Frontier Tech Leaders Programme Global Cohort 1

Sentiment Analysis on Hotel Reviews Using Machine Learning Techniques

Anas ALHARDI
Taha Mohammadyar

December 2023

Outline

- Concept note and implementation plan:
  - Background
  - Objectives
  - SDG Relation
- Data
  - Data Collection
  - Exploratory Data Analysis (EDA) and Feature Engineering
- Model Selection and Training
  - Model Evaluation and Hyperparameter Tuning
  - Model Refinement and Testing
- Results
- Deployment
- Future Work

Concept note and implementation plan

Background

- **Project Overview:** Analyzing sentiments in hospitality through machine learning for efficient processing of online hotel reviews.
- **Background:** Manual review analysis is inefficient due to the overwhelming volume of online feedback. Our solution leverages machine learning to automate sentiment classification.
- **Importance:** Critical for the hospitality industry, our project addresses the need for scalable and automated sentiment analysis, providing businesses with valuable insights for continuous improvement.

Objectives

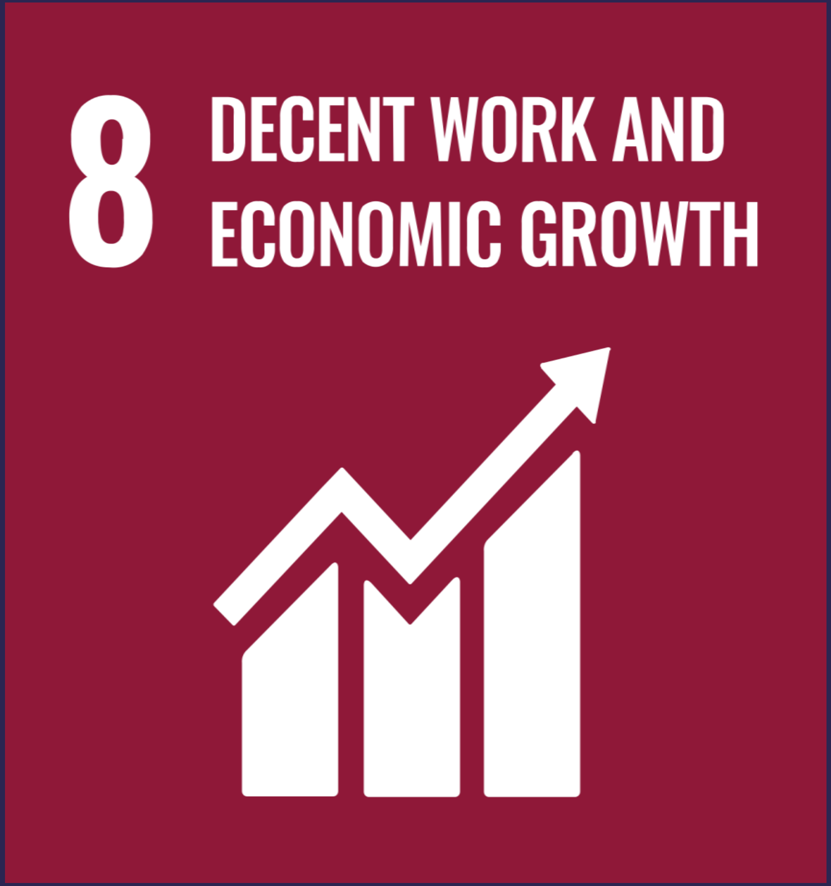
- Our project aims to achieve the following:

1. Develop a machine learning model for sentiment analysis on hotel reviews.
2. Classify reviews into positive, negative, or neutral categories.
3. Compare the performance of different machine learning algorithms.

By achieving these objectives, our project contributes to addressing the challenge of efficiently processing and understanding customer sentiments in the hospitality industry, ultimately supporting businesses in improving their services and making data driven decisions.

SDG Relation

This project aligns with SDG 8, which promotes decent work and economic growth. By automating the analysis of hotel reviews, businesses can gain valuable insights into customer sentiment, allowing them to make informed decisions that improve customer satisfaction, attract more guests, and ultimately boost their economic growth



Data

Data Collection

- **Data Source:**
- Sourced from Kaggle, our TripAdvisor Hotel Reviews dataset comprises 20,491 English reviews. With "Review" for textual feedback and "Rating" for star ratings, this structured data streamlines efficient analysis, facilitating insightful extraction.

Variable	Explanation	
Data Source	Hotel reviews dataset from TripAdvisor website	
Data Size	20,491 hotel reviews	
Variables	Review	Written text of the review
	Rating	Rating of the hotel from 1 to 5 stars

Data Collection

- **Preprocessing steps:** Prior to data analysis, a data cleansing process was performed including:

Converting text to lowercase.

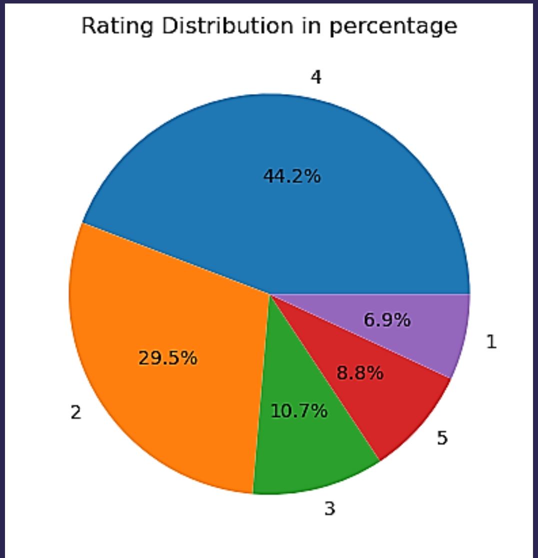
Removal of non-numeric characters and symbols.

Remove stop words (like "and," "for," "with").

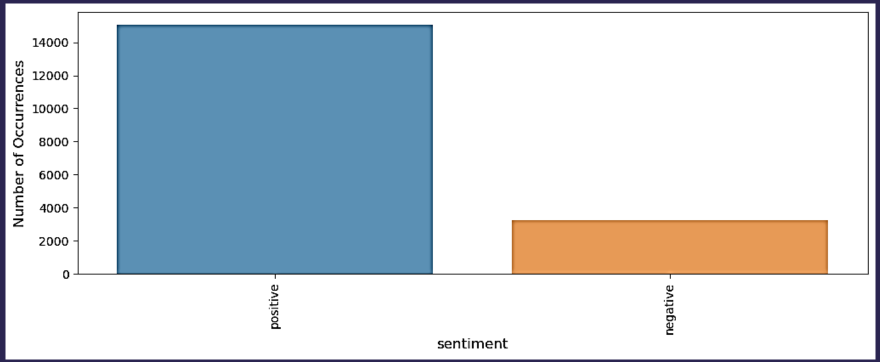
Transforming words into their roots by applying the word stemming process (Lemmatization).

Exploratory Data Analysis (EDA)

- TripAdvisor hotel reviews dataset generally indicates a dominance of 5 and 4 ratings, reflecting predominantly positive sentiments. When examining the data, it can be observed that the combined ratings of 5 and 4 constitute more than 70% of the total evaluations.



- Hotel reviews are classified as positive if the hotel has a rating of 3 stars or higher, and negative if it has a rating of 2 stars or lower.



Feature Engineering

- Created a new feature, 'clean_text,' through a multi-step process, including text cleaning, case conversion, stop-word removal, stemming, and lemmatization.
- Introduced the 'Sentiment' feature, categorizing reviews into positive (5, 4) and negative (1, 2, 3) sentiments based on ratings.

	clean_word	Rating
0	nice hotel expens park got good deal stay hote...	4
1	ok noth special charg diamond member hilton de.	2
2	nice room experi hotel monaco seattl good hote...	3
3	uniqu great stay wonder time hotel monaco loca..	5
4	great stay great stay went seahawk game awesom	5
….	ok just looks nice modern outside, desk staff ...	2

	clean_word	Rating	sentiment
0	nice hotel expens park got good deal stay hote...	4	1
1	ok noth special charg diamond member hilton...	2	0
2	nice room experi hotel monaco seattl good hote..	3	1
3	uniqu great stay wonder time hotel monaco loca.	5	1
…..	……	…	…
20490	great stay great stay went seahawk game ...	5	1

Model

Model Selection and Training

- **Rationale for choosing a specific model:**
- **Decision Tree Classifier** was chosen for sentiment analysis on hotel reviews. Decision trees are suitable for classification tasks, and their interpretability makes them useful for understanding feature importance.
- **Logistic Regression** was selected for its simplicity and effectiveness in binary classification problems, making it a good baseline model.
- **Naive Bayes Classifier** was included due to its efficiency, simplicity, and effectiveness in text classification tasks

Model Selection and Training

- **The strengths and weaknesses:**
- Decision Tree, Logistic Regression, Naive Bayes.

Model	Strengths	Weaknesses
Decision Tree	Handles non-linear relationships.	Prone to overfitting.
	Easy to interpret.	Sensitive to small variations in the data.
Logistic Regression	Simple and interpretable.	Assumes a linear relationship between features and the log-odds of the response.
	Works well for linearly separable data.	
Naive Bayes	Efficient, works well with high-dimensional data.	Assumes independence between features.
	Handles missing data and imbalanced datasets.	

Model Selection and Training

- **Comparison of Models**

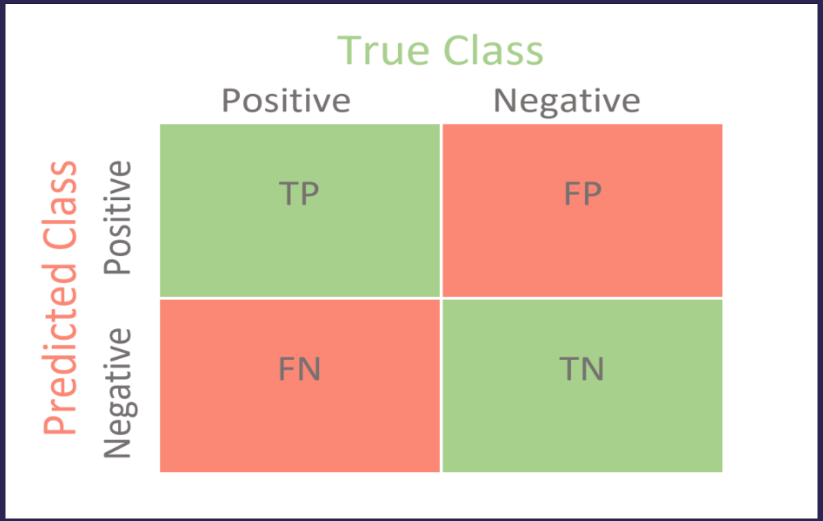
Feature	Decision Tree	Logistic Regression	Naive Bayes
Accuracy	Moderate	Moderate	Moderate
Overfitting	High	Low	Low
Interpretability	High	High	High
Data preprocessing	Low	Moderate	Low
Computational cost	Low	Low	Low
Non-linear relationships	Yes	No	No

Model Evaluation and Hyperparameter Tuning

- **Evaluation metrics and visualizations**

- **Evaluation Metrics:**
    - Accuracy, Precision, Recall, and F1 Score.
- **Visualizations:**
    - Confusion Matrix.

	Precision	Recall	F1-score	Support
positive				
negative				
Accuracy				
Macro ang				
Weighted avg				



Models Performance

- **Model's Performance Overview**
- **Decision Tree:**
  - Testing Accuracy: 0.775
  - Precision, Recall, and F1 Score metrics.
- **Logistic Regression:**
  - Testing Accuracy: 0.8871
  - Precision, Recall, and F1 Score metrics.
- **Naive Bayes:**
  - Testing Accuracy: 0.6279
  - Precision, Recall, and F1 Score metrics.

Model Evaluation and Hyperparameter Tuning

- **Details on Hyperparameter Tuning**
- **Decision Tree:**
  - Default hyperparameters; no significant tuning impact.
- **Logistic Regression:**
  - Tuned learning rates and regularization parameters.
  - Impact on accuracy and convergence speed.
- **Cross-Validation:**
  - Demonstrated impact on generalization.

Model Refinement Overview

- **Refinement Phase Highlights :**
- **Objective:**
  - Enhance initial model performance iteratively.
- **Challenges Addressed:**
  - Overfitting, precision, and recall improvements.

Techniques for Model Improvement

- **Refinement Techniques**

1. **Alternative Algorithms:**
   - Random Forest and Support Vector Machines (SVM).
2. **Ensemble Methods:**
   - Voting Classifier combining various models.
3. **Hyperparameter Tuning:**
   - Learning rates, regularization, batch sizes.
4. **Cross-Validation:**
   - Adjusted folds and applied stratified sampling.

Techniques for Model Improvement

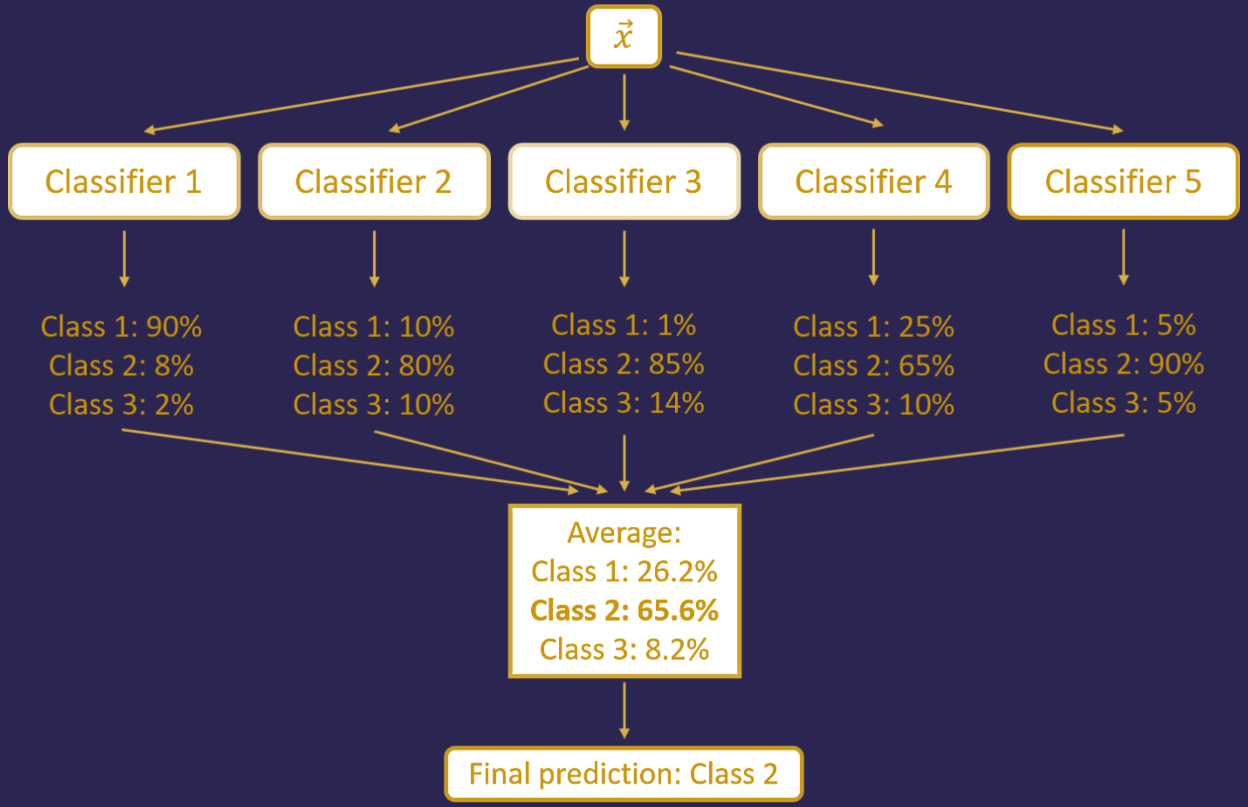
- we employed Random Forest and Support Vector Machines (SVM) to enhance model performance.

Random Forest	Support Vector Machines (SVM)
• Improved accuracy over a single decision tree. • Robust to overfitting.	• Effective in high-dimensional spaces. • Versatile with different kernel functions.

Techniques for Model Improvement

- **VotingClassifier:**

VotingClassifier is an algorithm that enhances classification accuracy by combining multiple classifiers. It has the potential to achieve high accuracy rates on both training and testing data.



Test Submission Overview

- **Test Submission Phase**
- **Dataset Preparation:**
  - Same preprocessing as training and validation datasets.
- **Model Application:**
  - Applied the trained model to a distinct test dataset.
- **Metrics and Results:**
  - Calculated accuracy, precision, recall, and F1 Score.

Metrics and Results on Test Dataset

- **Test Dataset Performance**
- **Metrics:**
  - Accuracy, Precision, Recall, F1 Score.
- **Comparison:**
  - Evaluate consistency with training and validation metrics.
- **Highlight Key Findings:**
  - How well the model generalizes to novel, unseen data.

Results

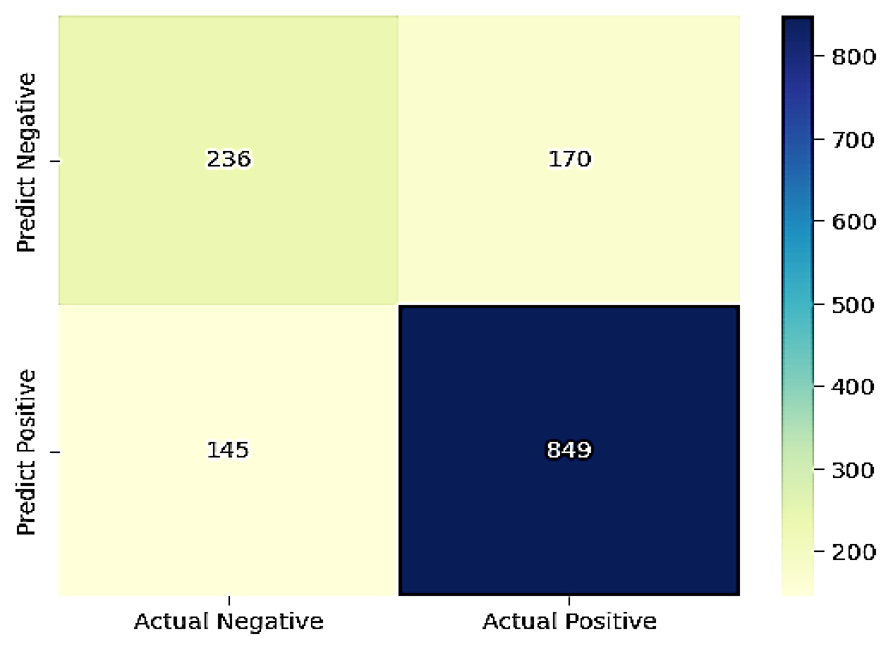
Results

- In our project, we found that TripAdvisor hotel reviews are mostly positive (86.74%) and a small portion of them are negative (13.26%). These reviews show that TripAdvisor services accurately reflect user satisfaction.

Evaluation Results

- **Decision Tree:**
  - In training, the model achieved perfect accuracy (1.0). However, its performance dropped to 0.775 on the test data, indicating oerfitting.

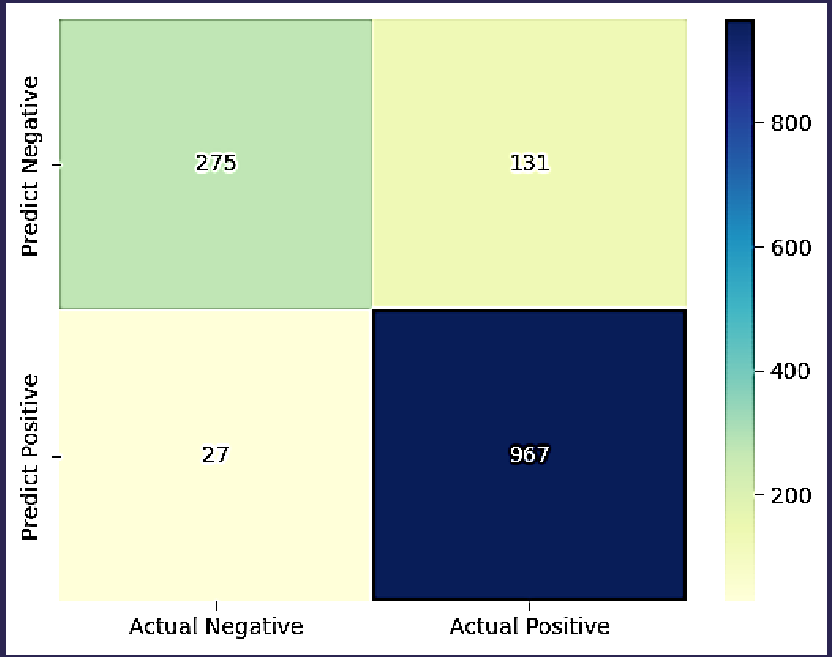
	Precision	Recall	F1-score	Support
positive	0.62	0.58	0.60	406
negative	0.83	0.85	0.84	994
Accuracy			0.78	1400
Macro ang	0.73	0.72	0.72	1400
Weighted avg	0.77	0.78	0.77	1400



Evaluation Results

- **Logistic Regression:**
  - The accuracy score in training is 0.9268, which means that the model has a good fit to the training data. In the test, it is 0.8871, which means that the model is also successful in the test data.

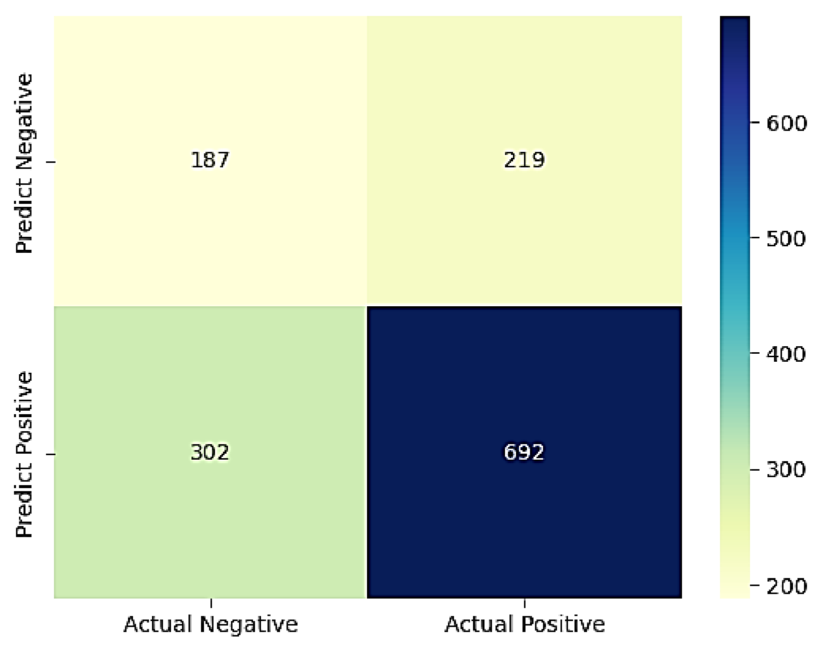
	Precision	Recall	F1-score	Support
positive	0.91	0.68	0.78	406
negative	0.88	0.97	0.92	994
Accuracy			0.89	1400
Macro ang	0.90	0.83	0.85	1400
Weighted avg	0.89	0.89	0.88	1400



Evaluation Results

- **Naive Bayes:**
  - The training accuracy is 86.45%, which means that the model successfully fit the training data. However, the test accuracy is 62.79%, which indicates that the model performs worse on the test data.

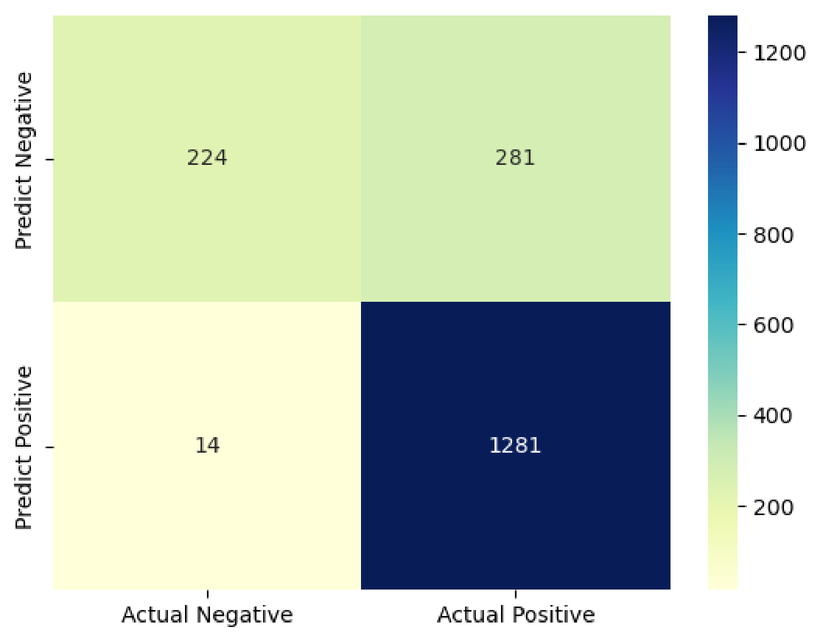
	Precision	Recall	F1-score	Support
positive	0.38	0.46	0.42	406
negative	0.76	0.70	0.73	994
Accuracy			0.89	1400
Macro ang	0.57	0.58	0.57	1400
Weighted avg	0.65	0.63	0.64	1400



Evaluation Results

- **Random Forest:**
  - Despite a perfect training accuracy of 1.0, the model's performance drops to 0.8361 on test data, suggesting overfitting and highlighting the need for further optimization to ensure better generalization.

	Precision	Recall	F1-score	Support
positive	0.94	0.44	0.60	505
negative	0.82	0.99	0.90	1295
Accuracy			0.84	1800
Macro ang	0.88	0.72	0.75	1800
Weighted avg	0.85	0.84	0.81	1800



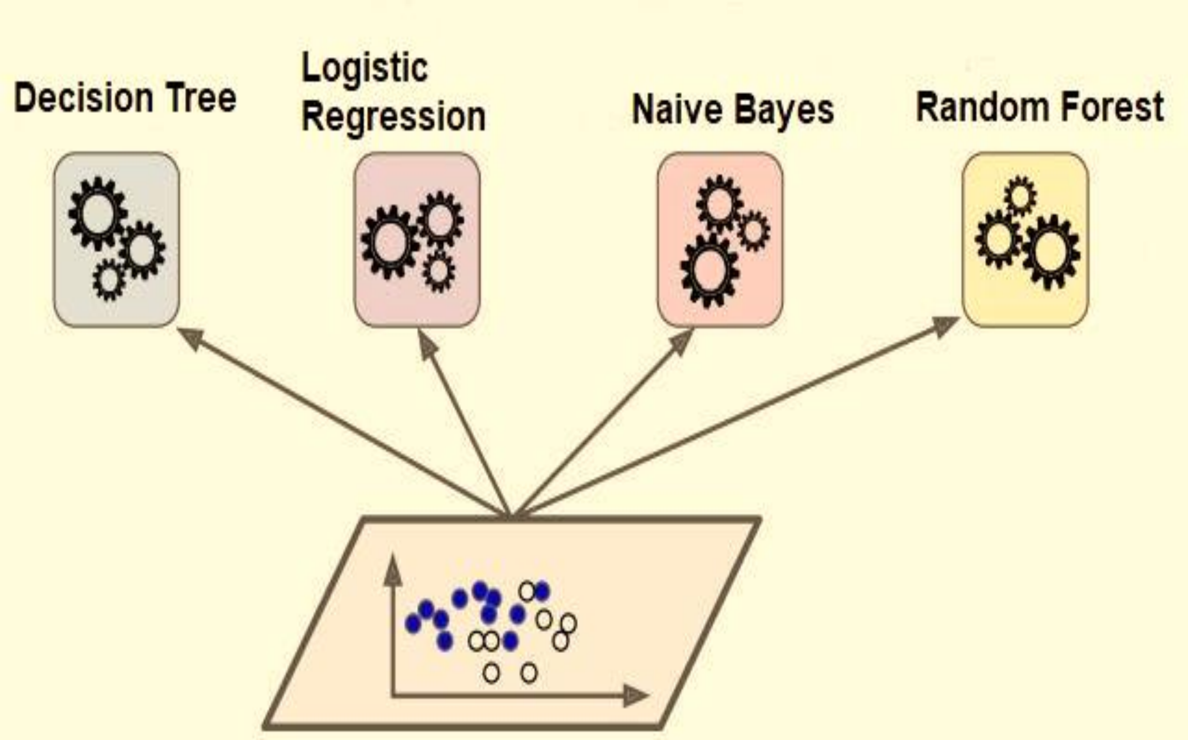
Evaluation Results

- **Support Vector Machine (SVM):**
  - Remarkable results were achieved with a training accuracy score of 0.9869 and a testing accuracy score of 0.8939 for the Support Vector Machine (SVM) algorithm.

Despite the impressive performance, this algorithm was excluded due to its excessively time-consuming training and testing processes, rendering it impractical for real-time application environments.

Evaluation Results

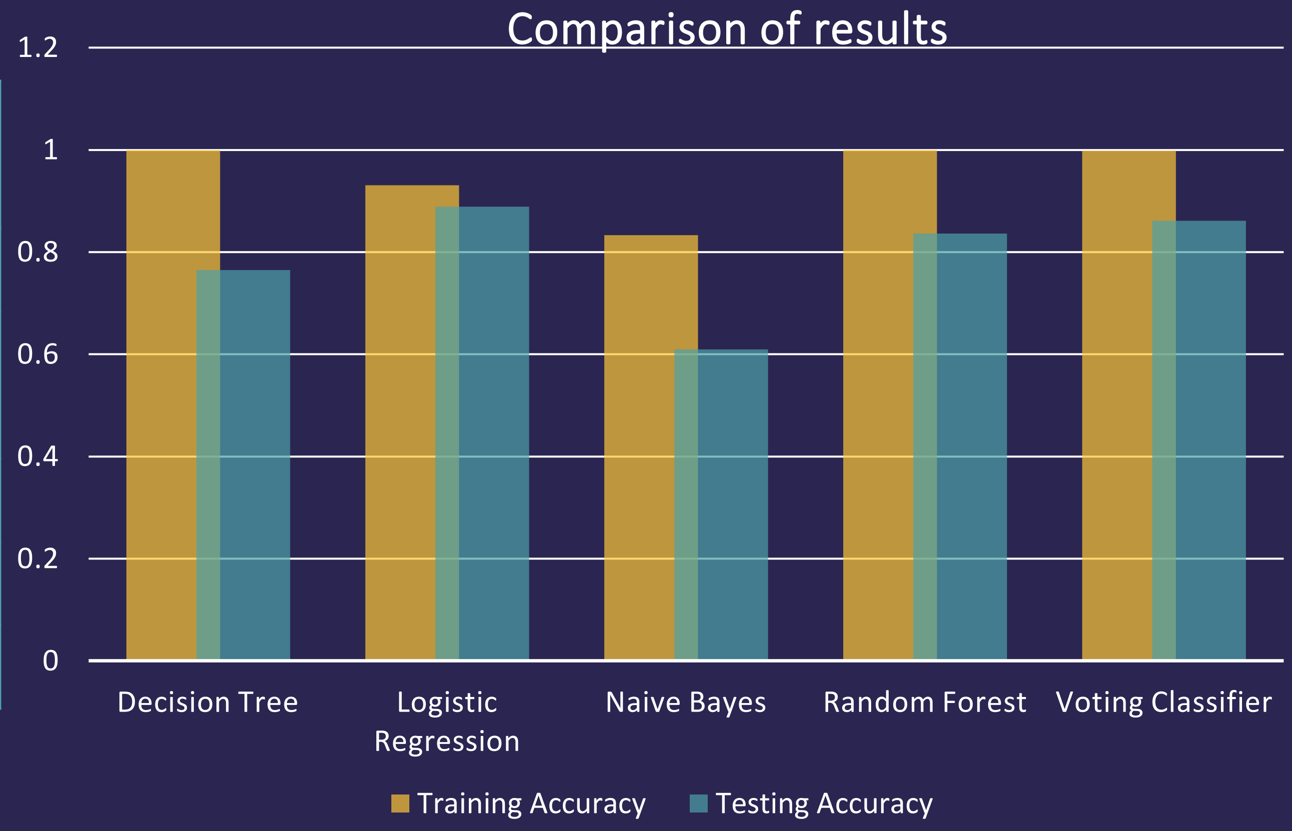
- **Voting Classifier:**
  - The model exhibits excellent learning from the training data, achieving a high accuracy score of 99.85%. Although the testing accuracy is slightly lower at 86.11%, the overall performance is still robust, indicating the model's capability to generalize well to unseen data. Further optimization and evaluation may enhance its effectiveness in real-world scenarios.



Comparison of results

- The results will be compared with the following table, which shows the difference between the training accuracy and test accuracy of each classification algorithm.

Algorithm	Training Accuracy	Testing Accuracy
Decision Tree	1.0	0.765
Logistic Regression	0.9306	0.8889
Naive Bayes	0.8333	0.6094
Random Forest	1.0	0.8361
Voting Classifier	0.9985	0.8611



Deployment

- Overview of the deployment phase
- Model serialization, serving, and API integration
- **(You can fill this part after deployment submission)**

Future Work

- To deepen our understanding of sentiment analysis, we aim to leverage various datasets, focusing on analyzing insights and emotions related to specific products. One avenue we are exploring involves extracting and analyzing a curated set of tweets from platforms like Twitter, gauging people's opinions and reactions towards issues, whether they be political, social, or emerging technological trends. We believe the abundance of data available in this realm presents an opportunity for comprehensive analysis. Our emphasis lies not just in organizing this data but also harnessing its value, providing accurate and meaningful insights.

Thank you!