Sentiment Analysis Model for Tweets Dataset

Aya Ragaa	7793
Nabeel Mohemed	7547

Dataset to be Used

The analysis was performed on the **Sentiment140 dataset**, which consists of 1.6 million tweets balanced between positive and negative sentiments. The dataset was split into training (70%), validation (15%), and test (15%) sets.

Data Preprocessing

Common preprocessing steps were applied across all models:

- Lowercasing
- Removal of usernames (@mentions)
- Removal of URLs and web links
- Removal of punctuation
- Removal of extra spaces
- Removal of stopwords

Model Implementations and Results

1. **Gaussian Naive Bayes**
    - **Implementation Details:**
        - Used TF-IDF vectorization with 1000 features
        - Simple and computationally efficient approach

- **Performance Metrics:**
  - Test Accuracy: 70.29%
  - Precision: 0.71 (weighted avg)
  - Recall: 0.70 (weighted avg)
  - F1-Score: 0.70 (weighted avg)

2. LSTM Model

- **Architecture:**
  - Embedding layer (5000 vocabulary size, 128 dimensions)
  - Two LSTM layers (128 and 64 units)
  - Layer normalization for stability
  - Dropout layers (0.4 and 0.2) for regularization
  - Binary classification output

- **Training Details:**
  - Early stopping with patience of 3 epochs
  - Model checkpoint to save best weights
  - Batch size of 64
  - Adam optimizer

- **Performance Metrics:**
  - Test Accuracy: 79.13%

- Precision: 0.79 (weighted avg)
- Recall: 0.79 (weighted avg)
- F1-Score: 0.79 (weighted avg)

3. BERT Model

- **Implementation Details:**
  - Pre-trained BERT base uncased model
  - Maximum sequence length of 15 tokens
  - Batch size of 128
  - Learning rate of 1e-5 with linear scheduler
  - Trained for 2 epochs
- **Performance Metrics:**
  - Test Accuracy: 80.78%
  - Precision: 0.81 (weighted avg)
  - Recall: 0.81 (weighted avg)
  - F1-Score: 0.81 (weighted avg)

Comparative Analysis

1. **Accuracy Comparison:**
    - **Naive Bayes**: 70.29%
    - **LSTM:** 79.13%
    - **BERT**: 80.78%

2. **Model Characteristics:**
    - **Naive Bayes**: Fastest to train and implement, but lowest performance
    - **LSTM**: Good balance between performance and computational requirements
    - **BERT**: Best performance but highest computational cost

3. **Class Balance Performance:**
    - **Naive Bayes** shows slight bias toward positive class (higher recall for positive)
    - **LSTM** shows balanced performance across both classes
    - **BERT** shows slightly better performance on negative class recognition

---

Key Findings

1. **Performance Hierarchy:**
    - BERT > LSTM > Naive Bayes in terms of accuracy and overall metrics

- Each step up in model complexity yielded approximately 9% improvement from Naive Bayes to LSTM, and 1.65% from LSTM to BERT

2. **Trade-offs:**

   - While BERT achieved the highest accuracy, the marginal improvement over LSTM (1.65%) may not justify the additional computational cost for some applications

   - LSTM provides a good balance between performance and resource requirements

   - Naive Bayes, despite lower accuracy, might be suitable for real-time applications where speed is crucial

3. **Model Selection Considerations:**

   - Resource-constrained environments: **Naive Bayes**

   - Balanced performance/resource needs: **LSTM**

   - Highest accuracy requirement: **BERT**