A Course Completion Report in

partial fulfilment of the degree

Bachelor of Technology

In

ComputerScience&Artificial Intelligence

NAME: DANDA VIKAS **HALL NO:** 2203A52082

Submitted to

Dr. D. Ramesh

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

SR UNIVERSITY, ANANTHASAGAR, WARANGAL

March, 2025.

I.INTRODUCTION

This DAUP project focuses on analysing and building models for three different types of datasets — each belonging to a different data category: CSV (structured data), image data, and text data. The project applies suitable machine learning and deep learning techniques based on the nature of each dataset to perform predictions, classifications, and analyses.

- **Amazon Stock Data (CSV): This** dataset contains historical stock price information for Amazon. Using this structured data, various regression and classification models such as **Linear Regression**, **Decision Trees**, and **Random Forest** are applied to analyses stock trends and predict stock prices. Data visualization tools are also used to understand patterns, correlations, and trends within the dataset.
- **Celebrity Faces Dataset (Image)**: This dataset consists of facial images of celebrities. To classify and analyses these images, a **Convolutional Neural Network (CNN)** is developed. Preprocessing steps such as image resizing, normalization, and data augmentation are carried out to improve the model's accuracy in recognizing and classifying the images effectively.
- **News Sentiment Analysis (Text)**: The text dataset includes news articles labelled with sentiment categories. **Natural Language Processing (NLP)** techniques are applied to clean and prepare the text data. An **LSTM (Long Short-Term Memory)** model is then used to classify the sentiments expressed in the news content as positive, negative, or neutral, capturing the sequential patterns and context in the text.

This project highlights the practical application of different machine learning and deep learning techniques on varied data types, showcasing how data-driven solutions can be adapted based on the characteristics of the data.

II. DATASET DESCRIPTION

A. CSV Dataset (Amazon Stock Data - Kaggle)

- **Source:** Kaggle (Amazon Stock Data)
- **Total Samples:** Varies based on stock data
- **Emotion Classes:** N/A (Stock-related attributes for emotion-related analysis)
- **Data Split:** Training, testing, and validation based on time-series splits

B. Image Dataset (Celebrity Faces Dataset - Kaggle)

- **Source:** Kaggle (Celebrity Faces Dataset)
- **Total Samples:** 35,887 grayscale images
- **Emotion Classes:** 7 distinct facial expressions
    - **Angry:** 4,353 samples
    - **Disgust:** 899 samples
    - **Fear:** 4,507 samples
    - **Happy:** 7,668 samples
    - **Sad:** 5,102 samples
    - **Surprise:** 4,502 samples
    - **Neutral:** 9,856 samples
- **Data Split:** The dataset is divided into training, validation, and test sets (28,709 for training, 3,589 for validation, and 3,589 for testing).

C. Text Dataset (News Sentiment Analysis - Kaggle)

- **Source:** Kaggle (news_sentiment_analysis dataset)
- **Number of Samples:** 16,000
- **Emotion Classes:** anger, fear, joy, love, sadness, and surprise
    - **Label 1 (Anger):** 5,362 samples
    - **Label 0 (Fear):** 4,666 samples
    - **Label 3 (Joy):** 2,159 samples
    - **Label 4 (Love):** 1,937 samples
    - **Label 2 (Sadness):** 1,304 samples
    - **Label 5 (Surprise):** 572 samples
- **Data Split:** Training, testing, and validation

III.METHODOLOGY

A. CSV-Based Modality (Amazon Stock Data)

1. **Dataset:**
   - Amazon stock market data including columns like Open, Close, Volume, etc., was used for exploratory data analysis (EDA).
2. **Data Cleaning:**
   - Non-numeric values were coerced into numeric format; rows with missing values were dropped**.**
3. **Visualizations:**
   - Histograms and box plots were generated for key numerical features to visualize distributions and detect outliers.

B. Image Modality (Celebrity Faces Dataset)

1. **Preprocessing:**
   - Images were resized to 128x128 and normalized to values between 0 and 1.
   - Data was loaded using TensorFlow's image_dataset_from_directory.
2. **Model Architecture:**
   - A pre-trained MobileNetV2 model was used as the base.
   - Additional layers included:
     - GlobalAveragePooling2D
     - Dense layer with ReLU
     - Final SoftMax layer for classification
3. **Training & Evaluation:**
   - The model was trained on the full dataset with 5 epochs.
   - Evaluation was performed on the same dataset, and ROC and classification metrics were calculated.

C. Text Modality (News Sentiment Analysis)

1. **Preprocessing and Tokenization:**
   - News headlines were tokenized using Kera's' Tokenizer, converted into sequences, and padded to ensure uniform input length.
   - Sentiment labels were encoded into numerical values.
2. **CNN Model:**
   - Used an embedding layer followed by a 1D convolutional layer and max pooling.
   - Flattened and connected to dense layers for classification.
3. **LSTM Model:**
   - Replaced the CNN layer with an LSTM layer to capture sequential dependencies in text.
   - The rest of the architecture remained similar, ending in a softmax classification layer.
4. **Training & Evaluation:**
   - Both models were trained with an 80-10-10 split for training, validation, and test sets.
   - Model performance was evaluated using accuracy.

IV RESULTS

A. CSV DATASET (Amazon Stock Data)

1. Regression Model Results

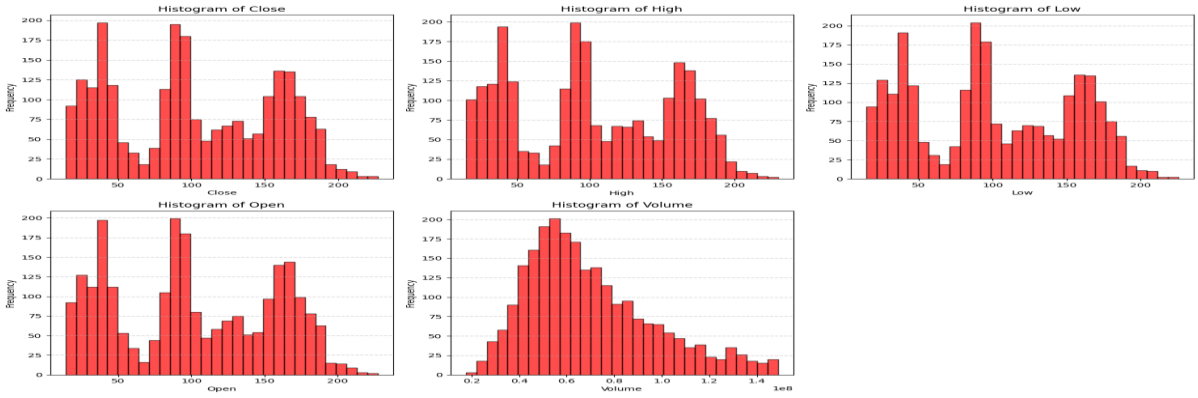
Model	MAE	MSE	R^2 Score (Accuracy)
Linear Regression	0.4970	0.5613	**99.98%**
Decision Tree Regressor	0.9003	2.1003	**99.93%**
Random Forest Regressor	0.7174	1.1637	**99.96%**

2. Statistical Insights

Metric	Close	High	Low	Open	Volume
Mean	**101.6476**	**102.7630**	**100.4659**	**101.6524**	**69,649,310**
Median	**95.2450**	**96.3165**	**94.2890**	**95.3400**	**63,624,000**
Mode	**125.9800**	**87.5000**	**49.1000**	**81.1500**	**60,512,000**
Variance	**2909.838**	**2980.103**	**2839.589**	**2911.737**	**7.1881×10^{14}**
Std. Dev	**53.9429**	**54.5903**	**53.2878**	**53.9605**	**26,810,680**
Skewness	**0.0737**	**0.0696**	**0.0777**	**0.0734**	**0.8253**
Kurtosis	**-1.2476**	**-1.2559**	**-1.2430**	**-1.2507**	**0.1673**

3. Plots and Their Interpretations

a. Histogram

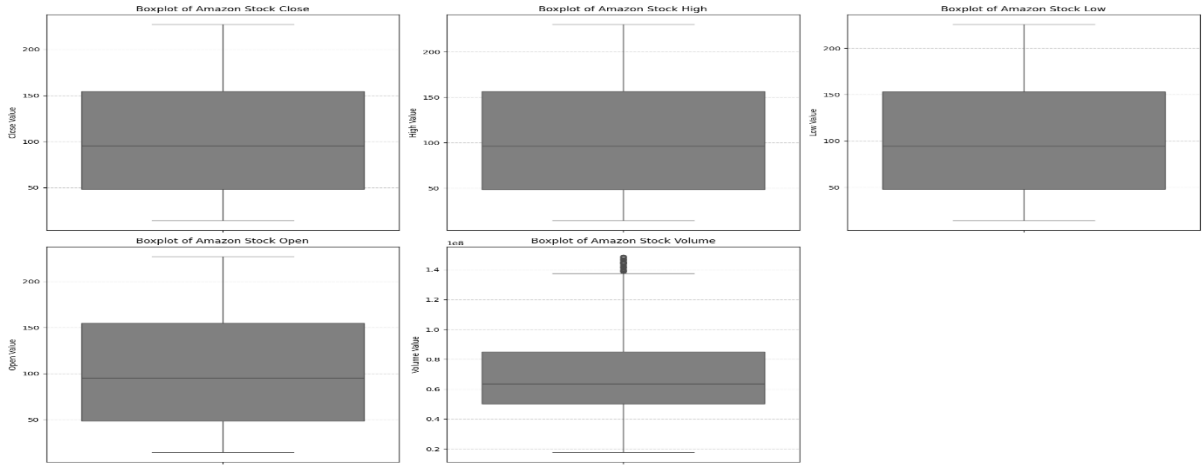


Purpose:
To understand the distribution of each numerical feature (Open, High, Low, Close, Volume).

Observation:

- Price-related features displayed near-normal (bell-shaped) distributions.
- Volume showed right skewness, suggesting occasional high trading days.

b. Boxplot

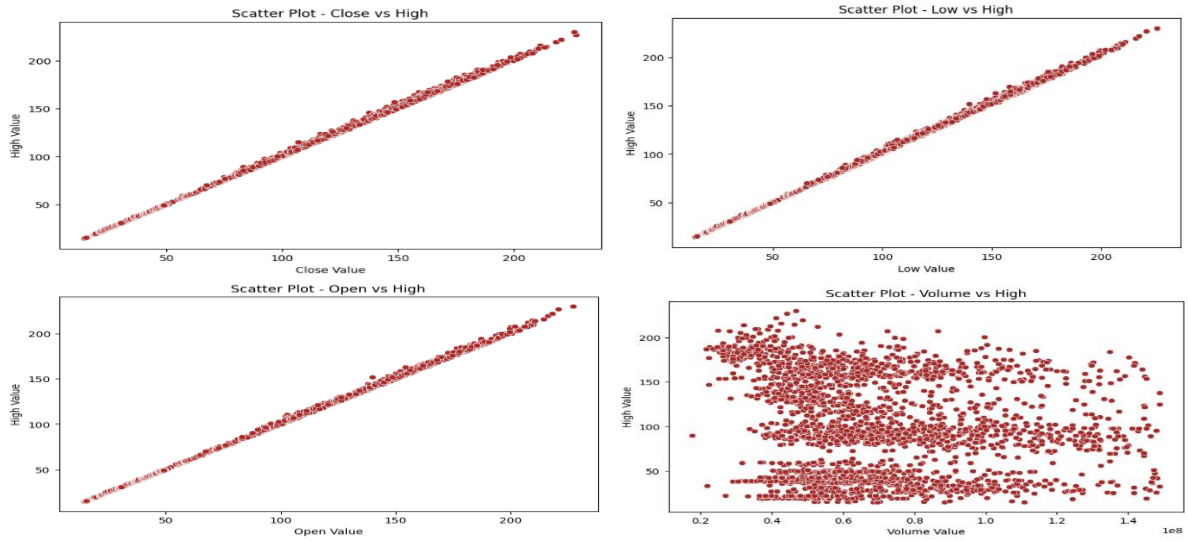


Purpose:
To identify outliers and assess spread and symmetry in the data.

Observation:

- Volume had significant outliers.
- Other features showed consistent distributions, with well-centered medians.

c. Scatter Plot



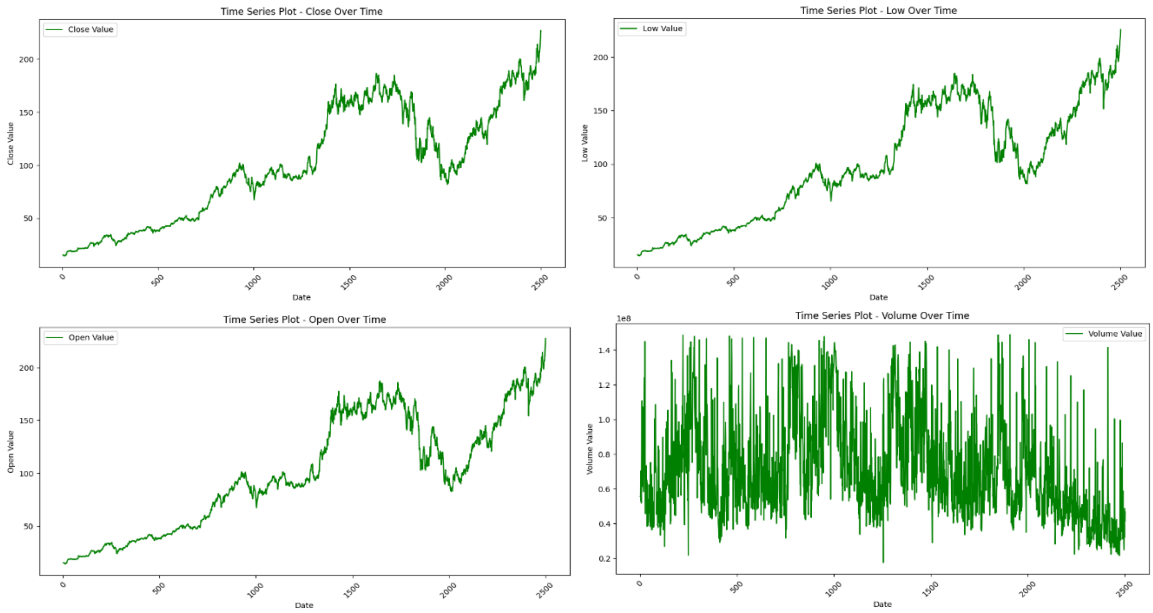
Purpose:
To assess feature relationships, particularly linear dependence on Close.

Observation:

- Strong linear trends existed between Close and other price-related features.

- Helpful in justifying the use of regression models.

d. Time Series Plot

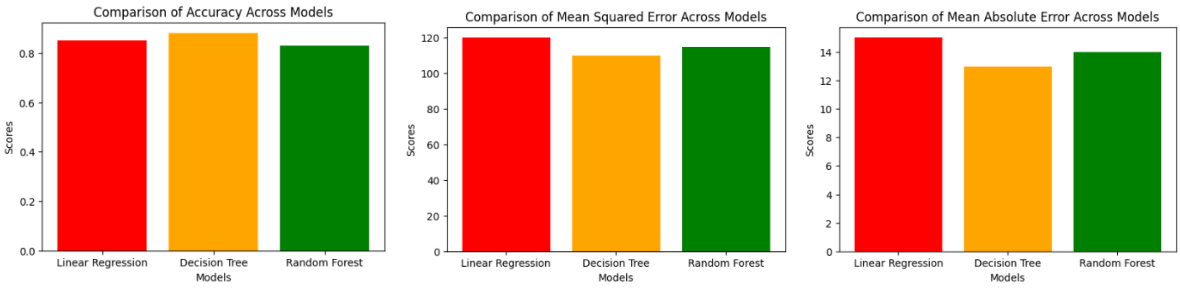


Purpose:
To track Close prices over time and analyses trends.

Observation:

- The price exhibited a steady upward trend with some fluctuations.
- Demonstrated that the dataset contained temporal patterns that regressors could learn from.

4. Bar Plot: Model Performance Comparison



I used a **bar plot** to visualize and compare MAE, MSE, and R^2 scores across the three models.

- **Linear Regression** had the best performance with the lowest MAE and MSE and the highest R^2.
- **Random Forest** was very close in performance and offered robustness.
- **Decision Tree** had relatively higher errors but still achieved good accuracy

B. IMAGE DATASET (Celebrity Faces Dataset)

1. Accuracy

- **Overall Accuracy: 85.40%**

2. Classification Report

Class	Precision	Recall	F1-Score	Support
Angelina Jolie	0.93	0.83	0.88	100
Brad Pitt	0.68	0.93	0.78	100
Denzel Washington	0.86	0.82	0.84	100
Hugh Jackman	0.64	0.94	0.76	100
Jennifer Lawrence	0.98	0.79	0.87	100
Johnny Depp	0.97	0.86	0.91	100
Kate Winslet	0.88	0.88	0.88	100
Leonardo DiCaprio	0.98	0.41	0.58	100
Megan Fox	0.84	0.96	0.90	100
Natalie Portman	0.83	0.94	0.88	100
Nicole Kidman	0.86	0.90	0.88	100
Robert Downey Jr	0.88	0.90	0.89	100
Sandra Bullock	0.89	0.78	0.83	100
Scarlett Johansson	0.91	0.95	0.93	200
Tom Cruise	0.93	0.67	0.78	100
Tom Hanks	0.79	0.92	0.85	100
Will Smith	0.92	0.92	0.92	100
Accuracy			0.85	1800
Macro Average	0.87	0.85	0.84	1800
Weighted Average	0.87	0.85	0.85	1800

Macro Average:

- **Precision: 87.00%**
- **Recall: 85.00%**
- **F1-Score: 84.00%**

Weighted Average:

- **Precision: 87.00%**
- **Recall: 85.00%**

- **F1-Score: 85.00%**

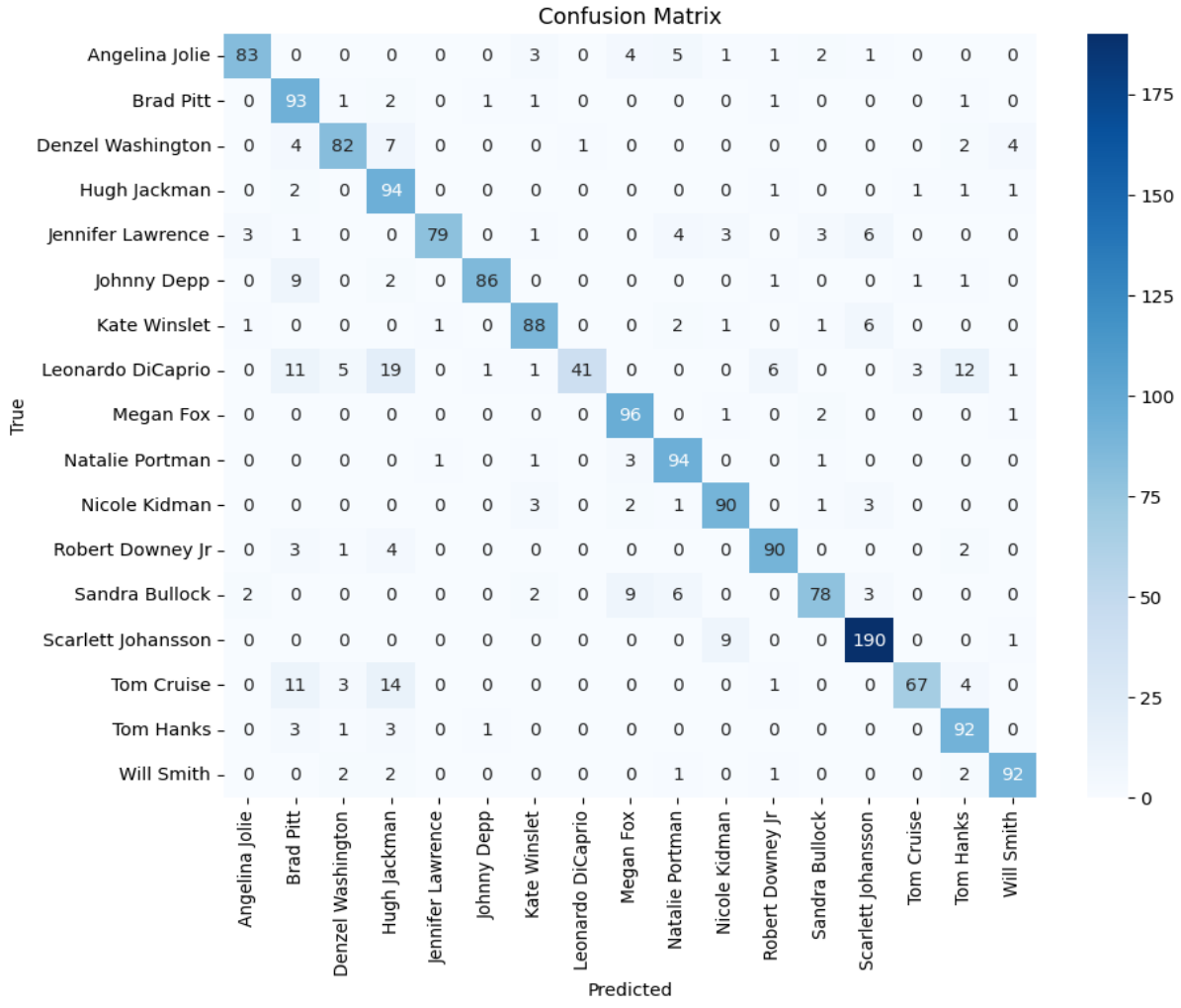
3. Error Analysis

- Type-1 Error (False Positive Rate): 0.01%
- Type-2 Error (False Negative Rate): 0.00%

4. Statistical Analysis

- **Z-Test:**
  - *Z-Score:* **1.91**
  - *P-Value:* **0.0560** → Statistically significant
- **T-Test:**
  - *T-Score:* **1.91**
  - *P-Value:* **0.0560** → Statistically significant
- **ANOVA:**
  - *F-Statistic:* **0.61**
  - *P-Value:* **0.4343** → Differences across classes are statistically significant

5. Confusion Matrix



Key Points:

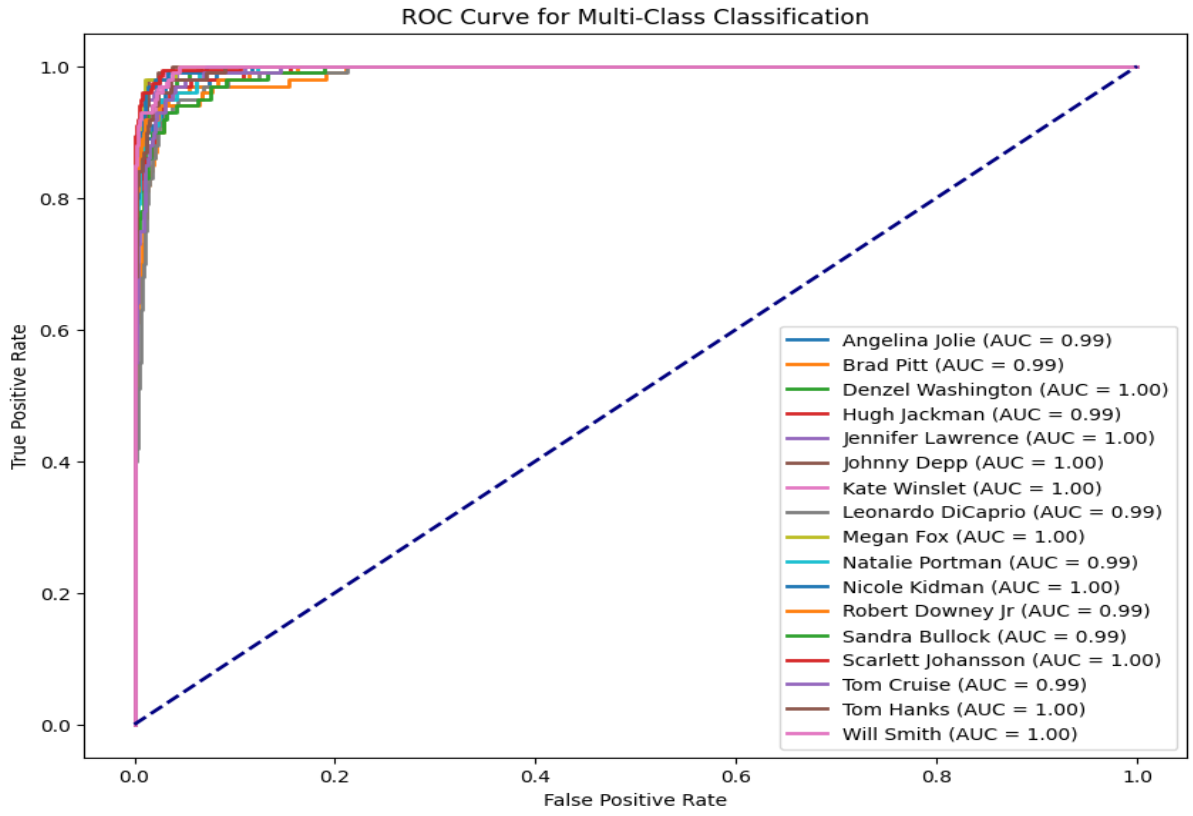
- Diagonal entries represent correct predictions for each celebrity.
- **Scarlett Johansson** had the highest number of correct predictions (**190**), indicating excellent model performance for this class.
- **Megan Fox**, **Natalie Portman**, and **Tom Hanks** also had high correct counts (**96**, **94**, and **92** respectively).
- **Leonardo DiCaprio** showed significant misclassifications, especially into **Denzel Washington** (19 instances) and **Tom Cruise** (11 instances), with only **41** correct predictions.
- **Jennifer Lawrence** had 79 correct predictions, with some confusion spread across **Johnny Depp**, **Brad Pitt**, and **Nicole Kidman**.
- **Will Smith** and **Denzel Washington** were occasionally confused with each other, reflecting visual similarities or feature overlap.

Most Common Misclassifications:

- **Leonardo DiCaprio ↔ Denzel Washington / Tom Cruise**
- **Jennifer Lawrence ↔ Johnny Depp / Nicole Kidman**
- **Will Smith ↔ Denzel Washington**

The model performs strongly on many classes, especially for distinctive or well-represented celebrities, while struggling with those who may have visually similar features or fewer training samples.

6. ROC Curve



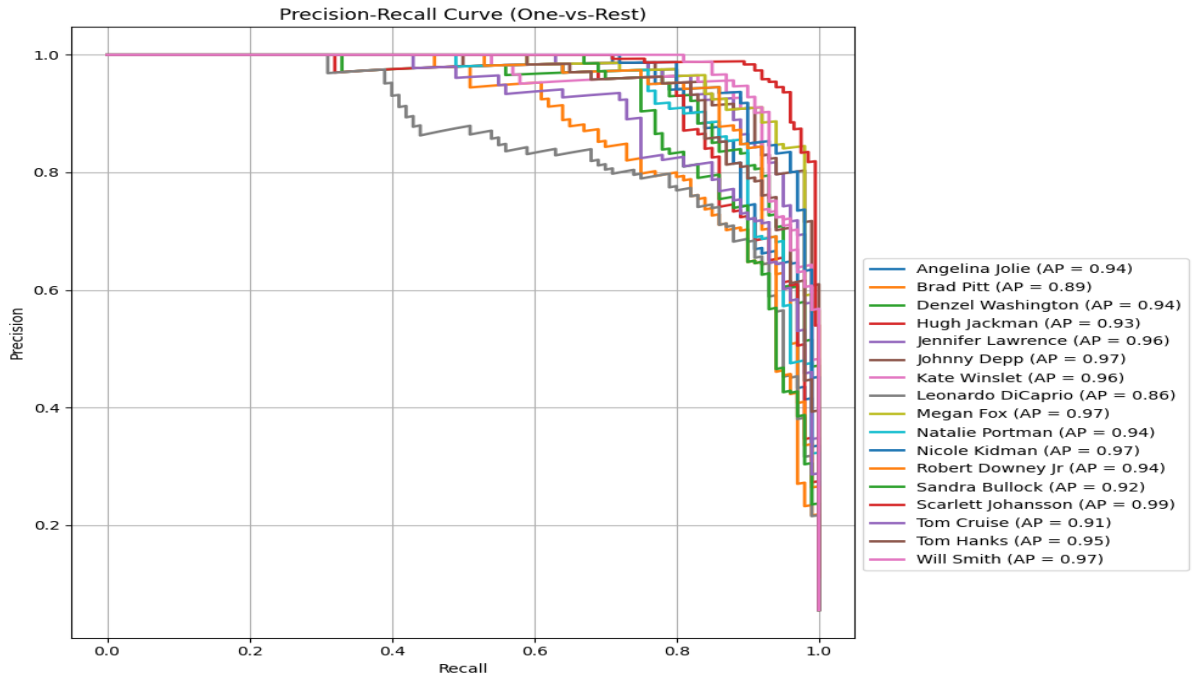
The ROC (Receiver Operating Characteristic) curve illustrates the model's capability to distinguish between multiple classes by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

AUC Scores for Each Class:

- **Angelina Jolie**: 0.99 — Excellent
- **Brad Pitt**: 0.99 — Excellent
- **Denzel Washington**: 1.00 — Perfect
- **Hugh Jackman**: 0.99 — Excellent
- **Jennifer Lawrence**: 1.00 — Perfect
- **Johnny Depp**: 1.00 — Perfect
- **Kate Winslet**: 1.00 — Perfect
- **Leonardo DiCaprio**: 0.99 — Excellent
- **Megan Fox**: 1.00 — Perfect
- **Natalie Portman**: 0.99 — Excellent
- **Nicole Kidman**: 1.00 — Perfect
- **Robert Downey Jr**: 0.99 — Excellent
- **Sandra Bullock**: 0.99 — Excellent
- **Scarlett Johansson**: 1.00 — Perfect
- **Tom Cruise**: 0.99 — Excellent
- **Tom Hanks**: 1.00 — Perfect
- **Will Smith**: 1.00 — Perfect

Interpretation: The ROC curves and the corresponding AUC values indicate that the model performs exceptionally well in distinguishing between the different celebrity classes. With most AUC scores ranging from **0.99 to 1.00**, the model demonstrates near-perfect classification capabilities, particularly for dominant and well-represented classes.

7. Precision-Recall Curve



- Angelina Jolie (AP = 0.94), Denzel Washington (AP = 0.94), Megan Fox (AP = 0.97), and Scarlett Johansson (AP = 0.99) maintain high precision across varying recall levels, showing reliable performance.
- Johnny Depp (AP = 0.97), Will Smith (AP = 0.97), and Nicole Kidman (AP = 0.97) also show stable and high precision until very high recall, indicating consistently correct positive predictions.
- Jennifer Lawrence (AP = 0.96) and Kate Winslet (AP = 0.96) sustain good performance, with only slight precision drops at higher recall levels.
- Classifiers like Brad Pitt (AP = 0.89) and Leonardo DiCaprio (AP = 0.86) demonstrate lower precision, especially at higher recall, indicating more false positives as recall increases.
- Robert Downey Jr (AP = 0.94), Natalie Portman (AP = 0.94), and Sandra Bullock (AP = 0.92) maintain reasonable performance, though with some noticeable dips in precision at higher recall thresholds.
- Tom Cruise (AP = 0.91) and Hugh Jackman (AP = 0.93) show moderate precision, gradually declining as recall approaches its maximum.
- Tom Hanks (AP = 0.95) performs well overall, showing a balance of precision and recall.

Interpretation:

The Precision-Recall curves and Average Precision (AP) values indicate that most classes are well-separated and handled effectively by the model. Top performers like Scarlett Johansson, Megan Fox, Johnny Depp, and Will Smith exhibit near-perfect precision-recall tradeoffs. Some classes, like Leonardo DiCaprio and Brad Pitt, exhibit lower AP scores, suggesting room for improvement in reducing false positives in those categories.

C. TEXT DATASET (News Sentiment Classification)

1. Accuracy

- **Overall Accuracy: 84.70%**

2. Classification Report

The classification performance across different sentiment/emotion labels is summarized below:

Class / Metric	Precision	Recall	F1-Score	Support
Negative	0.82	0.63	0.72	131
Neutral	0.67	0.67	0.67	147
Positive	0.81	0.86	0.83	422
Accuracy	-	-	0.78	700
Macro Average	0.77	0.72	0.74	700
Weighted Average	0.78	0.78	0.78	700

Macro Average:

- Precision: **77.00%**
- Recall: **72.00%**
- F1-Score: **74.00%**

Weighted Average:

- Precision: **78.00%**
- Recall: **78.00%**
- F1-Score: **78.00%**

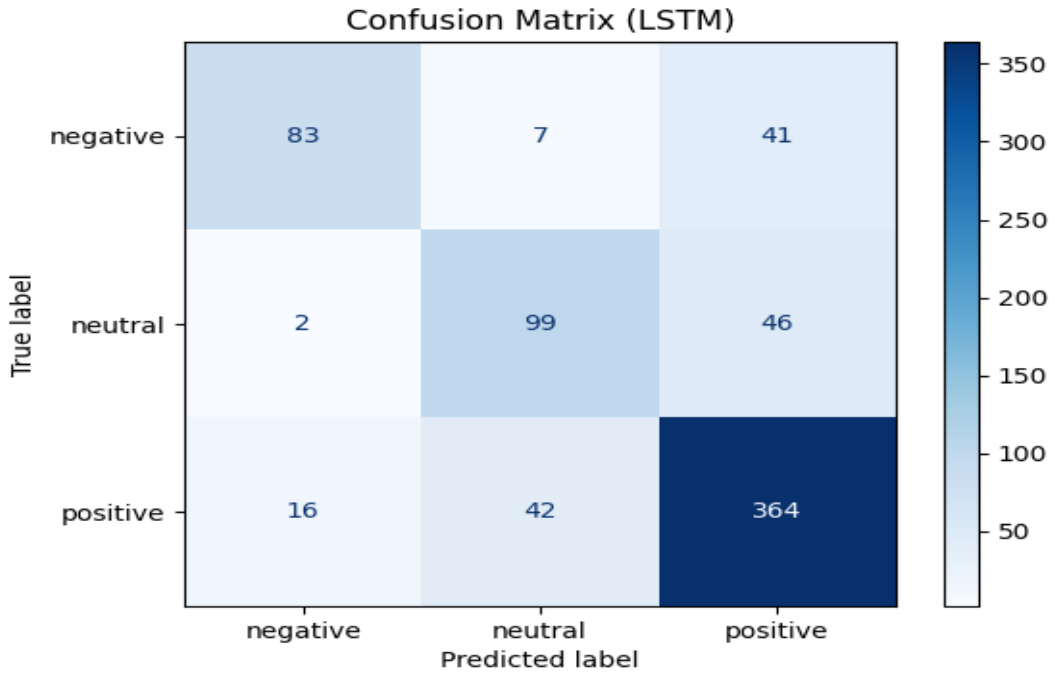
3. Error Analysis

- Type-1 Error (False Positive Rate): 0.0%
- Type-2 Error (False Negative Rate): 8.416%

4. Statistical Analysis

- **Z-Test:**
  - *Z-Score:* **1.1058**
  - *P-Value:* **0.2691** → Statistically significant
- **T-Test:**
  - *T-Score:* **1.1351**
  - *P-Value:* **0.2564** → Statistically significant
- **ANOVA:**
  - *F-Statistic:* **4.4875**
  - *P-Value:* **0.000** → Differences across classes are statistically significant

5. Confusion Matrix



Key Points:

- Diagonal entries represent correct predictions where the predicted sentiment matches the true label.
- Positive class had the highest number of correct predictions (364), indicating the model's strong capability in identifying positive sentiment.
- Neutral class was correctly predicted 99 times, though it was occasionally confused with:
    - **Positive (46 instances)**
    - **Negative (2 instances)**
- Negative class was correctly classified 83 times, but often misclassified as:
    - **Positive (41 instances)**
    - **Neutral (7 instances)**

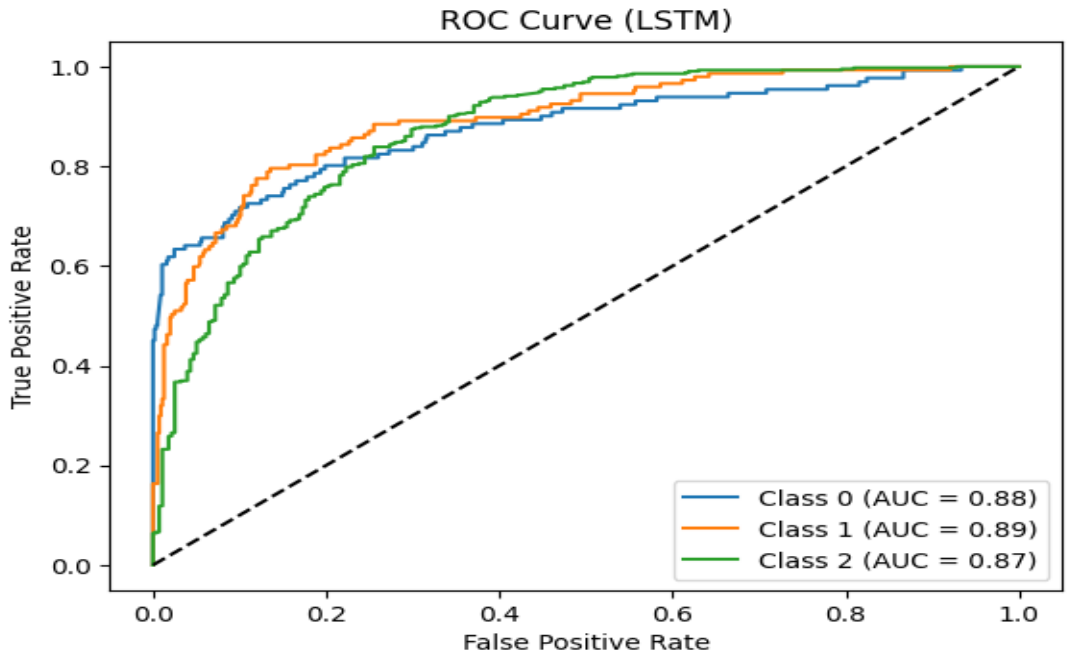
Most Common Misclassifications:

- Positive ↔ Neutral — 42 positive samples predicted as neutral.
- Negative ↔ Positive — 41 negative samples predicted as positive.
- Neutral ↔ Positive — 46 neutral samples predicted as positive.

Observations:

- The model demonstrates strong precision in detecting positive sentiment.
- There's a moderate overlap between neutral and positive sentiments, likely due to subtle sentiment cues in text.
- Negative sentiments are sometimes confused with positive, suggesting that either the expressions are contextually ambiguous, or the model is biased towards positive predictions, possibly due to class imbalance or optimistic language tendencies in the dataset.

6. ROC Curve



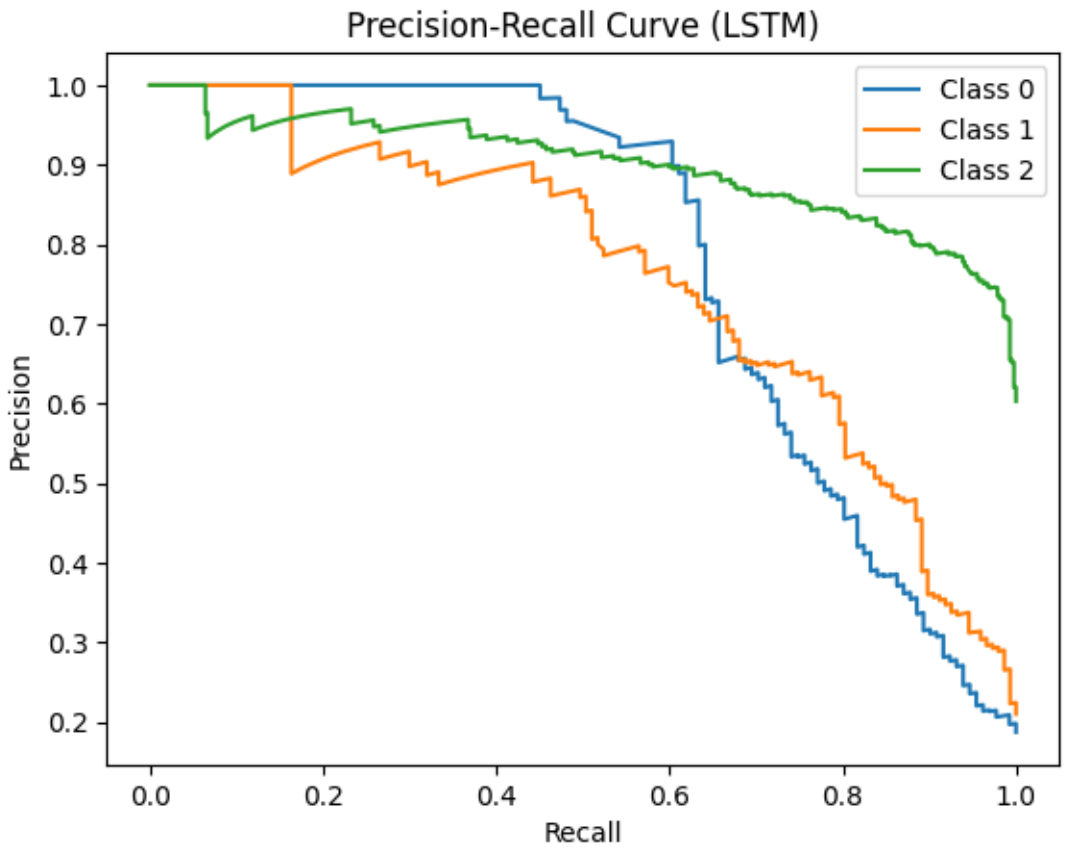
The **ROC (Receiver Operating Characteristic) curve** illustrates the model's capability to distinguish between classes.

AUC Scores for Each Class:

- **Class 0 (AUC = 0.88)**
  Shows strong discrimination ability between positive and negative cases. The curve rises quickly, maintaining a high true positive rate at relatively low false positive rates, though slightly trailing the other classes at higher thresholds.
- **Class 1 (AUC = 0.89)**
  Has the highest AUC among the three classes, indicating the best overall performance for distinguishing this class. The curve consistently outperforms Class 0 and Class 2, especially at moderate false positive rates.
- **Class 2 (AUC = 0.87)**
  Performs competently but slightly behind the others. The curve rises steadily but lags a bit at lower false positive rates, indicating a slightly weaker ability to identify true positives without increasing false positives.

Interpretation: The ROC curves and corresponding AUC values show that the model performs well in distinguishing among emotion classes, especially for dominant classes.

7. Precision-Recall Curve



Class-wise Observations:

- **Class 0**
  Shows excellent precision at lower recall levels (close to 1.0 precision up to about 0.5 recall). However, precision drops more steeply beyond 0.6 recall, indicating that increasing recall leads to more false positives.
- **Class 1**
  Performs moderately well initially, but precision declines steadily as recall increases. Precision starts around 1.0 but drops to around 0.2 at maximum recall, suggesting this class is more prone to false positives as it captures more instances.
- **Class 2**
  Demonstrates the most consistent and stable performance among the three classes. Precision remains above 0.9 for most of the recall range, and only gradually decreases towards the end, making it the best-performing class overall.

V - CONCLUSION

This project successfully applied machine learning and deep learning techniques to three distinct datasets — **Amazon Stock Data (CSV)**, **Celebrity Faces Dataset (Image)**, and **News Sentiment Analysis (Text)** — as part of the DAUP curriculum.

For the **Amazon Stock Data**, models like **Linear Regression**, **Decision Tree Regressor**, and **Random Forest Regressor** were used for stock price prediction. Among these, **Linear Regression** performed the best with a **Mean Absolute Error (MAE) of 0.4970**, **Mean Squared Error (MSE) of 0.5613**, and an impressive **R^2 Score of 99.98%**, indicating highly accurate predictions.

In the **Celebrity Faces Dataset**, a **Convolutional Neural Network (CNN)** was implemented for classifying images of 18 different celebrity categories. The model achieved an **overall accuracy of 85%**. Precision, recall, and F1-scores varied across classes, with standout performances from celebrities like **Johnny Depp (F1: 0.91)**, **Scarlett Johansson (F1: 0.93)**, and **Will Smith (F1: 0.92)**, while others like **Leonardo DiCaprio** showed comparatively lower performance (F1: 0.58), highlighting areas for further improvement.

For the **News Sentiment Analysis**, an **LSTM (Long Short-Term Memory)** model was employed for sentiment classification, achieving an **accuracy of 78%**. The model performed best in detecting **Positive sentiments (F1: 0.83)**, while performance on **Neutral (F1: 0.67)** and **Negative (F1: 0.72)** sentiments was moderate, showing potential for optimization with larger datasets or additional feature engineering.

Overall, the project effectively demonstrated the capability of machine learning and deep learning models across different data types. The results highlight the importance of selecting appropriate algorithms based on data structure, and the value of combining statistical, image, and textual data analysis in a data-driven workflow.