# A
# Project Report
# On
# "Sentiment Analysis of Movie Review"

**Prepared by**

20CS004 Vrushabh Amrutiya
20CS019 Disney Javiya

**Under the guidance of**

Assistant Prof. Hemang Thakar

**Submitted to**

Charotar University of Science & Technology

Degree of Bachelor of Technology

in Computer Science & Engineering

CS357: Software Group Project-IV

Of 6th Semester of B.Tech

**Submitted at**


**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**Faculty of Technology & Engineering, CHARUSAT**

**Chandubhai S. Patel Institute of Technology**

**At: Changa, Dist: Anand – 388421**

**April 2023**

# A
# Project Report
# On
# "Sentiment Analysis of Movie Review"

(CS357 - Software Group Project-IV)


**Prepared by**

Vrushabh Amrutiya (20CS004)

Disney Javiya (20CS019)


**Under the Supervision of**

Assistant Prof. Hemang Thakar

**Submitted to**

Charotar University of Science & Technology (CHARUSAT)
for Software Group Project IV
in Department of Computer Science & Engineering (CSE)
for B.Tech Semester 6

**Submitted at**

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

**Accredited with Grade A+ by NAAC**

Chandubhai S Patel Institute of Technology


**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**
**Chandubhai S. Patel Institute of Technology (CSPIT)**
**Faculty of Technology & Engineering (FTE), CHARUSAT**
**At: Changa, Dist: Anand, Pin: 388421.**


**April 2023**

# DECLARATION BY THE CANDIDATES

We hereby declare that the project report entitled "**Sentiment Analysis of Movie Review**" submitted by us to Chandubhai S. Patel Institute of Technology, Changa in partial fulfilment of the requirements for the award of the degree of **B. Tech Computer Science & Engineering**, from Department of Computer Science & Engineering, CSPIT, FTE, is a record of bonafide CS357 Software group Project-IV carried out by us under the guidance of **Assistant Prof. Hemang Thakar**. We further declare that the work carried out and documented in this project report has not been submitted anywhere else either in part or in full and it is the original work, for the award of any other degree or diploma in this institute or any other institute or university.

**Signature of the candidate**
**(Vrushabh Amrutiya – 20CS004)**

**Signature of the candidate**
**(Disney Javiya – 20CS019)**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Assistant Prof. Hemang Thakar
Department of Computer Science & Engineering,
Chandubhai S Patel Institute of Technology (CSPIT)
Faculty of Technology (FTE)
Charotar University of Science and Technology (CHARUSAT) - Changa.

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

# CERTIFICATE

This is to certify that the report entitled "**Sentiment Analysis of Movie Review**" is a bonafied work carried out by **Vrushabh Amrutiya (20CS004)** under the guidance and supervision of **Assistant Prof. Hemang Thakar** for the subject **Software Group Project - IV (CS357)** of 6th Semester of Bachelor of Technology in **Computer Science & Engineering** at Faculty of Technology & Engineering (C.S.P.I.T.) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Under the supervision of,

**Assistant Prof. Hemang Thakar**
Dept. of Computer Science & Engineering
C.S.P.I.T., CHARUSAT-Changa.

Dr. Amit Thakkar
Head,
Dept. of Computer Science & Engineering
C.S.P.I.T., CHARUSAT-Changa.

**Chandubhai S Patel Institute of Technology (C.S.P.I.T.)**

**Faculty of Technology & Engineering, CHARUSAT**

At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

# CERTIFICATE

This is to certify that the report entitled "**Sentiment Analysis of Movie Review**" is a bonafied work carried out by **Disney Javiya (20CS019)** under the guidance and supervision of **Assistant Prof. Hemang Thakar** for the subject **Software Group Project - IV (CS357)** of 6th Semester of Bachelor of Technology in **Computer Science & Engineering** at Faculty of Technology & Engineering (C.S.P.I.T.) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Under the supervision of,

**Assistant Prof. Hemang Thakar**
Dept. of Computer Science & Engineering
C.S.P.I.T., CHARUSAT-Changa.

Dr. Amit Thakkar
Head,
Dept. of Computer Science & Engineering
C.S.P.I.T., CHARUSAT-Changa.

**Chandubhai S Patel Institute of Technology (C.S.P.I.T.)**

**Faculty of Technology & Engineering, CHARUSAT**

At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

# ACKNOWLEDGEMENT

With immense pride and dedication, I would like to deliver the project brief. The development of this project has given me wide opportunity to think, implement and interact with various aspects of Machine Learning skills as well as the new emerging technologies. Every work that one completes successfully stands on the constant encouragement, good will and support of the people around. I hereby avail this opportunity to express my gratitude to number of people who extended their valuable time, full support and cooperation in developing the project. I express deep sense of gratitude towards our Principal of the Institute, Dr. Y. P. Kosta, our Head of the Department, Dr. Amit Thakkar, my internal guide Assistant Prof. Hemang Thakar for the support during the whole session of study and development. It is because of them, that I was prompted to do hard work, adopting new technologies. They altogether provided me favorable environment, and without them it would not have been possible to achieve my goal.

Thanks,

Vrushabh Amrutiya (20CS004)

Disney Javiya (20CS019)

# ABSTRACT

Sentiment analysis of movie reviews involves using natural language processing techniques to analyze the opinions expressed in written or spoken feedback about movies. The goal is to classify these opinions as positive, negative, or neutral based on the language used and the overall sentiment conveyed. The analysis involves several steps, including text preprocessing, feature extraction, and classification. The results of sentiment analysis can be used to understand audience reactions to a particular movie, inform marketing and promotional strategies, and help filmmakers and studios make informed decisions about future projects.

**TABLE OF CONTENTS**

<table>
<tr><td colspan="3"><u>CHAPTER 3 Methodology</u></td><td></td></tr>
<tr><td colspan="3"><u>CHAPTER 4 Implementation</u></td><td></td></tr>
<tr><td></td><td>4.1</td><td>Algorithm</td><td></td></tr>
<tr><td></td><td>4.2</td><td>Dataset Details</td><td></td></tr>
<tr><td></td><td>4.3</td><td>Implementation Details</td><td></td></tr>
<tr><td colspan="4"><u>CHAPTER 5 –Results and Discussion on Results</u></td></tr>
<tr><td colspan="4">5.1 Results<br><br>5.2 Visual Representations of Results</td></tr>
<tr><td colspan="4">CHAPTER 6 - Conclusion</td></tr>
<tr><td colspan="4">CHAPTER 7 - Future Extensions</td></tr>
<tr><td colspan="4">Bibliography……</td></tr>
<tr><td></td><td colspan="2">(Book / Web-link / Journal /Articles/Magazine/online course/online video lectures/ References)<br><br>Referencing in the thesis should be as per any one of the internationally accepted Referencing System (Harvard System/APA System/Vancouver System/Oxford/Footnote.....etc)</td><td></td></tr>
<tr><td colspan="3">Anti-Plagiarism Report</td><td></td></tr>
</table>

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER:1 INTRODUCTION

## 1.1 INTRODUCTION:

Sentiment analysis, which aims to extract subjective information from written reviews, is one of the most significant fields of machine learning.[1] Text mining and natural language processing are closely linked to sentiment analysis. The review's general polarity or the reviewer's viewpoint on a variety of topics can be ascertained using this technique. We can tell whether a reviewer was "positive," "negative," etc. while offering their input by using sentiment analysis.

## 1.2 RESEARCH DEFINITION:

Sentiment analysis is a rapidly expanding area of machine learning that concentrates on analyzing attitudes and expressions in sentences using NLP techniques. Its most well-known use is in movie review analysis, where researchers can learn what the audience thinks of a specific movie. Both for marketing and to raise the caliber of movies produced, the film business can benefit from this information. The main objective is to look into the use of sentiment analysis in movie evaluations. In order to categorize the sentiment of movie reviews as either positive or negative, we will examine how well different machine learning algorithms, including Naive Bayes, decision trees, K nearest neighbors, logistic regression, support vector machines, and random forests, perform.

## 1.3 MOTIVATION:

The objective is to increase knowledge of sentiment analysis and its potential applications in the film business. Our goal is to gain understanding into the benefits and drawbacks of various approaches and create efficient methods for sentiment analysis in this field by comparing various machine learning algorithms on movie reviews. In the end, our goal is to advance the body of knowledge on this topic by evaluating accuracy of different models.

## 1.4 SCOPE AND OBJECTIVES

**Identifying sentiment polarity:** Sentiment analysis can be used to determine whether a movie review is positive, negative, or neutral. This information can be used to provide an overall sentiment score for the movie.

**Identifying key themes:** Sentiment analysis can identify the key themes that are mentioned in the movie reviews, such as the plot, characters, or special effects. This information can be used to gain insights into what aspects of the movie are most popular or least popular among viewers.

**Identifying influencers:** Sentiment analysis can identify influential reviewers who have a significant impact on the overall sentiment of the movie. This information can be used to target these reviewers and encourage them to promote the movie to their followers.

# CHAPTER: 2 BACKGROUND STUDY

## 2.1 EVOLUTION OF THE TOPIC:

Sentiment analysis of movie reviews has come a long way since its inception in the early 2000s. Initially, sentiment analysis focused on analyzing the polarity of individual words in a text document and assigning a positive, negative, or neutral label to the document based on the overall sentiment score.

As machine learning algorithms became more sophisticated, sentiment analysis evolved to include more advanced techniques such as natural language processing (NLP) and deep learning. NLP allows sentiment analysis models to analyze the context and syntax of the text, while deep learning models use neural networks to extract and classify sentiment features from the text.

In recent years, sentiment analysis has expanded to include aspects such as irony detection, sarcasm detection, and emotion analysis. This has allowed sentiment analysis to become more nuanced and accurate in its analysis of movie reviews, taking into account the complex range of emotions and language that can be present in a single review.

## 2.2 THEORETICAL AND MATHEMATICAL BACKGROUND:

The theoretical and mathematical background of sentiment analysis of movie review systems is based on natural language processing (NLP), machine learning, and statistical modeling.

NLP involves the computational analysis of human language, allowing sentiment analysis models to process and analyze textual data. This includes techniques such as part-of-speech tagging, named entity recognition, and sentiment analysis.

Machine learning algorithms, such as supervised learning, unsupervised learning, and deep learning, are used to train sentiment analysis models on large datasets of annotated movie reviews. These algorithms learn to identify patterns and features in the text that correspond to positive, negative, or neutral sentiment.

Statistical modeling is also used in sentiment analysis of movie reviews.[6] One common approach is to use bag-of-words models, where the text is represented as a set of words and their frequency of occurrence. These models allow sentiment analysis models to identify which words and phrases are most strongly associated with positive or negative sentiment.

## 2.3 REVIEW PREVIOUS RESEARCH FINDINGS

There have been many research studies conducted on sentiment analysis of movie reviews over the years, and the findings have been generally positive. Here are some examples of previous research findings:

**Accuracy:** A common measure of performance in sentiment analysis is accuracy. In a study by Pang and Lee (2008), several machine learning algorithms were evaluated on their ability to classify movie reviews as positive or negative. The best-performing algorithm achieved an accuracy of 87.4%, indicating that sentiment analysis models can achieve high levels of accuracy in classifying movie reviews.

**Comparative analysis**: In a comparative analysis of different sentiment analysis methods for movie reviews, Thelwall and Buckley (2013) found that machine learning algorithms outperformed other approaches such as lexicon-based methods and rule-based methods. They also found that feature selection and dimensionality reduction techniques improved the performance of machine learning models.

**Domain adaptation:** Sentiment analysis models trained on general datasets may not perform as well on specific domains such as movie reviews. To address this issue, researchers have explored domain adaptation techniques, where sentiment analysis models are trained on a small amount of domain-specific data to improve their performance.

**Emotion analysis:** Some researchers have gone beyond simple positive/negative sentiment analysis to explore the more complex aspects of emotion in movie reviews. In a study by Baly et al. (2018), an emotion classification model was trained on a dataset of movie reviews. The model was able to identify six basic emotions (anger, fear, joy, sadness, surprise, and disgust) and achieved an accuracy of 64.7%.

## 2.4 COMPARATIVE STUDY:

A comparative study of sentiment analysis of movie reviews typically involves comparing different approaches or techniques used to analyze the sentiment of movie reviews. Some of the common techniques used for sentiment analysis of movie reviews include:

Rule-based systems:[7] These systems use a set of predefined rules to identify positive and negative sentiment in the text. These rules are often based on linguistic features such as the presence of specific words or phrases that are commonly associated with positive or negative sentiment.

Machine learning-based systems: These systems use statistical models to automatically learn patterns in the data and predict the sentiment of the text. These models are typically trained on a large dataset of labeled examples, where each example is tagged with a positive or negative sentiment.

Hybrid systems: These systems combine the strengths of rule-based and machine learning-based systems to improve the accuracy of sentiment analysis. For example, a hybrid system may use a rule-based approach to identify the sentiment of individual words or phrases and then use a machine learning algorithm to classify the overall sentiment of the text.

## 2.5 RESEARCH GAPS:

**Domain-specific sentiment analysis**: Most existing sentiment analysis systems are designed to analyze sentiment in general text, rather than specific domains such as movie reviews.

**Handling sarcasm and irony:** Movie reviews often contain sarcasm and irony, which can make it difficult to accurately identify the sentiment. Current sentiment analysis systems may struggle to accurately identify and interpret sarcasm and irony, resulting in inaccurate sentiment analysis.

**Multilingual sentiment analysis:** Most existing sentiment analysis systems are designed for English language text. However, movie reviews are often available in multiple languages. Developing multilingual sentiment analysis systems could enable accurate sentiment analysis of movie reviews in different languages.

**Handling temporal changes:** The sentiment expressed in movie reviews may change over time due to factors such as changing societal norms or cultural shifts. Developing sentiment analysis systems that can adapt to these temporal changes could improve the accuracy of sentiment analysis in movie reviews.

# CHAPTER: 3 METHODOLOGIES

Here, commonly occurring words were first identified using the bag of words model. Our sample contained a lot of HTML tags, bad language, and improper punctuation. We preprocess the dataset to remove all the unwanted data. The information must first be cleared of any URLs that start with www. Next, we removed every single html element, including. We also noticed that our sample contained the most br tags. We then got rid of all the and other scars.

The content is then tokenized using the Twitter tokenizer. The division of a text into smaller characters is known as tokenization. The word was then dissected into its constituent components using word stemming. Eliminating prefixes and suffixes is known as stemming.

The final stage is to remove the stopwords. Stopwords were once used to eliminate superfluous words so that program could focus on the essential ones.

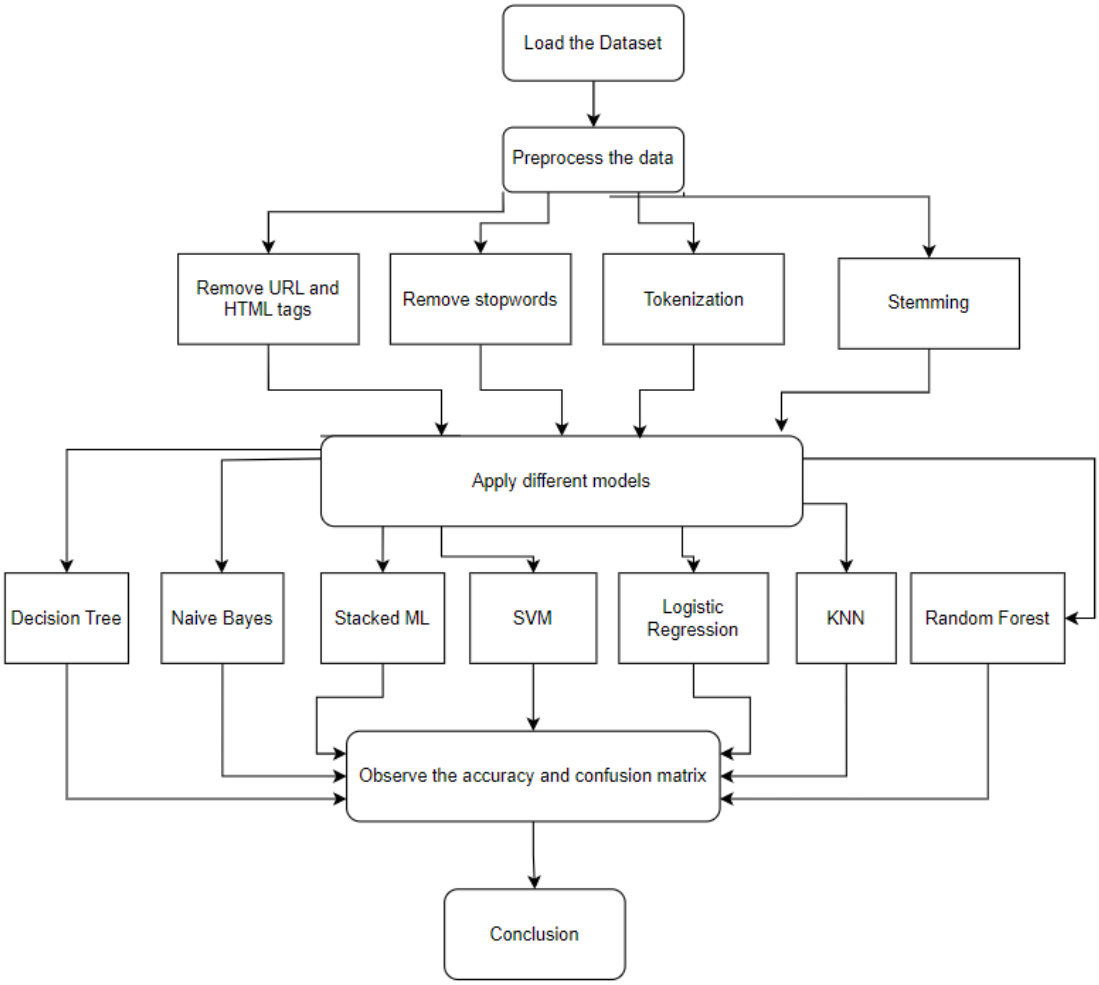
We'll apply the bag of words method once more to find words that commonly appear in sentences. We observed that all the pointless tags and phrases had been removed, leaving us with a strong set of terms.

The information will be divided into a training set and a testing set after being cleaned and prepared. After that, the training set will be further split into a smaller training set and a certification set. Using label encoding, we will next translate the labels into numerical numbers. The words that appear most frequently in the movie evaluations will also be visualised. Finally, we will use a statistical method known as TF-IDF, which is frequently used in information retrieval and natural language processing to determine the significance of a word in a document relative to a group of documents.

Term Frequency (TF) is the ratio of the number of times a term occurs in the document to the number of terms overall.

IDF = log (number of documents in the corpus/number of documents in the corpus containing the word)

We will continue to apply different methods as we train our model. After fitting the model, we will evaluate the accuracy of each algorithm and select the one with the highest accuracy. Furthermore, we will visualize the model and inspect the confusion matrix to gain more insights.


Load the Dataset
Preprocess the data
Remove URL and HTML tags
Remove stopwords
Tokenization
Stemming
Apply different models
Decision Tree
Naive Bayes
Stacked ML
SVM
Logistic Regression
KNN
Random Forest
Observe the accuracy and confusion matrix
Conclusion

# CHAPTER: 4 IMPLEMENTATIONS

## 4.1 ALGORITHMS

Different algorithms have been used like Naïve Bayes, Decision Tree, K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Stacked ML, Random Forest.

### 1) Naïve Bayes

Naïve Bayes is a supervised machine learning algorithm commonly used for text classification. It is a probabilistic classifier that makes predictions based on the probability of an object. The term "naïve" refers to the algorithm's assumption that the occurrence of a certain feature is independent of other features. The algorithm is named after the Bayes theorem, which is used as a fundamental principle.

Bayes Theorem:

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

P(A|B) is the Posterior probability, P(B|A) is Likelihood probability, P(A) is Prior Probability and P(B) is Marginal Probability.

### 2) Decision Trees:

The decision tree is a supervised machine learning algorithm that is commonly used for classification and regression.[2] Although it can be used for both purposes, it is typically used for classification. The algorithm is represented as a tree-like structure consisting of nodes, branches, and leaves. Nodes represent features, branches represent decision rules, and leaf nodes represent output. The decision tree is a graphical representation of all possible solutions to a problem based on given conditions. The algorithm starts at the root node and moves to other branches based on decision rules. The attribute of each node is selected based on measures such as information gain and entropy.

### 3) K- Nearest Neighbors:

The K-NN algorithm categorizes new data based on its similarity to existing data, placing it in the category that is most similar to the available categories. The algorithm stores all available data and uses it to classify new data points based on similarity. Therefore, when new data appears, it can be easily categorized using the K-NN algorithm. Because the algorithm stores data and uses it when in use, it is also known as a "lazy learner" algorithm.

- Steps:
  - select K.
  - calculate the distance of k number of neighbors.
  - Take k nearest neighbor based on distance.
  - From these K Neighbors, count the number of data points in each category.
  - Allocate new data points to the majority of the class.

### 4) Logistic Regression

Logistic regression is a statistical model used to analyze the relationship between a categorical dependent variable and one or more independent variables.[4] The model predicts the probability of an outcome belonging to a particular category based on the values of the independent variables. The logistic regression model is based on the logistic function, also called the sigmoid function, which maps any real-valued input to a probability value between 0 and 1.

### 5) SVM (Support Vector Machine)

Support Vector Machines (SVMs) are a type of supervised machine learning algorithm used for classification and regression.[5] SVMs work by finding the hyperplane that best separates the two classes in the feature space. The hyperplane is chosen to maximize the margin between the two closest data points of different classes. These closest data points are called support vectors. The SVM algorithm tries to find the hyperplane with the largest margin, which reduces the generalization error and improves the performance of the model on unseen data. The SVM algorithm can be extended to handle non-linear decision boundaries by using kernel functions. A kernel function maps the original data to a higher-dimensional feature space where a linear

hyperplane can separate the classes. This is called the kernel trick, and it allows SVMs to model complex, non-linear decision boundaries without explicitly computing the higher-dimensional feature space.

**6) Stacked ML**

Stacking is a method of combining classification or regression models by creating a two-layered model. The first layer includes multiple baseline models used to make predictions on the test data. The second layer includes a Meta-Classifier or Regressor that takes the predictions from the first layer as input to produce new predictions.

**7) Random Forest**

Random forest is a popular machine learning algorithm used for both classification and regression tasks. It is an ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and prevent overfitting. The algorithm works by randomly selecting a subset of features and a subset of data to train each decision tree. This introduces variation and reduces correlation between the trees, which helps to prevent overfitting. Once the trees are built, predictions are made by aggregating the outputs of all the trees.

## 4.2 DATASET DETAILS:

(https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews)

The IMDB Dataset of 50K Movie Reviews is a dataset containing reviews of movies from IMDB's website. This dataset is often used as a benchmark for natural language processing and sentiment analysis tasks. It consists of 50,000 reviews, with 25,000 labeled as positive and 25,000 labeled as negative.

The dataset was created by Stanford AI researchers Andrew Maas, Raymond Daly, Peter Pham, Dan Huang, Andrew Ng, and Christopher Potts. The reviews were scraped from IMDB's website

and labeled as positive or negative based on the rating that the reviewer gave the movie. The dataset was released in 2011 and has since been used in many research papers and projects.

The reviews in the dataset are in plain text format and range in length from a few words to several paragraphs. They cover a wide variety of movies, from classic films to recent releases. The dataset has become a popular benchmark for sentiment analysis models because it contains a balanced number of positive and negative reviews, making it a good test of a model's ability to identify sentiment.

Overall, the IMDB Dataset of 50K Movie Reviews is a valuable resource for researchers and developers working on natural language processing and sentiment analysis tasks.

## 4.3 IMPLEMENTATION DETAILS:

We discovered that the Stacked ML algorithm had the highest accuracy after implementing all of those algorithms. We discovered the accuracy of the train and validation sets and compared it to the accuracy of the test set.

**Classifer 7:- Stacked ML CLassifer (Naive Bayes, K Nearest Neighbor, Support Vector Machine)**

```
tic = time.time()
#Create a stack of these estimators
estimators = [
    ("nb",model_nb), #stack naive bayes
    ("knn",model_knn), #stack K nearest neighor
    ('svm', model_svm) #stack support vector machine
]

#Train a stacked model
# Details: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html
from sklearn.ensemble import StackingClassifier
model_stacked = StackingClassifier(estimators=estimators, final_estimator=model_lr) #use logisitc regression as the final estimat
model_stacked.fit(feature_matrix_train,  label_matrix_train)
toc = time.time()
y_pred_stacked = model_nb.predict(feature_matrix_valid)
accuracy_train_set = model_stacked.score(feature_matrix_train, label_matrix_train)  #get accuracy on train set
accuracy_valid_set = model_stacked.score(feature_matrix_valid, label_matrix_valid)  #get accuracy on valid set
print("Stacked ML Model, Accuracy (Train Set) : ", accuracy_train_set)
print("Stacked ML Model, Accuracy (Valid Set) : ", accuracy_valid_set)
print("Time Taken: " + str(((toc-tic)*1000)) + " ms")
```

```
Stacked ML Model, Accuracy (Train Set) :  0.9726111111111111
Stacked ML Model, Accuracy (Valid Set) :  0.8966666666666666
Time Taken: 82781.24141693115 ms
```

Fig 4.1

# CHAPTER: 5 RESULTS AND DISCUSSION ON RESULTS

## 5.1 RESULTS

| Model Name | Accuracy (Test Set) | Precision (Test Set) | Recall (Test Set) | F1_Score (Test Set) |
|---|---|---|---|---|
| Naive Bayes | 0.8477 | 0.8579 | 0.8316 | 0.8500 |
| Decision Trees | 0.7415 | 0.6892 | 0.8308 | 0.7321 |
| K Nearest Neighbors | 0.7686 | 0.729 | 0.839 | 0.7688 |
| Logistic Regression | 0.8888 | 0.8723 | 0.8929 | 0.8836 |
| Support Vector Machine | 0.8846 | 0.8716 | 0.8962 | 0.8846 |
| Stacking (NB, KNN, SVM, LR) | 0.8944 | 0.8824 | 0.9048 | 0.8944 |
| Random Forests (Base=Decision Tree) | 0.8363 | 0.8106 | 0.8709 | 0.8372 |

Table 5.1

# 5.2 VISUAL REPRESENTATION OF RESULTS

## 5.2.1 Confusion Matrix

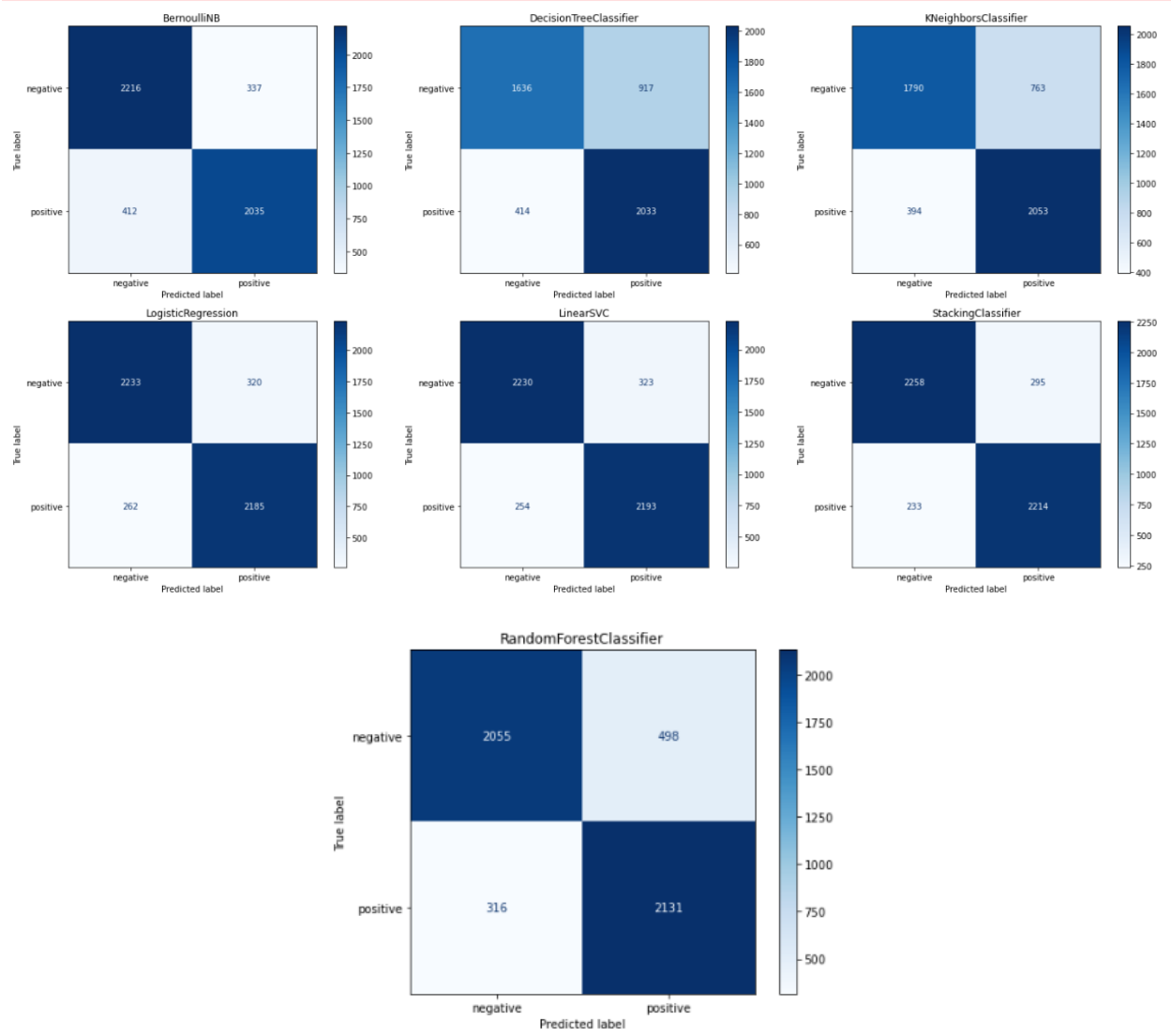



Fig 5.1

### 5.2.2 ROC Curve:

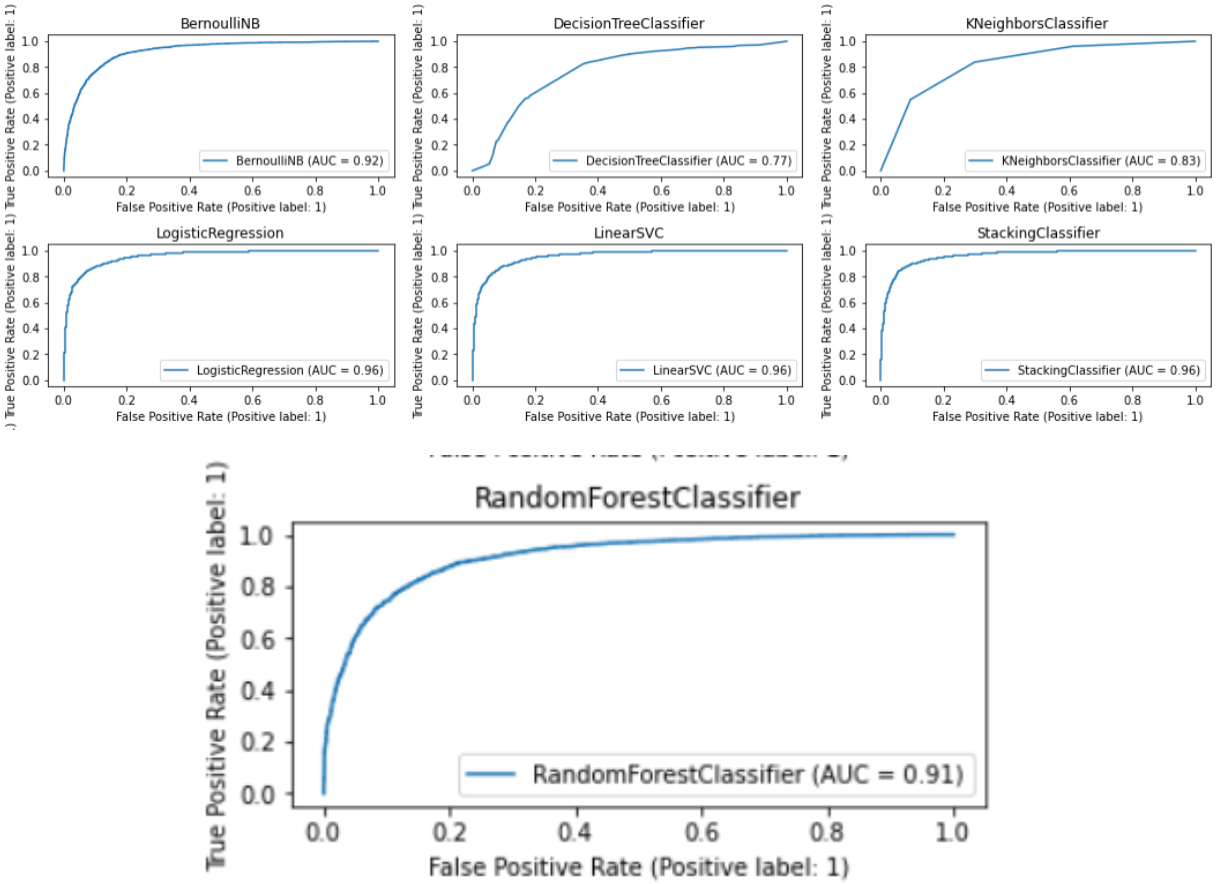



Fig 5.2

# CHAPTER:6 CONCLUSION

Based on the results obtained, it can be concluded that the Stacked ML model and SVM model have the highest accuracy among all the algorithms tested. The Stacked ML model showed the highest accuracy of 0.8944, while the SVM model showed an accuracy of 0.8846.

The Naive Bayes model showed an accuracy of 0.8477, which is relatively lower than the other models. The Decision Tree model showed an accuracy of 0.7415, indicating that it may not be the best algorithm for this particular problem. The K Nearest Neighbor model showed an accuracy of 0.7686, which is good but not as high as the Stacked ML or SVM models.

The Logistic Regression model showed a high accuracy of 0.8888, which is comparable to the Stacked ML model's accuracy. The Random Forest model showed an accuracy of 0.8363, which is relatively lower than the Stacked ML and SVM models.

Overall, the results suggest that Stacked ML and SVM models are suitable for this particular problem and could be used for further analysis and prediction. It is important to note that the performance of these models may vary depending on the specific dataset and problem being addressed, and thus it is recommended to thoroughly evaluate multiple models before selecting the most suitable one.

# CHAPTER: 7 FUTURE EXTENSIONS

In the future, we want to analyze reviews not just in terms of positive and negative aspects, but also in terms of the feelings that underlie them—emotions like wrath, fear, sadness, etc. Another goal might be to spot irony and sarcasm in the evaluation.

# BIBLIOGRAPHY

[1]https://link.springer.com/chapter/10.1007/978-981-15-0135-7_31#:~:text=Sentiment%20analysis%20of%20a%20movie,training%20and%20testing%20the%20data

[2] https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96

[3] https://www.ijraset.com/research-paper/sentiment-analysis-of-imdb-movie-reviews

[4] https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python

[5] https://www.geeksforgeeks.org/support-vector-machine-algorithm

[6] https://www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp/

[7] https://www.geeksforgeeks.org/rule-based-classifier-machine-learning/

# Sentiment Analysis of Movie Review

ORIGINALITY REPORT

| 10% | 4% | 9% | % |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Md.Thoufiq Zumma, K.B.M Tahmiduzzaman, Raihan Khan, Nashed Shah Roni, A.A.M Rahat-Bin-Rafique, Abdul Hady Akash. "Sentimental Analysis of Movie Review using Machine Learning Approach", 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET), 2022<br>Publication | 2% |
| 2 | Lecture Notes in Computer Science, 2015.<br>Publication | 1% |
| 3 | Argha Ghosh, A. Senthilrajan. "Comparison of machine learning techniques for spam detection", Multimedia Tools and Applications, 2023<br>Publication | 1% |
| 4 | "Artificial Intelligence and Speech Technology", Springer Science and Business Media LLC, 2022<br>Publication | 1% |
| 5 | scholarworks.uark.edu<br>Internet Source | 1% |

| | | |
|---|---|---|
| 6 | mspace.lib.umanitoba.ca<br>Internet Source | 1 % |
| 7 | Zhiqi Lee, Sumin Qi, Chongchong Fan, Ziwei Xie, Jing Meng. "RA V-Net: deep learning network for automated liver segmentation", Physics in Medicine & Biology, 2022<br>Publication | 1 % |
| 8 | Abhay Dadhwal, Meenu Gupta. "Analysis and Prediction of Drugs using Machine Learning Techniques", 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021<br>Publication | 1 % |
| 9 | www.mdpi.com<br>Internet Source | 1 % |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | < 20 words |
| Exclude bibliography | On | | |