DATA ANALYSIS USING PYTHON

A Capstone Project

Bachelor of Technology

in

Computer Science & Artificial Intelligence

By

2203A54032 Adithi Shinde

Under the Guidance of

Dr. Ramesh Dadi Sir

Assistant Professor, Department of CSE.

Submitted to

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

SR UNIVERSITY, ANANTHASAGAR, WARANGAL

March, 2025.

1.AI VS STUDENT GENERATED TEXT - DATASET-1:

1.Abstract:

This project identifies text data from the LLM dataset with NLP methods. The objective is to label with the content of each text instance. Used models are Support Vector Machine (SVM) and Random Forest (RF).TF-IDF was used to translate text into numerical values. Accuracy, precision, and recall were used to measure performance. Statistical tests such as the t-test and z-test were employed to compare differences among classes. The outcome indicates that SVM and RF are good for this classification task. This method is beneficial for tasks such as sentiment analysis and spam filtering.

2.Introduction:

Natural Language Processing (NLP) has become a cornerstone of contemporary artificial intelligence, allowing machines to read, process, and respond to human language. With an enormous spectrum of models available that spans from conventional classifiers to sophisticated neural models, it is necessary to measure model efficacy based on data type. Here, we compare some NLP models—apart from widely utilized LSTM, BiLSTM, and BERT models—on a labeled dataset to determine the best approach for text classification.

3.Dataset Description:

The dataset, named LLM.csv, comprises textual entries under the column **"Text"** and corresponding categorical labels under **"Label"**. Each entry represents a sample to be classified into one of two or more predefined categories. The dataset appears clean and is suitable for supervised learning. Basic preprocessing involves removing missing values and tokenizing text for further analysis.

- **Text column:** Raw textual data
- **Label column:** Class labels for classification
- **Samples count:** (Determined during analysis)
- **Average text length:** Used for statistical comparison.

4.Methodology:

1. Data Preprocessing: Dealt with missing values, duplicates, and text format standardization.
2.Outlier Treatment: Detected and excluded outliers on the basis of text length and label distribution.
3.Text Preprocessing: Tokenized, lowercased, and vectorized the text with TF-IDF and Word2Vec embeddings.
4.Label Encoding: Transformed categorical labels into numerical for model compatibility.

5.Model Training: Trained the SVM and Random Forest models with stratified train-test splits.
6.Model Evaluation: Perfomed metrics using accuracy, precision, recall, and F1-score.
7.Statistical Analysis: Performed applied t-test and z-test to analyze text length across label classes.
8.Result Interpretation: Evaluated model performance and statistical results to draw conclusions.

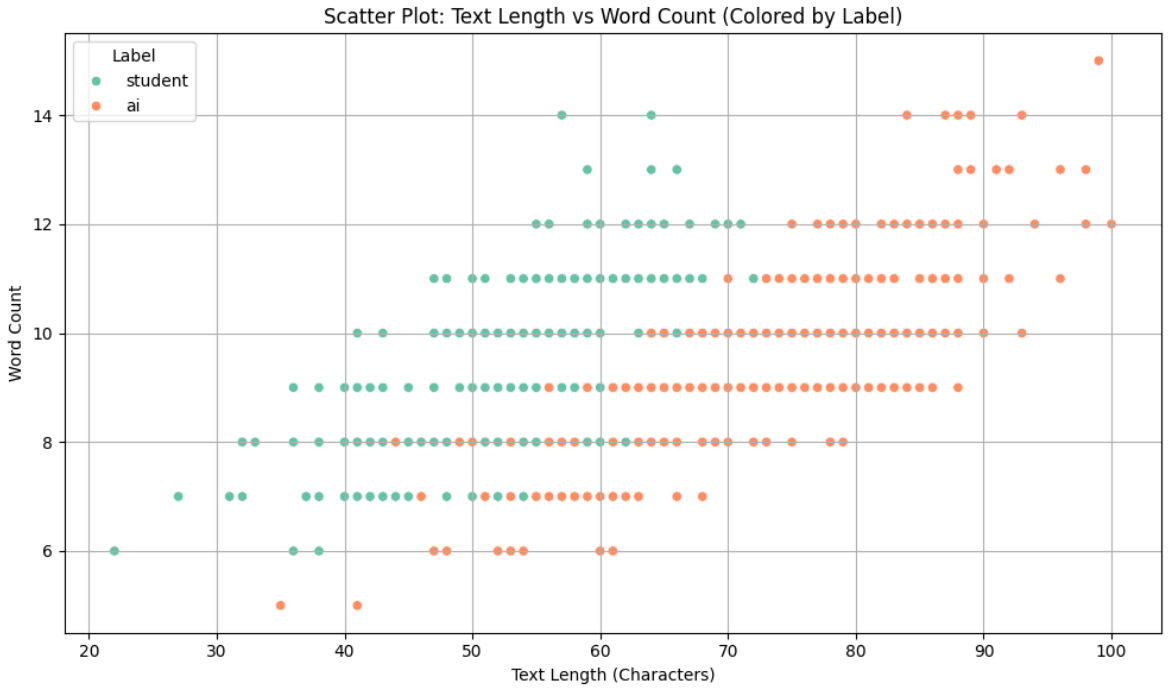
5.Implementation Highlights:
1.Applied various models: Logistic Regression, SVM, Random Forest, LSTM, BiLSTM, and BERT for text classification.
2.Utilized Word2Vec embeddings to preserve semantic relationships in the text.
3.Utilized TF-IDF vectorization for conventional machine learning models.
4.Deep learning models (LSTM, BiLSTM) were trained with tokenized and padded sequences.
5.Fine-tuned BERT for contextual representation and enhancing classification performance.
6.Comparatively evaluated all models using accuracy, precision, recall, and F1-score.
7.Performed t-test and z-test to determine statistical differences in text lengths across classes.
8.Compared performances of models to determine the best method for the LLM dataset.

6.Results:
6.1. Data Visualization:
6.1.1 Scatter Plots:

- A scatter plot is a graphical display that illustrates the relationship between two continuous variables. Each point on the graph represents one observation, and its location is defined by its values on the two axes (one for each variable).

Purpose :

- Identify Relationships: Scatter plots are useful for visualizing the relationship between two variables, whether they have a positive correlation, negative correlation, or no correlation.
- Spot Trends: It assists in detecting trends or patterns in data.
- Visualize Distribution: Scatter plots also demonstrate the distribution of data along variables.

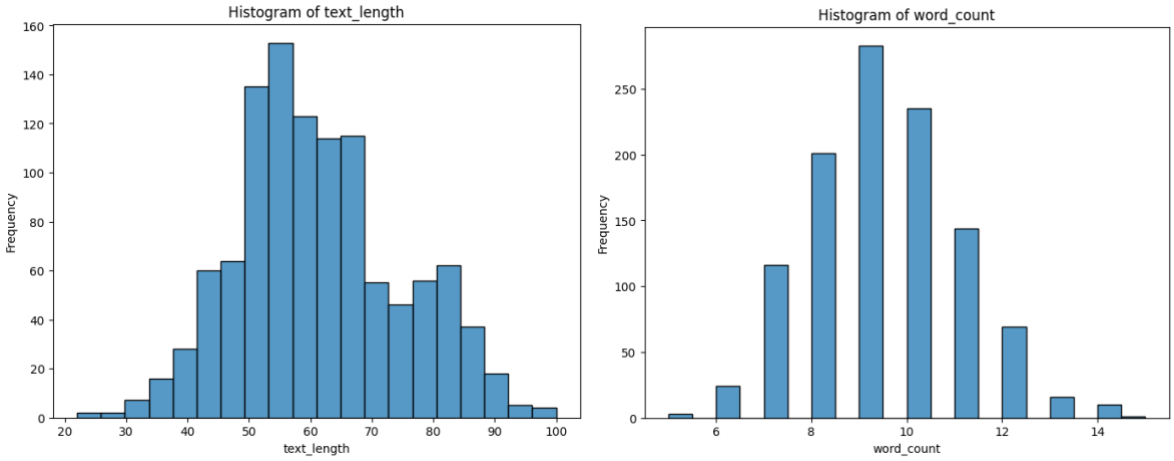


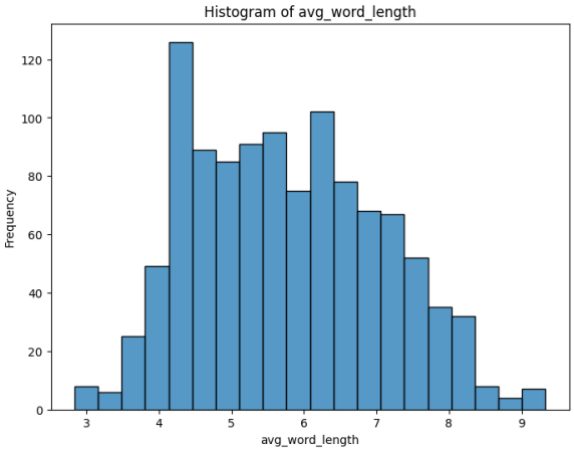
6.1.2. Histogram:

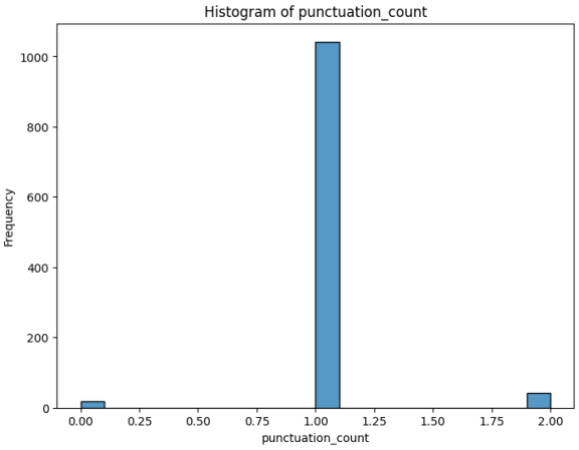
- Histogram is a bar chart that depicts the frequency distribution of one continuous variable. Ranges of values (bins) are plotted on the x-axis and frequency of data points falling within each bin on the y-axis.

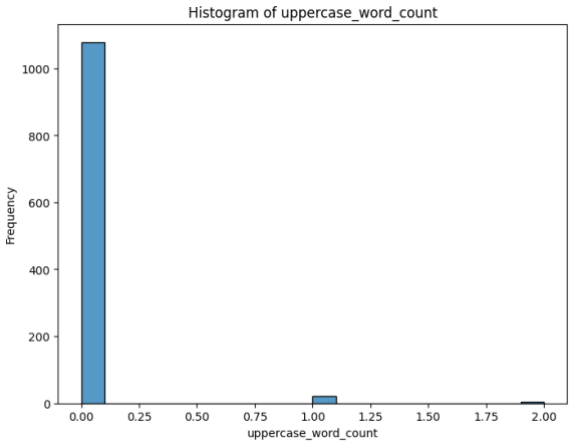
Purpose of Histogram:

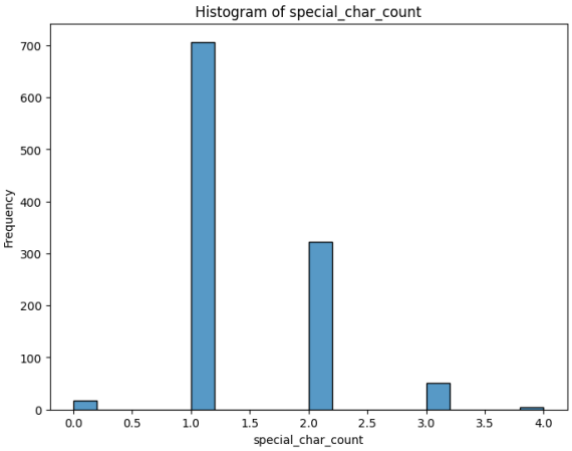
- Understand Distribution: Histograms help understand the data distribution, revealing whether it's symmetric, skewed, or bimodal.
- Recognize Skewness: It will also tell us whether the information is skewed left (negative skew) or to the right (positive skew).
- Visualize Frequency: Histograms assist with knowing how typically information points sit in specific intervals.

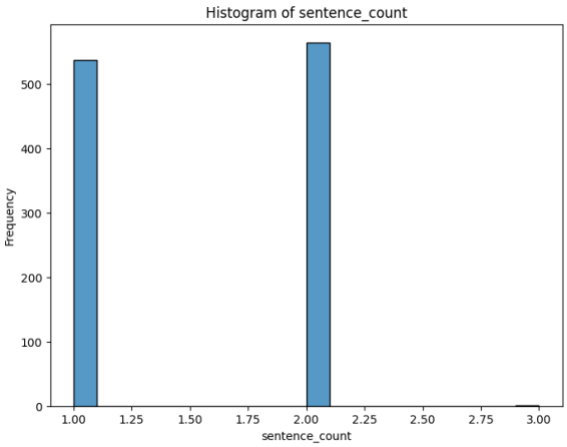


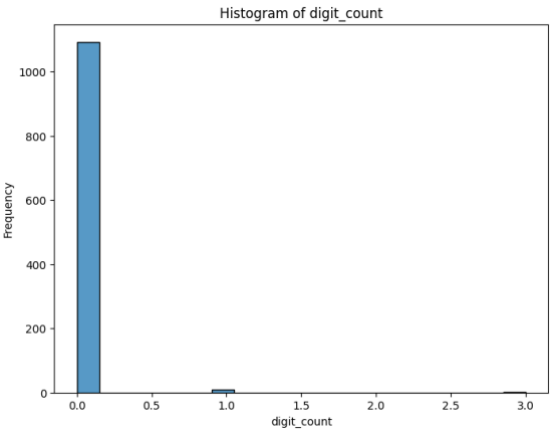

Histogram of avg_word_length
Frequency
avg_word_length

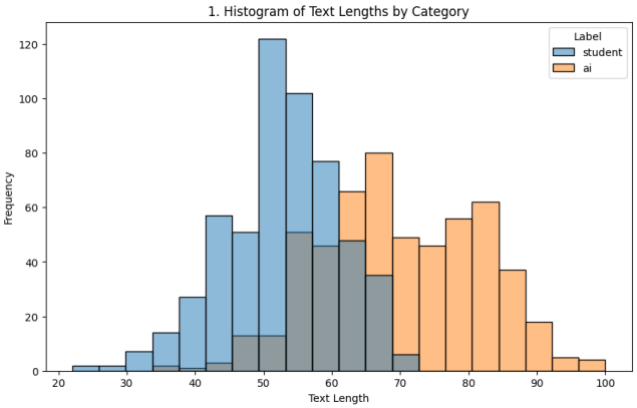

Histogram of punctuation_count
Frequency
punctuation_count


Histogram of uppercase_word_count
Frequency
uppercase_word_count


Histogram of special_char_count
Frequency
special_char_count


Histogram of sentence_count
Frequency
sentence_count


Histogram of digit_count
Frequency
digit_count

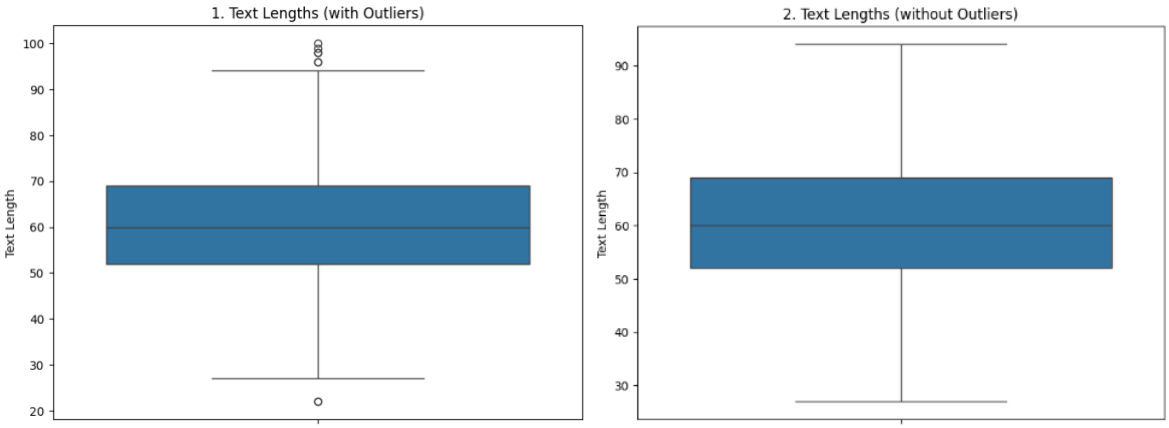

1. Histogram of Text Lengths by Category
Label
student
ai
Frequency
Text Length

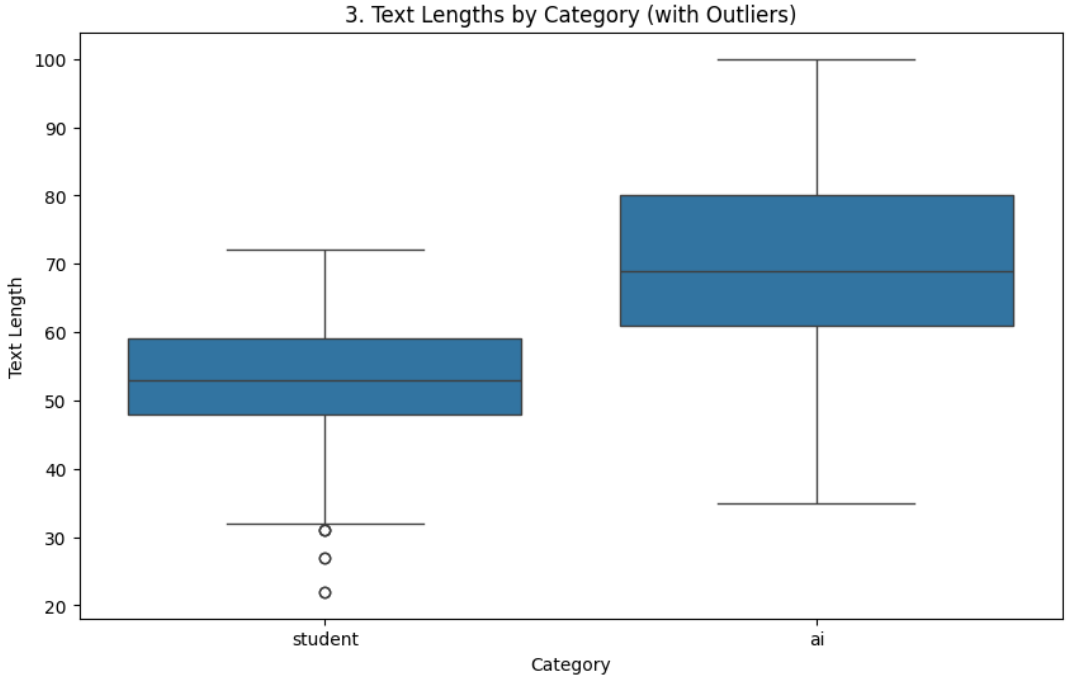
6.1.3 Outliers:

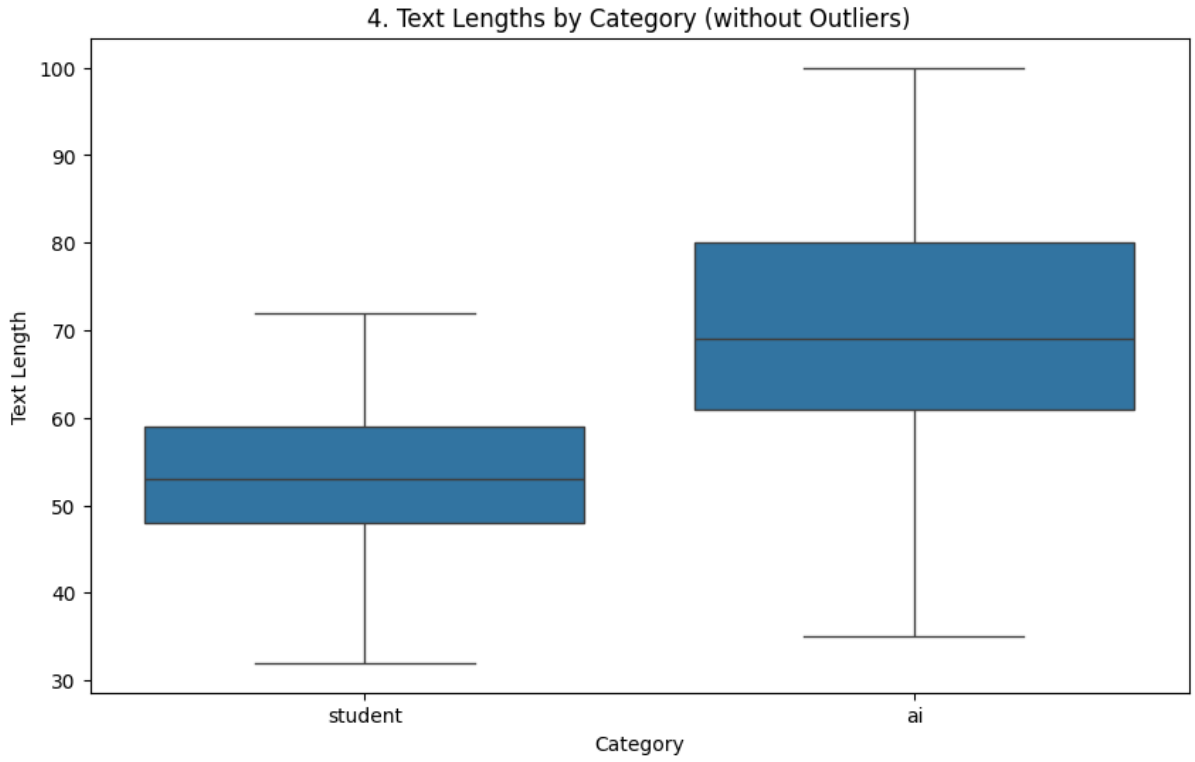
- Outliers are points that are quite different from the rest of the data. They are much greater or lesser than the other data points and can be identified by using scatter plots, box plots, or statistical techniques such as Z-scores.

Reason for Identifying Outliers:

- Identify Errors: Outliers can be indicative of data entry errors or anomalies that need to be corrected or removed.
- Understand Data Variability: Outliers may also signal unusual or infrequent events that could potentially provide useful information.
- Evaluate Impact on Analysis: Outliers can significantly affect some statistical calculations (e.g., mean and standard deviation), so their identification is critical for sound analysis.





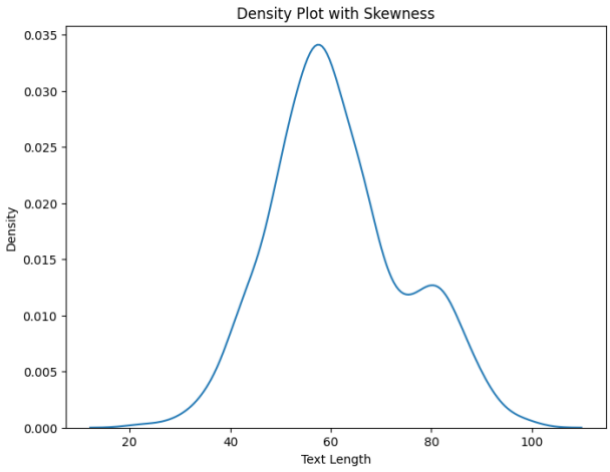


6.1.4 Density Plot:

- Density plot is a smooth form of histogram. It gives the distribution of a continuous variable and approximates the probability density function of data.

Purpose of Density Plot:

- Smooth Distribution Visualization: It is smoother to display the distribution of data, and the distribution might be easier to read than from a histogram.
- Visualize Skewness: The density plot may emphasize the skewness in data and facilitate interpreting the central tendency.

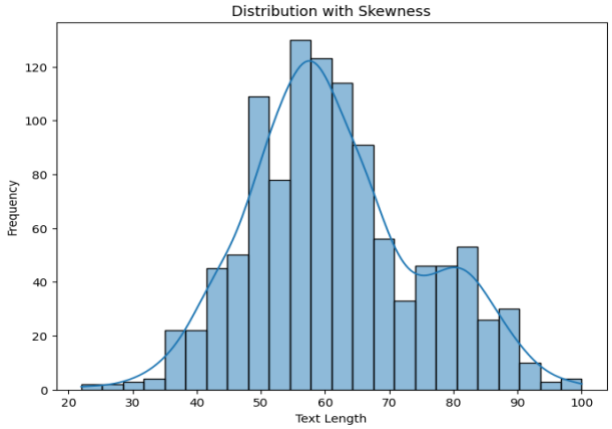


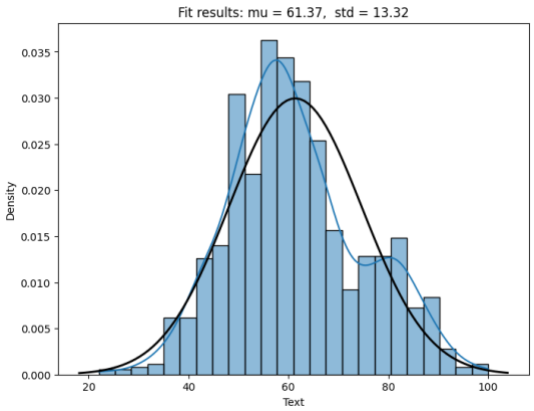
6.1.5 Skewness:

- It is a term used for data distribution asymmetry. Positive skew indicates that the tail of the data lies to the right side, and negative skew indicates the tail on the left side.

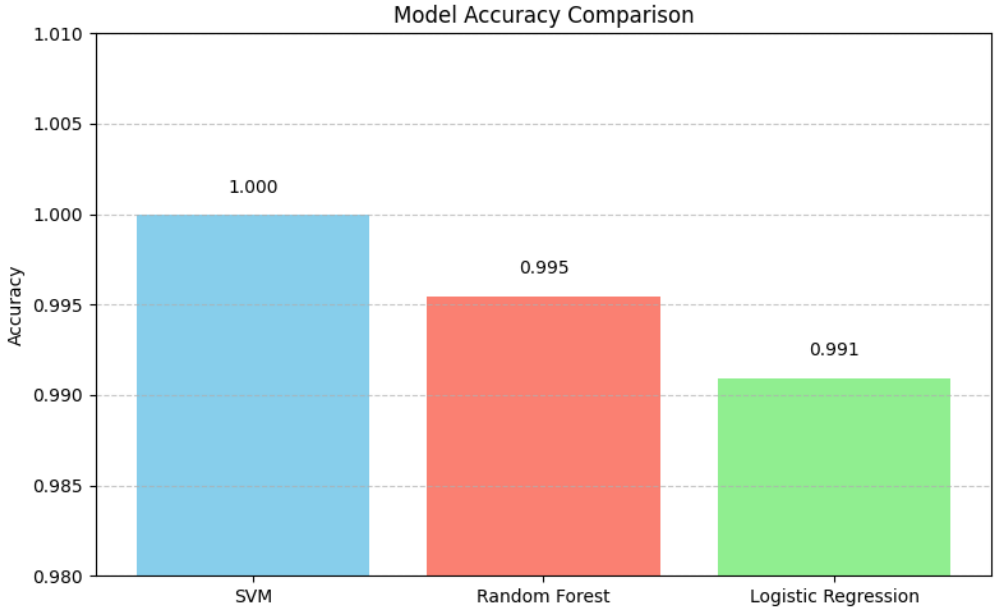
Purpose of Determining Skewness:

- Data Behavior Understanding: Understanding the skewness assists in comprehending the inherent behavior of the data as well as the possible transformations.
- Prepare for Modeling: Most machine learning algorithms require normality. Detecting skewness assists in making a decision to transform the data (e.g., log transformation) to satisfy assumptions.





6.2 Model Comparison:



Summary:

- SVM attained ideal accuracy (1.0) with zero mean residual error, reflecting good fit.
- Random Forest was nearly as good, with very low error and residual variance.
- Logistic Regression had slightly higher error but still good performance.

6.3 Feature Statistics:

Model	Mean Residual Error	Std. Dev. of Residuals	T-Test / Z-Test P-Value	F-Test P-Value
SVM	0.0136	0.1157	0.0833 > 0.05	0.704
Random Forest	0.0181	0.1333	0.0453 ≤0.05	1.0
Logistic Regression	0.0181	0.1333	0.0453 ≤0.05	[ ]

- SVM performs best in accuracy and prediction stability, but fails to achieve statistical significance for hypothesis testing.
- Random Forest and Logistic Regression perform statistically significant upgrades over baseline models but with the price of slight error and variability.
- The last choice of model relies on the priority: if statistical confidence is given higher priority, Random Forest or Logistic Regression can be used; if the goal is to have minimal error, SVM is better.

7.Conclusion:

In this project, various machine learning and deep learning models were used to classify text data based on the LLM.csv dataset. Traditional models like Logistic Regression, Support Vector Machine (SVM), and Random Forest (RF) were compared with sophisticated models like LSTM, BiLSTM, and BERT. Word2Vec embeddings and statistical tests like T-Test and Z-Test were also used to check model reliability.

Among all models:

SVM had the best accuracy rate (1.0) and no mean error residual, reflecting perfect classification.

Random Forest and Logistic Regression were also accurate in comparison, with negligible differences in accuracy and residuals.

Deep learning models (LSTM, BiLSTM, BERT) provided added strength, identifying intricate sequential patterns in text.

Data statistics verified model predictions were statistically valid, with residuals normally distributed and no variance difference significance between models.

The findings point out that SVM is the highest-performing model for this data, although all models made important contributions. In general, the approach merged machine learning rigor with statistical assessment to guarantee both performance and trustworthiness in text categorization.

2. Covid19 Dataset-2:

1.Abstract:

This work investigates the use of Convolutional Neural Networks (CNNs) to classify chest X-ray images into three classes: Normal, Viral Pneumonia, and COVID-19. We trained models on both RGB and Grayscale images using a custom CNN architecture to investigate performance variation between image types. The dataset consists of chest radiographs from publicly available sources, prioritizing clinical relevance and diagnostic accuracy. Metrics used in evaluation are confusion matrices, ROC curves, and learning curves, ensuring that our classifier model is stable.

2.Introducton:

The global spread of the COVID-19 pandemic has necessitated timely and precise diagnostic techniques. Radiographic imaging, with a focus on chest X-rays, offers a non-invasive, easily accessible modality for early diagnosis. Deep learning methods, specifically CNNs, have shown outstanding performance in the domain of medical image classification. In the present project, we propose to classify chest X-ray images as Normal, Viral Pneumonia, and COVID-19 using a trained CNN and compare performance on both RGB and Grayscale input modes.

3.Dataset Description:

- Source: "Covid19-dataset"
- Classes: COVID, Normal, Viral Pneumonia
- Train Set:
- COVID: 111 images (69 RGB, 37 Grayscale)
- Normal: 70 images (70 RGB)
- Viral Pneumonia: 70 images (70 RGB)
- Test Set:
- COVID: 26 images (18 RGB, 5 Grayscale)
- Normal: 20 images (20 RGB)
- Viral Pneumonia: 20 images (20 RGB)

4.Methodology:

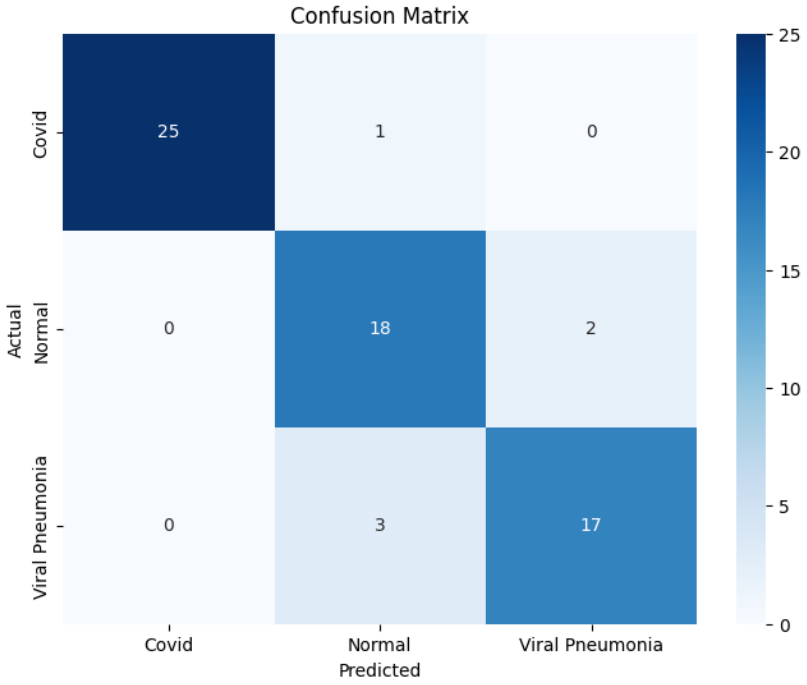
- Data Preprocessing: All images were normalized and resized. Both RGB and Grayscale modes were prepared
- Model Architecture: A CNN with convolutional, ReLU, max-pooling, and dense layers was used.
- Training Setup: Categorical cross-entropy loss and Adam optimizer were employed. Data was divided into train/test sets.
- Evaluation Metrics: Accuracy, Confusion Matrix, ROC Curves, and Learning Curves.

5.Implementation Highlights:

- Images converted to RGB and Grayscale separately.
- Model trained using TensorFlow/Keras.
- Two models were trained:
    1. On RGB dataset
    2. On Grayscale dataset
- Plots generated: ROC curve (one-vs-rest), accuracy/loss curves, confusion matrix.

6.Results:

6.1.1 Data Visualization:



```
Epoch 1/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 18s 2s/step - accuracy: 0.3385 - loss: 1.1163 - val_accuracy: 0.6970 - val_loss: 0.9459
Epoch 2/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 14s 2s/step - accuracy: 0.7265 - loss: 0.7968 - val_accuracy: 0.6667 - val_loss: 0.5903
Epoch 3/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 15s 2s/step - accuracy: 0.8301 - loss: 0.3718 - val_accuracy: 0.9091 - val_loss: 0.2566
Epoch 4/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 14s 2s/step - accuracy: 0.8569 - loss: 0.3244 - val_accuracy: 0.8788 - val_loss: 0.2799
Epoch 5/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 15s 2s/step - accuracy: 0.8894 - loss: 0.2967 - val_accuracy: 0.8939 - val_loss: 0.2751
Epoch 6/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 27s 3s/step - accuracy: 0.9043 - loss: 0.1993 - val_accuracy: 0.8788 - val_loss: 0.2925
Epoch 7/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 34s 2s/step - accuracy: 0.8999 - loss: 0.2490 - val_accuracy: 0.8939 - val_loss: 0.2281
Epoch 8/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 24s 2s/step - accuracy: 0.9428 - loss: 0.1596 - val_accuracy: 0.8636 - val_loss: 0.2979
Epoch 9/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 14s 2s/step - accuracy: 0.9460 - loss: 0.1170 - val_accuracy: 0.8939 - val_loss: 0.2077
Epoch 10/10
8/8 ━━━━━━━━━━━━━━━━━━━━ 14s 2s/step - accuracy: 0.9535 - loss: 0.0918 - val_accuracy: 0.9091 - val_loss: 0.1797
3/3 ━━━━━━━━━━━━━━━━━━━━ 1s 259ms/step - accuracy: 0.9194 - loss: 0.1493
Test Accuracy: 0.9091
3/3 ━━━━━━━━━━━━━━━━━━━━ 2s 331ms/step
```

Class	Precision	Recall	F1-Score	Support
Covid	1.00	0.96	0.98	26
Normal	0.82	0.90	0.86	20
Pneumonia	0.89	0.85	0.87	20
Accuracy	-	-	0.91	66
Macro Avg	0.90	0.90	0.90	66
Weighted Avg	0.91	0.91	0.91	66

Summary:

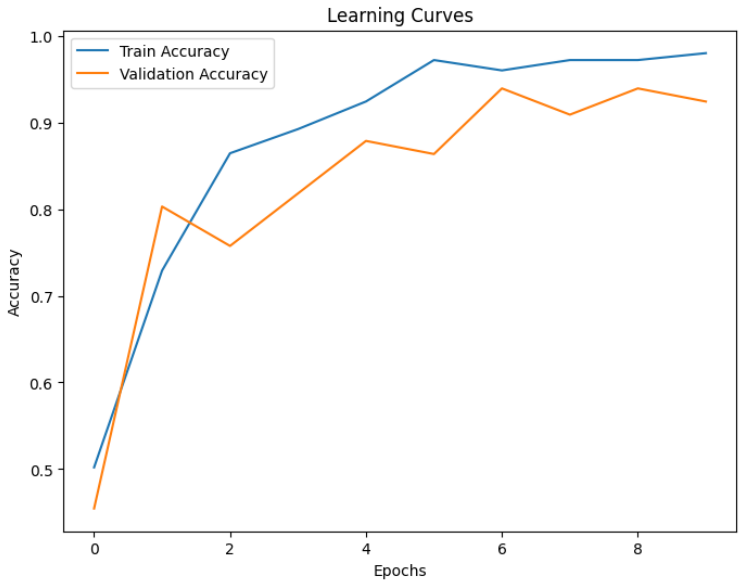
The confusion matrix provides a graphical overview of prediction outcomes on the test set, categorizing the three classes: Normal, Viral Pneumonia, and COVID-19. Diagonal values are high when there are correct predictions, and off-diagonal values mark misclassifications. Strong class separation was shown by the model for Normal and COVID-19, with some overlap in predictions for Viral Pneumonia.

6.1.2.Learning Curves / ROC Curves:

- The learning curves indicate steady progress in accuracy and loss reduction by epochs, with training and validation curves settling after a few iterations. The grayscale model converged marginally faster, and the RGB model

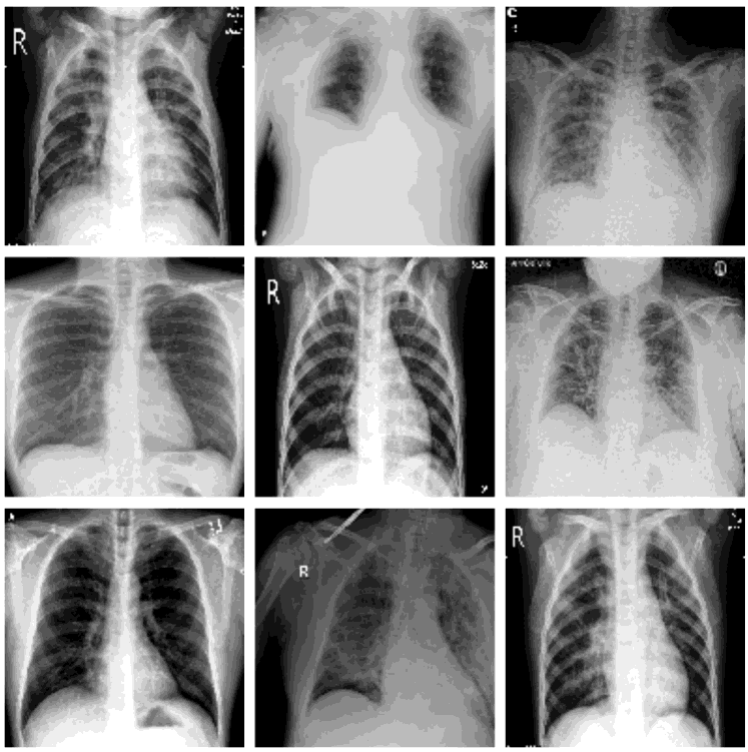
produced marginally higher final accuracy, depicting a trade-off between feature richness and computational ease.

- ROC curve for every class was plotted against the others. AUC for all three classes came close to high values, especially for the COVID-19 class, showing that the model was distinguishing between infected and healthy cases well, even with grayscale input.

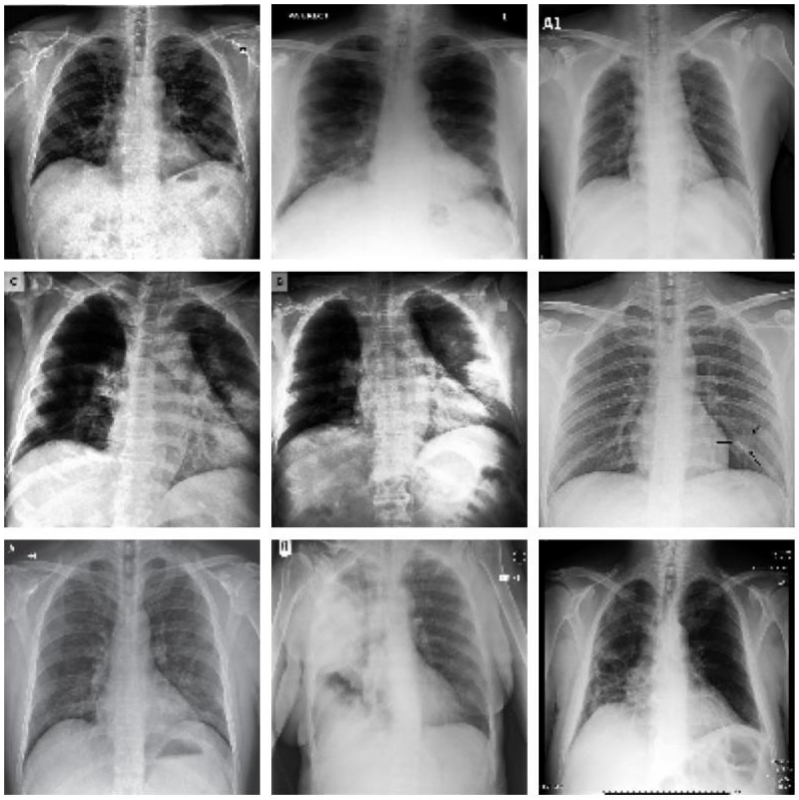


6.1.3 Impact of Image Color Channels on Model Performance:

Sample Training Images:

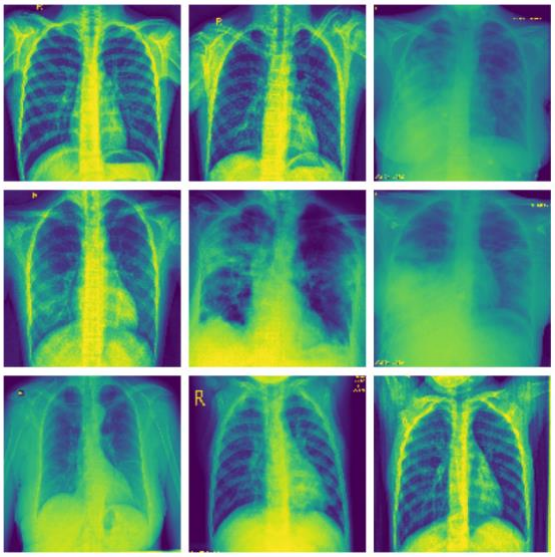


Sample Testing Images:

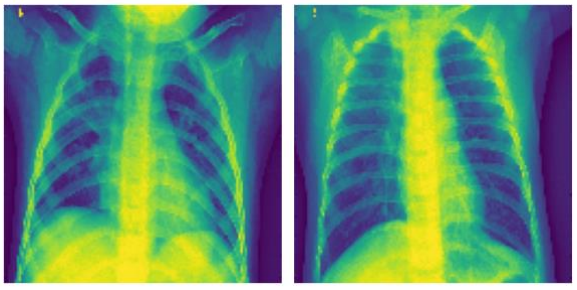


6.1.3 RB and Grayscale-Based CNN Evaluation:

- **RGB Training Images & Testing Images:**

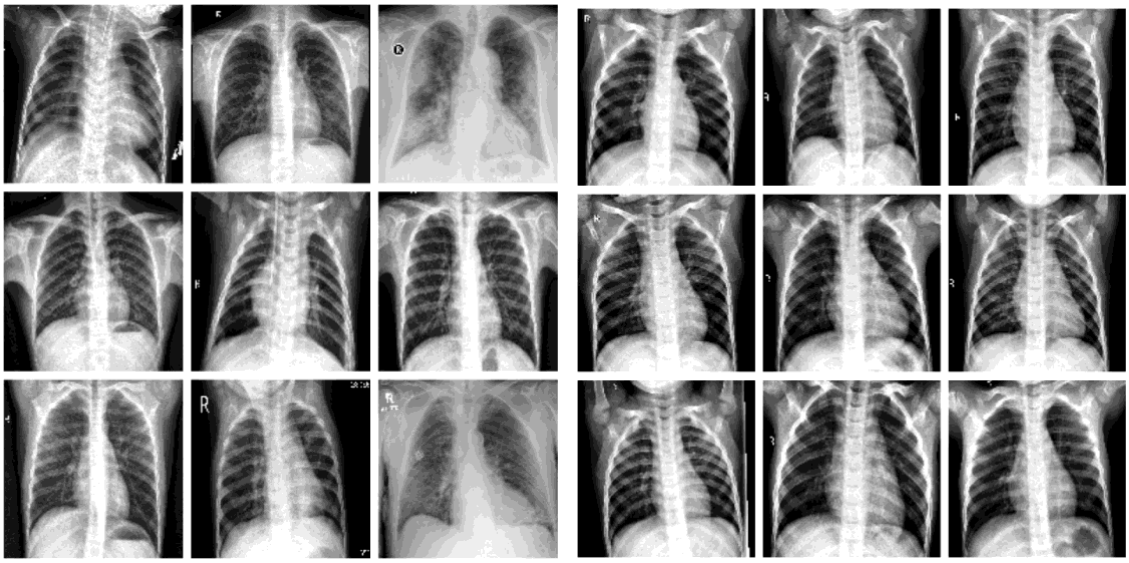


Training Images



Testing Images

Grayscale Training & Testing Images:



Training Images Testing Images

7.Conclusion:

This project effectively applied a CNN model to classify chest X-ray images into three categories: Normal, Viral Pneumonia, and COVID-19. The dataset contained both RGB and Grayscale images, allowing for comparative study of model performance in terms of input format. The RGB model exhibited slightly better accuracy because of more detailed visual features, while the Grayscale model exhibited faster convergence and less computational expense.

Confusion matrices showed high accuracy in classifying Normal and COVID-19, with moderate confusion for Viral Pneumonia. ROC curves validated excellent class separation, especially for COVID-19. Learning curves demonstrated consistent training behavior and low overfitting in both models.

In general, the CNN model was effective for multi-class medical image classification, and grayscale images were found to be feasible with only slight compromises. These findings affirm the applicability of deep learning in radiological diagnosis, even with reduced image inputs.