**1. INTRODUCTION:**

**1.1 OVERVIEW:**

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. Researchers in the fields of natural language processing (NLP) and machine learning (ML) have explored a variety of methods to implement the process with highest accuracy possible. Huge textual data is available on sites like Amazon, IMDB, and Rotten Tomatoes on movies and analysing 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 the process of using natural language processing, text analysis, and statistics to analyse customer sentiment. Sentiment analysis can be defined as analysing the positive or negative sentiment of the customer in text. The contextual analysis of identifying information helps businesses understand their customers’ social sentiment by monitoring online conversations.

As customers express their reviews and thoughts about the brand more openly than ever before, sentiment analysis has become a powerful tool to monitor and understand online conversations. Analysing customer feedback and reviews automatically through survey responses or social media discussions allows you to learn what makes your customer happy or disappointed. Further, you can use this analysis to tailor your products and services to meet your customer’s needs and make your brand successful.

The tools assist businesses in extracting information from unstructured and unorganized text found on the internet, such as emails, blog posts, support tickets, web chats, social media channels, forums, and comments. To replace manual data processing, algorithms use rule-based, automatic, or hybrid techniques. Automatic systems learn from data using machine learning techniques, whereas rule-based systems execute sentiment analysis based on predetermined, lexicon-based rules. Both methodologies are combined in hybrid sentiment analysis.

Sentiment analysis is one class of computational techniques which automatically extracts and summarizes the opinions of such an immense volume of data which the average human reader is unable to process.

Sentiment analysis of a movie review can rate how positive or negative a movie review is and hence the overall rating for a movie. Therefore, the process of understanding if a review is positive or negative can be automated as the machine learns through training and testing the data.

The First step of sentiment analysis is to evaluate the predictions of sentiment classifiers, additional feature extractions to boost performance are discussed with practical results.

This may explain why sentimental analysis and theoretical analysis are often used interchangeably; however, we believe that seeing emotions as emotionally charged ideas is more accurate. The aim of this paper is to categorize movie reviews by examining the magnitude (good or bad) of each category in the review. We devised an approach to describe the polarity of the movie using these sentiment words because of the abundant usage of sentiment words in the study.

**1.2 STATEMENT OF PROBLEM:**

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative. Advanced, “beyond polarity” sentiment classification looks, for instance, at emotional states such as “angry”, “sad”, and “happy”.

In this, we have to predict the number of positive and negative reviews based on sentiments by using different classification models.

The goal of sentiment analysis is to determine the attitude of a speaker or writer based on the language they use regarding a specific topic or product. Various platforms can be used to analyse whether a sentiment is positive, negative or neutral.

**1.3 OBJECTIVES:**

* The objective of sentiment analysis is to accurately extract people's opinions from a large number of unstructured review texts and classifying them into sentiment classes, i.e., positive, negative. Sometimes “highly positive” and “highly negative” are also considered.
* It can be used to determine the attitude of the reviewer with respect to various topics or the overall polarity of review. Using sentiment analysis, we can find the state of mind of the reviewer while providing the review and understand if the person was “happy”, “sad”, “angry” and so on.
* We analyse and study the features that affect the sentiment score of the movie review text. Also, we use the state of the art classification algorithms for the evaluation of performance and accuracy of the approach used. Also, we not only study the approach but try to have a deeper understanding of the problem domain.
* In today’s world, lot of data created every day in the form of messages, posts, and reviews. All these raw texts can be extracted or collected in various ways for understanding the sentiments and conduct sentiment analysis.
  1. **LIMITATIONS OF THE STUDY:**

**1. Tone:** Tone can be difficult to interpret verbally, and even more difficult to figure out in the written word. Things get even more complicated when one tries to analyse a massive volume of data that can contain both subjective and objective responses.

**2. Sarcasm:** People use irony and sarcasm in casual conversations and memes on social media. The act of expressing negative sentiment using backhanded compliments can make it difficult for sentiment analysis tools to detect the true context of what the response is actually implying.

**3. Idioms:** Machine learning programs don’t necessarily understand a figure of speech. For example, an idiom like “not my cup of tea” will boggle the algorithm because it understands things in the literal sense. Hence, when an idiom is used in a comment or a review, the sentence can be misconstrued by the algorithm or even ignored.

**4. Negations:** Negations, given by words such as not, never, cannot, were not, etc. can confuse the ML model. For example, a machine algorithm needs to understand that a phrase that says, “I can’t not go to my class reunion”, means that the person intends to go to the class reunion.

**5. Comparative Sentences:** Comparative sentences can be tricky because they may not always give an opinion. Much of it has to be deduced.

**1.5 OPERATIONAL DEFINATION OF TERMS:**

* **NLP:** stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence. It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages. It helps developers to organize knowledge for performing tasks such as translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction, and topic segmentation.
* **Data mining:** is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.
* **Machine Learning:** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.
  1. **ORGANIZATION OF REPORT:**

Chapter I: Gives a brief overview about the project in terms of its, Statement of the problem, Aim and Objective of the study, Limitation of the study, and the approach that is used to achieve the goal. It also provides definitions and terms that are widely used throughout this framework.

Chapter 2: Literature Survey in this section which shows the various analysis and research made in the fields of one’s interest and the result analysis and research made in the fields of one’s interest and the result already published, taking into account the various parameters of the project and extent of the project.

Chapter 3: Methodology it includes overview and the architecture of the project system Overview consists of algorithms and their details.

Chapter 4: Experiments and results. The experiments and results include the Screenshots of some important aspects of the project.

Chapter 5: Provides the conclusion drawn from testing. It also states the work that can be done in future in order to further enhance the proposed system.

**2. LITERATURE SURVEY:**

**2.1 OVERVIEW:**

In recent years, several studies have applied machine learning algorithms on different movie review datasets to classify the positive and negative movie reviews. These algorithms show good classification results and encourage many researchers to apply these kinds of algorithms to solve challenging tasks. A list of some literature studies related to this method is presented below.

**2.2 LITERATURE REVIEW:**

A Large number of works have been carried out previously on opinion mining and sentiment analysis. Nagamma P et al proposed various data mining techniques for classification of movie audits and it likewise predicts the box office collection for the movie.

The online movie review data collected from IMDB dataset, the box office collection and the success or failure of the movie is predicted based on the reviews. Pang et al applied the machine learning technique for classification of reviews present on IMDB movie reviews database, by forming the list of 14 keywords which are useful in finding the baseline for classification accuracy.

J. Erman et al. studied three types of clustering algorithms namely K-Means, DBSCAN and Auto Class algorithm for the classification of network traffic problem. This study depends on the capacity of every algorithm for forming clusters having higher predictive power of a single traffic class and for deciding the capacity of every algorithm to produce small number of clusters that has numerous associations. The Auto Class algorithm is compared with DBSCAN and K-Means algorithm and the result indicates that both K-Means and DBSCAN work faster than Auto Class algorithm. Turney et al. studied the unsupervised learning algorithm for sentiment classification process. They determined the similarity of words with help of NEAR operation and developed a classifier for finding polarity result.

‘The authors proposed to apply eight classifiers on the IMDb movie reviews dataset. These are - Naive Bayes, Decision Tree, Random Forest, Ripple Rule Leaning, K- Nearest Neighbors, Support Vector Classifier, Bayes Net, Stochastic Gradient Descent. In their work Ripple Rule Learning was found give the worst results, whereas Random Forest outperformed other classifiers. Performance of these eight classifiers were measured by five different evaluation metrics, namely, Accuracy, Area Under Curve (AUC), F1- measure, Recall, and Precision. Authors of were of the thought that reviews of movies shared on social media platforms and other web portals are important factors in a movie’s financial success. The results showed that positive sentiment is more efficient for a movie domain with a small number of existing reviews, which indicated that sentiment alone is not the only factor. Rather, sentiment could perform better in combination with other factors such as movie genre and festive season etc.

**2.3 BASIC TERMINOLOGIES AND CONCEPTS:**

**Naïve Bayes:** It is a classification algorithm, primarily used for text classification involving high dimensional training data sets. Example spam filtering, sentiment analysis etc. This algorithm learns the probability of an object with certain features belonging to a particular

Class. It is a probabilistic classifier. This algorithm is called Naive Bayes because it makes a naive assumption that occurrence of certain features is independent of each other which in reality is not the case.

**Logistic Regression:** Logistic regression is quite similar to linear regression but here, instead of fitting a line to our data we try to fit ‘S’ shaped logistic function(Sigmoid Function). Although it’s name contains regression, on the contrary it is used for classification purposes. Logistic regression’s capability to classify data using continuous and discrete measurements makes it a popular machine learning algorithm. Logistic regression uses something called maximum likelihood to fit data. It can be used to classify samples and can use different kinds of data to classify samples. It can also be used to assess what variables are useful for classifying samples.

**Support Vector Classifier:**  SVC, or Support Vector Classifier, is a supervised machine learning algorithm typically used for classification tasks. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyper plane that divides the data into two classes.

**Decision Tree Classification**: Decision Tree is a supervised learning techniquethat can be used for both classification and Regression problems, it is a tree-structured classifier, whereinternal nodes represent the features of a dataset, branches represent the decision rulesandeach leaf node represents the outcome*.* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions*.* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

**3. METHODOLOGY:**

**3.1 OVERVIEW:**

In this research, many algorithms are used for analysis. Hardware and software requirements are listed below.

**3.2 HARDWARE AND SOFTWARE REQUIREMENTS:**

**3.2.1 HARDWARE:**

**Hardware requires to develop the research is listed below:**

1. Processor: Intel CORE I3
2. RAM: 4GB
3. Hard Disk Space: 250 GB

**3.2.2 SOFTWARE:**

**Software requires to develop the research is listed below:**

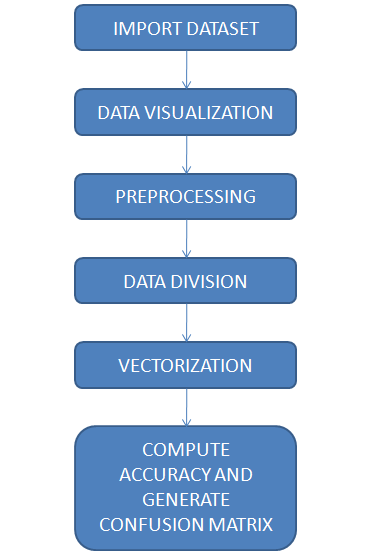
1. Operating System: Windows 10
2. Language: Python

**3.3 LANGUAGE USED:**

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Python sentiment analysis is a methodology for analysing a piece of text to discover the sentiment hidden within it. It accomplishes this by combining machine learning and natural language processing (NLP). Sentiment analysis allows you to examine the feelings expressed in a piece of text.

**3.4 ARCHITECTURE:**

****

1. **IMPORT DATASET:**

For this paper a data set containing 50k movie reviews from IMDb, created by Andrew Maas is utilized. The data have already been splitted into 25k reviews for training purposes while the other 25k is intended for testing the classifier. In addition, both sets contain 12.5k positive and negative reviews. The reviews are classified into positive and negative in reference to the IMDB rating system. It allows viewers to rate on a scale from 1 to 10, and according to dataset creator anything ≤ 5 stars is labelled negative and ≥ 6 stars is marked as positive. Reviews with ratings out of the above ranges are not included. There are at most 30 reviews for each movie. The average number of words per review is 234.76 with a standard deviation of 172.91 words. Collectively, the dataset contains 88585 different words.

1. **DATA VISUALIZATION:**

Data visualization is a field in [data analysis](https://www.simplilearn.com/data-analysis-methods-process-types-article) that deals with visual representation of data. It graphically plots data and is an effective way to communicate inferences from data. Python offers several plotting libraries, namely [Matplotlib](https://www.simplilearn.com/tutorials/python-tutorial/matplotlib), [Seaborn](https://www.simplilearn.com/tutorials/python-tutorial/python-seaborn) and many other such data visualization packages with different features for creating informative, customized, and appealing plots to present data in the most simple and effective way. Visualization packages with different features for creating informative, customized, and appealing plots to present data in the most simple and effective way.

1. **DATA DIVISION:**

It is common to divide the dataset into training and testing vectors .The training vector is the set of data that trains the considered classifier. The validation vector is a subset of the training vector that does not necessarily train, but is used to give some insight on the classifier performance. Test data is to evaluate the model accuracy. The split of the training and validation, testing, or both can occur in many ways. However, there is a rule-of-thumb that training gets the most data. A recurrent ratio encountered in various eclectic ML settings is the 80-20 split which gives 80% to the training and 20% to the testing. In order to avoid any biasness an equal representation of positive and negative reviews is employed. 80% of this data is used for the training purpose and the 20% is used for the testing purpose.

1. **PREPROCESSING:**

Data that are not well cleaned and organized might lead to false identifications. Hence, data pre-processing is a crucial task in the data mining process. It refers to cleaning up the data from useless information that will not help in the training process and might cause confusion during the classification process. For the IMDB dataset, several data pre-processing steps are utilized. Firstly, all the symbols such as “?”, “!” are removed. Secondly, all letters in the dataset are converted into lowercase letters. Thirdly, all hybrid links are removed from the text. Fourthly, stop words such as, me, you, and we are evicted. Finally, stemming techniques are applied on the text to present the word in its original form after removing prefixes and suffixes.

1. **VECTORIZATON:**

Vectorization or text embedding is the process of extracting features from text and pass it as an input to the classifier. Python is a general purpose object-oriented high level programming language. Due to the extensive libraries and frameworks dedicated to ML which facilitates the development process and save time, it is considered and used in this paper. Each movie review is encoded “vectorized” into a numeric value. It is achieved by utilizing the genism, a Python library for topic modelling and NLP.

1. **COMPUTE ACCURACY:**

The classification accuracy is used to measure that how well the devised model is able to automatically identify the data. It is the percentage of labels that have been correctly classified. The mathematical formulation for accuracy is given in Equation below. Where, TP, TN, FP, and FN respectively denote the true positives, true negatives, false positives, and false negatives in the predicted labels.

Accuracy = TP+TN x 100%

TP+TN+FN+FP

**3.5 PACKAGES USED:**

**PANDAS:**

Pandas is an open-source library that is made mainly for working with relational or labelled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

**MATPLOTLIB:**

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc.

**SEABORN:**

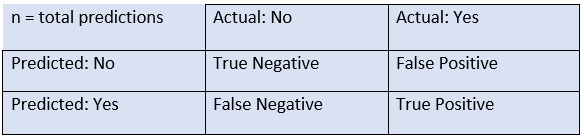
Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.  
Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

**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.

**3.6 CONFUSION MATRIX:**

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**. Some features of Confusion matrix table looks like:



The above table has the following cases:

* **True Negative:** Model has given prediction No, and the real or actual value was also No.
* **True Positive:** The model has predicted yes, and the actual value was also true.
* **False Negative:** The model has predicted no, but the actual value was Yes, it is also called as **Type-II error**.
* **False Positive:** The model has predicted Yes, but the actual value was No. It is also called a **Type-I error.**

**Need for Confusion Matrix in Machine learning:**

* It evaluates the performance of the classification models, when they make predictions on test data, and tells how good our classification model is.
* It not only tells the error made by the classifiers but also the type of errors such as it is either type-I or type-II error.
* With the help of the confusion matrix, we can calculate the different parameters for the model, such as accuracy, precision, etc.

**4. EXPERIMENTS AND RESULTS:**

**4.1 DATASET USED:**

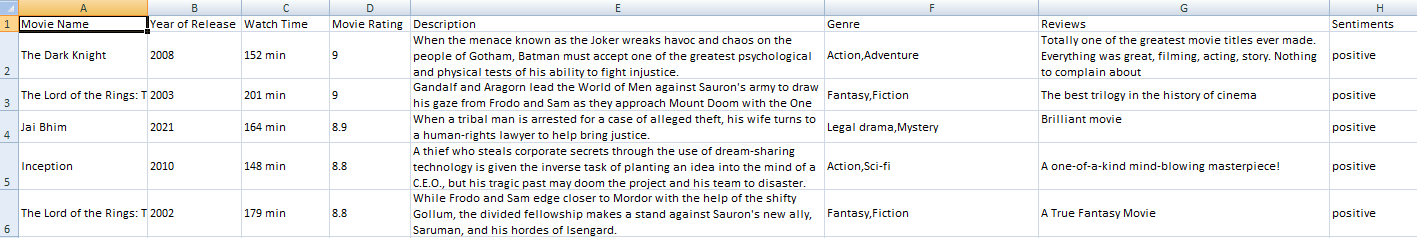
**IMDB DATASET:**

In our project we use “IMDB Movie dataset” for analysis of sentiment on movie review for positive and negative accuracy results.

**4.1.1 ATTRIBUTE INFORMATION:**

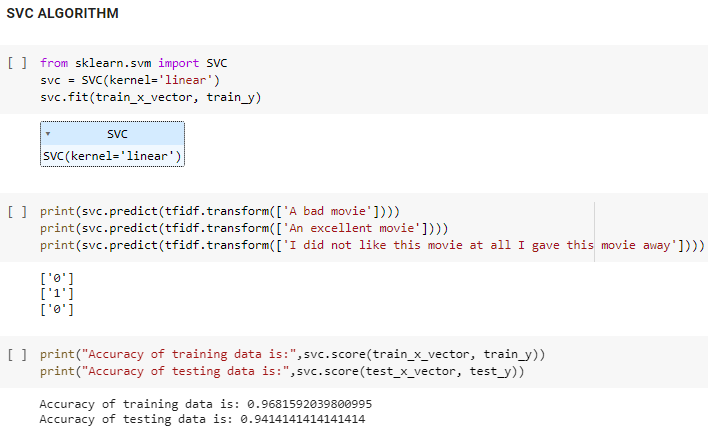
1. Movie Name: This attribute provides the details of name of the movie.
2. Year of release: Attribute gives the information of year in which movie released.
3. Watch time: The duration of movie runtime is specified in this column.
4. Movie Rating: The ratings given by the viewers are provided.
5. Description: Movie descriptions are specified in this column.
6. Genre: Contains Category of the movie.
7. Reviews: This column provides the details regarding the reviews of a movie.
8. Sentiment: Feelings expressed by the reviewers are specified in this attribute.

**DATASET:**

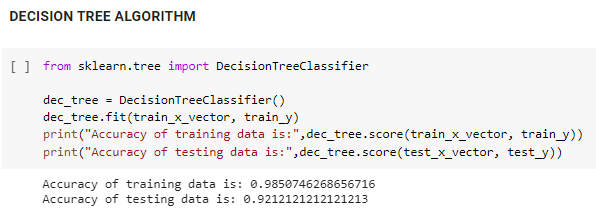


**COMPUTED ACCURACY:**

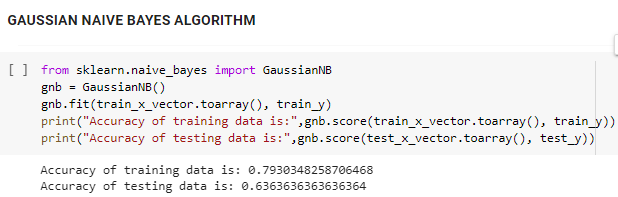
**ALGORITHM 1: SVC ALGORITHM**



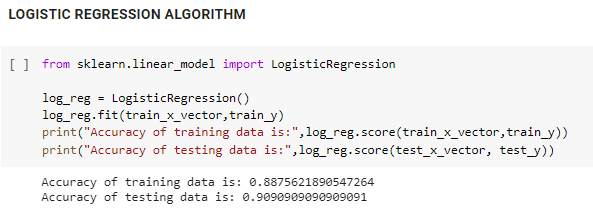
**ALGORITHM 2: DECISION TREE ALGORITHM**



**ALGORITHM 3: GAUSSIAN NAIVE BAYES ALGORITHM**



**ALGORITHM 4: LOGISTIC REGRESSION ALGORITHM**



**5. CONCLUSION AND FUTURE WORKS:**

**5.1 CONCLUSION:**

In this work, we extracted new features that have strong impact on determining the polarity of the movie reviews and applied computation methods for pre-processing of the data.

Sentiment analysis also referred to as opinion mining is the process of extracting opinions from text data and classifies it into positive, negative or neutral ones.

**5.2 FUTURE SCOPE:**

In future, we would like to evaluate the effectiveness of the proposed sentiment classification features and techniques for other tasks, such as sentiment classification. We would like to apply in-depth concepts of NLP for better prediction of the polarity of the document. We would also like to extend this technique on other domains of opinion mining like’s newspaper articles, product reviews, political discussion Forums etc. This research can be expanded to include the analysis of online feedback from social media platforms and other sophisticated algorithms.

**5.3 REFERENCES:**

[1] Saeed Mian Qaisar-Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory,College of Engineering, Effat University, 21478, Jeddah, Saudi Arabia.

[2] Abhimanyu Singh -Sentiment Analysis of IMDB Movie Reviews, SUNY Binghamton.

[3] KamilTopal-Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews, Electrical Engineering and Computer Science Case Western Reserve University

Cleveland.

[4] Tirath Prasad Sahu- Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms, National Institute of Technology Raipur, India

**5.4 ONLINE REFERENCES:**

1. <https://www.geeksforgeeks.org/ml-logistic-regression-using-python/>

2. <https://www.javatpoint.com/confusion-matrix-in-machine-learning>

3. <https://www.kaggle.com/lakshmi/sentimentanalysisimdbmoviereviews>

4. <https://www.kaggle.com/code/slayomer/eda-on-imbb-film-dataset>

5. https://www.kaggle.com/datasets/harshit/imdbdatasetmovi/code

6. <https://www.kaggle.com/code/gabriel/imdb-exploratory-data-analysis>

7. https://www.kaggle.com/code/vamsi/eda-of-1000-movies-dataset