KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

A

Project Report

on

Project Title

SENTIMENT ANALYSIS OF IMDB MOVIE REVIEWS

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2022-23

in

ELECTRONICS AND COMMUNICATION ENGINEERING

By

PRAKHAR SINGHANIA (1900290310100)

MAHAK SINGH (1900290310077)

Under the supervision of

MR. ABHISHEK

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2023

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

SIGNATURE :

NAME: PRAKHAR SINGHANIA

ROLL NO : 1900290310100

DATE:

SIGNATURE:

NAME: MAHAK SINGH

ROLL NO : 1900290310077

DATE:

CERTIFICATE

This is to certify that Project Report entitled "**SENTIMENT ANALYSIS ON IMDB MOVIE REVIEWS**"which is submitted by **PRAKHAR SINGHANIA & MAHAK SINGH** in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Electronics & Communication Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date:

MR. ABHISHEK POKHRIYAL

(Asst. Prof. ECE DEPARTMENT)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Mr. Abhishek Pokhriyal, Asst. Prof., Department of Electronics & Communication Engineering, KIET Group of Institutions, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day. We also take the opportunity to acknowledge the contribution of Dr. Vibhav Kumar Sachan, HoD, Electronics and Communication Engineering Department, KIET Group of Institutions, Ghaziabad, for his full support and assistance during the development of the project. We also donot like to miss the opportunity to acknowledge the contribution of all the faculty members ofthe department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

SIGNATURE:

NAME: PRAKHAR SINGHANIA

ROLL NO. : 1900290310100

DATE:

SIGNATURE:

NAME: MAHAK SINGH

ROLL NO.: 1900290310077

DATE:

ABSTRACT

From the last 10 years, popularity of social media has increased at an alarming rate. Everyone is utilizing technology at higher rates than earlier. People are now sharing their emotions and opinions on social media sites allowing others to know what they think about a particular thing. Many companies are utilizing the data from various websites to generate meaningful information out of it which can be further used for business purposes. Huge textual data is available on sites like Amazon, IMDB, Rotten Tomatoes on movies and analyzing such massive data manually is a tedious task. So, to speed up the process, programmers use certain techniques to extract out public opinion. One of which is using sentiment analysis.

Sentiment analysis is a sub-module of opinion mining where the analysis focuses on the extraction of text and opinions of the people on a particular topic. We are making use of IMDB reviews on movies to predict how the users have rated the movies and predict the movies that have a positive or negative review. Sentiment analysis is a new research area in which large amounts of data are evaluated to give helpful insights on a certain issue. It is a powerful tool that can help governments, companies, and even consumers. Text emotion recognition plays an important part in this framework. Natural language processing (NLP) and machine learning (ML) researchers have explored a variety of methods for implementing the procedure with the maximum accuracy feasible. It is based on the Logistic Regression algorithm. To improve post clarification performance, the data is efficiently pre-processed and partitioned. The classification performance is studied in terms of accuracy. It validates the feasibility of incorporating the provided method into contemporary text-based sentiment analyzers.

LIST OF FIGURES

CHAPTER 1
INTRODUCTION
SENTIMENT ANALYSIS

1.1 INTRODUCTION

The interpretation and categorization of emotions within text data using text analysis techniques is known as sentiment analysis. Sentiment analysis models consider not just polarity (positive, negative, or neutral), but also sentiments and emotions (angry, joyful, sad, and so on), as well as intents (e.g. interested v. not interested). Many large corporations are putting their resources on predicting the outcomes of their enterprises. Tokenization , word filtering, stemming, and classifications are all part of the sentiment analysis process. Tokenization requires text to be split into units such as words, numbers, and punctuation. Next step is stemming, which is the act of removing prefixes and affixes from a word in order to convert it to its stem. After preprocessing, we analyze the dataset using Nave Bayes, Support Vector Machine, and Logistic Regression. In this step, we choose the best model based on accuracy. As a result, we research and study the aspects that influence the ratings of our review text before classifying the film as favorable or bad. Sentiment analysis allows businesses to identify customer sentiment towards products, brands or services in online conversations and feedback. Sentiment analysis models focus on polarity (positive, negative, neutral) but also on feelings and emotions (angry, happy, sad, etc) and even on intentions (e.g. interested vs not interested). Sentiment analysis has become a hot topic and many big companies are investing their resources to predict the results for their businesses. Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. It describes the process of understanding natural language–the way that humans communicate–based on meaning and context.

On a variety of online platforms, such as review sites, blogs, as well as social services such as Twitter, Facebook, Instagram, Internet users produce vast amount of opinionated text of movie reviews, travel experiences, product reviews, opinions about news and others. Automatic opinion mining is the ability to produce large amounts of opinionated text information from online sources without human interference. Therefore, it is necessary to have an efficient way to predict user sentiments about a product or service. For example, if someone has to buy tickets for a movie, then rather than manually going through all the reviews, a sentiment classifier can predict the overall sentiment of the movie. Based on positive or negative sentiment, a decision can be taken whether to buy tickets or not. Our end goal with this dataset is to analyze sentiments of reviews that each user gave to the movie, and evaluating its performance.

Fig.1: Emotions of people

1.2. PROJECT DESCRIPTION

This project mainly deals with the IMDB Dataset. The proposed framework for our model includes data cleaning, data preprocessing, applying classifiers on the data and finally comparing the results from the different classification models we used.

1.2.1. OVERVIEW OF PROJECT

The working principle of sentiment analysis includes tokenization, word filtering, stemming and classifications. In tokenization, text needs to be segmented into units such as words/ numbers or punctuations. Next step stemming which is the process of removing prefixes and affixes to convert a particular word into its stem. After preprocessing, we analyze the dataset by performing classification using Naïve Bayes, Support Vector Machine and Logistic Regression. Here, we determine the best model based on accuracy. Hence, We analyze and study the features that affect the scores of our review text and finally classify the movie as positive or negative.

1.2.2. PROBLEM STATEMENT

Extracting opinions about a particular entity is very important. Trying to go through such a vast amount of information to understand the general opinion is impossible for users just by the sheer volume of the data. Hence, the need of a system that differentiates between good reviews and bad reviews is required. Further, labeling these documents with their sentiment would provide a succinct summary to the readers about the general opinion regarding an entity.

If there was no such system like sentiment analysis, people would have been unclear and be in constant dilemma whether to watch a particular movie or not. Without the sentiment analysis, a person have no clue whether the movie will be hit, super hit or flop. It not only affect the general audience but also the movie critics and ultimately the whole movie industry.

1.2.3. MOTIVATION OF THE PROJECT

In the completely changing digital world, word of mouth in form of reviews and ratings is playing a prominent role in success of any product released into the market. Most of these ratings are found online after a product's initial release. What if we have expected ratings before a product's release? If we have something like that available, anything releasing in the market could build some positive hype around it's release with this expected rating. This positive hype could lead to a product's success within days of release. Movies are also one type of product being released in the entertainment market. Our main idea is to create a positive hype for much deserving movies using this predicted expected ratings. Our predicted rating could play a crucial role in the success of the movie.

Movie review sites that were available online now a days shows the average rating for movies calculated using user ratings. Here, the user's reviews for a movie are not taken into consideration. We have seen recently, how some communities are using these rating sites for review-bombing the movies by giving a bad rating and hence creating a bad impression in audience for those movies. Keeping these in mind to avoid such things to happen, we thought of considering user's reviews for movie average rating instead of user ratings. Assigning the rating for each user review is the main task here. Using sentiment analysis on user reviews will make our task easier in assigning ratings to reviews.

1.2.4. OBJECTIVES OF THE PROPOSED WORK

The primary objective of this project is to compare the accuracy percentage of different machine learning models which are trained and tested on IMDB datasets to conclude which machine learning algorithm gives the best accuracy in determining the sentiments analysis.

1.2.5. OUTCOMES OF THE PROJECT

ML algorithms used classifies the sentiments as positive and negative. Different ML algorithms have been used to compare the accuracy %. Finally, the one with the highest accuracy % is recommended for the sentiment analysis purposes.

1.2.6. OVERVIEW OF THE PROJECT

The report is organized as follows- Chapter 1 provides a brief introduction about the sentiment analysis. Chapter 2 describes about the existing literature surveys conducted regarding sentiment analysis. Chapter 3 gives detailed methodology and system architecture of the project. It also covers detailed information about python and machine learning. Chapter 4 discusses about. The wholesome results and about its consequences. Chapter 5 concludes the work by summarizing the results.

CHAPTER 2
LITERATURE REVIEW

2.1. STUDY ON FEATURE SELECTION & CLASSIFICATION ALGORITHMS:

The present era of net has become enormous Cyber information that hosts large quantity of information that is formed and consumed by the users. The information has been growing at associate exponential rate giving rise to a replacement business crammed with it, during which users specific their opinions across channels like Facebook, Twitter, Rotten Tomatoes, and Foursquare. Opinions that area unit being expressed within the type of reviews offer a chance for brand spanking new explorations to seek out collective likes and dislikes of cyber community. One such domain of reviews is that the domain of picture reviews that affect everybody from audience, film critics to the assembly company. The picture show reviews being denoted on the websites don't seem to be formal reviews however square measure rather terribly informal and square measure unstructured kind of descriptive linguistics. An opinion expressed in picture show reviews provides a terribly true reflection of the feeling that is being sent. The presence of such a good North American nation of sentiment words to precise the review impressed us to plot Associate in Nursing approach to classify the polarity of the picture show exploitation these sentiment words. Sentiment Analysis could be a technology, which will be important within the next few years. With opinion mining, we are able to distinguish poor content from prime quality content. With the technologies on the market, we will able to apprehend if a pic has additional smart opinions than dangerous opinions and realize the explanations why those opinions are positive or negative. Much of the early research in this field was centered around product reviews, such as reviews on different products on Amazon.com, defining sentiments as positive, negative, or neutral. Most sentiment analysis studies are now focused on social media sources such as IMDB, Twitter and Facebook, requiring the approaches being tailored to serve the rising demand of opinions in the form of text.

2.2. RESEARCH PAPER BASED LITERATURE REVIEWS:

Vr, Nithin & Babu Pb, Sarath adopted many applications of machine learning such as linear regression and logistic regression [1]. The accuracy of their model using linear regression was 51%. While using logistic regression, they obtained about 42.2% accuracy. Sequel of movies cannot not be predicted from this model. Predicting only on the basis of one attribute that is gross revenue but success of movie should be relative and therefore cannot be predicted by using only one attribute.

Akshay Amolik and Dr.M.Venkatesan [2] in their paper used two machine learning classifie-rs Naïve Bayes and Support vector machine. Naïve Bayesian and Support vector machine performs well and also provide higher accuracy. The results show that they got 75 % accuracy form SVM and 65% accuracy form Naïve Bayesian classifier. They concluded that the accuracy of classification can be increased by increasing the training data.

Nagamma P and Pruthvi H.R [3] in their paper focused on classification using clustering and it resulted in good accuracy. Using clustering with a classification model reduces the data used for prediction. Hence the classification with or without clustering gave the same result as the data used for prediction is small. They concluded that the effect of using clustering with classification model could be seen predominantly if the dataset used is large.

As people spend more time watching movies through streaming services, the need also increases for efficient assessment of which movies to recommend. One proven method is sentiment analysis of movie reviews, which have become more available to the public through developments in online media. In the following article "A difference of multimedia consumer's rating and review through sentiment analysis" published by Lee, Jiang, Kong, and Liu in 2020, authors address a strong need for review-based sentiment analysis in the consumer world. "This is so because services are difficult to predict until they are experienced" [4].

Many of the challenges associated with sentiment analysis have been outlined in the surveys in [5]. In particular, a review may contain both positive and negative sentiments, but the review itself may be either positive, negative, or neutral. Further, a review may contain many words associated with positive sentiment, but the review may be, for example, negative overall. An early study by Peng et al. [6] showed that the use of a human-derived list of keywords to determine sentiment at the document level performed worse than a list of keywords chosen with simple statistics. Moreover, some of the keywords chosen with simple statistics would likely not be chosen by a human to determine sentiment.

According to Liu [7], sentiment computing was domain of study which analysed people's opinions, sentiments, appraisals, emotions, attitudes and evaluation towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes.

Verma et al (2009) [8] have proposed an interesting view, by predicting the sentiments via the structure of sentences, i.e. objectivity and subjectivity in the sentences. They considered the negation of some adjectives and phrases; they also assigned the maximum sentiment subjectivity value to the whole sentence. The sentiment subjectivity values were extracted from SentiWordNet, which contains polarity scores for sets of synonyms in English language. Later, and based on the sentiment vectors, they have created additional TF-IDF vectors and then they have calculated information gain. The maximum accuracy they have achieved was 83% after applying the information gain based pruning.

In 2004, a study by Pang and Lee utilized ML to analyze sentiment to select subjective sentences and use them to categorize text. Using a Naive Bayes model, they achieved an accuracy of 86.4% in determining sentence polarity [9]

The Sentiment analysis [10] is also understood as a task of determining the sentiment orientation of a given textual unit distinguished into two or more classes. Hence, the task of sentiment classification has also been implemented for different number of classes such as binary (e.g. positive/negative classification), ternary (e.g. positive/negative/neutral), n-ary (e.g. 1-5 star labeling).

Aurangzeb Khan, 2011 [11] proposed a rule-based technique in which SentiWordNet is used to obtain more accuracy than a pure lexicon-based technique for sentiment analysis for customer reviews and software reviews. The proposed system has 91% of accuracy at the document level and 86% of accuracy at the sentence level.

Lei Zhang et al., 2010 [12] proposed a ranking and extracting product features in opinion documents algorithm. Initially, they have reviews of users and it was difficult to determine by the machine to differentiate between positive reviews and negative reviews. They used the associated rule mining technique for extracting product features.

Rafeeque Pandarachalil et al., 2014 [13] proposed a method for Twitter sentiment analysis using an unsupervised learning approach. They determined the Polarity of tweets is evaluated by using three sentiment lexicons-SenticNet, SentiWordNet, and SentislangNet. They used parallel python framework to implement this method.

CHAPTER 3

PROPOSED METHODOLOGY

3.1. IN-DEPTH SENTIMENT ANALYSIS

A movie review is a work that reflects the opinions of the authors on and criticizes a particular movie positively or negatively, allowing everyone to appreciate the general idea of the film and decide to choose whether or not to watch it. A film analysis will influence the entire crew involved in the film. An analysis shows that the popularity or failure of a film depends in certain ways on its ratings. A significant issue, however, is to be captured, retrieved, quantified, and analyzed more accurately by classifying movie reviews. The rating of movie reviews in favorable or negative opinions is linked to terms in the review text as well as the climate is shown positively or negatively previously. These considerations help to improve the process of assessment comprehension using SA where SA is now the portal to customer understanding.

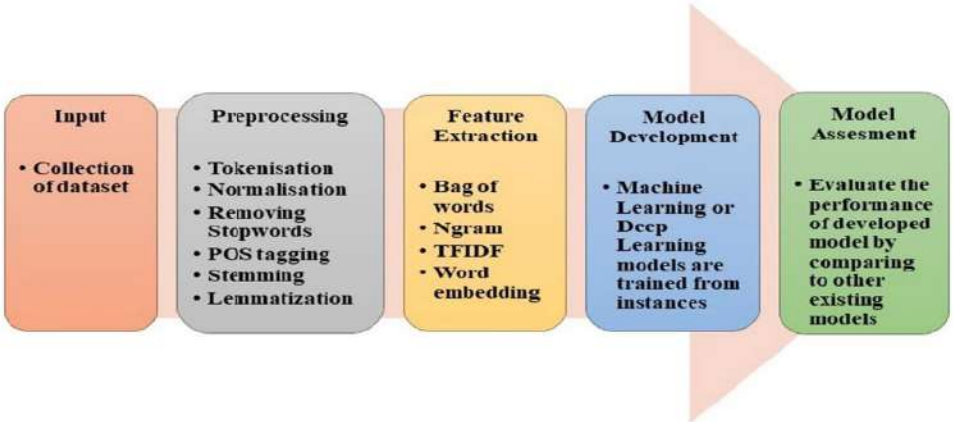

Fig.2 : Workflow of sentiment analysis

Sentiment Analysis refers to the use of natural language processing (NLP) and text analysis to schematically study, identify, extract and evaluate opinions within text. It builds systems that try to identify whether the statement makes a positive, negative impact or doesn't create an impact at all i.e., neural impact. Sentiment Analysis also called as Opinion Mining, besides identifying the opinion and its emotion, also extract attributes of the statement such as **Polarity:** the opinion expressed by speaker is positive or negative, **Opinion Holder:** who is expressing the opinion (the person or the entity), **Subject:** what is the product that is being discussed. Since SA has many practical applications, there has been a tremendous growth of interest in research and development of various Analysis and Prediction Techniques. The input fort this mechanism is the large number of texts expressing opinions which are available publicly and privately from various review sites, forums, blogs, and social media platforms such as Facebook, Twitter, LinkedIn, Reddit, Quora etc.

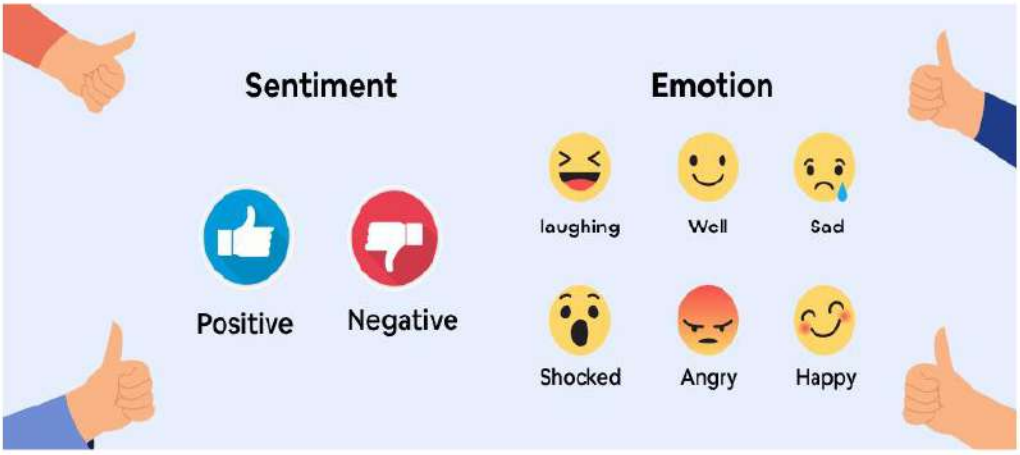

Fig.3 : Difference between sentiment and emotions

Sentiment analysis is a kind of data mining where you measure the inclination of people's opinions by using NLP (natural language processing), text analysis, and computational linguistics. We perform sentiment analysis mostly on public reviews, social media platforms, and similar sites.

3.1.1. TYPES OF SENTIMENT ANALYSIS

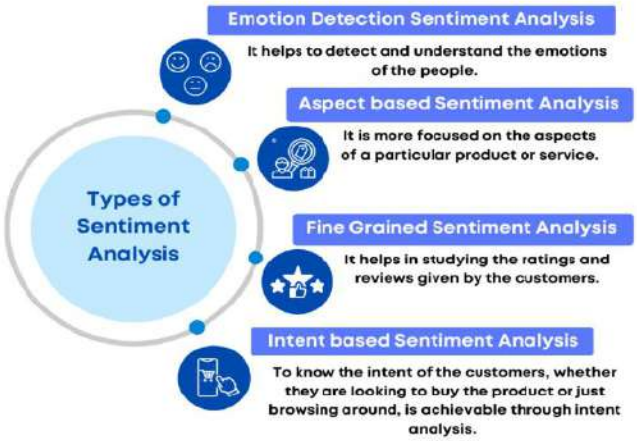

Fig.4 : Types of Sentiment Analysis

Fine-grained

Fine-grained sentiment analysis gives precise results to what the public opinion is about the subject. It classified its results in different categories such as: Very Negative, Negative, Neutral, Positive, Very Positive.

Detecting Emotion

This kind of sentiment analysis identifies emotions such as anger, happiness, sadness, and others. Many times, you'll use lexicons to recognize emotions. However, lexicons have drawbacks too, and in those cases, you'd need to use ML Algorithms.

Based on Aspect

In aspect-based sentiment analysis, you look at the aspect of the thing people are talking about. Suppose you have reviews of a smartphone, you might want to see what the people are talking about its battery life or its screen size.

Intent based Sentiment Analysis

Intent-based analysis distinguishes between facts and opinions in a text. An online comment indicating dissatisfaction with changing a battery, for example, can motivate customer service to contact you to remedy the problem.

All the algorithms that are being used into Sentiment Analysis systems may be broadly classified into following three types:

To identify the subjectivity, polarity, or the topic of an opinion, a **rule-based system** follows manually crafted rules in some kind of scripting language.Unlike the other approaches, the rule-based approach is quite easy to comprehend. It basically counts the total number of negative and positive words present in the data set. Following this, if the result indicates that the number of positive words is more than the number of negative words, then the sentiment is positive, and vice versa.

Automatic systems that emphasize on ML algorithms to learn from data. Contrary to rule-based systems, automatic systems do not use manual rules. Instead those systems implement the ML techniques. The data set is initially trained, following which predictive analysis is done. After completion of this stage, words are extracted from the text. The task of SA can be modeled into classification problem, which in turn can be solved using ML classifiers, like K-Nearest Neighbors, Logistic Regression, Support Vector or Machine, Decision Tree & Random Forest.

Hybrid Systems combine the features from both rule-based systems and automatic systems in order to improve the precision and accuracy.

3.1.2. LEVELS OF SENTIMENT ANALYSIS

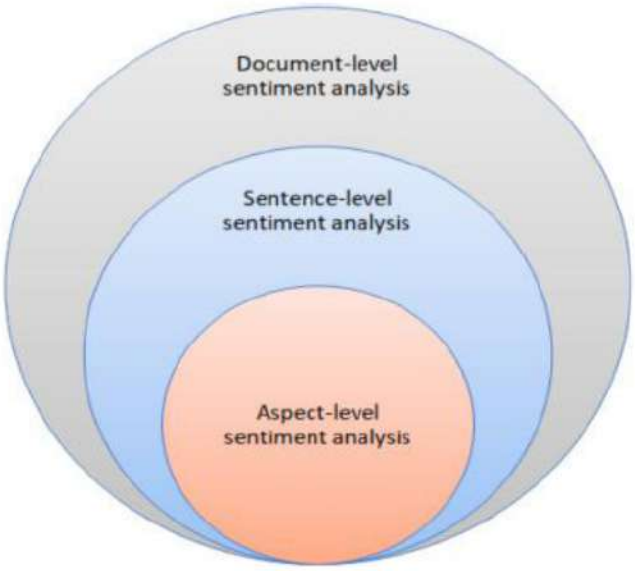

Fig.5 : Levels of Sentiment Analysis

Document-level

This includes examination of the document stage used for the whole document. A single subject paper is included with the level of classification. The customer has an attitude towards comparing two or two subjects in the study of the level of documents. The necessity of document-level sentiment analysis is to extract global sentiment from long texts that contain redundant local patterns and lots of noise. The most challenging aspect of document-level sentiment classification is taking into account the link between words and phrases and the full context of semantic information to reflect document composition. It necessitates a deeper understanding of the intricate internal structure of sentiments and dependent words

Sentence-level

The classification of subjectivity is directly linked to the study of sentence level. Examination of sentence-level is to locate the word from the sentence as good, negative, or neutral. For sentence-level SA, all classifiers in the text level analysis are used.

Aspect-level

The SA for the aspect level is used to detect sentiments about the aspect of these individuals. Let's take this instance. "My car has decent dealing but it is a little bit." There is an opinion in this case of a vehicle that the handling of a cat is good but the car is negative. The competitive comment is part of the SA on the aspect stage.

3.1.3. USE-CASES OF SENTIMENT ANALYSIS

Sentiment analysis tools are used in nearly every industry for a variety of applications:

i) Social media monitoring, a key strategy that tracks customer sentiments across social media platforms, such as Facebook, Instagram and Twitter.
ii) Monitoring brand awareness, reputation and popularity at a specific moment or over time.
iii) Analyzing consumer reception of new products or features to identify possible product improvements.
iv) Evaluating the success of a marketing campaign.
v) Pinpointing a target audience or demographic.
vi) Conducting market research, such as emerging trends and competitive insights.
vii) Categorizing customer service requests and automating customer service.
viii) Customer support analysis to assess the effectiveness of customer support and monitor trending issues.

3.1.4. BENEFITS OF SENTIMENT ANALYSIS

The benefits of sentiment analysis include the following:

i) Collecting large amounts of unstructured data from various sources.
ii) Tracking real-time customer feedback and sentiment about an organization's brand, products and services.

iii) Providing feedback on ways to improve products, services and customer experience.

iv) Getting data and feedback on problems with products and services.

v) Gathering data and feedback that keeps customer support staff up to date on customer issues and improves their ability to respond.

vi) Tracking the effectiveness of customer support through support tickets and other online feedback.

vii) Automating customer service by identifying customers' sentiments and automatically sending them to relevant FAQ responses for resolution.

viii) Identifying emerging marketing trends, and understanding and improving what marketing strategies resonate with customers.

ix) Gaining competitive insights by monitoring comments about competitors.

x) Establishing consistent criteria for evaluating sentiment instead of relying on subjective human analysis.

xi) Identifying and reacting to emerging negative sentiments before they escalate.

xii) Freeing employee time and energy for other tasks.

Sentiment analysis is an important way for organizations to understand how customers perceive and experience their products and brands. Increasingly, customer feedback is given online through a variety of unconnected platforms, such as Amazon product reviews and posts on social platforms. Organizations typically don't have the time or resources to scour the internet and read and analyze every piece of data relating to their products, services and brand. Instead, they use sentiment analysis algorithms to automate this process and provide real-time feedback. Organizations use this feedback to improve their products, services and customer experience. A proactive approach to incorporating sentiment analysis into product development can lead to improved customer loyalty and retention.

3.2. SYSTEM ARCHITECHTURE

System Architecture describes "the overall structure of the system and the ways in which the structure provides conceptual integrity".

The proposed framework for our model includes:

i) Data Collection
ii) Data Pre-processing
iii) Feature Extraction
iv) Applying Classifiers on the data
v) Comparing the results from the different classification models we used

3.2.1. DATA COLLECTION

Internet Movie Database (IMDb)

Fig.6 : IMDb (Internet Movie Database)

Large number of relevant pre-release movie attributes over a very large set of movies. General movie ratings on a scale of 0.0 to 10.0 based on a specified number of user votes. Although the weighted voting system isn't fully disclosed in order to reduce the so called vote stuffing, the voting system is according to the Internet Movie Database (no date) meant to reflect the overall user rating of a movie. The data is available for non-commercial use through plain text data files,

to be used within the terms of their copying policy. The Internet Movie Database provides a dataset consisting of a very large number of movie ratings, in most cases based on many thousand votes, as well as a wide selection of relevant attributes. As the dataset is not regarded freely distributable the experiment will be conducted according to the copyright policy of IMDb.com, inc. Information courtesy of IMDb (https://www.imdb.com)

We have gathered data from [1000] which includes a dataset that has 50000 reviews from IMDB which is equally divided into 25000 for training and testing. There are only 30 reviews per movie as reviews for the same movie tend to have correlated ratings. Furthermore, the train and test sets contain a disjoint set of movies so memorizing a particular movie terms and their associated labels would have no significance. A negative review is given a score of <=4 out of 10 while a positive one holds a score of >=7 and a neutral review has scores from >4 and <7.

3.2.2. DATA PRE-PROCESSING

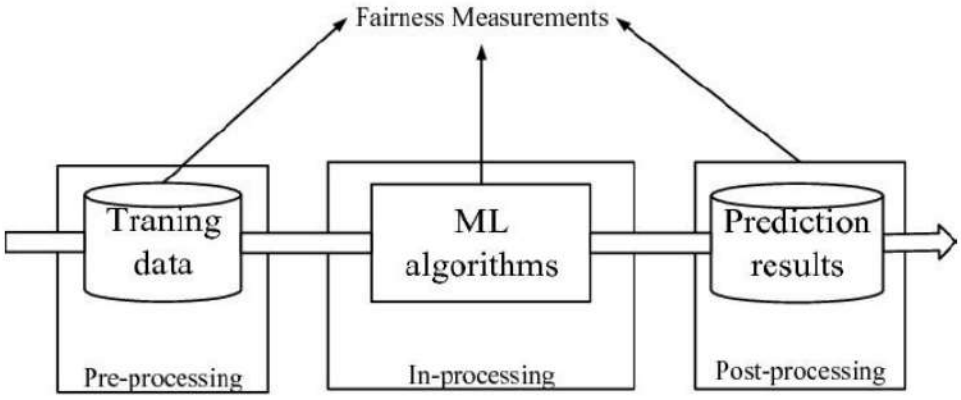

Fig. 7 : Machine Learning Pre-Processing, In-Processing and Post-Processing steps

Reviews are messy, people love to throw in attempts at expressing themselves more clearly by adding extravagant punctuation and spelling words incorrectly. Before training the model, the movie review data needs to be preprocessed as it contains a lot of unwanted data like stop words, punctuation, etc. To preprocess the data, we have used NLP techniques and regular expression

libraries. Preprocessing the data includes: **removal of HTML tags, removal of stop word & special characters, lemmatization and text tokenization.**

Removing HTML tags: The dataset has some unnecessary html tags which might affect the accuracy of our model. Hence, we have used regex to remove the tags.

```
def rmvhtmltags(text):
    remreg = re.compile('<.*?>')
    cleartext = re.sub(remreg, '', text)
    return text

def remove_urls (vTEXT):
    vTEXT = re.sub(r'(https|http)?:\/\/(\w|
\.|\/|\?|\=|\&|\%)*\b', '', vTEXT, flags=re
.MULTILINE)
    return(vTEXT)
```

Fig. 8 : Code Snippet of removing HTML tags

Lemmatization: It is a process of converting the given word to its root word. The main objective of lemmatization is to get proper morphological meaning of a word by referring it to the dictionary which is incorporated in the library. We have used wordnet and porter stemmer lemmatization.For example, words: o Smile o Smiling o Laugh o Laughing All the above words can be reduced to the root word "smile".

```
def lemmatize_words(text):
    lemmatized_words = [lemmatizer.lemmatiz
e(word, 'v') for word in text.split()]
    return('  '.join(lemmatized_words))
```

Fig. 9 : Code snippet of Lemmitization

Removing Stop words and special characters: The data consist of stop words like "a", "the", "this", "that", etc. These words mostly appear in a lot of reviews and are unimportant.

```
def rmvspclcharacter(text):
    clearspcl = re.sub(r'[^A-Za-z0-9\s.]',
r'', str(text).lower())
    clearspcl = re.sub(r'\n', r' ', text)

    clearspcl = " ".join([word for word in
text.split() if word not in stopWords])
    return text
```

Fig. 10 : Code snippet of removing stop words and special characters

Text Tokenization: Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of a token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph. For tokenization we have used NLTK library. It consists of many languages like German, English, Spanish, French, etc. trained with NLTK. In NLTK word tokenization is a wrapper function that utilizes treebank tokenization and splits the punctuations other than periods.

Any classification algorithm works well and produces better output only when the data is consistent. So pre-processing step in sentimental analysis plays a major role in training and testing of data. If data is not processed or partially processed, the accuracy may differ. This pre-processed data from the above methods is used to train the model for predictions. Text pre-processing is crucial to keep the dimensionality of the text low to improve the performance of the machine learning classifier . Thus, it is highly recommended to remove the noise as much as possible and to properly preprocess the text in this pre-analysis stage.

3.2.3. FEATURE EXTRACTION

Feature extraction is the process of converting a word into a matrix form. We have used the following approaches for feature extraction:

Bag of Words Approach:

In dataset, documents are represented as a vector and every word is converted into a number. The number can be binary (0 or 1). The BOW model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier. If a word appears in a document it gets a score 1 and if the word does not appears it gets a score 0.

Bag of Words model is the representation of the text data in the form of vectors which are machine understandable. It calculates total number of occurrences of each word in the sentence.It represents the words and their corresponding occurrences in each sentence of the data in the form of vectors in a matrix. This model emphasizes more on words rather than focusing on context in which the words are spoken. This method of study contains a sort of "Dictionary" which consists of words that add weight which is referred to as sentiment in this context. The textual records consists of tokens which have a specific "weight" when mentioned in the context. The sentiment valuation is simply the result of the addition of the weights derived from all the textual records. This model offers no importance to grammar, language essentials and thus lacks in the filed of Human Computer Interaction(HCI).

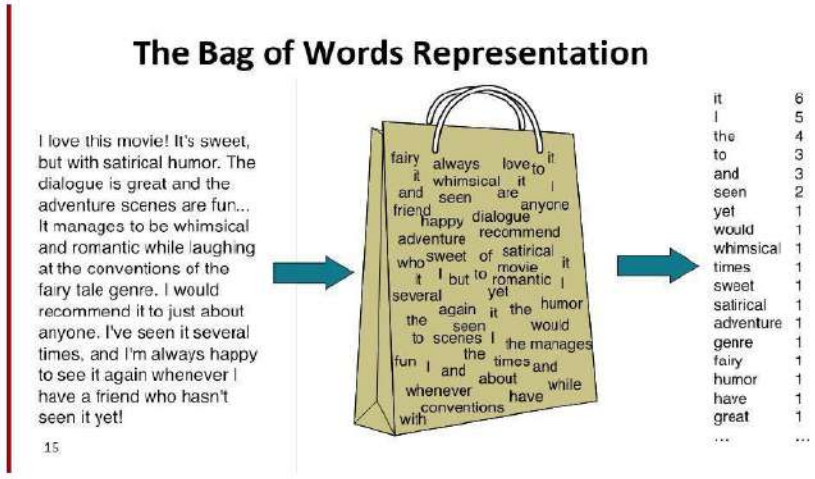

Fig.11 : BoW Representation

The advantage of this technique is its easy implementation but has significant drawbacks as it leads to a sparse matrix, loses the order of words in the sentence, and does not capture the meaning of a sentence. For example, to represent the text "are you enjoying reading" from the pre-defined dictionary I, Hope, you, are, enjoying, reading would be (0,0,1,1,1,1).

The bag-of-words (BOW) model is a representation that turns arbitrary text into fixed-length vectors by counting how many times each word appears. This process is often referred to as vectorization.

Vectorization:

Machines cannot understand the human language words, but it can understand numbers. To make machine understandable, we need to convert our text words to numbers. This can be done by vectorization process. There are many ways to vectorize the text. We have used following method for vectorization:

Count Vectorizer:

The process includes a blend of tokenizing a collection of documents from the datasets and then building a set of vocabulary for those words. The result of this is a length of vocabulary words and an integer value assigned for words according to how many times they appear. Words that do not occur may possess zeros as value and are defined as sparse. To show you how it works let's take an example:

```
text = ['Hello my name is james, this is my python notebook']
```

The text is transformed to a sparse matrix as shown below.

	hello	is	james	my	name	notebook	python	this
0	1	2	1	2	1	1	1	1

Fig.12 : Sparse Matrix of CountVectorizer of example 1

We have 8 unique words in the text and hence 8 different columns each representing a unique word in the matrix. The row represents the word count. Since the words 'is' and 'my' were repeated twice we have the count for those particular words as 2 and 1 for the rest.

Countvectorizer makes it easy for text data to be used directly in machine learning and deep learning models such as text classification. Let's take another example, but this time with more than 1 input:

```
text = ['Hello my name is james' , 'this is my python notebook']
```

I have 2 text inputs, what happens is that each input is preprocessed, tokenized, and represented as a sparse matrix. By default, Countvectorizer converts the text to lowercase and uses word-level tokenization.

	hello	is	james	my	name	notebook	python	this
0	1	1	1	1	1	0	0	0
1	0	1	0	1	0	1	1	1

Fig. 13 : Sparse Matrix of CountVectorizer of example 2

3.2.4. APPLYING OF CLASSIFIERS ON THE DATA

The entire dataset is divided into two parts for training and testing purposes: a training dataset and a testing dataset. The training dataset is the information used to train the model by supplying the characteristics of different instances of an item. The testing dataset is then used to see how successfully the model from the training dataset has been trained.

Now, we have our training and testing data. The next step is to identify the possible training methods and train our models.

In our experiment we have made use of Naïve Bayes, Logistic Regression, and Support Vector Machine. We have trained our model on the above classifiers to predict the movie polarity as positive or negative and discuss the accuracy percentage by comparing them.

The detailed overview of these used models are mentioned in section 3.6.9.

3.2.5. RESULT COMPARISON

After carrying out the abobe mentioned steps in system architecture, we got the results and after comparison we found out that **Logistic Regression** gave us the maximum accuracy and it outperforms the other two classifiers.

The detailed result analysis is mentioned in Chapter 4.

3.3. NATURAL LANGUAGE PROCESSING

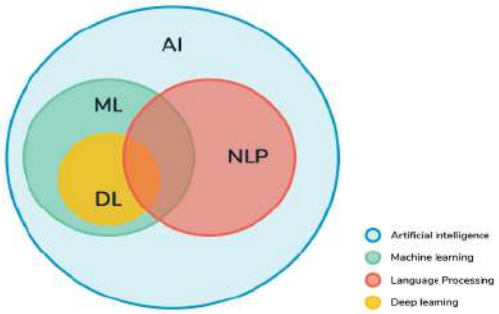

Fig.14 : NLP and its occupancy with AI, ML and DL

Natural Language Processing (NLP) is a branch of artificial intelligence that assists computers in understanding and explaining human language text or speech that contains a large amount of structured, semi-structured, and unstructured data. NLP defines the important parts of human language to computers and allows computers to interact with humans in their language, which involves a lot of processing. NLP can be applied to translation, search engine optimization, and filtering. The available search suggestions when using the built-in search bar in most of the applications are based on NLP content categorization and topic modeling. The emails are categorized due to the use of Bayesian spam filtering and a statistical model that compares the subject of the email with spam words to identify spam mails. Customer feedback and reviews about any application or organization can be determined using sentiment analysis, which predicts the user's feelings about them by extracting information from various sources.

NLP techniques include:

Named Entity recognition

This is a popular technique of NLP that is used in information extraction. This approach takes the sentence of the text as input and identifies all the nouns present in the input text. It is widely used in applications like news content categorization, search engines to retrieve information easily.

Tokenization

Tokenization is the process of splitting the data into tokens like characters, words, sentences, numbers. Tokenization is used for effective storage space for the data and decreases search degree. It is applied in most information retrieval systems.

Stemming and Lemmatization

Stemming is the process of decreasing all the words in the input text into their base or root form by removing the suffixes and making it easy for the model training. Lemmatization is used to obtain the proper vocabulary word for each word in the input text by transforming them to root form by understanding the meaning, parts of speech of the word.

Bag of Words

The bag of words model is used in text pre-processing to extract the features from the data to train the machine learning models. It counts the occurrence of each word in the sentence, and they can be represented in the form of vectors using vectorization. The bag of words model is used in several applications like email filtering, document classification. This concept from the field of Artificial Intelligence gives us way to understand the context, string of words and sentence structures. The machine need to understand the grammar principles. Tagging parts of speech, named entities are some of the techniques to perform the task of Natural Language Processing. Both the process has acquired fair results when applied to the different tasks. But the area of expertise may have an impact on the accuracy with which the task is done. The "Bag of Words" model requires massive amounts of machine learning concepts to be built in. Algorithms such as Support Vector Machine(SVM), Naïve Bayes Classifier, Maximum Entropy(MaxEnt) recognize patterns and add the weights. It is very common practice to use Bag of Words representation in Natural Language Processing for obtaining better results. The Bag of Words representation is obtained by using Naïve Bayes Classifier and the processing is done in Natural Language Tool Kit(NLTK). This method of processing help us to obtain promising results.

3.4. CHALLENGES

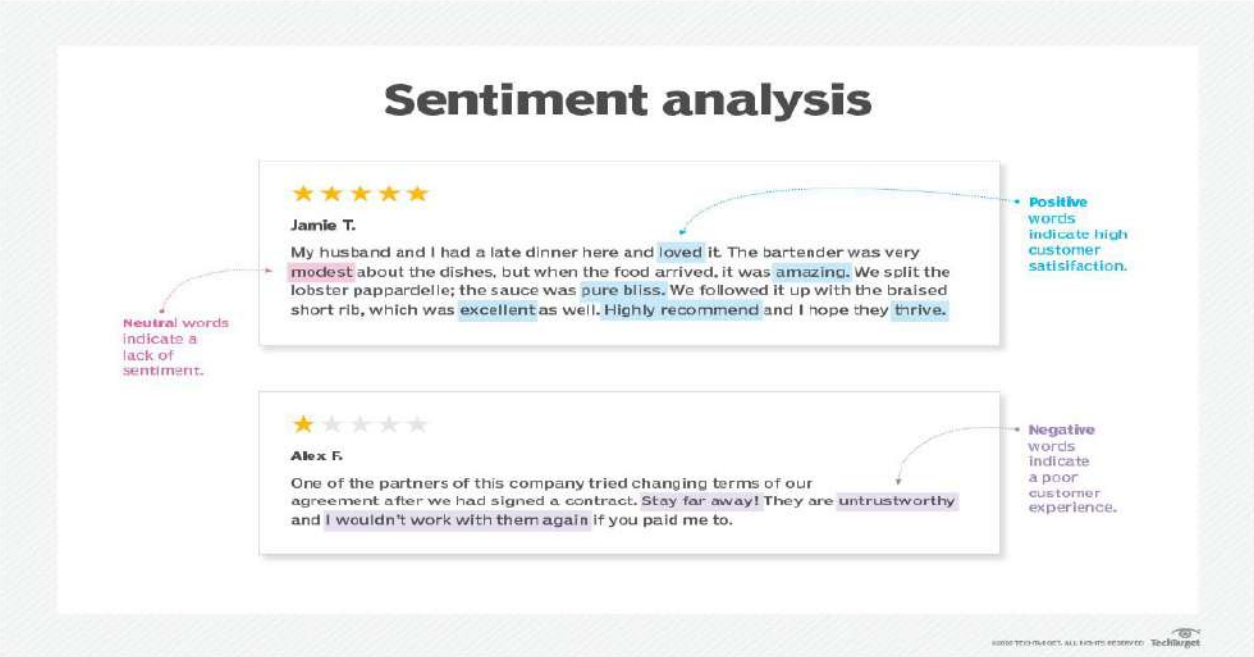

Fig.15 : Challenges of Sentiment Analysis

In the Internet era, people are generating a lot of data in the form of informal text. Social networking sites present various challenges, which includes spelling mistakes, new slang, and incorrect use of grammar. These challenges make it difficult for machines to perform sentiment and emotion analysis. Sometimes individuals do not express their emotions clearly. For instance, in the sentence "Y have u been soooo late?", 'why' is misspelled as 'y,' 'you' is misspelled as 'u,' and 'soooo' is used to show more impact. Moreover, this sentence does not express whether the person is angry or worried.

Therefore, sentiment and emotion detection from real-world data is full of challenges due to several reasons:

Sarcasm Detection

In sarcastic text, people express their negative sentiments using positive words. This fact allows sarcasm to easily cheat sentiment analysis models unless they're specifically designed to take its possibility into account. Sarcasm occurs most often in user-generated content such as Facebook comments, tweets etc. Sarcasm detection in sentiment analysis is very difficult to accomplish without having a good understanding of the context of the situation, the specific topic, and the environment.

Negation Detection

In linguistics, negation is a way of reversing the polarity of words, phrases, and even sentences. Researchers use different linguistic rules to identify whether negation is occurring, but it's also important to determine the range of the words that are affected by negation words. There is no fixed size for the scope of affected words. For example, in the sentence "The show was not interesting," the scope is only the next word after the negation word. But for sentences like "I do not call this film a comedy movie," the effect of the negation word "not" is until the end of the sentence. The original meaning of the words changes if a positive or negative word falls inside the scope of negation—in that case, opposite polarity will be returned.

Word Ambiguity

People can be contradictory in their statements. Most reviews will have both positive and negative comments. This situation can be managed by analyzing sentences one at a time. However, sentences that contain two contradictory words, also known as contrastive conjunctions, can confuse sentiment analysis tools. For example, "The packaging was terrible but the product was great." Word ambiguity is another pitfall you'll face working on a sentiment analysis problem. The problem of word ambiguity is the impossibility to define polarity in advance because the polarity for some words is strongly dependent on the sentence context. Lexicon-based sentiment analysis approaches are popular among existing methods. An opinion lexicon contains opinion words with their polarity value. There are some public opinion lexicons available on the internet: SentiWordNet, General Inquirer, and SenticNet, among others. Because word polarity varies in different domains, it is impossible to develop a universal opinion lexicon that has a polarity for every word.

For example:

“The story is unpredictable.”

“The steering wheel is unpredictable.”

These two examples show how context affects opinion word sentiment. In the first example, the word polarity of “unpredictable” is predicted as positive.

In the second, the same word’s polarity is negative.

Multipolarity

Sometimes, a given sentence or document—or whatever unit of text we would like to analyze—will exhibit multipolarity. In these cases, having only the total result of the analysis can be misleading, very much like how an average can sometimes hide valuable information about all the numbers that went into it. Picture when authors talk about different people, products, or companies (or aspects of them) in an article or review. It’s common that within a piece of text, some subjects will be criticized and some praised.

Neutral sentiments

Comments with a neutral sentiment tend to pose a problem for systems and are often misidentified. For example, if a customer received the wrong color item and submitted a comment, "The product was blue," this could be identified as neutral when in fact it should be negative.

Unclassifiable language

Computer programs have difficulty understanding emojis and irrelevant information. Special attention must be given to training models with emojis and neutral data so they don't improperly flag texts.

Small data sets

Sentiment analysis tools work best when analyzing large quantities of text data. Smaller data sets often won't provide the insight needed.

Fake reviews

Algorithms can't always tell the difference between real and fake reviews of products, or other pieces of text created by bots.

3.5. APPLICATIONS

Sentiment analysis has many applications in various industries. As it helps in understanding public opinion, companies use sentiment analysis in doing market research and figuring out if their customers like a particular product (or service) or not. Then, according to the findings of the sentiment analysis, the organization can modify the respective product or service and achieve better results. Applications of Sentiment Analysis-

Online Commerce

E-commerce practices are most often used for opinion analyses. Websites allow their users to share their shopping & product quality experience. The consumer and various characteristics of the product are described by giving assessments or grades. Customers can read reviews and recommendations on all products and individual product features quickly.

Voice of the Market (VOM)

VOM determines what consumers know about rivals' goods or services. Precise and appropriate market intelligence contributes to the strategic edge and production of the new products.

Voice of the Customer (VOC)

The Customer's voice is concerned with what each customer says about goods or services. It requires the analysis of consumer reviews and comments. VOC is a central factor in the management of user experience.

Brand Reputation Management

BRM is concerned with the business reputation. Customers' or some other party's opinions will harm or improve your credibility. BRM is not only a consumer but a product and organization. Now, a whole series of discussions are held online at a fast pace.

Government

By evaluating public perceptions, SA makes the government evaluate its strengths and disadvantages. For instance, "What do you expect from the truth if that is the state? The MP who examines the 2g racket is profoundly unethical." This post shows pessimistic government sentiments.

3.6. MACHINE LEARNING

3.6.1. HISTORY & THE FUTURE OF ML

Machine learning was first conceived from the mathematical modeling of neural networks. A paper by logician Walter Pitts and neuroscientist Warren McCulloch, published in 1943, attempted to mathematically map out thought processes and decision making in human cognition.

In 1950, Alan Turning proposed the Turing Test, which became the litmus test for which machines were deemed "intelligent" or "unintelligent." The criteria for a machine to receive status as an "intelligent" machine, was for it to have the ability to convince a human being that it, the machine, was also a human being. Soon after, a summer research program at Dartmouth College became the official birthplace of AI.

From this point on, "intelligent" machine learning algorithms and computer programs started to appear, doing everything from planning travel routes for salespeople, to playing board games with humans such as checkers and tic-tac-toe.

Intelligent machines went on to do everything from using speech recognition to learning to pronounce words the way a baby would learn to defeating a world chess champion at his own game. The info-graphic below shows the history of machine learning and how it grew from mathematical models to sophisticated technology.

In the late 1970s and early 1980s, artificial intelligence research focused on using logical, knowledge-based approaches rather than algorithms. Additionally, neural network research was abandoned by computer science and AI researchers. This caused a schism between artificial intelligence and machine learning. Until then, machine learning had been used as a training program for AI.

The machine learning industry, which included a large number of researchers and technicians, was reorganized into a separate field and struggled for nearly a decade. The industry goal shifted from training for artificial intelligence to solving practical problems in terms of providing services. Its focus shifted from the approaches inherited from AI research to methods and tactics used in probability theory and statistics. During this time, the ML industry maintained its focus on neural networks and then flourished in the 1990s. Most of this success was a result of Internet growth, benefiting from the ever-growing availability of digital data and the ability to share its services by way of the Internet.

The future of machine learning is expected to be characterized by continued advances in algorithms, computing power and data availability. As machine learning becomes more widely adopted and integrated into various industries, it has the potential to greatly impact society in a number of ways.

Some of the key trends and developments in the future of machine learning include:

Increased automation: As machine learning algorithms progress, they will be able to automate a larger range of jobs, requiring less human input and boosting productivity.

More personalized experiences: Machine learning algorithms will have the capacity to assess and make use of enormous volumes of data to deliver highly individualized experiences, such as personalized suggestions and adverts.

Enhanced judgment: As machine learning algorithms get better at making complicated judgments and predictions, numerous businesses will benefit from more precise and efficient decision-making.

AI ethical advancements: As machine learning becomes more common, there will be a growing emphasis on ensuring that it is developed and utilized ethically and responsibly, with a focus on safeguarding privacy and eliminating biases in decision-making.

Interdisciplinary collaboration: Machine learning will increasingly be used in collaboration with other fields, such as neuroscience and biology, to drive new discoveries and advancements in those areas.

Overall, the future of machine learning holds great promise and is expected to continue transforming a wide range of industries, from finance to healthcare, in the coming years.

3.6.2. MACHINE LEARNING AT PRESENT

Machine learning is now responsible for some of the most significant advancements in technology. It is being used for the new industry of self-driving vehicles, and for exploring the galaxy as it helps in identifying exoplanets. Recently, Machine learning was defined by Stanford University as "the science of getting computers to act without being explicitly programmed." Machine learning has prompted a new array of concepts and technologies, including supervised and unsupervised learning, new algorithms for robots, the Internet of Things, analytics tools, chatbots, and more. Listed below are seven common ways the world of business is currently using machine learning:

Analyzing Sales Data: Streamlining the data

Real-Time Mobile Personalization: Promoting the experience

Fraud Detection: Detecting pattern changes

Product Recommendations: Customer personalization

Learning Management Systems: Decision-making programs

Dynamic Pricing: Flexible pricing based on a need or demand

Natural Language Processing: Speaking with humans

Machine learning models have become quite adaptive in continuously learning, which makes them increasingly accurate the longer they operate. ML algorithms combined with new computing technologies promote scalability and improve efficiency. Combined with business analytics, machine learning can resolve a variety of organizational complexities. Modern ML models can be used to make predictions ranging from outbreaks of disease to the rise and fall of stocks.

Google is currently experimenting with machine learning using an approach called instruction fine-tuning. The goal is to train an ML model to resolve natural language processing issues in a generalized way. The process trains the model to solve a broad range of problems, rather than only one kind of problem.

3.6.3. INTRODUCTION

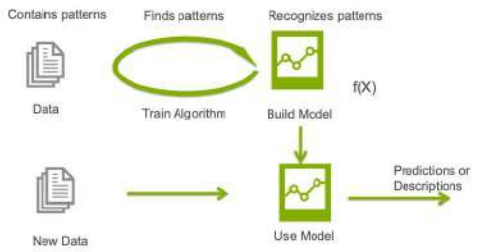

Fig.16 : Working of Machine Learning

As previously noted, machine learning makes it possible to discover trends, patterns or anomalies in a set of data, providing an effective method of analyzing large and complex datasets. Although the concept of machine learning have been established within the area computer science for several decades, the explosion of data in recent years have made the technology highly interesting for mining or extracting information and knowledge from large amounts of data.

Machine learning is a subset of artificial intelligence that aims on training computers to learn from the data and develop with the knowledge of data. Machine learning applications become more efficient when they have more data to learn and improve with their use. Machine learning algorithms are used to train the model to develop patterns and correlations between the features in large datasets and to make predictions based on the knowledge of training.

3.6.4. FEATURES

Machine learning has become one of the most important technological advancements in recent years and has significantly impacted a broad range of industries and applications. Its main features are:

Predictive modeling: Data is used by machine learning algorithms to create models that forecast future events. These models can be used to determine the risk of a loan default or the likelihood that a consumer would make a purchase, among other things.

Automation: Machine learning algorithms automate the process of finding patterns in data, requiring less human involvement and enabling more precise and effective analysis.

Scalability: Machine learning techniques are well suited for processing big data because they are made to handle massive amounts of data. As a result, businesses can make decisions based on information gleaned from such data.

Generalization: Algorithms for machine learning are capable of discovering broad patterns in data that can be used to analyze fresh, unexplored data. Even though the data used to train the model may not be immediately applicable to the task at hand, they are useful for forecasting future events.

Adaptiveness: As new data becomes available, machine learning algorithms are built to learn and adapt continuously. As a result, they can enhance their performance over time, becoming more precise and efficient as more data is made available to them.

3.6.5. APPLICATIONS

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications of Machine Learning:

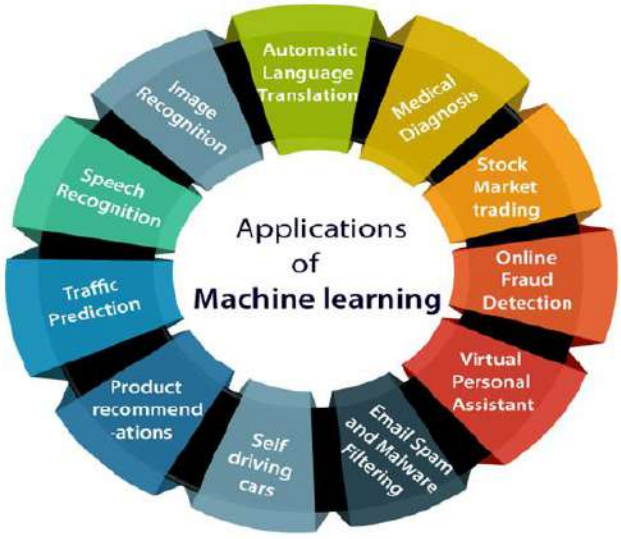

Fig.17 : Applications of Machine Learning

Image Recognition: Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion: Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm. It is based on the Facebook project named "Deep Face", which is responsible for face recognition and person identification in the picture.

Speech Recognition: While using Google, we get an option of "Search by voice" it comes under speech recognition, and it's a popular application of machine learning. Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition". At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

Traffic prediction: If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions. It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

Real Time location of the vehicle form Google Map app and sensors

Average time has taken on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

Product recommendations: Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for

the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest. As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

Self-driving cars: One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

Email Spam and Malware Filtering: Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail: Content Filter, Header filter, General blacklists filter, Rules-based filters, Permission filters. Some machine learning algorithms such as Multi-Layer Perceptron, Decision tree and Naïve Bayes classifier are used for email spam filtering and malware detection.

Virtual Personal Assistant: We have various virtual personal assistants such as Google assistant, Alexa, Cortana, Siri. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.
These virtual assistants use machine learning algorithms as an important part. These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

Online Fraud Detection: Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and steal

money in the middle of a transaction. So to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction.

For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

Medical Diagnosis: In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain. It helps in finding brain tumors and other brain-related diseases easily.

Automatic Language Translation: Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation. The technology behind the automatic translation is a sequence to sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

3.6.6. CHARACTERISTICS

The ability to perform automated data visualization:

A massive amount of data is being generated by businesses and common people on a regular basis. By visualizing notable relationships in data, businesses can not only make better decisions but build confidence as well. Machine learning offers a number of tools that provide rich snippets of data which can be applied to both unstructured and structured data. With the help of user-friendly automated data visualization platforms in machine learning, businesses can obtain a wealth of new insights in an effort to increase productivity in their processes.

Automation at its best:

One of the biggest characteristics of machine learning is its ability to automate repetitive tasks and thus, increasing productivity. A huge number of organizations are already using machine learning-powered paperwork and email automation. In the financial sector, for example, a huge number of repetitive, data-heavy and predictable tasks are needed to be performed. Because of this, this sector uses different types of machine learning solutions to a great extent. The make accounting tasks faster, more insightful, and more accurate. Some aspects that have been already addressed by machine learning include addressing financial queries with the help of chatbots, making predictions, managing expenses, simplifying invoicing, and automating bank reconciliations.

Customer engagement like never before:

For any business, one of the most crucial ways to drive engagement, promote brand loyalty and establish long-lasting customer relationships is by triggering meaningful conversations with its target customer base. Machine learning plays a critical role in enabling businesses and brands to spark more valuable conversations in terms of customer engagement. The technology analyzes particular phrases, words, sentences, idioms, and content formats which resonate with certain audience members. You can think of Pinterest which is successfully using machine learning to personalize suggestions to its users. It uses the technology to source content in which users will be interested, based on objects which they have pinned already.

Accurate data analysis:

Traditionally, data analysis has always been encompassing trial and error method, an approach which becomes impossible when we are working with large and heterogeneous datasets. Machine learning comes as the best solution to all these issues by offering effective alternatives to analyzing massive volumes of data. By developing efficient and fast algorithms, as well as, data-driven models for processing of data in real-time, machine learning is able to generate accurate analysis and results.

Business intelligence at its best:

Machine learning characteristics, when merged with big data analytical work, can generate extreme levels of business intelligence with the help of which several different industries are making strategic initiatives. From retail to financial services to healthcare, and many more – machine learning has already become one of the most effective technologies to boost business operations.

Whether you are convinced or not, the above characteristics of machine learning have contributed heavily toward making it one of the most crucial technology trends – it underlies a huge number of things we use these days without even thinking about them.

3.6.7. METHODS

The different categories of machine learning algorithms are supervised learning, unsupervised learning, reinforcement learning.

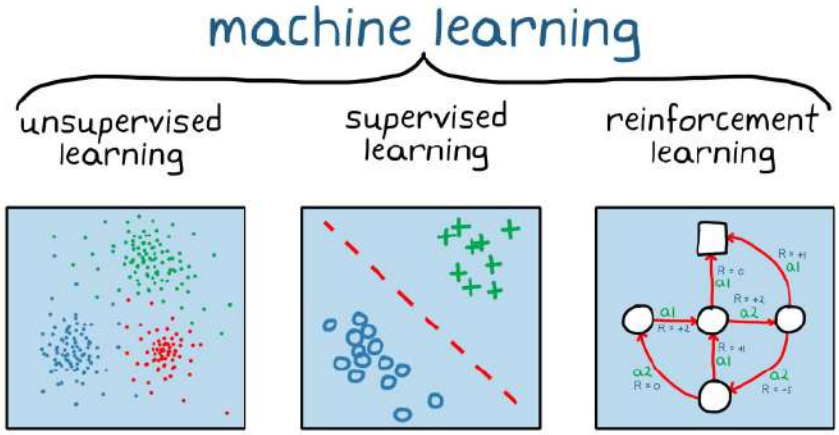

Fig.18 : Methods of Machine Learning

Supervised learning

Supervised learning is a method of training the model with labeled datasets and is used to predict the test data based on the trained datasets. In supervised learning, the model is trained by feeding the model with input data and as well as output data, the model learns from the training data and after training, the algorithm predicts new data based on the learning from training. The goal of supervised learning is to develop a pattern or a procedure that predicts the new test data based on the analysis of training data that already has the class label. In supervised learning, the input labeled data acts as the reference to predict the test data correctly. Supervised learning is classified into two types such as classification, regression. In classification, the algorithm predicts the class of the new test data and in regression, the algorithm predicts the real number of the new test data. Supervised learning is used in many real-world applications such as image classification, spam detection, risk assessment.

The supervised learning techniques use machine learning on a previously classified to be almost accurate. These pre-classified datasets are often domain specific, therefore the model it generate can work only for a particular domain. These datasets are first converted into intermediate models where documents are represented as vectors and then the intermediate representations are fed to the machine learning algorithm.

As previously described supervised learning is based on the notion of being able to predict an unknown attribute of an object based on the attributes we actually know. A supervised learning algorithm could for example be used to predict the value of a house using information that we know, such as the number of rooms, age and location. This type of predictions are made possible by training the supervised machine learning algorithm on data where our target attribute is known, as illustrated in Figure.

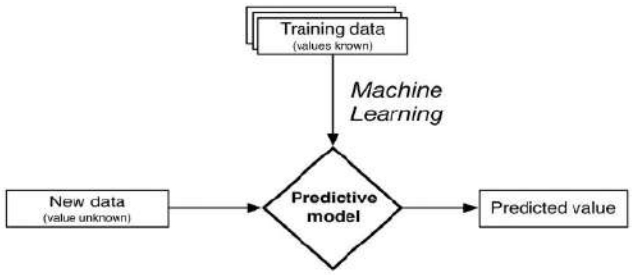

Fig.19 : Supervised Learning

Unsupervised learning

Unsupervised learning is a method of training the model using algorithms to analyze and cluster unlabeled datasets without any reference. It is like learning new things with the human brain. In unsupervised learning, only input data is provided to the model at the time of training but not the output data. Unsupervised learning aims to learn the structure of the dataset and predict the test data based on the similarities and characterize the data in a unique format. Unsupervised learning can be used, when there is no prior labeled data set for training and in more complex task processing. It requires minimum human supervision compared to supervised learning. Unsupervised learning is categorized into two types such as clustering, association. Unsupervised learning is used in many areas that include market basket analysis, pattern recognition, identifying accident-prone areas, and many business models.

The unsupervised learning techniques mainly use lexicon based approach where they use existing lexical resources like WordNet and language specific sentiment seed words to construct and update sentiment prediction. Although unsupervised learning algorithms do not require a corpus of previously classified data and generates a general sentiment, they fail to capture context or domain specific information of the document.

Reinforcement learning

Reinforcement learning is an approach in which the machines learn by communicating with the environment. In this approach, the machine learns by performing different operations in which there will be rewards and punishment for each step. During the training of the model, there will

be a reward for every appropriate action and a punishment for every inappropriate action. The main goal of reinforcement learning is to find the strategy such that it maximizes the number of rewards. In reinforcement learning, the machine works on its own without any supervision. Various applications of reinforcement learning include industrial automation in robotics, development of games, marketing, and advertising.

3.6.8. CLASSIFICATION

Classification is the method of supervised learning that is used in the prediction of discrete data. In classification, the machine learns from the labeled data and classifies the data into different classes. Classification can be binary or multi-class classification based on the classes in the training dataset. In classification, the trained data is grouped into different target classes and the test data is predicted based on the target classes. Classification algorithms are applied in different real-world applications such as sentiment analysis, spam classification, and document classification.

3.6.9. ALGORITHMS

SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) algorithm is a supervised learning algorithm that can be used for both regression and classification problems. The main aim of the Support Vector Machine algorithm is to find the best possible hyperplane that uniquely classifies all the data points plotted in the N-dimensional space. The hyperplane can be in different dimensions, which is based on the number of independent features in the dataset and the best hyperplane is chosen considering the largest margin or separation between the data classes. Support Vector Machine can be used for linear and non-linear classification problems by using different kernel functions. The division of the hyperplane is represented in figure.

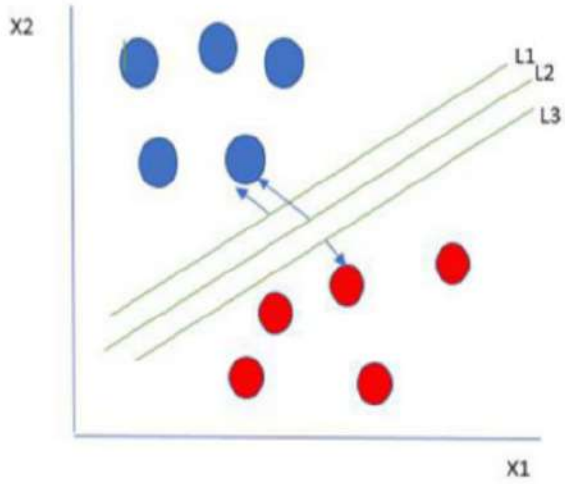

Fig. 20 : SVM

The major advantages of support vector machines is effectiveness in high dimensional spaces. Also it uses a subset of training points in the decision function called support vectors, so it is also memory efficient. They achieve high generalization by maximizing the margin, and that they support an efficient learning of nonlinear functions using kernels. This is also confirmed by recent studies on the usage of SVM for machine learning applications, where the algorithm yield good prediction performance in comparison to other well respected algorithms such as artificial neural networks and random forests. As such, the SVM should be able to effectively handle very high dimensional and continuous data. In sci-kit library, different types of kernels such as linear, RBF and polynomial are provided.

The one drawback in SVM is when training data is highly unbalanced, resulting model tends to perform well on majority data but perform bad on minority data.

SVM is described as one of the most robust and accurate methods among all well-known machine learning algorithms, and even though it was originally intended for classification it has been noted that the algorithm could easily be extended to perform numerical predictions in the form of support vector regression (SVR), as well as time series predictions. The support vector algorithm dates back to the 1960s, but was in its current form mostly developed during the 1990s

for the purpose of optical character recognition (OCR), in converting printed or written characters into machine encoded text. Even though SVM during the last decades have been extended for multi-class classification as well as regression, ranking and time series prediction, it is worth noting that the algorithm was originally developed for binary classification, where the output is categorized into one of two groups as either positive or negative.

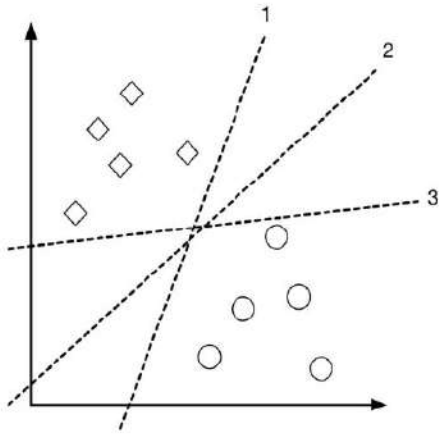

Fig.21 : Illustration of linear binary SVM classifiers in 2-D

Figure above show a simplified representation of such a scenario, where the ten data points are classified into one of the two groups consisting of either diamonds or circles, as each datapoint consist of a two dimensional vector representing its position in the coordinate system. Just as illustrated, the two groups would have been correctly separated and thus correctly classified by all three linear classifiers, represented by dashed lines and numbers 1-3 in the figure. In order for better generalizations and thereby better predictions on unseen data, the SVM tries to obtain the maximum separation between the two groups by creating a line or hyperplane with the largest possible margin to the nearest data points from both groups.

As the SVM can only separate data linearly using a flat line or hyperplane, separating non-linear data isn't directly possible. By using a kernel function to map the non-linear input data into a different, higher dimensional feature space, a linear hyperplane could however still be used for successfully separating the non-linear dataset into the two classes. Figure 3 below show a

simplified illustration of how a two dimensional dataset could be represented by a three dimensional feature space, making it possible to separate the circle objects from the diamonds by the means of a linear hyperplane, despite the objects being linearly inseparable within the original two dimensional input space. One of the most widely used kernels for the SVM is the radial basis function (RBF), mapping data to an infinite dimensional space.

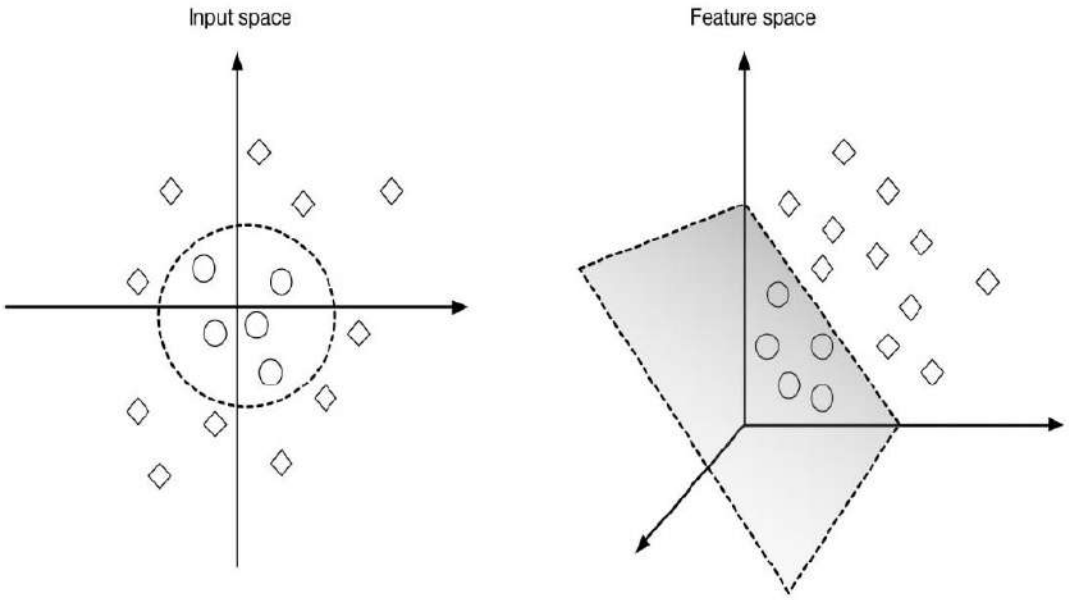

Fig.22 : 2-D dimentional input space mapped into a 3-D feature space

Like most other machine learning algorithms and prediction models the SVM require a correct setup in terms of its configuration parameters in order to perform well. There are two relevant parameters to consider for an SVM model based on the RBF kernel: the error penalty parameter *C* and the kernel coeffificient *gamma*. As the error penalty parameter *C* controls the margin of misclassification allowed within the prediction model, a value too large will result in the hyperplane being tuned to classify every training example correctly, thus risking overfitting the data causing low generalizability, while a *C* value too small will result in underfitting causing an error prone prediction model. As the *gamma* parameter is used to tune the the mapping

between input space and feature space within the RBF kernel, it also require considerable tuning in order to get an optimal prediction model. Support vector regression works in mostly the same manner as the classifying SVM, utilizing the same margin optimization and kernel functions for nonlinear input data. Support vector regression is the most common application form of all support vector machines, which considering that SVM is regarded as one of the most popular machine learning algorithms give the algorithm an important role in many applications.

NAIVE BAYES

Naive Bayes classifier is a classification algorithm that depends on the Bayes theorem with an assumption of independence between the data points. It assumes that the presence of one data point is not correlated to the presence of another data point. To make it as simple as possible to understand, in general terms we can say that "a classifier which makes the use of Naïve Bayes algorithm presuppose that the occurrence of a precise feature in a class is not related to the occurrence of any other feature". For illustration, a fruit which is black in colour, oval in shape and having a diagonal length about 3cm may be considered to be an grape. There are 3 types of models are available for Naïve Bayes classifier in scikit-learn library:

Gaussian Naive Bayes approach is applied for the classification problems which is assumed to be a normal distribution. Gaussian Naive Bayes distribution function is:

$$P(x_i y) = 1/\sqrt{2\pi\sigma^2} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

where sigma y and mu y are the assumptions of likelihood.

Multinomial approach is applied for discrete data.

Bernoulli is used for classifying when the feature vectors are binary (i.e. consist of zeros and ones). One application would be to classify the text with 'bag of paragraph' model where the 0s

& 1s are "word is a sub-string of the sentence" and "word is not a sub-string of the sentence" respectively.

Multinomial model is used under Naïve Bayes classifier.

Bayes theorem provides a way of calculating posterior probability:

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

Where $P(c|x)$ = Posterior Probability
$P(c)$ = Class Prior Priority
$P(x|c)$ = Likelihood
$P(x)$ = Predictor Prior Probability
$P(c|x) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$

Naive Bayes techniques are applied in real-world applications such as text classification, recommendation systems. It is a highly scalable algorithm which requires many of the parameters to be linear in the problem to be solved.Naive Bayes model is easy & simple to build and is highly useful for classifying text in very large data sets. Along with its simplicity, it is also known to perform even better than highly advanced &sophisticated classification algorithms.

LOGISTIC REGRESSION

Logistic Regression algorithm is used in the classification to find the category of the dependent feature with the set of independent features. It is categorized into different types such as binary, multinomial, and ordinal logistic regression based on the dependent features. It predicts the likelihood of the test data and classifies it into one of the classes of dependent features. Only discrete data can be predicted using Logistic Regression which differs from Linear Regression, which can be used to predict the continuous data. Logistic Regression can be applied in many applications like predicting spam classification, the effect of the disease(low/high/medium). The logistic curve is represented in figure.

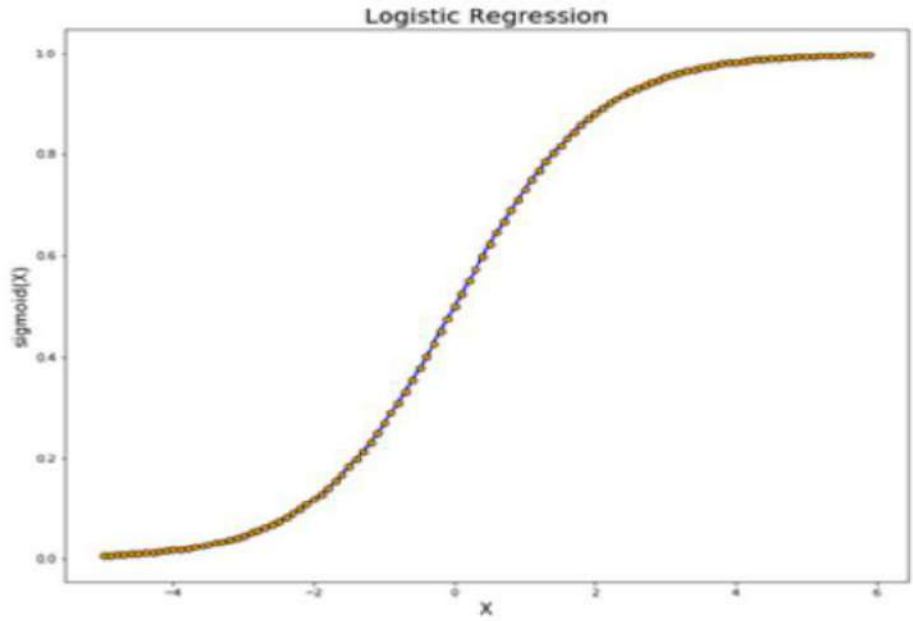

Fig.23 : LR (Logistic Regression Curve)

Despite its name, it is a linear model for classification rather than regression. It is also known in the literature as logit regression, log-linear classifier and maximum-entropy classification (MaxEnt). In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. A logistic function or logistic curve is a common S shape (sigmoid curve), with equation:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Logistic regression can be binomial, ordinal or multinomial.

Binomial,otherwise known as binary logistic regression deals with situations where the observed outcome for a dependent variable has only two possible types, "0" and "1" (which may represent, "yes" vs. "no" or "true" vs. "false").

In order to represent the binary outcome or categorical outcome, we make use of certain dummy variables.

Hypothesis => Z = WX + B

hΘ(x) = sigmoid (Z)

Sigmoid(t) = 1/(1+e-t)

Cost Function

$$Cost(h_\theta(x), Y(actual)) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases}$$

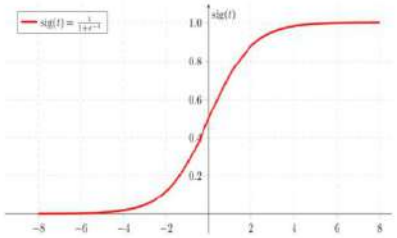

Fig.24 : Sigmoid Function

The second type, Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., "Movie A" vs. "Movie B" vs. "Movie C"). The third type, Ordinal logistic regression deals with dependent variables that are ordered. A logistic function is used to determine the relationship between categorically dependent and one or more independent variables. Advantages of Logistic Regression: It is much more robust to correlated features. If two features f1 and f2 are perfectly correlated, regression will simply assign half the weight to w1 and half to w2. It is discriminative. It works well on large datasets when compared with naïve bayes.

3.7. PYTHON

3.7.1. INTRODUCTION

Fig.25 : Python

Python is an interpreted, high-level, general-purpose programming language. Python is simple and easy to read syntax emphasizes readability and therefore reduces system maintenance costs. Python supports modules and packages, which promote system layout and code reuse. It saves space but it takes slightly higher time when its code is compiled. Indentation needs to be taken care while coding. Python does the following:

i) Python can be used on a server to create web applications.
ii) It can connect to database systems. It can also read and modify files.
iii) It can be used to handle big data and perform complex mathematics.
iv) It can be used for production-ready software development.
v) Python has many inbuilt library functions that can be used easily for working with machine learning algorithms. All the necessary python libraries must be pre-installed using "pip" command.
vi) It is often described as a "batteries included" language due to its comprehensive standard library.
vii)It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming.

3.7.2. HISTORY

Python is a widely-used general-purpose, high-level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. In the late 1980s, history was about to be written. It was that time when working on Python started. Soon after that, Guido Van Rossum began doing its application-based work in December of 1989 at Centrum Wiskunde & Informatica (CWI) which is situated in the Netherlands. It was started firstly as a hobby project because he was looking for an interesting project to keep him occupied during Christmas. The programming language in which Python is said to have succeeded is ABC Programming Language, which had interfacing with the Amoeba Operating System and had the feature of exception handling. He had already helped to create ABC earlier in his career and he had seen some issues with ABC but liked most of the features. After that what he did was really very clever. He had taken the syntax of ABC, and some of its good features. It came with a lot of complaints too, so he fixed those issues completely and had created a good scripting language that had removed all the flaws. The inspiration for the name came from BBC's TV Show – 'Monty Python's Flying Circus', as he was a big fan of the TV show and also he wanted a short, unique and slightly mysterious name for his invention and hence he named it Python! He was the "Benevolent dictator for life" (BDFL) until he stepped down from the position as the leader on 12th July 2018. For quite some time he used to work for Google, but currently, he is working at Dropbox.

The language was finally released in 1991. When it was released, it used a lot fewer codes to express the concepts, when we compare it with Java, C++ & C. Its design philosophy was quite good too. Its main objective is to provide code readability and advanced developer productivity. When it was released it had more than enough capability to provide classes with inheritance, several core data types exception handling and functions.

Python consistently ranks as one of the most popular programming languages. Since 2003, Python has consistently ranked in the top ten most popular programming languages in the TIOBE Programming Community Index.

3.7.3. PYTHON LIBRARIES

The Python Standard Library contains the exact syntax, semantics, and tokens of Python. It contains built-in modules that provide access to basic system functionality like I/O and some other core modules. Most of the Python Libraries are written in the C programming language. The Python standard library consists of more than 200 core modules. All these work together to make Python a high-level programming language. Python Standard Library plays a very important role. Without it, the programmers can't have access to the functionalities of Python. But other than this, there are several other libraries in Python that make a programmer's life easier. Let's have a look at some of the commonly used libraries:

NumPy: NumPy is a general-purpose array-processing package. It provides a high performance multidimensional array object, and tools for working with these arrays.It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

i) A powerful N-dimensional array object
ii) Sophisticated (broadcasting) functions
iii) Tools for integrating C/C++ and Fortran code
iv) Useful linear algebra, Fourier transform, and random number capabilities
v) Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data.

Pandas: Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is fast and it has high-performance & productivity for users. It provides high-performance and is easy-to-use data structures and data analysis tools for the Python language. It is used to create data frames and perform operations on data frames. It is used in a wide range of fields including academic and commercial domains including economics, Statistics, analytics, etc.

SkLearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It is an open-source Python library that implements a range of machine learning, pre-processing, cross-validation and visualization algorithms using a unified interface. Sklearn provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

NLTK: It is used to pre-process the data provided by humans into machine understandable language. It is the main platform that has the modules to perform human language related operations.

matplotlib - It is a comprehensive python library for creating interactive and animated visualizations in python.

3.7.4. FEATURES

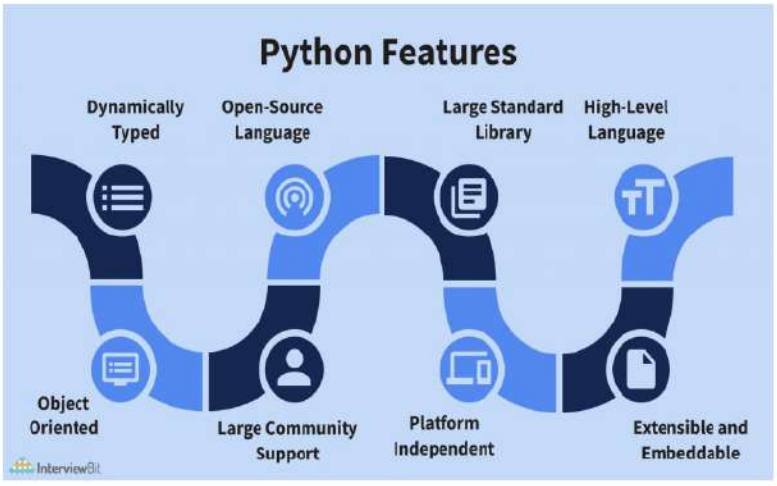

Fig.26 : Python Features

Interpreted Language

Python is an interpreted language (an interpreted language is a programming language that is generally interpreted, without compiling a program into machine instructions. It is one where the instructions are not directly executed by the target machine, but instead, read and executed by some other program known as the interpreter) and an IDLE (Interactive Development Environment) is packaged along with Python. It is nothing but an interpreter which follows the REPL (Read Evaluate Print Loop) structure just like in Node.js. IDLE executes and displays the output of one line of Python code at a time. Hence, it displays errors when we are running a line of Python code and displays the entire stack trace for the error.

Dynamically Typed Language

Python is a dynamically typed language. In other words, in Python, we do not need to declare the data types of the variables which we define. It is the job of the Python interpreter to determine the data types of the variables at runtime based on the types of the parts of the expression. Though it makes coding easier for programmers, this property might create runtime errors. To be specific, Python follows duck typing. It means that "If it looks like a duck, swims like a duck and quacks like a duck, it must be a duck."

Open Source And Free

Python is an open-source programming language and one can download it for free from Python's official website. The community of Python users is constantly contributing to the code of Python in order to improve it.

Large Standard Library

One of the very important features because of which Python is so famous in today's times is the huge standard library it offers to its users. The standard library of Python is extremely large with a diverse set of packages and modules like itertools, functools, operator, and many more with

common and important functionalities in them. If the code of some functionality is already present in these modules and packages, the developers do not need to rewrite them from scratch, saving both time and effort on the developer's end. Moreover, the developers can now focus on more important things concerning their projects. Also, Python provides the PyPI (Python Package Index) which contains more packages that we can install and use if we want even more functionality.

High-Level Language

A high-level language (HLL) is a programming language that enables a programmer to write programs that are more or less independent of a particular type of computer. These languages are said to be high-level since they are very close to human languages and far away from machine languages. Unlike C, Python is a high-level language. We can easily understand Python and it is closer to the user than middle-level languages like C. In Python, we do not need to remember system architecture or manage the memory.

Object Oriented Programming Language

Python supports various programming paradigms like structured programming, functional programming, and object-oriented programming. However, the most important fact is that the Object-Oriented approach of Python allows its users to implement the concepts of Encapsulation, Inheritance, Polymorphism, etc. which is extremely important for the coding done in most Software Industries as objects map to entities in real-world and lot of real-world problems can be solved using the Object-Oriented Approach.

Platform Independent

Platform independence is yet another amazing feature of Python. In other words, it means that if we write a program in Python, it can run on a variety of platforms, for instance, Windows, Mac, Linux, etc. We do not have to write separate Python code for different platforms.

Extensible and Embeddable

Python is an Embeddable language. We can write some Python code into C or C++ language and also we can compile that code in C/C++ language. Python is also extensible. It means that we can extend our Python code in various other languages like C++, etc. too.

Graphical User Interface (GUI) Support

Yet another interesting feature of Python is the fact that we can use it to create GUI (Graphical User Interfaces). We can use Tkinter, PyQt, wxPython, or Pyside for doing the same. Python also features a large number of GUI frameworks available for it and various other cross-platform solutions. Python binds to platform-specific technologies.

Frontend and Backend Development

With a new project py script, you can run and write Python codes in HTML with the help of some simple tags <py-script>, <py-env>, etc. This will help you do frontend development work in Python like javascript. Backend is the strong forte of Python. It's extensively used for this work cause of its frameworks like Django and Flask.

3.7.5. APPLICATIONS

Web Applications

We can use Python to develop web applications. It provides libraries to handle internet protocols such as HTML and XML, JSON, Email processing, request, beautifulSoup, Feedparser, etc. One of Python web-framework named Django is used on Instagram. Python provides many useful frameworks, and these are given below:

i) Django and Pyramid framework (Use for heavy applications)

ii) Flask and Bottle (Micro-framework)

iii) Plone and Django CMS (Advance Content management)

Desktop GUI Applications

The GUI stands for the Graphical User Interface, which provides a smooth interaction to any application. Python provides a Tk GUI library to develop a user interface. Some popular GUI libraries are: Tkinter or Tk, wxWidgetM, Kivy (used for writing multitouch applications), PyQt or Pyside.

Console-based Application

Console-based applications run from the command-line or shell. These applications are computer program which are used commands to execute. This kind of application was more popular in the old generation of computers. Python can develop this kind of application very effectively. It is famous for having REPL, which means the Read-Eval-Print Loop that makes it the most suitable language for the command-line applications.

Python provides many free library or module which helps to build the command-line apps. The necessary IO libraries are used to read and write. It helps to parse argument and create console help text out-of-the-box. There are also advance libraries that can develop independent console apps.

Software Development

Python is useful for the software development process. It works as a support language and can be used to build control and management, testing, etc.

SCons is used to build control.

Buildbot and Apache Gumps are used for automated continuous compilation and testing.

Round or Trac for bug tracking and project management.

Scientific and Numeric

This is the era of Artificial intelligence where the machine can perform the task the same as the human. Python language is the most suitable language for Artificial intelligence or machine learning. It consists of many scientific and mathematical libraries, which makes easy to solve complex calculations.

Implementing machine learning algorithms require complex mathematical calculation. Python has many libraries for scientific and numeric such as Numpy, Pandas, Scipy, Scikit-learn, etc. If you have some basic knowledge of Python, you need to import libraries on the top of the code. Few popular frameworks of machine libraries are: SciPy, Scikit-learn, NumPy, Pandas, Matplotlib.

Business Applications

Business Applications differ from standard applications. E-commerce and ERP are an example of a business application. This kind of application requires extensively, scalability and readability, and Python provides all these features.

Oddo is an example of the all-in-one Python-based application which offers a range of business applications. Python provides a Tryton platform which is used to develop the business application.

Audio or Video-based Applications

Python is flexible to perform multiple tasks and can be used to create multimedia applications. Some multimedia applications which are made by using Python are **TimPlayer, cplay,** etc. The few multimedia libraries are: Gstreamer, Pyglet, QT Phonon.

3D CAD Applications

The CAD (Computer-aided design) is used to design engineering related architecture. It is used to develop the 3D representation of a part of a system. Python can create a 3D CAD application

by using the functionalities like Fandango (Popular), CAMVOX, HeeksCNC, AnyCAD, RCAM.

Enterprise Applications

Python can be used to create applications that can be used within an Enterprise or an Organization. Some real-time applications are OpenERP, Tryton, Picalo, etc.

Image Processing Application

Python contains many libraries that are used to work with the image. The image can be manipulated according to our requirements. Some libraries of image processing are: OpenCV, Pillow, SimpleITK.

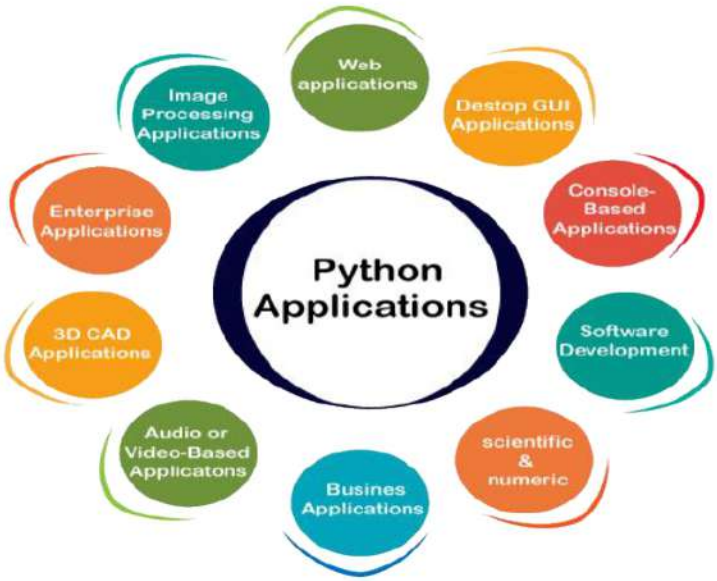

Fig. 27 : Python Applications

3.8. MOVIES

3.8.1. BRIEF

Cinematography is the illusion of motion by recording many still photographs and then showing them on a screen as quickly as possible. Originally a product of 19th-century scientific endeavour, cinema has become a medium of mass entertainment and communication, and today it is a multi-billion-pound industry.

3.8.2. HISTORY

The **history of film** chronicles the development of a visual art form created using film technologies that began in the late 19th century. The advent of film as an artistic medium is not clearly defined. However, the commercial, public screening of ten of the Lumière brothers' short films in Paris on 28 December 1895, can be regarded as the breakthrough of projected cinematographic motion pictures. There had been earlier cinematographic results and screenings by others, like the Skladanowsky brothers, who used their self-made Bioscop to display the first moving picture show to a paying audience on 1 November 1895, in Berlin, but they had neither the quality, financial backing, stamina, or luck to find the momentum that propelled the cinématographe Lumière into worldwide success. Those earliest films were in black and white, under a minute long, without recorded sound, and consisted of a single shot from a steady camera. The first decade of motion pictures saw film move from a novelty to an established mass entertainment industry, with film production companies and studios established all over the world. Conventions toward a general cinematic language also developed, with editing camera movements and other cinematic techniques contributing specific roles in the narrative of films.

Popular new media, including television (mainstream since the 1950s), home video (mainstream since the 1980s), and the internet (mainstream since the 1990s), influenced the distribution and consumption of films. Film production usually responded with content to fit the new media, and

with technical innovations (including widescreen (mainstream since the 1950s), 3D, and 4D film) and more spectacular films to keep theatrical screenings attractive.

Systems that were cheaper and more easily handled (including 8mm film, video, and smartphone cameras) allowed for an increasing number of people to create films of varying qualities, for any purpose (including home movies and video art). The technical quality was usually lower than that of professional movies, but improved with digital video and affordable, high-quality digital cameras.

Improving over time, digital production methods became more and more popular during the 1990s, resulting in increasingly realistic visual effects and popular feature-length computer animations. Various film genres emerged and enjoyed variable degrees of success over time, with huge differences among, for instance, horror.

3.8.3. FILM CRITICISM

The first film critiques came soon after the dawn of film media in the early 1900s. As films became more popular, newspapers began hiring professional critics to write more serious analysis of the films to add more than just entertainment value. New styles of film analysis developed over time and eventually became a standard feature for prominent magazines. In more modern times, film critique was additionally made popular through television media. Established critics Roger Ebert and Gene Siskel were notable for developing the show "Siskel & Ebert At the Movies" in the 1980s that would not only review films, but also conduct interviews with film actors. The main task for most review media is to explain a film's premise in addition to its artistic or entertainment merits. Film summaries are often expressed through a rating system such as numeric scales, grades, image representations, or "thumbs" in the case of Siskel and Ebert.

3.8.4. DEVELOPMENT OF ONLINE FILM CRITICISM

Online blogs were one of the first internet media to be used for film criticism, allowing any person to write their opinion of a film for others to read. However, audience size was limited by

the popularity of amateur writers and the sites they used. Using more modern platforms such as YouTube function in a similar fashion but provide access to a wider audience and interests with the use of videos, cut-scenes, animations, and actors to express film critiques. Specialized websites were also developed to provide a direct source for film critiques and reviews. Specific types of criticism have developed within online media that focus on particular aspects such as scientific realism, plot holes, and theories on possible sequels. Other sites may be specifically tailored to offer analysis on aspects such as content advisories, for parents concerned with their children watching the film. Several sites have dedicated their use to providing a source for the general public to express their views on films. These typically incorporate a written commentary from the user that can vary greatly in length depending on depth and breadth. Additionally, a scaled rating system is commonly included that is then used to calculate an average rating and rank to compare with other movies. The modern film criticism industry has been shown to exhibit some bias, particularly toward gender. Often it is the case that reviews are more male dominated with fewer representations of women.

CHAPTER 4
RESULTS AND DISCUSSION

We have compared the results of the classification models based on accuracy.

Accuracy: It is simply the ratio of correctly predicted observations to the total number of observations.

We can say that the higher the accuracy, the better the model. The accuracy is given by-

Accuracy = TP + TN / TP + TN + FP + FN.

The analysis of accuracies of all the models is given below:

svm_cv = SVM with Count Vectorizer = 85.29

lr_cv = Logistic Regression with Count Vectorizer = 86.89

nb_cv = Naive Bayes with Count Vectorizer = 85.48

Accuracy Graph

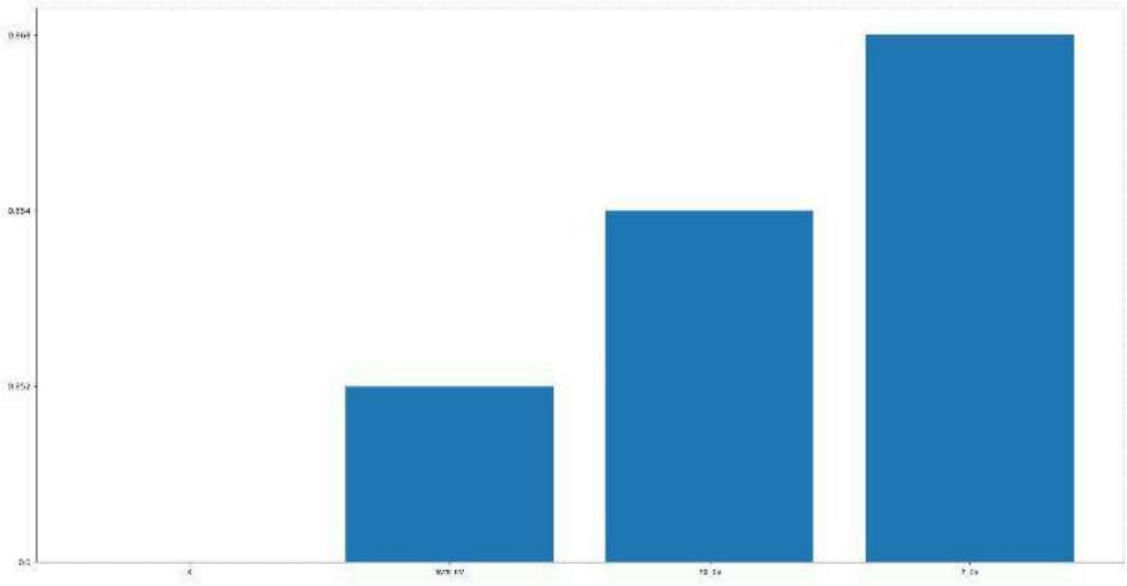

Fig. 28 : Accuracy Graph

```
Training NB model using bag of words
Accuracy on testing dataset is 0.8548
Accuracy on training dataset is : 0.847
```

```
Training Logistic regression model using bag of words
Accuracy on testing dataset is 0.8689333333333333
Accuracy on training dataset is : 0.868
```

```
Training SVM model using bag of words
Accuracy on testing dataset is 0.8529333333333333
Accuracy on training dataset is : 0.847
```

Fig.29 : Code snippet of the training and testing accuracies of all the three classifiers

This study makes an attempt to classify the IMDB movie reviews dataset using logistic regression, SVM and Naive Bayes. Usage of different feature extraction methods have been done; it is observed that usage of TF-IDF approach generally gives a higher accuracy. The accuracy for Bag of Words approach also increases with increase in the n-gram range but the computation time for extracting features increases drastically.

Unarguably, sentimental analysis techniques are among the utmost significant bases in the decision-making process. A lot of people depend on sentimental analysis for achieving efficient results of services or products. It is an undeniable fact that human languages are relatively complex to be understood by the machine, which leads to conditions where a negatively said word has a positive association and vice versa. So, a sentimental analysis of movie reviews is a challenging task.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

The main motive behind this project was to construct a sentiment analysis model that will help us to get a better understanding of movie reviews that we have collected. We compared the results of the 3 classifiers - Naive Bayes, Logistic Regression and Support Vector Machine (SVM).

For Evaluation, we observed the accuracy provided by each model. By evaluating themodels, we found out that Logistic Regression gives us the highest accuracy score of **86.89%**.

Social media Monitoring has been growing very rapidly so there is a need for various organizations to analyze customer behavior or attitude of particular product or any movie review. So, the concepts of sentiment analysis have been introduced. Text analytics and sentiment analysis can help organization to derive valuable business insights. Attitude can be calculated based on polarity check. Sentiment analysis on Online review are done by forming dictionary which shows that it is easier to build dictionary on phrases but complex in case of Twitter as tweets consist of short hands as online review were written in more clear way as compared to Tweets. So, form hidden relationship between different keywords and a dictionary of the words on the basis of different categories of comments & tweets. Future work include to determine their features for the movie in detail i.e. make polarity check on different features such as actors, directors, scripts, music etc. and make the dictionary for them.

Future work on sentiment analysis of IMDb movie reviews can focus on several aspects to improve the accuracy, coverage, and applicability of the analysis.

Here are some potential areas for future research:

Dataset Expansion:

To enhance the performance and generalizability of sentiment analysis models, future work can involve collecting and incorporating larger and more diverse datasets. Expanding the dataset to include a wider range of movie genres, languages, and cultural contexts can help capture a more comprehensive understanding of audience sentiment.

Enhanced Feature Representation:

Improving the feature representation of movie reviews can contribute to better sentiment analysis results. Future work can explore advanced techniques such as contextual word embeddings, domain-specific embeddings, or deep learning-based approaches like transformers to capture the semantics and contextual information more effectively.

Aspect-based Sentiment Analysis:

Going beyond overall sentiment classification, future work can focus on aspect-based sentiment analysis, where the sentiment of specific aspects or entities within movie reviews is identified. This can provide deeper insights into audience sentiment towards different elements like acting, direction, music, or cinematography.

Deep Learning Architectures:

Exploring advanced deep learning architectures, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), or attention mechanisms, can enhance the sentiment analysis performance. These architectures can effectively capture long-term dependencies, local patterns, and the importance of different words or phrases in determining sentiment.

Sentiment Analysis for Streaming Platforms:

With the rise of streaming platforms, future work can focus on sentiment analysis specific to these platforms. Analyzing sentiment on IMDb movie reviews in the context of streaming platforms can provide insights into how users perceive and react to movies in a streaming environment.

Multi-modal Analysis:

Integrating multi-modal data, such as text, images, and audio, can enrich sentiment analysis of IMDb movie reviews. Future work can explore the fusion of textual and visual features from movie posters, trailers, or user-generated content like memes to capture a more comprehensive understanding of sentiment.

Cross-Lingual Sentiment Analysis:

Extending sentiment analysis to multiple languages can enhance the applicability and reach of the analysis. Future work can focus on developing cross-lingual sentiment analysis models that can effectively analyze sentiment in reviews written in different languages, allowing for a more global perspective.

By addressing these areas of future work, sentiment analysis on IMDb movie reviews can continue to evolve and provide valuable insights into audience sentiment, aiding decision-making processes in the film industry.

REFERENCES

[1] Vr, Nithin & Babu Pb, Sarath. (2014). Predicting Movie Success Based on IMDB Data

[2] Amolik, A., Jivane, N., Bhandari, M., & Venkatesan, M. (2015). Twitter sentiment analysis of movie reviews using machine learning techniques. International Journal of Engineering and Technology, 7(6), 2038-2044.

[3] Nagamma, P., Pruthvi, H. R., Nisha, K. K., & Shwetha, N. H. (2015, May). An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In Computing, Communication & Automation (ICCCA), 2015 International Conference on (pp. 933- 937). IEEE.

[4] Lee, Sung-Won ; Jiang, Guangbo ; Kong, Hai-Yan ; Liu, Chang(2020), " A difference of multimedia consumer's rating and review through sentiment analysis", Multimedia tools and applications.

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.

[6] B.Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86.

[7] Liu, "Sentiment analysis and opinion mining," in Synthesis Lectures on Human Language Technologies, G. Hirst, Ed. San Rafael, CA, USA: Morgan & Claypool, 2012, pp. 1–167.

[8] B. P. Verma S, "Incorporating semantic knowledge for sentiment analysis. Proceedings of ICON," 2009.

[9] 2004, Pang and Lee " ML based sentiment analysis of text"

[10] Huaxia Rui, Yizao Liu, and Andrew Whinston. 2013. Whose and what chatter matters? the effect of tweets on movie sales. Decision Support Systems 55(4):863–870.

[11] A. Khan, B. Baharudin, K. Khan; "Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure" ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer, pp. 317-331, 2011.

[12] Lei Zhang University of Illinois, Chicago "Extracting and Ranking Product Features" Coling 2010: Poster Volume, pages 1462–1470, Beijing, 2010.

[13] Rafeeque Pandarachalil Govt. College of Engineering, Kannur "Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach", Springer, 2014.