that computes the probability of generation of a word sequence. The Passive-Aggressive algorithms are second which consists of a family of margin based online learning algorithms for binary classification. Third, to predict the polarity of a review. Apart from classifying reviews in two broader categories, there also exists a term polarity degree to measure the strength of opinions, as in is the opinion strongly positive, mildly positive, highly negative etc.

**2.5.5 Combining lexicon and learning based approaches for concept-level sentiment analysis. (Mudinas *et al.* 2014)** says product of sentiment value and occurring frequency gives measurement of sentiments. Psenti approach calculates the overall sentiment of stated opinionated text like customer reviews and scales them as a real score between -1 and +1, which can then be easily transformed positive/negative classification or into a scale of 1-5 stars. Creating candidate list using POS tagging with removal of stop word leads to aspect identification. The aspect having less than 5 comments on it, is removed from the candidate list.

**2.5.6 A joint model of feature mining and sentiment analysis for product review rating de (Albornoz *et al.*2016)**, the product review was translated into a Vector of Feature Intensities (VFI). A VFI is a vector of N+ 1 value, each one representing a different product feature and the other features. The authors addressed the problem of analyzing multiple related opinions in a text and presented an algorithm that jointly learns ranking models for individual aspects by modeling the dependencies between assigned ranks.

**2.5.7 Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches (Lina and Pimwadee, 2015).** The authors investigated movie review mining using machine learning and semantic orientation. Supervised classification and text classification techniques were used in the proposed machine learning approach to classify the movie review. A corpus is formed to represent the data in the documents and all the classifiers were trained using this corpus. Thus, the proposed technique was more efficient. Though, the machine learning approach uses supervised learning, the proposed semantic orientation approach uses “unsupervised learning” because it does not require prior training in order to mine the data. Experimental results showed that the supervised approach achieved 84.49% accuracy in three-fold cross validation and 66.27% accuracy on hold-out samples. The proposed semantic orientation approach achieved 77% accuracy of movie reviews. Thus, the study concludes that the supervised machine learning is more efficient but requires a considerable amount of time to train the model. On the other hand, the semantic orientation approach was slightly less accurate but was more efficient to use in real time applications. The results confirmed that it is practicable to automatically mine opinions from unstructured data.

**2.5.8 Aspect-Based Opinion Polling from Customer Reviews, Zhu et al., (2011**) proposed aspect based opinion polling from free form textual customers’ reviews. The aspect related terms used for aspect identification was learnt using a multi-aspect bootstrapping method. A proposed aspect-based segmentation model, segments the multi aspect sentence into single aspect units which was used for opinion polling. Using an opinion polling algorithm, they tested on real Chinese restaurant reviews achieving 75.5 percent accuracy in aspect-based opinion polling tasks. This method is easy to implement and are applicable to other domains like product or movie reviews.

**2.5.9 Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques, (Yi et al., 2013).**  The authors proposed a Sentiment Analyzer to extract opinions about a subject from online data documents. Sentiment analyzer uses natural language processing techniques. The Sentiment analyzer finds out all the references on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes. Online product review articles for digital camera and music were analyzed using the system with good results.

**2.5.10 Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified (Alekh Agarwal et al. 2015).** The authors proposed a machine learning method incorporating linguistic knowledge gathered through synonymy graphs, for effective opinion classification. This approach shows the degree of influence among relationships of documents have on their sentiment analysis. This was brought about by the use of graph-cut technique and opinion words got through synonymy graphs of WordNet. The proposed approach also improves the accuracy of predictions in classification task. Experiments using the system have given results with an accuracy of over 90% said the authors, with an added advantage of reduction in processing time, with minimal difference in final accuracies. The proposed methodology from the authors resulted in the following conclusions:

1. Automated mining of linguistic information is possible, so demonstrated with the structure of links in WordNet.
2. Generic method of using graph-cut technique for efficient opinion classification.

**2.5.11 Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums Ahmed, (Abbasi et al. 2014).** The authorsproposed novel sentiment analysis methods to classify web forum opinions in multiple languages. The proposed sentiment analysis method utilized the function of stylistic and syntactic features to evaluate the sentiment in English and Arabic content. The Entropy weighted Genetic Algorithm was incorporated to enhance the performance of the classifier and achieve the true assessment of the key features. Experiments were conducted using movie review data set and the results demonstrated that the proposed techniques were efficient.

**2.5.12 Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews (Anidya et al. 2017).** Theranked the product reviews based on customer-oriented and manufacturer ranking mechanism. The expected helpfulness of the review was used for the ranking and also ranking was based on the expected effect on sale. The proposed methods identify the reviews which have the most impact. For feature based products, the authors reviews that confirm the information contained in the product description were used, and reviews with subjective point of view were useful for experience goods. Econometric analysis with text mining techniques and with subjectivity analysis was used in the proposed method. Product prices and sales ranking publicly available on amazon website were used to compile the data set. The product and sales data were the two sets of information collected for each product. Products such as audio and video players, digital cameras were used to form the data set. The empirical analysis was performed using the compiled data set.

**2.5.13 Identifying Noun Product Features that Imply Opinions, (Michael et al. 2013)**. The authors presented „Pulse‟ a prototype system for mining topics and sentiment orientation from free text customer feedback. Blogs, newsgroups, feedback email from customers, and web sites that collect product reviews are all source of free text customer feedback. The proposed system was designed to handle the free form information of the customer feedbacks as the sources of information are less structured than traditional surveys. A clustering technique and machine learned sentiment classifiers were used in the proposed method. Sentiment and topic detections were performed at the sentence level not at the document level. The Pulse was evaluated using car reviews database, and the sample data contains 4, 06,818 customer car reviews written over a four-year period. The data set contained almost 900,000 sentences in total. Sentiment analysis was performed using 3000 randomly selected sentences. Each sentence was classified as positive, negative and others. The other category contained both positive and negative sentiment and sentences with no complex sentiments. Training of the sentiment classifier was done using 2500 sentences and the remaining 500 sentences were reserved for test set. Results reflect the efficiency of the proposed system.

**2.5.14 Opinion Word Expansion and Target Extraction through Double Propagation, (Qui et al. 2011).** The authors analyzed the problems related to opinion mining such as opinion lexicon expansion and opinion target extraction. Opinion targets are entities and their attributes on which opinions have been expressed. The list of opinion words such as good, bad, excellent, poor used to indicate positive and negative sentiments is Opinion lexicon. The links between the opinion words and targets Syntactic relations were identified using dependency parser based on bootstrapping. The process uses semi-supervised methods; opinion word seeds were used in the initial opinion lexicon. Bootstrapping process was started using the initial opinion lexicon. Double propagation method was used as information was propagated back and forth between opinion words and targets.

**2.5.15 A Holistic Lexicon-Based Approach to Opinion Mining, (Xiaowen Ding, 2013).** The authors proposed a holistic lexicon-based approach which uses external indications and linguistic conventions of natural language expressions to determine the semantic orientations of opinions. Advantage of this approach was that opinion words which are context dependent are easily handled. The algorithm used linguistic patterns to deal with special words, phrases. Researchers built a system called Opinion Observer based on this technique. Experiments using product review dataset was highly effective. It was shown that multiple conflicting opinion words in sentence are also dealt with efficiently. This system shows better performance when compared to existing methods.

**CHAPTER THREE**

**SYSTEM ANALYSIS AND DESIGN**

**3.1 INTRODUCTION**

The methodologies to be used in this research are Naive Bayes and Decision Tree which are to be used for data training and classification.

**3.2 Dataset and its Features:** The first step includes dataset collection and preprocessing of the dataset.

This dataset consists of 3150 Amazon customer reviews (input text), star ratings, date of review, variant and feedback of various amazon Alexa products like Alexa Echo, Echo dots, Alexa Firesticks etc. It includes six features as explained in Table 3.1.

Table 3.1: Feature of review data

|  |  |  |
| --- | --- | --- |
| **S/No** | **Feature** | **Description** |
| 1 | Reviews | Input text |
| 2 | Start rating | Star rating from the customer |
| 3 | Date | Date of the review |
| 4 | Variation | Different categories of Alexa products |
| 5 | Feedback | Feedback from the customer |

**3.3 Approach**

The experimental dataset is collected from Amazon e-commerce website. Each dataset is in the Comma Separated Values (CSV) file format and available as supplement. In the second step, data are pre-processed to remove stop words, punctuation marks, whitespaces, digits and special symbols. In the third step, feature selection is performed to extract relevant features from the data set. In the given data set out of the six features, only three features, i.e, Product Name, Brand Name and Reviews would be considered. In the fourth step, sentiment orientation of the reviews would be determined. In the fifth step, ‘Pos/Neg’ tags are appended to the dataset to corresponding to each review to conduct supervised learning. The sixth step involves training and testing the classified data using Naïve Bayes and SVM models.

**3.3.1 Pre-processing:** The dataset is unstructured; it may contain repetitive words, large number of words that are not at all needed in summarizing of opinions. Pre-processing involves removal of stop words such as ‘and’, ‘or’, ‘that’ etc. followed by porter stemming which involves simplifying target words to base words by removal of suffixes such as – ed, ate, ion, ional, ment, ator, ssess, es, ance or conversion from ator to ate etc. For example, “replacement” is stemmed to replac; “troubled” to trouble; “happy” to happi ; “operator” to operate. The raw data is pre-processed to improve data quality.

**3.3.2 Feature Extraction**: Features in reviews are extracted so that it helps customer to know which feature has positive comment and which one has negative. Since, overall conclusion about product is much needed but there is also situation where customer requirements come into the scenario. Use of adjectives is done to classify opinions as positive or negative using unigram model. For example, “the Samsung camera I bought was good; it has got great touch screen, awesome flashlight.” The feature extracted out of it would be like: Domain: Mobile; Product: Samsung; Feature: Camera; Adjective: Good.

**3.3.3 Training and classification:** Supervised learning generates a function which maps inputs to desired outputs also called as labels because they are training examples labeled by human experts. Naïve Bayes and Decision Tree techniques were adopted to carry out supervised learning on the dataset fetched.

**Naive Bayes and support Vector Machine:** Naive Bayes classifiers work on the principle that the value of a particular feature is independent of the value of any other feature. For example, Samsung phone will be considered as phone if it has basic call function, touch screen and camera. A naive Bayes classifier considers each of these features to contribute independently to the probability that this Samsung phone is a mobile, regardless of any possible correlations between the cameras, call function and touch screen. Assign to a given document d a, the class c \* = arg maxc P(c | d)

Class c\* is assigned to product review d, where, f represents a feature and ni(d) represents the count of feature fi found in product review d. There are a total of m features. Parameters P(c) and P (f|c) are obtained through maximum likelihood estimates which are incremented by one for smoothing. Dataset after being preprocessed and after extracting features, is input to train through naïve bayes, hence providing polarity for reviews. For example: “the phone is great” provides positive opinion regarding the product.

**Decision Tree:** Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

There are two popular attribute selection measures:

1. Information Gain
2. Gini Index

For information gain which is measure of the changes in entropy, that is its uses a node in a decision tree to partition the training instances into smaller subsets the entropy changes, suppose S is a set of instances, A is an attribute, Sv is the subset of S with A = v, and Values (A) is the set of all possible values of A, then:



For entropy which measure of uncertainty of a random variable, suppose S is a set of instances, A is an attribute, Sv is the subset of S with A = v, and Values (A) is the set of all possible values of A, then:

**The pseudocode for process is as shown:**

**Input**- Dataset of product reviews

**Output**- Classification of these reviews as positive and negative.

Step1: Preprocess the data

Removal of special characters

Removal of stop words

Stemming the word.

Step2: Get feature list

If word in stop word list

Removal word

Return file

Else append word to file

Step3: Extract feature list

Match every word in preprocessed list

If word matches adjective in base list

Display word

If word matches feature in base list

Display feature

Step4: combine both feature and preprocessed list

Step5: Use machine learning algorithms

Compute probability

Step6: classify opinion as positive, negative.

Browser or input

dataset

Pre-process dataset

Select pre-processed

dataset

Feature extraction

Train using machine learning approaches

Naive Bayes

Classify opinion

Plot result in form

of graph

Compute

probability score

Decision Tree

Figure 3.1 Proposed methodology flow chart

Figure 3.1 depicts the proposed methodology for the product-based sentiment analysis. The first step uploading the dataset into machine learning environment, the dataset will be pre-processed, feature extraction will be performed on the pre-processed dataset. The dataset will be trained using machine learning approaches, the approach would be Naïve Bayes and Decision Tree. Afterwards, accuracy score is calculated to show which of the algorithm give is better.

**CHAPTER FOUR**

**SYSTEM IMPLEMENTATION AND TESTING**

**4.1 SYSTEM IMPLEMENTATION**

The implementation of this research was done using Python programming language; the justification for chosen Python is because of its speed, flexibility, dynamic in nature and high accuracy. Naïve Bayes and Decision Tree algorithms were used for dataset training and prediction.

**4.2. SYSTEM REQUIREMENT**

The system requirements for the developed Cryptography Ciphertext Using Chaos that will be expected at a minimum level for the software to work are stated below:

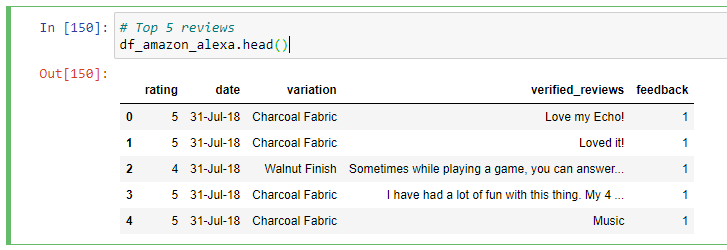
* + 1. **HARDWARE REQUIREMENT**

1. Processor 1.5MHZ or higher
2. Hard disk 150GB or higher
3. RAM 2GB or higher

**4.2.2 SOFTWARE REQUIREMENT**

1. Windows 7 or higher
2. Python 2.7 or higher version
3. Anaconda Navigator

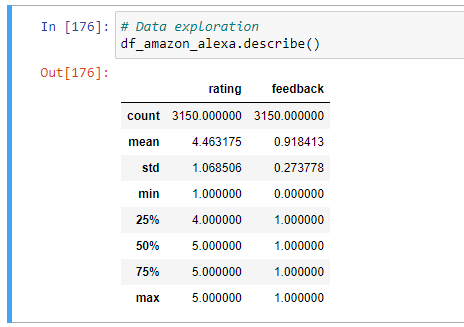
## DATA UNDERSTANDING

The dataset used has 3150 Amazon Alexa customer comments (text) on amazon Alexa products. Figure 4.1 shows the header of the dataset. There are 5 columns: “rating” column is the comment scores; “date” is the comment date; the “variation” column is various Alexa products like Charcoal Fabric, Alexa Echo, Echo dots, Alexa Fire sticks etc.; “verified\_reviews” column is the review text; in “feedback” column, if the comment score is greater than or equal to 3 it's one (1), and if it's less than three it's zero (0).

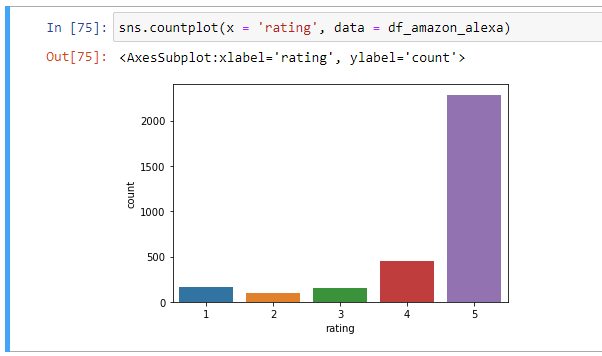
**Figure 4.2: Dataset Head**

## 4.4 Data exploration

After successfully importing the needed Python libraries and the dataset, the next thing is data exploration to generally understanding the dataset. Hence the Python ***describe()*** gives a detail description of the dataset as depicted in Figure 4.2. It can be observed that the average rating (that is the mean) is about 4.46 with standard variance of 1.06, and the feedback is about 0.92 with standard variance 0.27, which implies the products are highly rated by the majority of the customers.

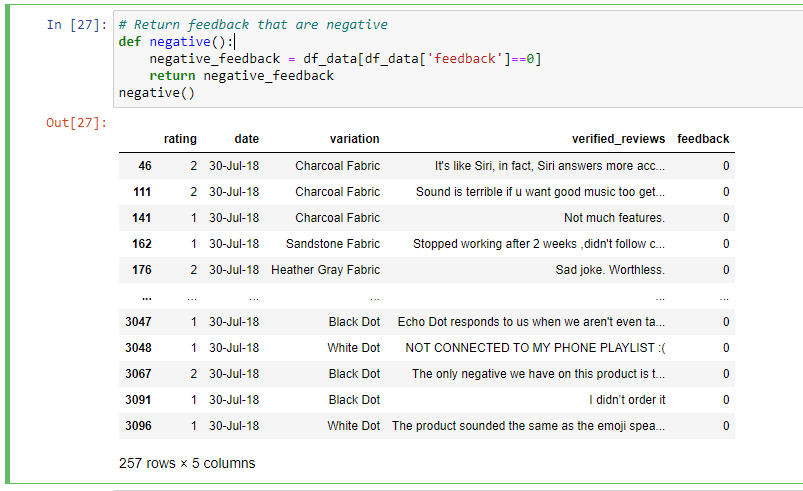


**Figure 4.2: Data Exploration**

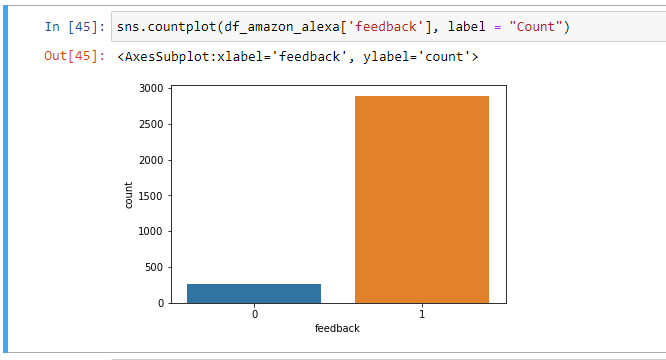
More sentiment details were revealed by plotting the distribution of the rating, which shows that more than 2000 reviews were rated 5 scores, 500 reviews rated 4 scores and less than 500 reviews were rated between 1-3 scores as depicted in Figure 4.3.

**Figure 4.3: Rating Distribution**

**Figure 4.4 Positive Sentiment**

As depicted in Figure 4.4, the positive sentiments are more than the negative sentiments. The positive sentiments from the dataset are 2893 while the negative sentiments are 257 as depicted in Figure 4.5.

**Figure 4.4 Negative Sentiment**

**Figure 4.5 Bar Chart for Feedback Sentiment**

It can be deduced from Figure 4.5 that the positive sentiment feedback is higher than the negative sentiment feedback.

**Figure 4.6 All Words in The Reviews**

As depicted in Figure 4.6, the Python library ***Wordcloud and Matplotlib*** make it easy to visualize the entire words in the review.

Since there are some words that have little or no importance with the sentiment of the comment, such as: use, device, time etc. Thus, there is need to perform a word count of certain key words that could describe the sentiment in a better way. Also, with the help Python library ***Wordcloud and Matplotlib***, the needed words for the sentiment analysis can be visualized as depicted in Figure 4.7.

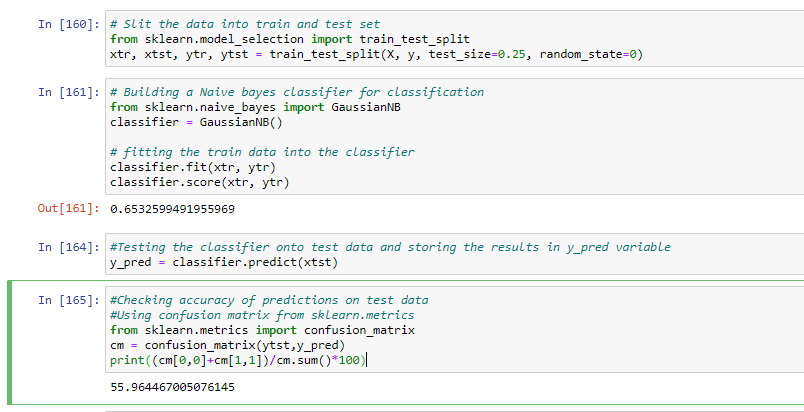


**Figure 4.7 Sentiment Words**

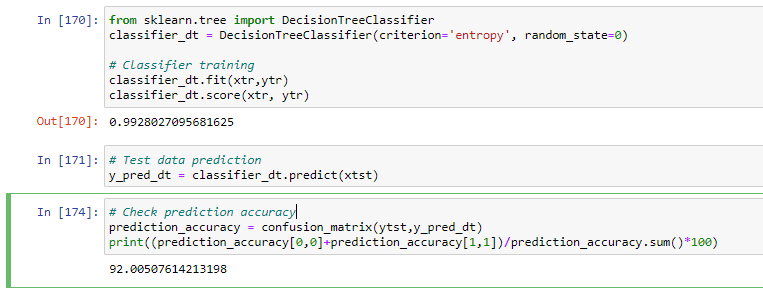
## 4.5 Applying the Machine Learning Model to Predict Sentiment

From the feature vector of the reviews, the dataset was used to establish machine leaning model to predict the rating or sentiment. Majorly, the dataset gives the feedback and ratings for each of the reviews, which of course was used to create a semimetal score; which is either positive or negative. This research considered the ratings 4 or 5 as positive and ratings less than 4 for negative. Then, the dataset was separated into two parts: training and test dataset. The training data is 70% while the testing data is 30%. However, the training dataset is the sample dataset for learning, while the test dataset is the sample dataset for performance evaluation or prediction.

Naive Bayes and Decision Tree were the classifiers that were used; however, Decision Tree give a better performance with 92% accuracy compared to its counterpart with 55% accuracy as depicted in Figure 4.8 and 4.9 respectively.



**Figure 4.8: Naïve Bayes Prediction Accuracy**



**Figure 4.8: Decision Tree Prediction Accuracy**

**CHAPTER FIVE**

**SUMMARY, CONCLUSION AND RECOMMENDATION**

**5.1 SUMMARY**

The sentiment analysis of customer feedback or product reviews, has recently become very popular in text mining and computational linguistics research. Sentiment could be referred to as an attitude, thought, or judgment prompted by feeling. Sentiment analysis, which is also commonly known as opinion mining, studies people’s sentiments towards certain entities. From a customers or user’s perspective, people are able to post their own content through various social media channels, such as forums, blogs, or online social networking sites. From a typical researcher’s perspective, many social media platforms release their Application Programming Interfaces (APIs), which allow data collection and analysis by researchers and developers. However, many of these online data are found to have several flaws that often hinder the effective process of sentiment analysis. The first flaw is that since people can freely and unguardedly post their own content, thus, the quality of their opinions cannot be guaranteed. Also, the real truth of such online data is not always available and can hardly be validated.

**5.2 CONCLUSION**

This research has done a sentimental analysis about the Amazon Alexa products reviews, and also trained the data, built a model to make prediction of the sentiment of the comment given the review text. Naïve Bayes and Decision Tree classifiers were used in which Decision Tree proved to be more efficient as it has 92% accuracy.

Because it is a text data, most times the distribution and characteristics of data involved and the choice of machine algorithm to used affect the accuracy of machine learning prediction. The word clouds are mostly used to check the high frequency of words in the reviews, but can hardly be used for more detail analysis.

**5.3. RECOMMENDATION**

Base on the analysis and results presented in this research, it is highly recommended that business owners and organizations leverage on sentimental analysis to get more insight on their customers opinions, this will go a long way in helping them to make informed decision, serve the customer better, which will eventually transform to more profit.

**REFERENCES**

Ahmed Abbasi, Hsinchun Chen, And Arab Salem, (2014), Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums, *ACM Transmission Information System*., Vol. 26, No. 3, pp. 1-34.

Alekh Agarwal and Pushpak Bhattacharyya, (2015). Sentiment Analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified, *In Proceedings of the International Conference on Natural Language Processing (ICON)*, pp 28-76.

Anindya Ghose, Panagiotis G. Ipeirotis, (2017), Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews, *Proceedings of the Ninth International conference on Electronic commerce ICEC07*, pp 303-310.

Cui, Hang, Vibhu Mittal, and Mayur Datar, (2017). Comparative experiments on sentiment classification for online product reviews. *Proceedings of the 21st national conference on Artificial* I*ntelligence (AAAI)*. Volume 2, pp 1265-1270.

De Albornoz, Jorge Carrillo, et al. (2016). A joint model of feature mining and sentiment analysis for product review rating. *Advances in information retrieval. Springer Berlin Heidelberg*, pp.55-66.

Ding, Xiaowen, Bing Liu, and Philip S. Yu, (2016). A holistic lexicon-based approach to opinion mining. *Proceedings of the International Conference on Web Search and Data Mining. ACM,* pp. 231-240.

Ding, Xiaowen, Bing Liu, and Philip S. Yu. (2014). A holistic lexicon-based approach to opinion mining. *Proceedings of International Conference on Web Search and Data Mining. ACM*, pp. 231-240.

Feldman, Ronen., (2013). Techniques and applications for sentiment Analysis. *Communications of the ACM 5.* Volume 56 Issue4, pp.82-89.

Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, Vol. 37, No. 1: 9.27. pp 1-57.

Irina Rish, (2017). An empirical study of the Naive Bayes classiﬁer. *In IJCAI workshop on Empirical Methods in Artiﬁcial Intelligence, IBM*. volume 3, pp 41–46.

Jason Brownlee. (2016). Supervised and unsupervised machine learning algorithms. Pp 7-23.

Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. (2014). Low-quality product review detection in opinion summarization. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp.203-265.

Lina Zhou, Pimwadee Chaovalit**, (**2015**).** Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches. *Proceedings of the 38th Hawaii International Conference on System Sciences*, pp 205-243.

Liu, Chien-Liang, et al., (2012). Movie rating and review Summarization in Mobile Environment. Systems, Man, and Cybernetics, Part C: *IEEE Transactions on Applications and Reviews,* Volume 42 issue 3, pp.397-407.

Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger, (2013). Pulse: Mining Customer Opinions from Free Text Natural Language Processing”, Microsoft Research, Redmond, WA 98052, USA. Pp 123-168.

Mudinas, Andrius, Dell Zhang, and Mark Levene, (2014). Combining lexicon and learning based approaches for concept-level sentiment analysis. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM,* pp 330-333.

Nello Cristianini and John Shawe-Taylor, (2016). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press. pp 38-98.

P´adraig Cunningham, Matthieu Cord, and Sarah Jane Delany. (2018). Supervised learning. In Machine Learning Techniques for multimedia, Springer, 2008 pp 21–49.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan, (2012) "Thumbs up? sentiment classification using machine learning techniques*. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics*, Volume 10, pp. 79-86.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan, (2014). Thumbs up? Sentiment Classification using Machine Learning techniques. *In Proceedings of the ACL-02 conference on Empirical methods in Natural Language Processing- Association for Computational Linguistics*, Volume 10, pp. 79-86.

Peter D Turney and Michael L Littman, (2012). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS),* 21(4):315–346.

Priyank Pandey, Manoj Kumar, and Prakhar Srivastava. (2016), Classification techniques for big data*: A survey. In Computing for Sustainable Global Development (INDIACom),* 2016 3rd International Conference on, pages 3625–3629.

Qiang Ye, Ziqiong Zhang, and Rob Law, (2015). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications,* 36(3): pp6527–6535.

Richard A Berk., (2016.). Statistical learning from a regression perspective. Springer. pp 5-19.

Saurav Kaushik, (2016) An introduction to clustering and different methods of clustering. pp 143-165.

Sepideh Paknejad, (2018). Sentiment classification on Amazon reviews using machine learning approaches*.* Degree project *Stockholm*, *Sweden.*

Xiaojin Zhu. (2015.) Semi-supervised learning literature survey. Pp 10-16.

Xiaowen Ding, (2013). A Holistic Lexicon-Based Approach to Opinion Mining. Palo Alto, California, USA.2008 ACM. Pp 302-367.

Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack: (2013), Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques, *In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003)*. Melbourne, Florida. pp 98-105.

Zeenia Singla, Sukhchandan Randhawa, and Sushma Jain, (2017). Statistical and sentiment analysis of consumer product reviews. *In Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on IEEE*, pp 1–6.

Zeenia Singla, Sukhchandan Randhawa, and Sushma Jain. (2017). Statistical and sentiment analysis of consumer product reviews. *In Computing, Communication and Networking Technologies (ICCCNT), 8th International Conference on, IEEE*. pp. 1-6.

Zhu, Jingbo Wang, Huizhen Zhu, Muhua Tsou, Benjamin K. Ma, Matthew, (2011). Aspect- Based Opinion Polling from Customer Reviews”, *IEEE Transactions on Affective Computing*, Volume: 2, Issue:1. pp 37.