NLP Emoji Prediction Task: Model Comparison Report

1. Data Loading and Preprocessing

- **Training Data:** Loaded training data from `train_emoji.csv`.
- **Test Data:** Loaded test data from `test_emoji.csv`.
- **Data Splitting:** Split the training data into `X_train` and `Y_train`, and the test data into `X_test` and `Y_test`.
- **Text Preprocessing:**
  - Converted text to lowercase.
  - Tokenized sentences into words.
- **One-Hot Encoding:** Converted labels (`Y_train` and `Y_test`) into one-hot vectors (`Y_train_oh` and `Y_test_oh`).

---

2. Word Embeddings and Preprocessing

- **GloVe Embeddings:** Loaded pre-trained GloVe word embeddings (`glove.6B.50d.txt`).
- **Word Indices:** Created word indices and mapped them to their respective GloVe vectors.
- **Sentence to Indices Conversion:** Implemented a function (`sentences_to_indices`) to convert sentences to indices based on the word embeddings.

---

3. Model Architecture

Model 1 (`Emojify_model`)

- **Embedding Layer:**
  - Used pre-trained GloVe embeddings.
- **LSTM Layers:**
  - First LSTM Layer:
    - Units: 64
    - Return Sequences: True
    - Dropout: 0.5
    - Regularization: L2 (0.05)
  - Second LSTM Layer:
    - Units: 64
    - Return Sequences: False
    - Dropout: 0.5
    - Regularization: None
- **Dense Layer:**

  - Units: 5 (Number of classes)
  - Activation: Softmax

Model 2 (`Emojify_modelv2`)

- **Embedding Layer:**
  - Used pre-trained GloVe embeddings.
- **Bidirectional LSTM Layers:**
  - First Bidirectional LSTM Layer:
    - Units: 128
    - Return Sequences: True
    - Dropout: 0.5
    - Regularization: L2 (0.01)
  - Second Bidirectional LSTM Layer:
    - Units: 128
    - Return Sequences: False
    - Dropout: 0.5
    - Regularization: None

---

4. Model Training

- **Text to Indices Conversion:** Applied the `sentences_to_indices` function to convert sentences to indices based on the word embeddings.
- **Model Training Data:** Trained both models on `X_train_indices` with corresponding labels (`Y_train`).
- **Validation Data:** Utilized early stopping with a validation set (`X_val` and `y_val`) to prevent overfitting.
- **Training Parameters:** 100 epochs, batch size of 32, Adam optimizer, and categorical crossentropy loss.

---

5. Model Evaluation

Model 1 Evaluation

- **Test Data:** Evaluated Model 1 on `X_test_indices` and `Y_test_oh`.
- **Metrics:** Obtained accuracy, precision, recall, and F1-score.

Model 2 Evaluation

- **Test Data:** Evaluated Model 2 on `X_test_indices` and `Y_test_oh`.
- **Metrics:** Obtained accuracy, precision, recall, and F1-score.

---

6. Model Comparison and Visualization

- **Comparison Metrics:** Accuracy, Precision, Recall, F1 Score.
- **Comparison Plot:** Bar chart comparing Model 1 and Model 2 across metrics.

---

7. Results

Model 1

- **Train Accuracy:** 0.9805
- **Validation Accuracy:** 0.9578
- **Test Accuracy:** 0.7809523940086365
- **Precision:** 0.7756
- **Recall:** 0.7238
- **F1 Score:** 0.7204

Model 2

- **Train Accuracy:** 0.9773
- **Validation Accuracy:** 0.9513
- **Test Accuracy:** 0.7714285850524902
- **Precision:** 0.7847
- **Recall:** 0.7524
- **F1 Score** 0.7507

Model Comparison

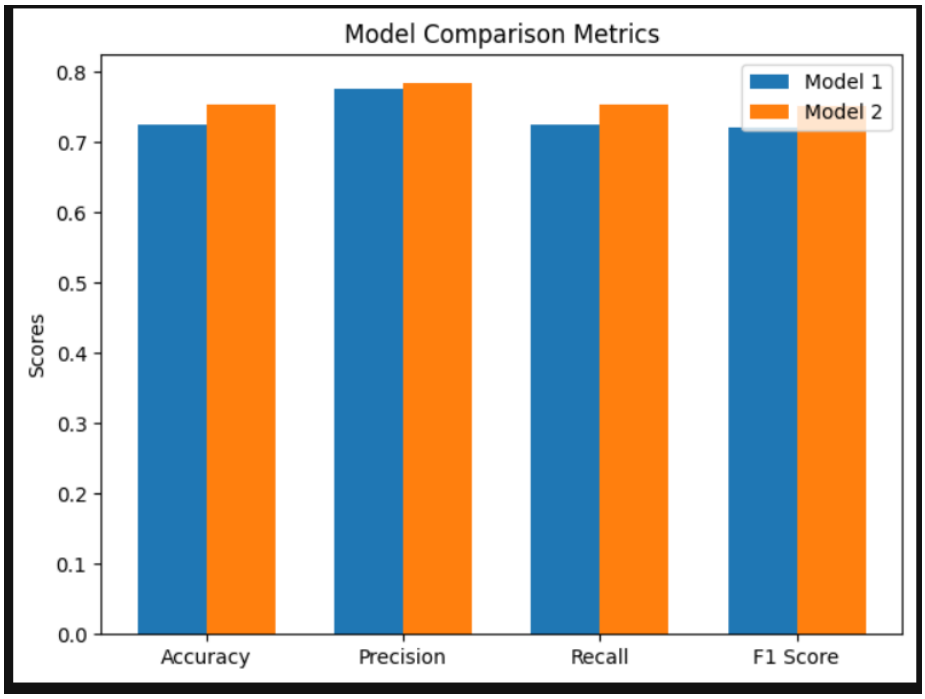

Figure 1: Model Comparison