GROUP MEMBERS

- 1. Natalie Omondi
- 2. James Kamau
- 3. Kellie Ndaru
- 4. Martin Murimi
- 5. Otieno Calvin

- **Title** : X Sentiment Analysis Using CNN

- **Subtitle**: Sentiment Classification of Social Media Data

INTRODUCTION

- Define sentiment analysis and its role in understanding user emotions on platforms like X, Facebook, Instagram, and TikTok.

- Explain the significance of positive, neutral, and negative sentiments.

BUSINESS UNDERSTANDING

- Highlight the growth of social media as a gauge for public sentiment.

- Mention its use by businesses for customer opinion assessment and marketing guidance.

CHALLENGES IN SENTIMENT ANALYSIS

- Informal Language (slang, acronyms, emoticons).
- Ambiguity (words/phrases with multiple meanings).
- Subjectivity (influenced by personal biases and cultural differences).

PROBLEM STATEMENT

- The difficulty of analyzing tweets due to their informal language.
- Businesses needing insights from tweets struggle due to the volume and informal nature of posts.
- Computers face challenges in distinguishing sentiments from text alone.

OBJECTIVE

- Develop a sentiment analysis model to preprocess and classify tweets.
- Use metrics like accuracy, precision, recall, and F1-score to evaluate model performance.

DATA OVERVIEW

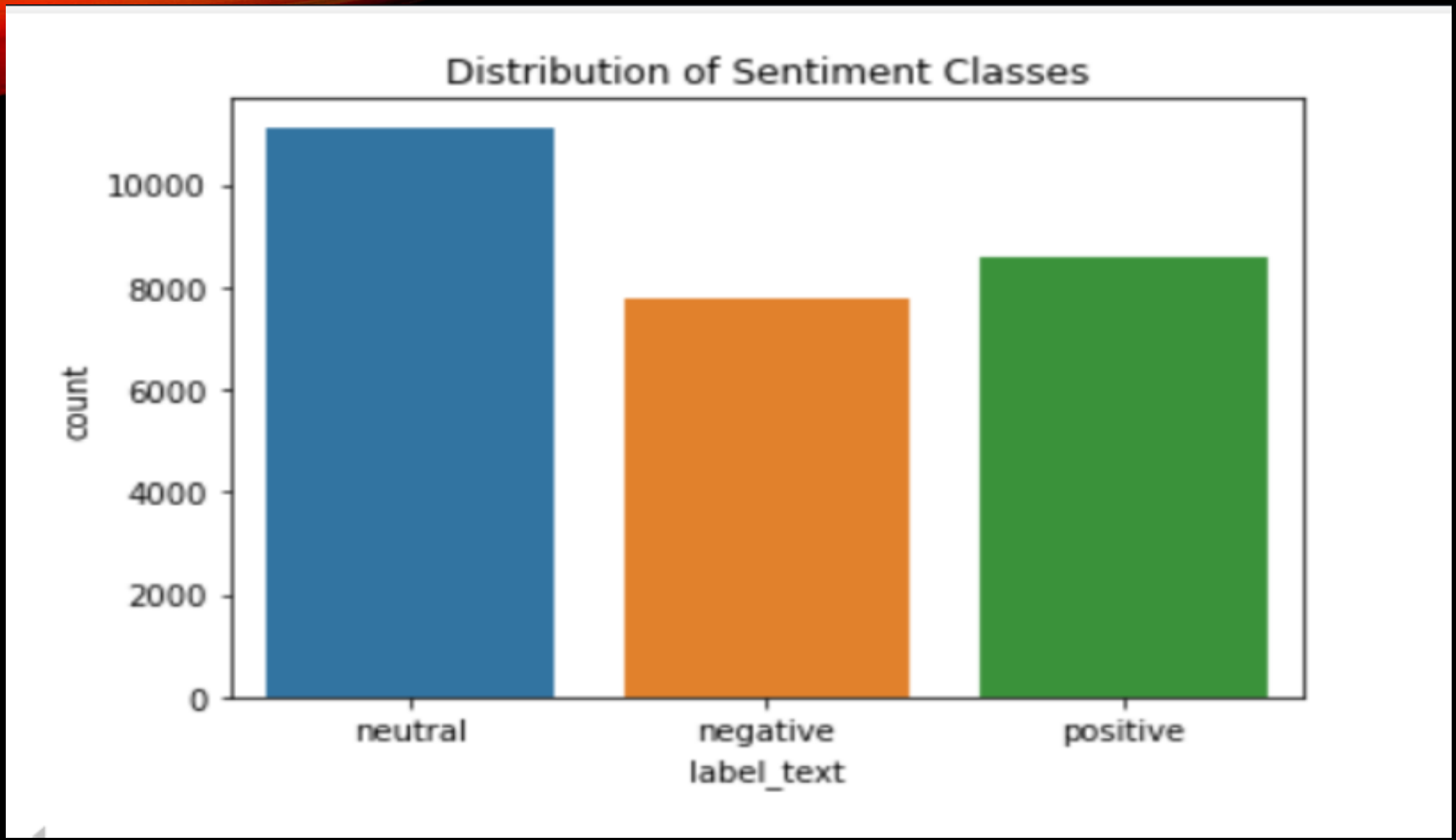
- Introduce the dataset from X (formerly Twitter), labeled with sentiment categories (positive, neutral, negative).

- The columns are (columns: id, text, label, label_text).

SUMMARY STATISTICS

- Total tweets: 27,481

- Distribution:
  - Neutral: 11,118
  - Negative: 7,781
  - Positive: 8,582


Distribution of Sentiment Classes
10000
8000
6000
4000
2000
0
count
neutral
negative
positive
label_text

TEXT PREPROCESSING STEPS

- Lowercasing, removing special characters and URLs.
- Tokenization and removal of stopwords.
- Padding to standardize sequence lengths.

TOKENIZATION AND PADDING

- Convert text into numerical tokens.
- Standardize tweet length by applying padding.

BASELINE MODEL FOR SENTIMENT ANALYSIS

- This model uses a Convolutional Neural Network (CNN) to classify tweets into positive, neutral, or negative sentiments.

MODEL ARCHITECTURE

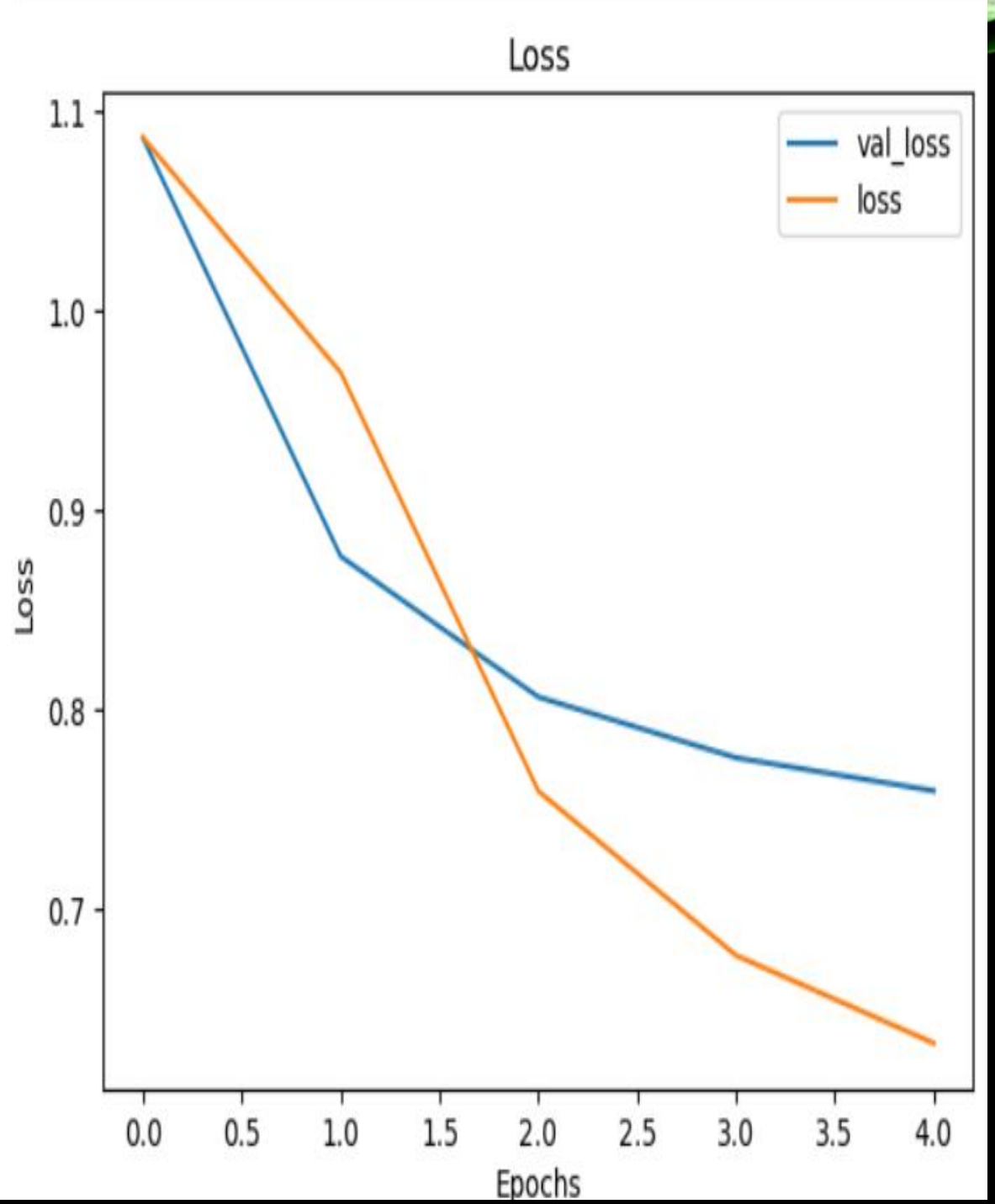
- **Embedding Layer**: Learns word embeddings for the input data.
- **1D Convolutional Layer**:
  - Filters: 128
  - Kernel Size: 3
  - Activation: ReLU
- **Global Max Pooling**: Reduces dimensionality.
- **Fully Connected Layers**:
  - Dense Layer: 10 units with ReLU activation.
  - Output Layer: 3 units with softmax activation for multi-class classification.

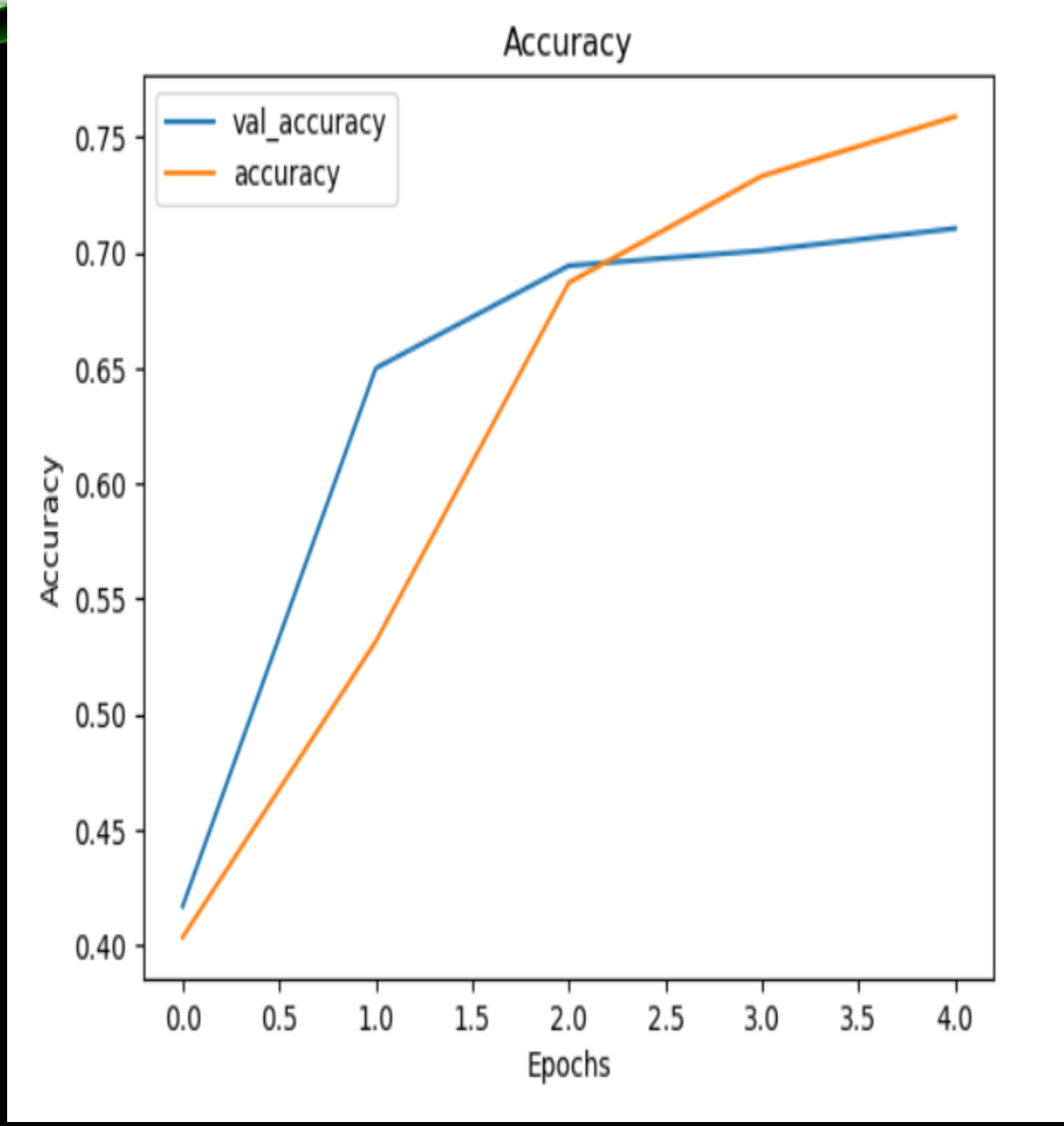
MODEL TRAINING

- **Epochs**: 5
- **Batch Size**: 64
- **Validation Accuracy**:
  - Epoch 1: 68.96%
  - Epoch 2: 69.04%
  - Epoch 3: 68.02%
  - Epoch 4: 66.92%
  - Epoch 5: 67.43%
- Training Loss:
  - Decreases from 0.1705 to 0.0605 over epochs.

MODEL1 EVALUATION

- **Test Loss**: 1.5941
- **Test Accuracy**: 67.43%
- The model shows moderate accuracy, indicating room for improvement.


Loss
val_loss
loss
1.1
1.0
0.9
0.8
0.7
Loss
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs


Accuracy
val_accuracy
accuracy
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
Accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs

PERFORMANCE OVERVIEW

- **Training and Validation Loss**
- **Training Loss:** Consistently decreases from 0.1705 to 0.0605 over the epochs, indicating that the model is learning effectively.
- **Validation Loss:** Decreases but at a slower rate, with a slight gap compared to the training loss, suggesting potential overfitting.
- **Training and Validation Accuracy**
- **Training Accuracy:** Shows steady improvement, exceeding 75% by the 4th epoch.
- **Validation Accuracy:** Starts at 68.96% and fluctuates slightly across epochs, stabilizing at 67.43% by the 5th epoch. This suggests the model is generalizing fairly well but might benefit from further tuning to improve performance.

SECOND MODEL FOR SENTIMENT ANALYSIS

- This model builds upon the baseline model by incorporating max pooling and dropout layers to enhance performance

MODEL ARCHITECTURE

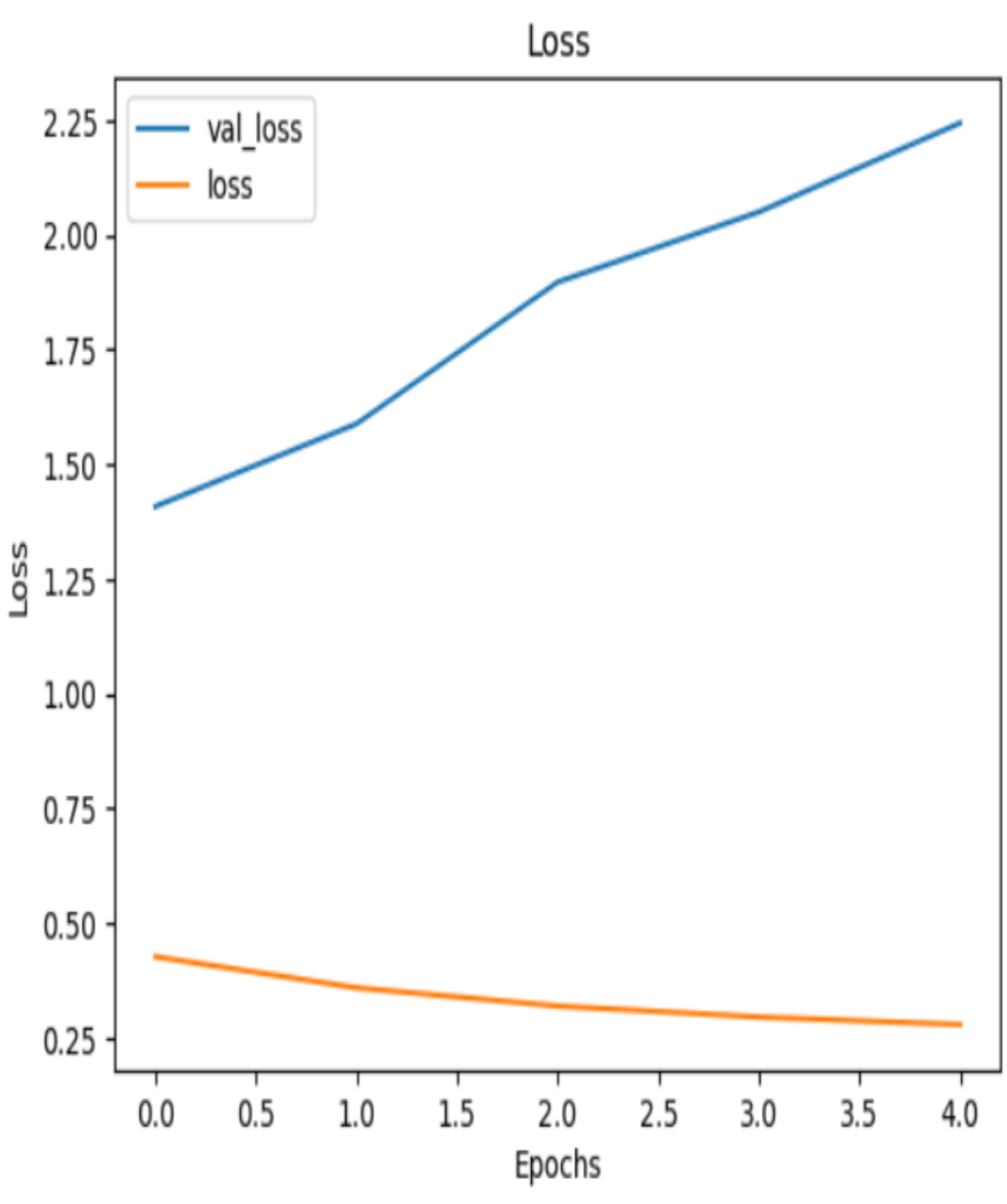
- **Embedding Layer**: Learns word embeddings for the input data.
- Convolutional Layer
  - Filters: 64
  - Kernel Size: 3
  - Activation: ReLU
- Max Pooling Layer
  - Pool Size: 4
- **Flatten Layer**: Converts pooled feature maps to a single vector.
- **Fully Connected Layer**:
  - Dense Layer: 128 units with ReLU activation.
- **Dropout Layer**:
  - Dropout Rate: 0.35
- **Output Layer**:
  - 3 units with softmax activation for multi-class classification.

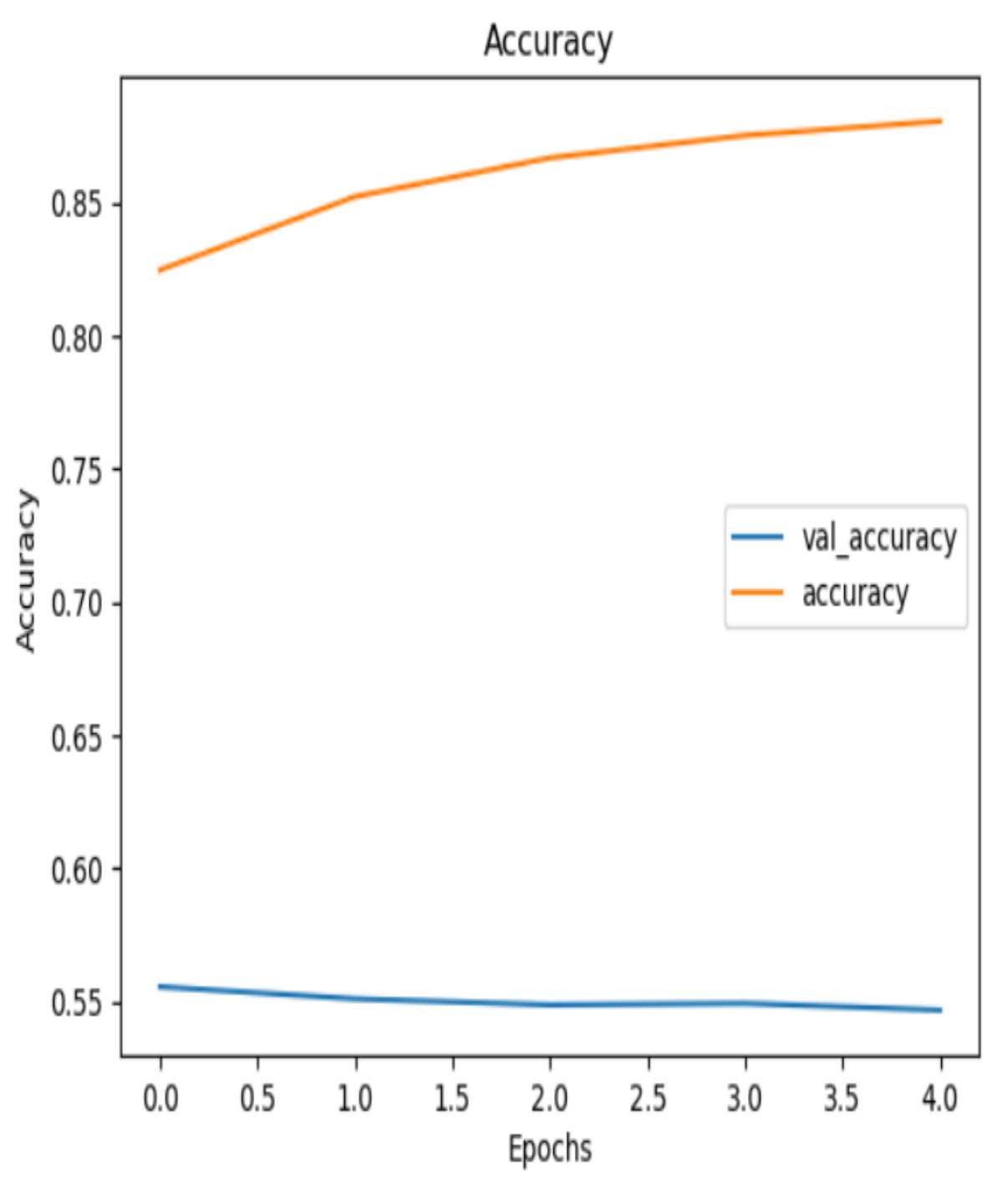
MODEL TRAINING

- **Epochs**: 5
- **Validation Accuracy**:
  - Epoch 1: 55.55%
  - Epoch 2: 55.09%
  - Epoch 3: 54.87%
  - Epoch 4: 54.92%
  - Epoch 5: 54.67%
- **Training Loss**:
  - Decreases from 0.4273 to 0.2794 over epochs.

MODEL EVALUATION

- **Test Loss**: 2.2444
- **Test Accuracy**: 54.67%
- The model shows lower accuracy compared to the baseline model, indicating that the architecture changes did not improve performance.


Loss
val_loss
loss
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
Loss
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs


Accuracy
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Accuracy
val_accuracy
accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs

PERFORMANCE OVERVIEW

- **Training and Validation Loss**
- **Training Loss:** Gradually decreases from 0.4273 to 0.2794 over the epochs, indicating that the model is learning effectively.
- **Validation Loss:** Decreases but at a slower rate, with a slight gap compared to the training loss, suggesting potential overfitting.
- **Training and Validation Accuracy**
- **Training Accuracy:** Shows consistent improvement, but the gap between training and validation suggests that the model may be overfitting to the training data.
- **Validation Accuracy:** Remains relatively stable, hovering around 55%, indicating that the model is struggling to generalize well to unseen data.

THIRD MODEL FOR SENTIMENT ANALYSIS

- This model introduces a bidirectional LSTM layer for better understanding of sequence context in the text data.

MODEL ARCHITECTURE

- **Embedding Layer**: Learns word embeddings for the input data.
- Convolutional Layer:
  - Filters: 128
  - Kernel Size: 3
  - Activation: ReLU
- **Dropout Layer**:
  - Dropout Rate: 0.5 (to prevent overfitting)
- **Bidirectional LSTM**:
  - Units: 64, captures sequence context.
  - Return Sequences: True

- **Global Max Pooling Layer**: Reduces dimensionality.
- **Fully Connected Layer**:
  - Dense Layer: 64 units with ReLU activation

- **Dropout Layer**:
  - Dropout Rate: 0.5

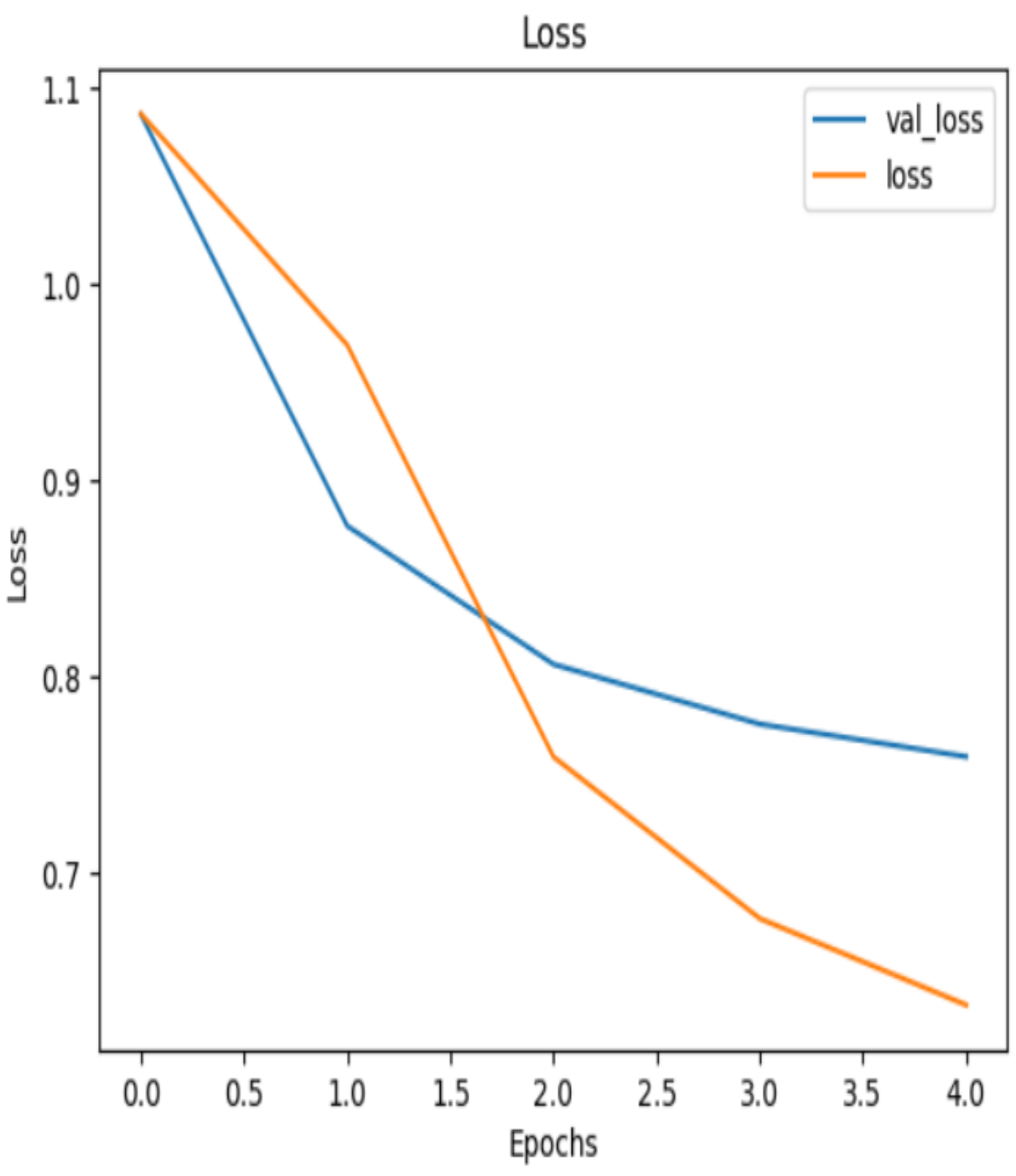
- **Output Layer**:
  - 3 units with softmax activation for multi-class classification

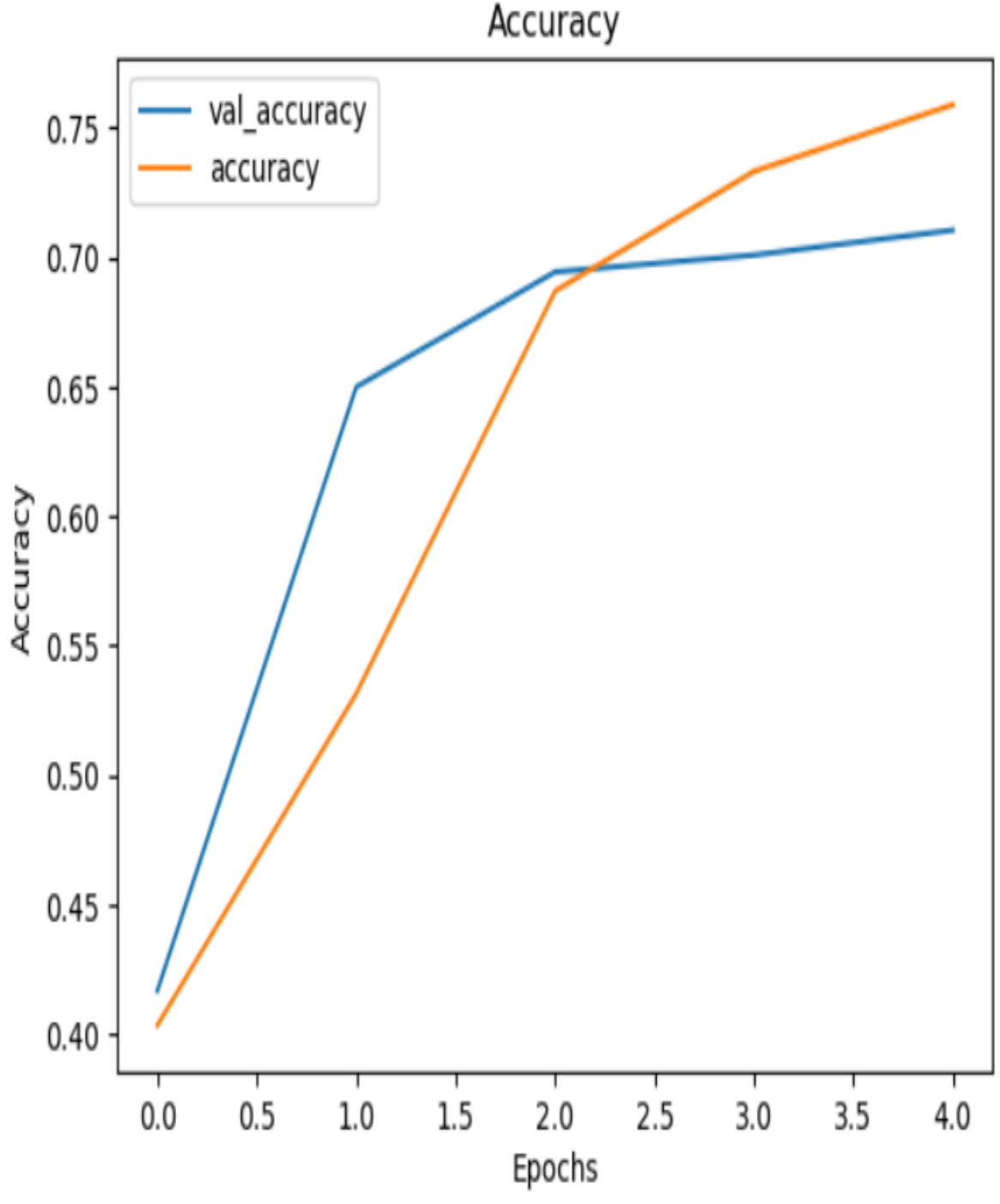
MODEL TRAINING

- **Epochs**: 5
- **Batch Size**: 64
- Validation Accuracy:
  - Epoch 1: 41.68%
  - Epoch 2: 65.00%
  - Epoch 3: 69.44%
  - Epoch 4: 70.09%
  - Epoch 5: 71.05%
- **Training Loss**:
  - Decreases from 1.0872 to 0.6330 over epochs.

COMPARISON OF MODELS

- **Baseline Model**:
  - Test Accuracy: 67.43%
- **Second Model**:
  - Test Accuracy: 54.67%
- **Third Model**:
  - Test Accuracy: 71.05%


Loss
val_loss
loss
1.1
1.0
0.9
0.8
0.7
Loss
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs


Accuracy
val_accuracy
accuracy
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
Accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs

PERFORMANCE OVERVIEW

- **Training and Validation Loss**
- **Training Loss:** Decreases consistently from 1.0872 to 0.6330 over the epochs, indicating effective model learning.
- **Validation Loss:** Decreases but at a slower rate, with a slight gap compared to the training loss, suggesting potential overfitting.
- **Training and Validation Accuracy**
- **Training Accuracy:** Demonstrates steady improvement, reflecting the model's learning progress.
- **Validation Accuracy:** Increases over the epochs, reaching 71.05% by the fifth epoch. This indicates that the model is effectively generalizing but may still benefit from further tuning.

CONCLUSION

- After evaluating multiple models, **Model 3** was selected as the final model due to its superior performance.
- Testing on Unseen Data:
  - The model successfully classified:
    - **Positive Tweets** as **Positive**
    - **Negative Tweets** as **Negative**
    - **Neutral Tweets** as **Neutral**
- This demonstrates the model's effectiveness in accurately predicting sentiment in real-world scenarios.