Opinion explanation first

Different languages on sentiment analysis

Online consumer reviews have changed the marketing reality in which advertising has traditionally operated as one-way communication from companies to consumers via mass communication channels (Campbell et al. 2011). Although brands still advertise in these traditional ‘paid’ media such as television, radio, and print, new advertising channels are now available to them (Edelman and Salsberg 2011). Companies can utilize ‘owned’ media to contact customers directly to, for example, compel them to share their experiences or visit the company’s website, or their ‘earned’ media by providing customers with space where they can promote the company to other consumers. These new communication channels have become especially important for companies, especially in the context of consumer endorsements being an important advertising strategy (Bernritter, Verlegh, and Smit, 2016; Lee, Park, and Han 2011).

Consumers do not rely solely on advertising messages anymore, but direct their attention to other sources of information, especially online reviews. Surveys show that more than half of consumers consults online reviews (CMA 2015; Mintel 2015). That is because the majority of consumers trust recommendations from others more than traditional forms of advertising (Nielsen, 2012). Online reviews help consumers make decisions such as purchasing products, watching movies, or joining a sports club. They have become a major driving force in marketing (Cui, Lui, and Guo 2012) and are a common feature on many websites. Information from other consumers, such as online reviews, is thought to be more persuasive and trustworthy, and is especially important in the online environment (Ba and Pavlou 2002; Willemsen, et al. 2011). Consumers buying products online have to rely on the information provided on the website and do not have the ability to try out the product (Lee 1998). Due to their alleged persuasiveness and prevalence, online reviews have attracted substantial attention from both researchers and practitioners.

However, due to a large number of reviews on e-commerce websites from customers, it is not possible to analyse the opinion manually. It is a rather complex task for a customer to identify significant details from the prevalent information available on the website. Therefore, sentiment analysis is a significant approach for opinion extraction.

Many researches for the understanding of consumer’s behaviour have been carried out in different and top most languages i.e. in Chinese language research is performed on “Personalized recommendation based on customer preference mining and sentiment assessment from a Chinese e-commerce website”, in English language “Aspect-level sentiment analysis on e-commerce data”, in Hindi language “Aspect based sentiment analysis: category detection and sentiment classification for Hindi, Deep learning for Hindi text classification: A comparison”, in Spanish language “Sentiment analysis in Spanish for improvement of products and services: a deep learning approach”, in Arabic language “Sentiment analysis in Arabic tweets”, in Malay language “Sentiment analysis for Malay language: systematic literature review”, in Russian language “Analysis of sentiment and emotion from text written in Russian language”, in Bengali language “Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words”, in Portuguese languages “A survey of sentiment analysis in the Portuguese language”.

In this study we are focusing on the English language reviews. In previous researches, sentiment analysis has been applied in many different fields i.e. Citizens’ Political Preferences [8], Supply Chain Intelligence [9], Set-Based Feature Selection For Arabic Sentiment Analysis [10], Spanish Text Transformations For Twitter Sentiment Analysis [11], Odia Language Using Supervised Classifier [12], Customer Satisfaction At Aspect-Level [13], Sentiment Analysis Algorithms And Applications [14], Aspect Based Sentiment Analysis To Evaluate Arabic News Effect On Readers [15], Sentiment Analysis On Social Media For Stock Movement Prediction [16], Opinion Mining And Sentiment Analysis: Tasks, Approaches And Applications [17], A Review And Comparative Analysis Of Web Services [19] Sentiment Analysis Using Revised Sentiment Strength Based On Sentiwordnet [20]. Sentiment analysis and its types are explained in the next section.

## Sentiment Analysis

Sentiment analysis is the study of people's opinions, sentiments, attitudes, and emotions expressed in written languages [1]. Sentiment analysis is a developing field in the research area. People are now using the web for business correspondence, e-commerce [2]. 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 it is also being used to predict the results of national events like elections, etc. [3].

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 [4]. Moreover, multiple styles and varieties of products are available in online shopping stores and consumers have a good choice to buy variations of products without going outside [5].

While e-commerce is performing a positive role for the convenience of the 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 [6]. To keep track the customer satisfaction level, it is very important to evaluate products of online shopping stores and to check the tendency of the sentiment of the customer towards the product. It also helps the organization for business growth and the reference for other consumers. Sentiment analysis for customer reviews is also defined as the process of systematically analysing the reviews and detect the feelings. This is also termed as opinion mining and text analysis [7].

Sentiment analysis allows the new customer to examine previous customer's suggestions and reviews about the product. Sentiment analysis is a basic viewpoint for the consumers when they start e-commerce. With the advancement of the internet throughout the world, many number of people share their feedback not only on the respective e-commerce website but also on the other platforms like social media. The reviews, written by the consumers, help both the other buyers to make informed decisions about the product(s) and for the manufacturers of products. They need to go through the reviews of buyers [2].

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 [20]. (Different language explain, English specific) 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 [21]. Information retrieval techniques are used to gather data from different Blogs and E-commerce websites where people share their opinion. [22]. Once the reviews are collected, then the next problem is to analyze the reviews. Multiple Data mining and Machine Learning approaches are present for the resolution of this problem [23]. From the bulk of reviews, some opinions are positive and some are negative. The negative and positive opinions represent the polarity of review, and the analysis of a large number of opinions 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 [24].

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 [25] [26] [27]. 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 [28] [29].

Moreover, the lexicon-based models are efficient and simple but manually sentiment lexicon creation is a time-consuming and labour-intensive job. Secondly, already static polarity is needed to be provided for every sentence. For this solution, some kinds of models that automatically generate sentiment lexicons have been proposed [68] [69].

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[70].

Sentiment Analysis

Document Level Sentiment Analysis

Sentence Level Sentiment Analysis

Aspect Level Sentiment Analysis

(Caption)

### Document Level Sentiment Analysis

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

### Sentence Level Sentiment Analysis

Sentiment analysis at the sentence level is considered as 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 [32, 33, and 34]

### Aspect Level Sentiment Analysis

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

## Machine Learning

In 1997 researcher defines machine learning as the feature of computer science that aims to gain knowledge from data [37]. 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 [38].

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 [39], 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.

### Research Types

Research has been divided into the following three types:

#### Descriptive Research:

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

#### Predictive Research:

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

#### Explanatory Research:

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

To solve the above research categories, machine learning has been divided into three types mentioned in figure 2 [Reference].

Machine Learning

Unsupervised Machine Learning

Supervised Machine Learning

Semi-Supervised Machine Learning

(Caption)

### 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 [71].

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.

The general goal of unsupervised learning is to find patterns in the data. The two most common types of unsupervised learning, cluster analysis, and principal component analysis are quite well-known in the social and behavioral sciences. Cluster analysis can be used to obtain qualitatively different groups of individuals; for instance, Drysdale et al. (2017) used this technique to identify different neurophysiological subtypes of patients based on distinct patterns of dysfunctional connectivity in the brain. Principal component analysis has been used to study large numbers of recorded neurons. Such unsupervised learning approaches are often used as a form of pre-processed data, to reduce the number of predictors in big data. Moreover, unsupervised learning can also be used for outlier detection (i.e., a case that does not fit the general pattern in the data).

### 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 [40] [41] [42].

### 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 [43-46]. 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 [47]

#### Classification

Classification is said to be supervised machine learning [48] because the labels are already given with the data in contradiction with unsupervised learning in which there are no predefined classes or labels inside the data. Each set of data that is used in supervised machine learning contains a set of features or attributes that may be continuous or categorical [49], [50].  Classification is said to be the process of creating the model with the help of training data set having labels 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:

##### 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 [49].  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 e decision tree explains feature and every branch represent the outcome of the test and every node explains the class label [5].

**Advantages**

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

**Disadvantages**

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