INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY

HYDERABAD

# PROJECT REPORT
# OF
# NATURAL LANGUAGE INFERENCE

**Team Name:** Team TwoSix

**Team Members:** Harshita Harshita (2023201002)
Shreyansh Shrivastava (2023201006)

**Course:** Introduction To NLP

**Session:** Spring 2024

# Natural Language Inference: Deep Dive into Textual Entailment Recognition

## Abstract:

This report outlines our project on Natural Language Inference (NLI), focusing on the crucial task of textual entailment recognition. NLI plays a pivotal role in understanding and processing human language, enabling machines to determine if the meaning of one text fragment can be inferred from another. Our project aims to explore innovative approaches to improve the accuracy and efficiency of textual entailment recognition, leveraging state-of-the-art machine learning techniques and comprehensive datasets such as MultiNLI, SICK, and SNLI.

## 1. Introduction:

Natural Language Inference (NLI) represents a fundamental challenge within the field of computational linguistics, focusing on determining the relationship between a pair of text fragments. These relationships are categorized into three distinct classes: entailment, contradiction, and neutral. The ability to accurately identify these relationships is crucial for advancing natural language understanding and has significant implications for various applications, including automated summarization, machine translation, and question-answering systems. Our project seeks to address the complexities involved in textual entailment recognition, aiming to enhance model performance and understanding in this domain.

## 2. Problem Statement:

The challenge of textual entailment recognition within NLI involves classifying the semantic relationship between two text fragments (a premise and a hypothesis) into entailment, contradiction, or neutral. This task underscores the variability and complexity of natural language, requiring models to possess deep semantic understanding and reasoning capabilities. Despite advancements in machine Learning and NLP, achieving high accuracy in textual entailment recognition remains a significant challenge, partly due to the nuanced nature of language and the diversity of linguistic expressions. Our project is dedicated to exploring and developing methodologies that can overcome these challenges, thereby advancing the field of NLI.

## 3. Datasets Overview

Our research will utilize three prominent datasets renowned for their role in NLI research:

**MultiNLI**

Description**:** A comprehensive dataset featuring a wide range of spoken and written text genres, enabling the development of models that can generalize across different contexts.

Utility: Provides a diverse set of linguistic scenarios, facilitating the exploration of model robustness and adaptability.

**SICK**

Description: The Sentences Involving Compositional Knowledge (SICK) dataset Constructed from sentence pairs derived from image descriptions, focusing on the evaluation of compositional distributional semantic models.

Utility: Offers unique challenges in understanding complex semantic relationships, essential for testing model comprehension abilities.

**SNLI**

Description:The Stanford Natural Language Inference (SNLI) is a large-scale dataset created from image captions, instrumental in fostering research on semantic relationship understanding between sentence pairs.

Utility: Its size and composition support extensive training and benchmarking efforts, contributing to model scalability and performance evaluation.

## 4. Literature Review:

In our review of current literature, we explored a variety of approaches to NLI, from traditional machine learning techniques to more recent deep learning innovations. Studies such as Bowman et al. (2015) and Williams et al. (2018) have provided foundational insights into dataset creation and baseline model performance. Recent advancements have shown the efficacy of transformer-based models, like BERT and GPT, in achieving state-of-the-art results in NLI tasks. Our project draws inspiration from these findings, aiming to integrate and build upon these methodologies to push the boundaries of current NLI models.

## 5. Methodology:

Our approach will involve a combination of traditional NLP techniques and advanced deep learning models. We plan to:

**Preprocess and Analyze Data:** Conduct thorough explorations of the MultiNLI, SICK, and SNLI datasets to understand their structure, content, and potential biases.
**Baseline Models:** Implement and evaluate baseline models as reported in seminal papers to establish performance benchmarks.
**Model Development:** Experiment with advanced neural network architectures, including transformer models, to explore their effectiveness in capturing semantic relationships.

## Approaches:

1. Logistic Regression
2. BiLSTM’s
3. Transformers - BERT , RoBERTa, ALBERT
4. Max Epoch Selection
5. Ensemble with Max Voting
6. Epoch Max and Ensemble Aggregation
7. RoBERTa with PEFT (LoRA)

**Evaluation Metrics:** Utilize accuracy, precision, recall, and F1 score to assess model performance, with a focus on ensuring balanced performance across all three classification categories.

## 1. <u>Logistic Regression</u>

**Baseline Comparisons**:

For benchmarking our approach, we selected well-established baselines in NLI tasks that use logistic regression with different feature representations. The chosen baselines are primarily from seminal papers in NLP that discuss the use of TF-IDF, and word embeddings for logistic regression models. These methods have been widely recognized for their simplicity and effectiveness in several text classification tasks

**Understanding and Interpretation of the Approach:**

The approach utilizes TF-IDF vectorization to transform textual data into a numerical format that highlights the importance of words based on their frequency and distribution across documents. Logistic regression, a statistical model that predicts the probabilities of different possible outcomes, is then applied. This model is particularly suited for binary or categorical dependent variables like the labels in NLI tasks.

**Implementation Summary**

**Data Preprocessing:**

The preprocessing step involved cleaning the data and converting the textual content of the premises and hypotheses into a suitable format for vectorization. This included filtering out any unlabelled instances and ensuring that the text data was free of encoding errors or irrelevant characters.

**Model Training:**

The TF-IDF vectorizer was first fitted to the training texts to create a vocabulary and transform the texts into feature vectors. The logistic regression model was then trained on these vectors, adjusting weights to minimize prediction error over the training iterations, set to a maximum of 10,000 to ensure convergence.

**Model Evaluation:**

The performance of the model was evaluated on validation datasets using metrics like accuracy, precision, recall, and F1-score. Detailed reports provided insights into the model's ability to classify each type of relationship (entailment, neutral, contradiction) correctly.

## Results and Analysis

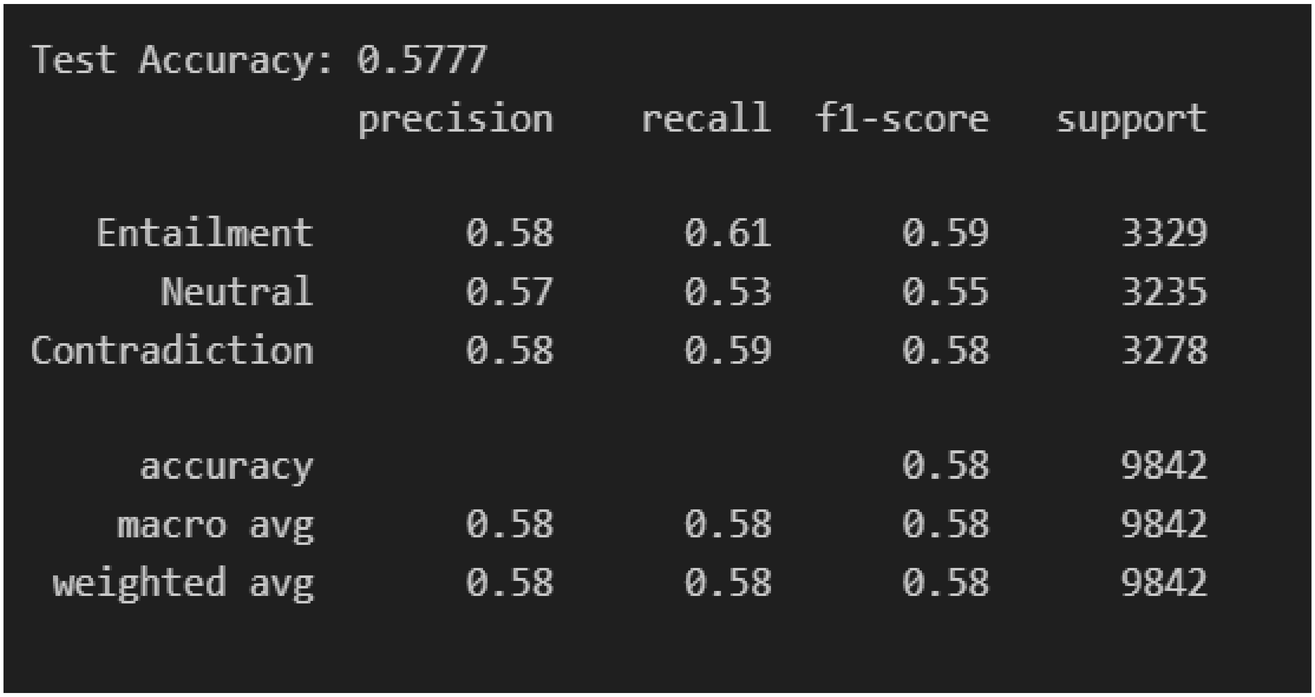
### 1. SNLI Dataset

Accuracy and Classification Report:

- Accuracy: 57.77%
- Precision, Recall, F1-Score:
    - Entailment: Precision 0.58, Recall 0.61, F1-Score 0.59
    - Neutral: Precision 0.57, Recall 0.53, F1-Score 0.55
    - Contradiction: Precision 0.58, Recall 0.59, F1-Score 0.58

Confusion Matrix :

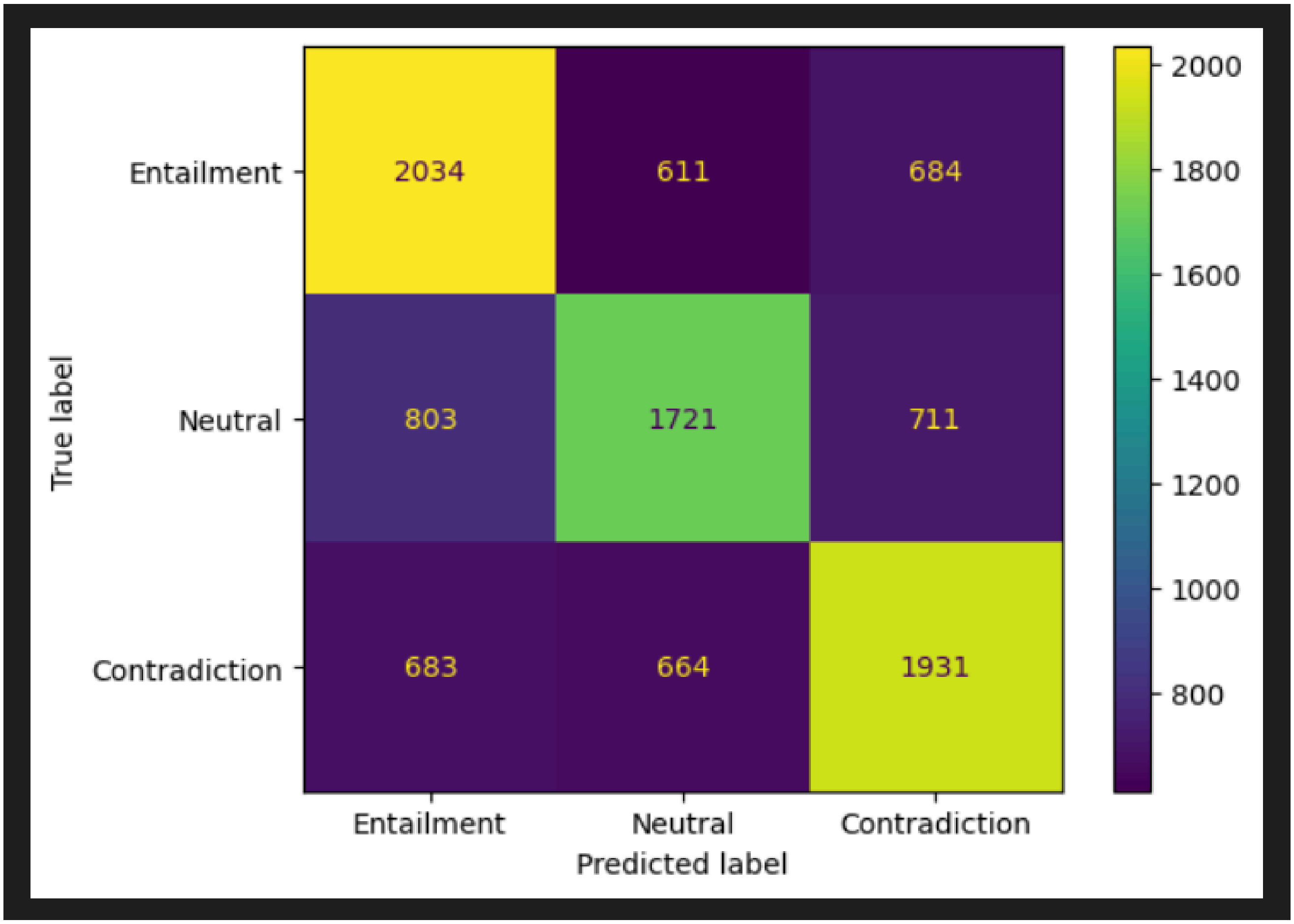
- True Positives: Entailment 2034, Neutral 1721, Contradiction 1931
- Misclassifications: Substantial between all categories, indicating overlap in textual cues



```
Test Accuracy: 0.5777
               precision    recall  f1-score   support

   Entailment       0.58      0.61      0.59      3329
      Neutral       0.57      0.53      0.55      3235
Contradiction       0.58      0.59      0.58      3278

     accuracy                           0.58      9842
    macro avg       0.58      0.58      0.58      9842
 weighted avg       0.58      0.58      0.58      9842
```



**Key Observations:**

- The classification accuracy and F1-scores indicate moderate performance with some balance across classes.
- Misclassification rates suggest the model struggles with overlapping linguistic features between the categories.

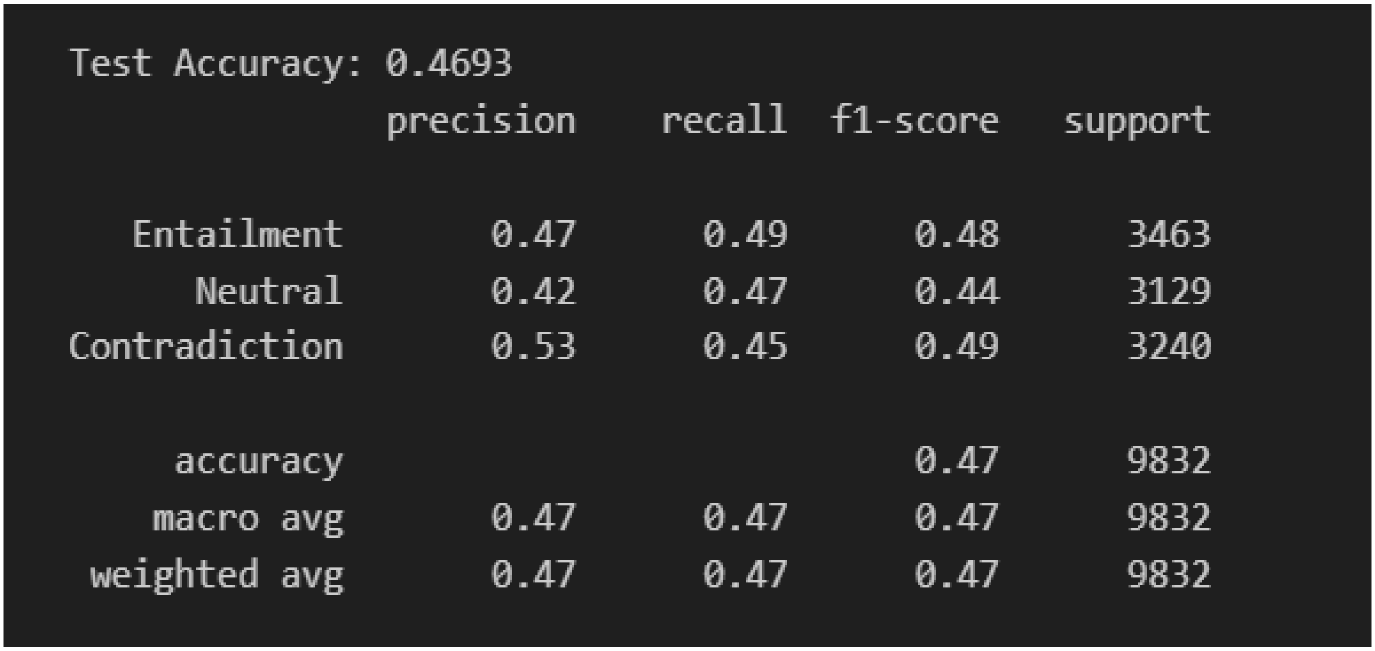
**2. MNLI Dataset**

Accuracy and Classification Report:

- Accuracy: 46.93%
- Precision, Recall, F1-Score:
  - Entailment: Precision 0.47, Recall 0.49, F1-Score 0.48
  - Neutral: Precision 0.42, Recall 0.47, F1-Score 0.44
  - Contradiction: Precision 0.53, Recall 0.45, F1-Score 0.49

Confusion Matrix :

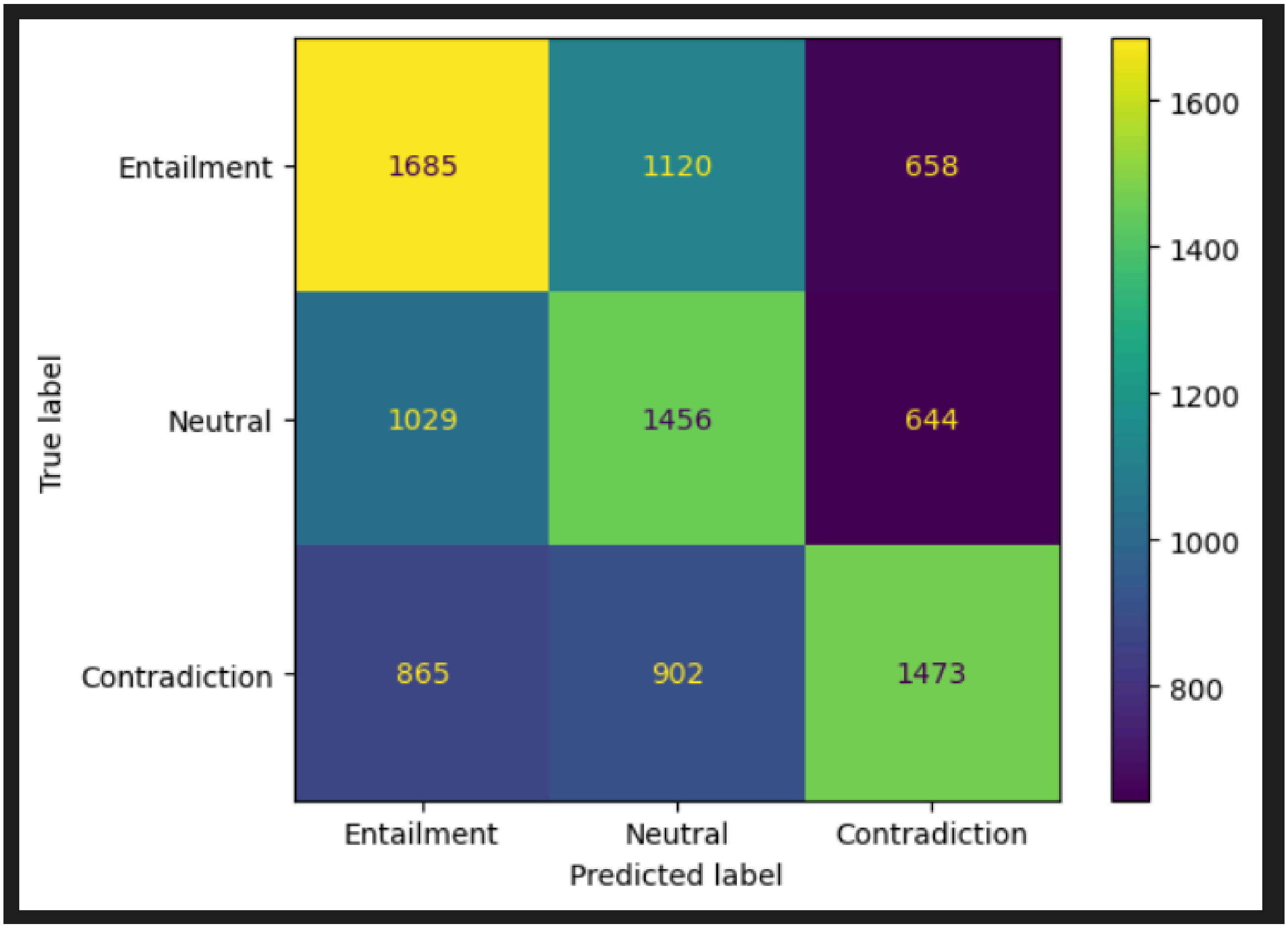
- True Positives: Entailment 1685, Neutral 1456, Contradiction 1473
- Misclassifications: High confusion between "Entailment" and "Neutral".



```
Test Accuracy: 0.4693
               precision    recall  f1-score   support

   Entailment       0.47      0.49      0.48      3463
      Neutral       0.42      0.47      0.44      3129
Contradiction       0.53      0.45      0.49      3240

     accuracy                           0.47      9832
    macro avg       0.47      0.47      0.47      9832
 weighted avg       0.47      0.47      0.47      9832
```



Key Observations:

- Despite the numeric values indicating reasonable performance in terms of the number of true positives for each category, the overall accuracy is quite low at 46.93%.
- There is significant misclassification across all labels, especially between "Entailment" and "Neutral," which suggests a potential overlap in the linguistic cues used to represent these classes.
- The precision and recall values indicate that while the model can sometimes identify the correct labels, it struggles to do so consistently across different instances, particularly struggling with "Neutral" where precision is the lowest.

### 3. SICK Dataset

Accuracy and Classification Report:

- Accuracy: 63.43%
- Precision, Recall, F1-Score:
  - Entailment: Precision 0.61, Recall 0.38, F1-Score 0.47
  - Neutral: Precision 0.64, Recall 0.85, F1-Score 0.73
  - Contradiction: Precision 0.66, Recall 0.27, F1-Score 0.38

Confusion Matrix :

- True Positives: Entailment 55, Neutral 240, Contradiction 19
- Misclassifications:Entailment misclassified as: Neutral 88, Contradiction 0,
  Neutral misclassified as: Entailment 31, Contradiction 10,
  Contradiction misclassified as: Entailment 4, Neutral 48

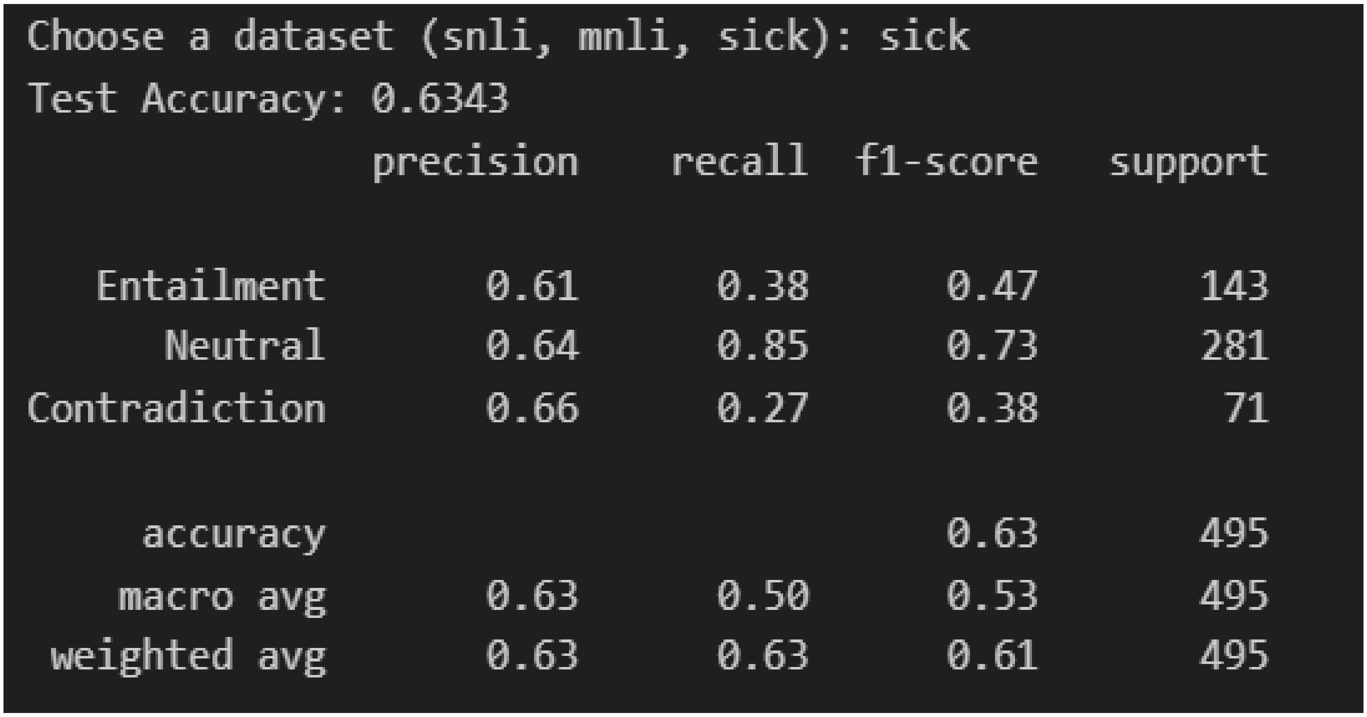
Key Observations:

- The accuracy appears good at first glance; however, the precision, recall, and F1-scores, particularly for "Entailment" and "Contradiction," indicate that the model struggles significantly with these categories.
- The "Neutral" category shows a relatively high degree of correct classification but also suffers from being a common misclassification target for the other two categories, suggesting a possible bias in feature representation towards this class.
- The absence of "Entailment" being misclassified as "Contradiction" and the very low number of "Contradiction" correctly identified indicates a specific weakness in recognizing and distinguishing the features of contradiction.

### Recommendations

Enhanced Feature Engineering: Using more sophisticated feature extraction methods such as embeddings could address the overlap and improve differentiation in text features.

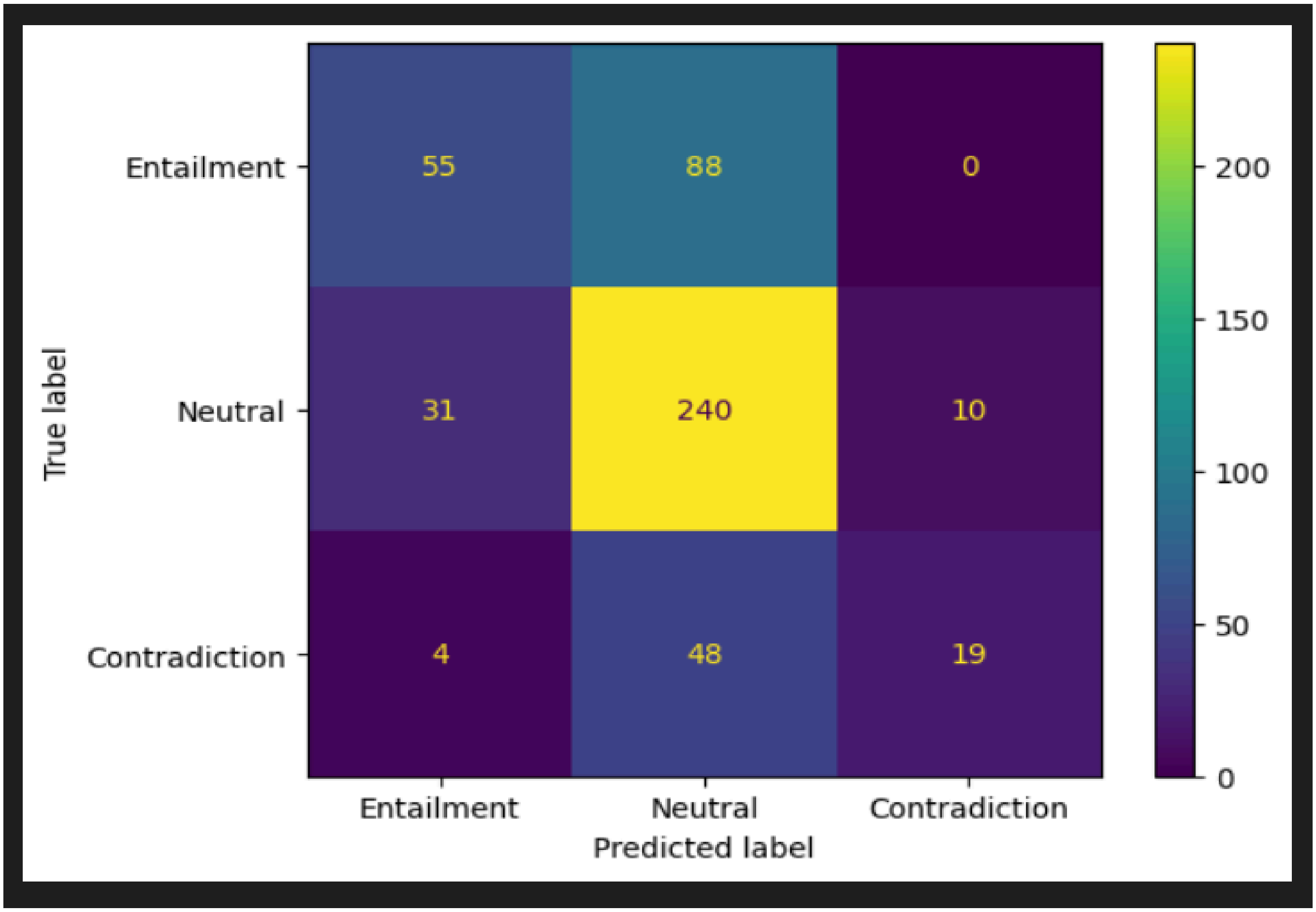
Model Adjustments: Considering more complex models or ensemble techniques.



```
Choose a dataset (snli, mnli, sick): sick
Test Accuracy: 0.6343
               precision    recall  f1-score   support

   Entailment       0.61      0.38      0.47       143
      Neutral       0.64      0.85      0.73       281
Contradiction       0.66      0.27      0.38        71

     accuracy                           0.63       495
    macro avg       0.63      0.50      0.53       495
 weighted avg       0.63      0.63      0.61       495
```

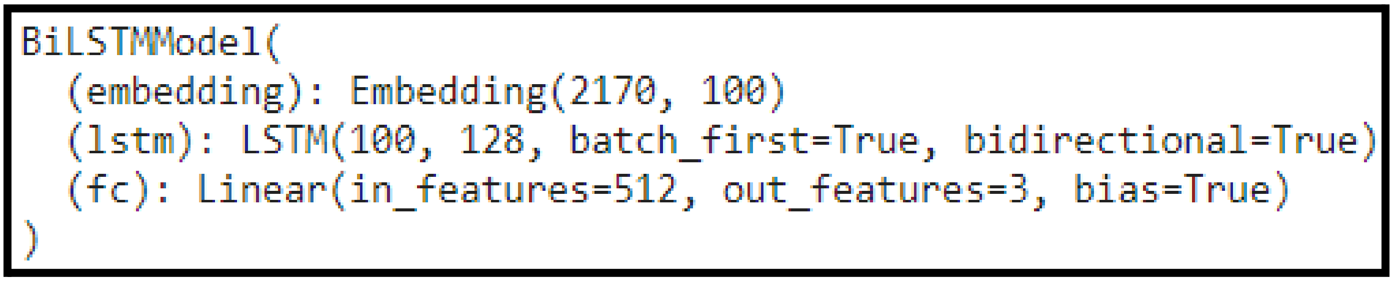


## 2. BILSTM’s

### Model Overview

The BiLSTMModel class defines a neural network for the Natural Language Inference (NLI) task, which determines the relationship between two text sequences (premise and hypothesis). The network architecture includes:

**Embedding Layer:** Maps words to vectors to capture semantic meanings.
**Bidirectional LSTM (BiLSTM)**: Processes text from both forward and backward directions to enhance the understanding of context and dependencies.
**Concatenation:** Combines hidden states from the BiLSTM for both the premise and hypothesis to create a comprehensive feature set.
**Fully Connected Layer:** Outputs the classification results based on the concatenated features.



```
BiLSTMModel(
  (embedding): Embedding(2170, 100)
  (lstm): LSTM(100, 128, batch_first=True, bidirectional=True)
  (fc): Linear(in_features=512, out_features=3, bias=True)
)
```

### Results and Analysis

### 1. SNLI Dataset

Overview:

- The results for the BiLSTM on the SNLI dataset show an overall accuracy of 68.08%. The confusion matrix and classification report provide a breakdown of performance across the three categories (0: Entailment, 1: Neutral, 2: Contradiction).

Confusion Matrix Analysis:

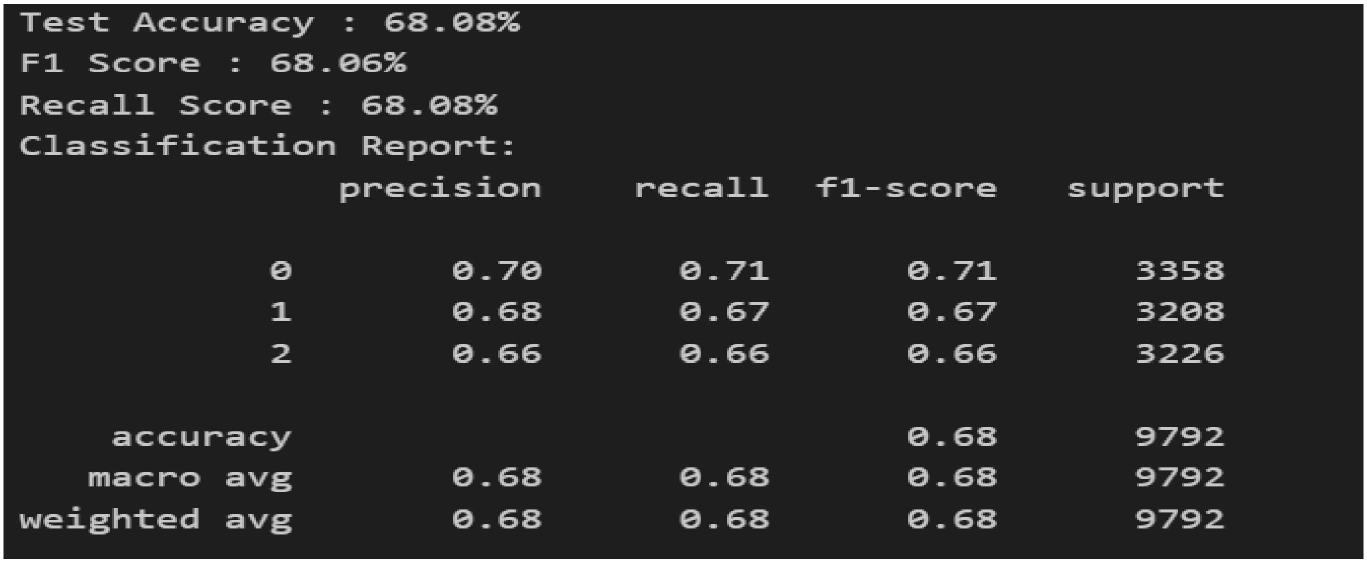
- Entailment (0): The model predicted 2396 correct labels out of 3358 actual samples, with a precision of 0.70 and a recall of 0.71. The model mainly confused entailment with contradiction, incorrectly predicting 512 instances as contradictions.
- Neutral (1): The model achieved 2138 correct predictions for neutral labels from 3208 samples, corresponding to a precision of 0.68 and a recall of 0.67. A significant number of misclassifications occurred between neutral and contradiction (564 instances).
- Contradiction (2): The contradiction label had 2132 correctly predicted outcomes from 3226 samples, with both precision and recall at 0.66. The model also confused contradictions with entailment in 536 cases.

Performance Metrics:

The F1 score for each category indicates a relatively balanced performance, with entailment showing slightly better results. The macro average of 0.68 for precision, recall, and F1-score suggests that the model performs uniformly across all categories.

Summary:

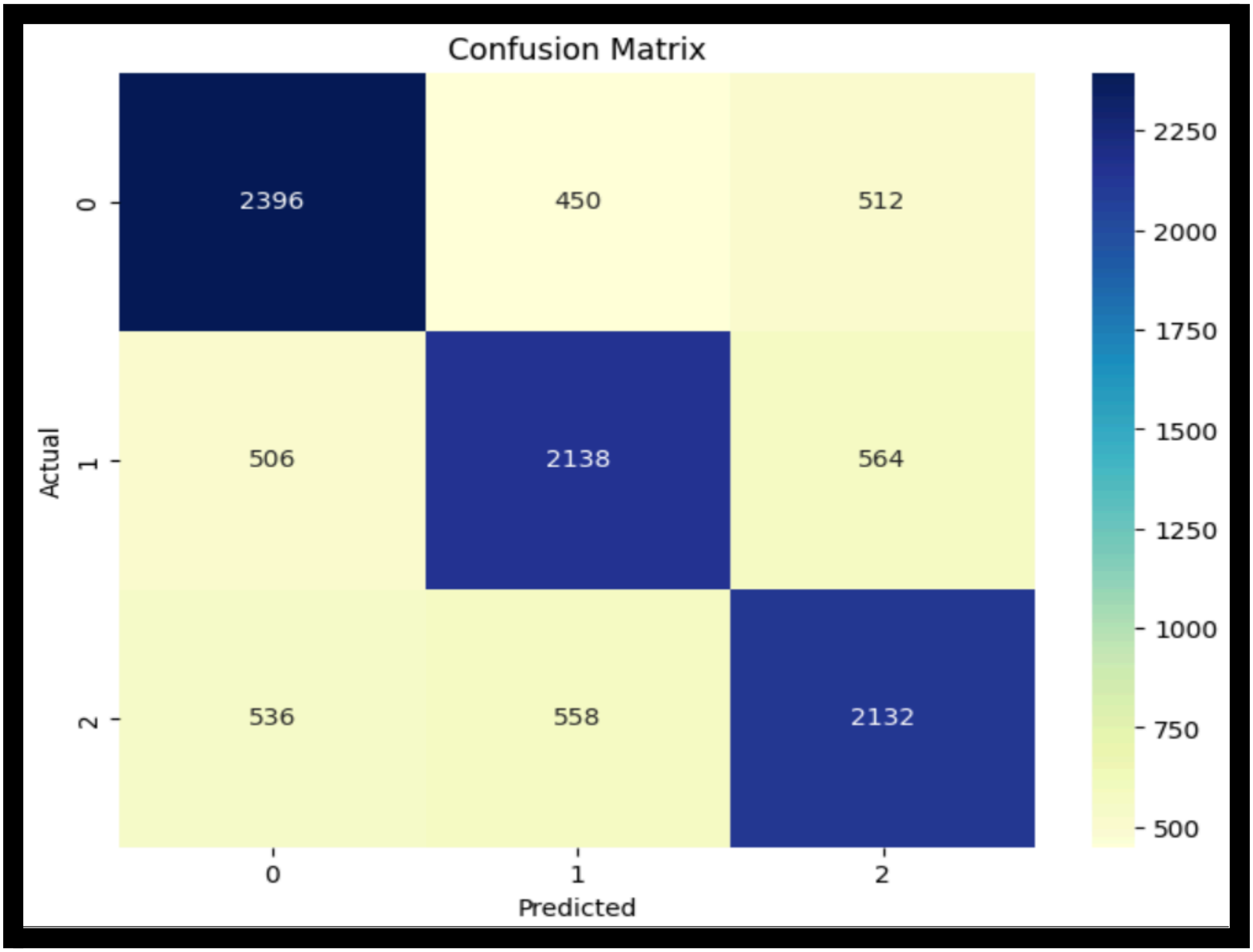
While the BiLSTM model shows a decent generalization ability on the SNLI dataset, the noticeable number of misclassifications between the categories, particularly between entailment and contradiction, and neutral and contradiction, indicates potential areas for improvement in distinguishing finer nuances between these categories.



```
Test Accuracy : 68.08%
F1 Score : 68.06%
Recall Score : 68.08%
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.71      0.71      3358
           1       0.68      0.67      0.67      3208
           2       0.66      0.66      0.66      3226

    accuracy                           0.68      9792
   macro avg       0.68      0.68      0.68      9792
weighted avg       0.68      0.68      0.68      9792
```



## 2. MNLI Dataset

### Overview:

The performance of the BiLSTM model on the MNLI dataset shows an overall accuracy of 53.43%. This analysis uses the attached confusion matrix to illustrate the distribution of predictions across the three classes: 0 (Entailment), 1 (Neutral), and 2 (Contradiction).

### Confusion Matrix Analysis:

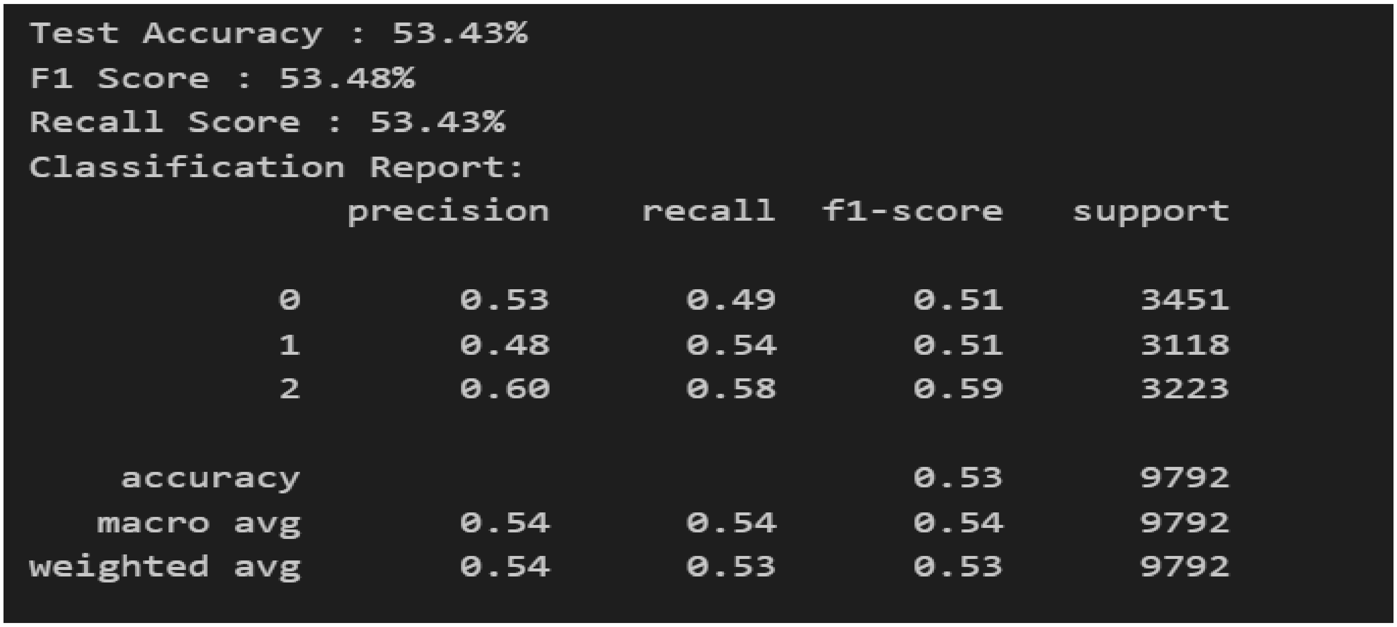
Entailment (0): Correctly predicted 1688 out of 3451 actual instances, with a precision of 0.53 and a recall of 0.49.
Neutral (1): Correctly predicted 1689 out of 3118 instances, achieving the highest recall of 0.54 but a lower precision of 0.48.
Contradiction (2): Showed relatively better precision at 0.60, with 1855 correct predictions out of 3223 instances and a recall of 0.58.

### Performance Metrics:

The F1 scores are relatively uniform across categories, with the highest for contradiction at 0.59, indicating slightly better performance in distinguishing contradictions compared to other categories.

### Conclusion:

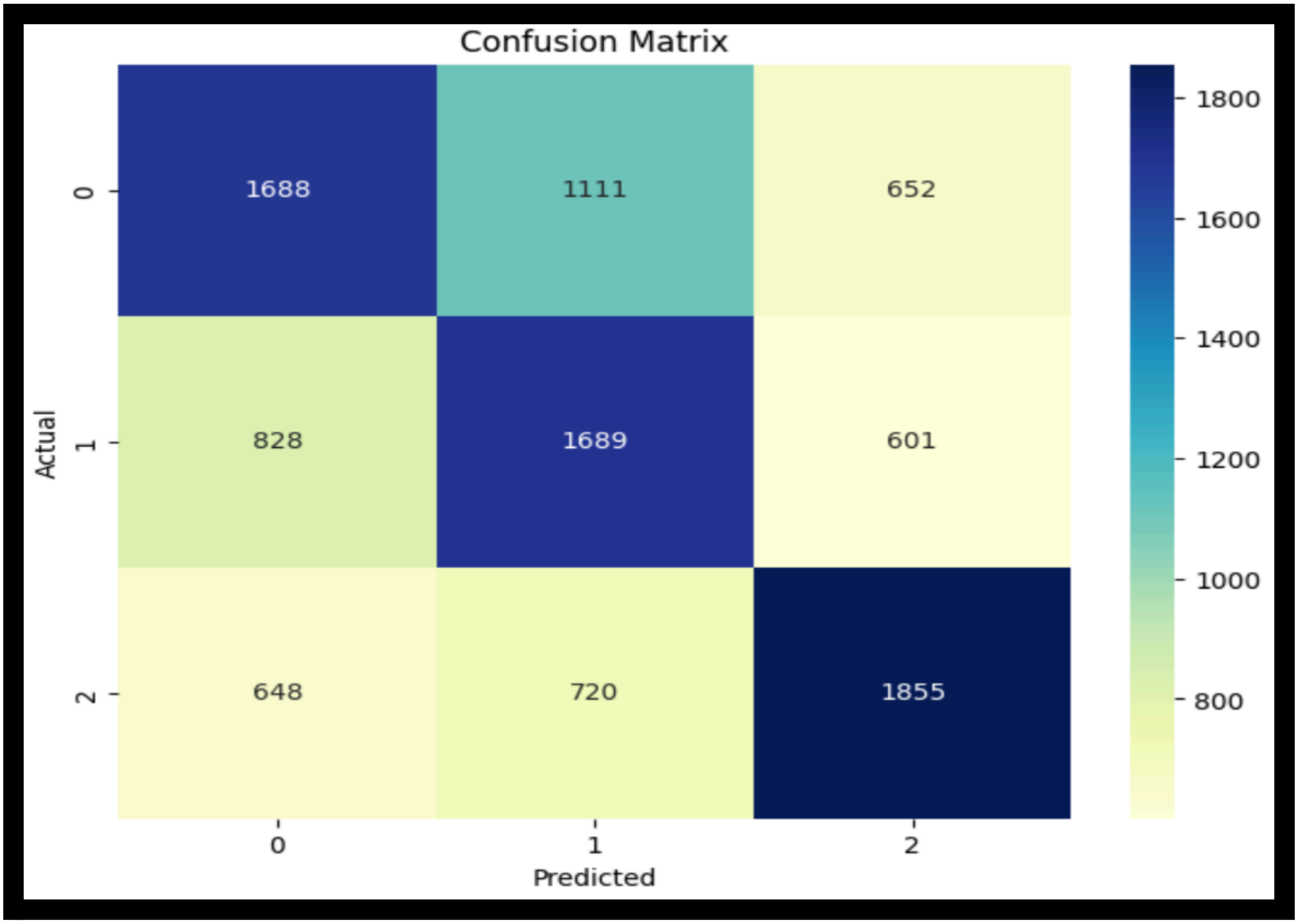
The model’s performance is modest, with significant room for improvement, particularly in increasing precision for neutral predictions and improving recall for entailments. Enhancements in model tuning and possibly integrating more complex features or attention mechanisms may help improve these metrics.



```
Test Accuracy : 53.43%
F1 Score : 53.48%
Recall Score : 53.43%
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.49      0.51      3451
           1       0.48      0.54      0.51      3118
           2       0.60      0.58      0.59      3223

    accuracy                           0.53      9792
   macro avg       0.54      0.54      0.54      9792
weighted avg       0.54      0.53      0.53      9792
```



### 3. SICK Dataset

**Overview:**

The BiLSTM model evaluated on the SICK dataset shows a total accuracy of 58.76%. The detailed confusion matrix provides insights into the model's ability to differentiate among the three classes: 0 (Entailment), 1 (Neutral), and 2 (Contradiction).

**Confusion Matrix Analysis:**

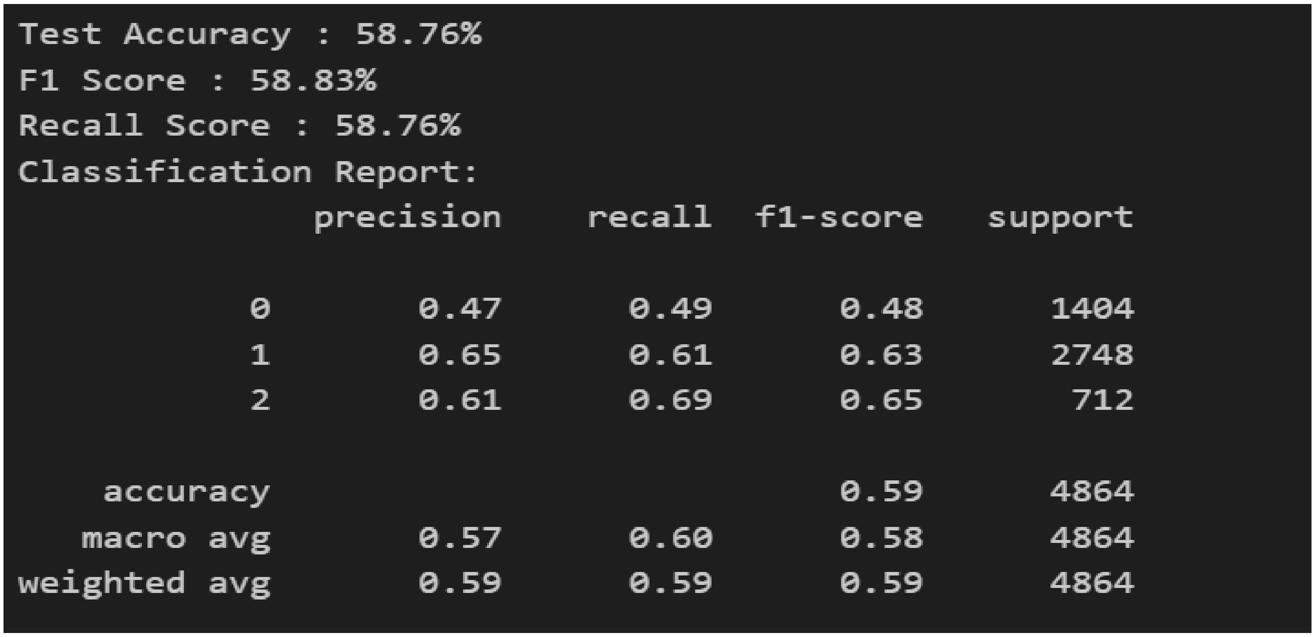
Entailment (0): Correct predictions made for 682 out of 1404 samples, showing a precision of 0.47 and a recall of 0.49.
Neutral (1): The model predicted 1684 correct results out of 2748, with a precision of 0.65 and recall of 0.61, marking it as the best-performing category.
Contradiction (2): The best recall of 0.69 was noted here, with 492 correct predictions out of 712 samples, and a precision of 0.61.

**Performance Metrics:**

Precision Average: 0.57
Recall Average: 0.60
F1 Score: 0.58

**Conclusion:**

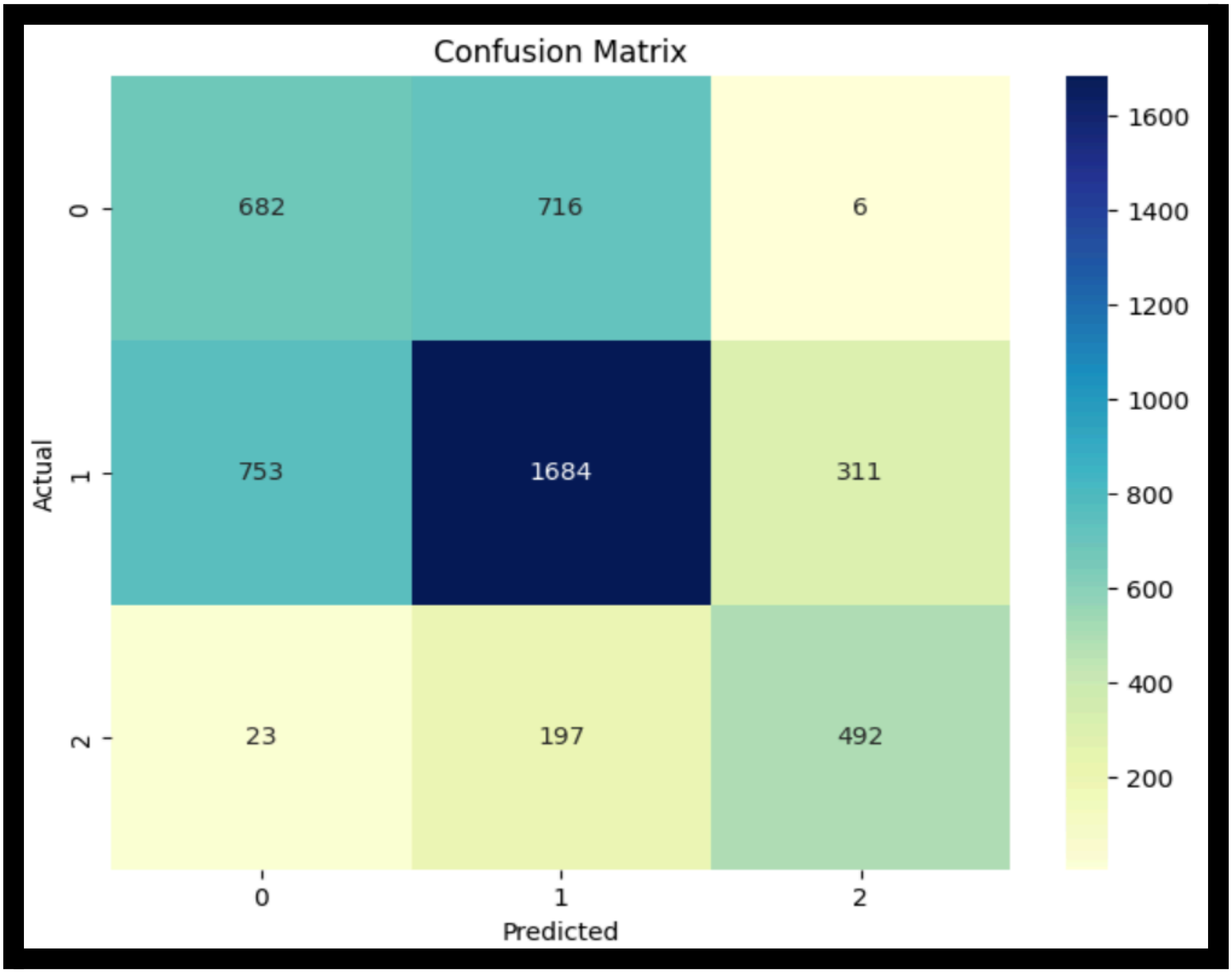
While the model performs decently across categories, especially in identifying contradictions, its lower precision in entailment predictions indicates a need for more refined feature extraction or model adjustments to improve overall classification accuracy.



```
Test Accuracy : 58.76%
F1 Score : 58.83%
Recall Score : 58.76%
Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.49      0.48      1404
           1       0.65      0.61      0.63      2748
           2       0.61      0.69      0.65       712

    accuracy                           0.59      4864
   macro avg       0.57      0.60      0.58      4864
weighted avg       0.59      0.59      0.59      4864
```



## 3. Transformers

Here's what the we have done :

**BERT Training and Evaluation:** We used a BERT model, training it, and evaluating its performance on the validation and test datasets. It prints out the accuracy of the model on each dataset.

**RoBERTa Training and Evaluation:** Similar to BERT, we then defined a RoBERTa model, trained it, and evaluated its performance on the validation and test datasets.

**ALBERT Training and Evaluation:** Likewise, we defined an ALBERT model, trained it, and evaluated its performance on the validation and test datasets.

**Ensemble Method**: After evaluating each model individually, we combined their predictions using a majority voting ensemble method. It calculates the accuracy of the ensemble model on the test dataset.

**Majority Voting Ensemble:**

In this method, predictions from multiple individual models (BERT, RoBERTa, and ALBERTa) are combined by taking the majority vote.

Each model independently predicts the label for a given input instance.

The final prediction for that instance is determined by selecting the label that receives the most votes among the predictions from all models.

This approach is simple and effective, especially when the models have diverse characteristics and may make different types of errors.

**Ensemble + Epoch Max:**

This method extends the majority voting ensemble by incorporating the max prediction probabilities from all models for all epochs.

Instead of just selecting the majority vote, the probabilities of each class predicted by each model are averaged across all models across all epochs.

The final prediction is then determined by selecting the class with the highest average probability.

This approach can provide more nuanced insights into the combined certainty of each model regarding its predictions.

It's particularly useful when the models' predictions are not only categorical but also probabilistic.

By combining predictions from multiple models using these ensemble methods, we aim to improve overall prediction accuracy and robustness by leveraging the strengths of each individual model and mitigating their weaknesses. Ensemble methods are widely used in machine learning to enhance model performance, especially in scenarios where multiple models trained on different architectures or datasets are available.

**Overall Results:**

| **MODELS** | **SNLI** | **MULTI - NLI** | **SICK** |
|---|---|---|---|
| **BERT** | 89.31% | 78.80% | 84.69% |
| **RoBERTa** | 90.50% | 84.29% | 88.44% |
| **ALBERT** | 89.64% | 81.20% | 86.91% |
| **Max Epoch Selection - BERT** | 90.28% | 80.64% | 84.12% |
| **Max Epoch Selection - RoBERTa** | 91.36% | 85.25% | 89.36% |
| **Max Epoch Selection - ALBERT** | 90.36% | 82.49% | 88.57% |

| **MODELS** | **SNLI** | **MULTI - NLI** | **SICK** |
|---|---|---|---|
| **EWMV -** Bert + Roberta + Alberta | 91.28% | 84.87% | 88.44% |
| **EWMV -** Bert + Alberta | 91.36% | 82.82% | 85.59% |
| **EWMV -** Bert + Roberta | 91.04% | 84.07% | 86.14% |
| **EWMV -** Roberta + Alberta | 90.97% | 84.97% | 89.30% |
| **EMEA -** Bert + Roberta + Alberta | 91.61% | 85.20% | 88.86% |
| **EMEA -** Bert + Alberta | 91.07% | 83.50% | 86.36% |
| **EMEA -** Bert + Roberta | 91.40% | 85.02% | 86.89% |
| **EMEA -** Roberta + Alberta | **91.63%** | **85.41%** | **89.56%** |

* EWMV - Ensemble with Max Voting
EMEA - Epoch Max and Ensemble Aggregation

### 1. SNLI Dataset

**Overview:**

Model Configuration: The ensemble combines predictions from BERT and ALBERT models, leveraging their strengths to enhance overall performance on the NLI task.

Performance Metrics: Achieved a high test accuracy of 91.62%, with an F1 score of 91.63% and a recall of 91.62%, indicating a highly effective model ensemble.

**Confusion Matrix Analysis:**

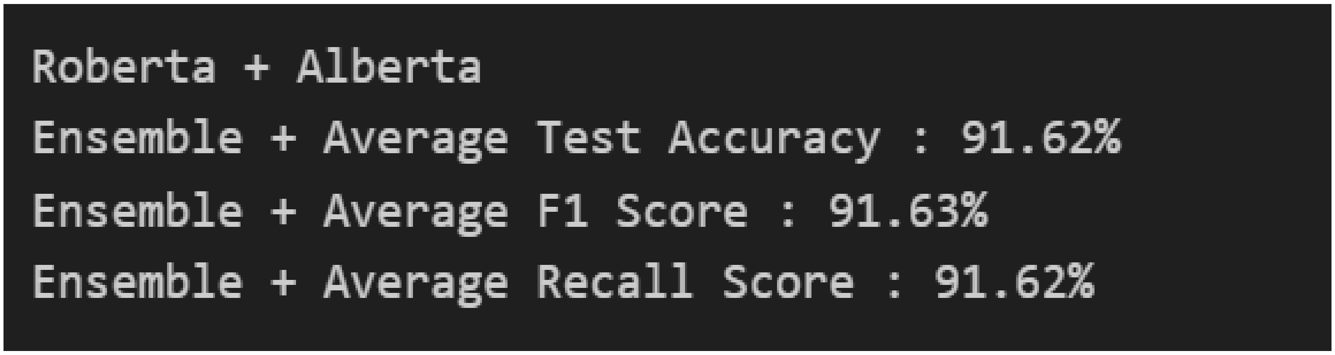
Class 0 (Entailment): High precision with 3149 correctly predicted out of 3368, a slight majority are accurately classified with few misclassifications.

Class 1 (Neutral): Exhibits robust performance with 2841 correctly identified out of 3219; shows good balance in handling neutral classifications.

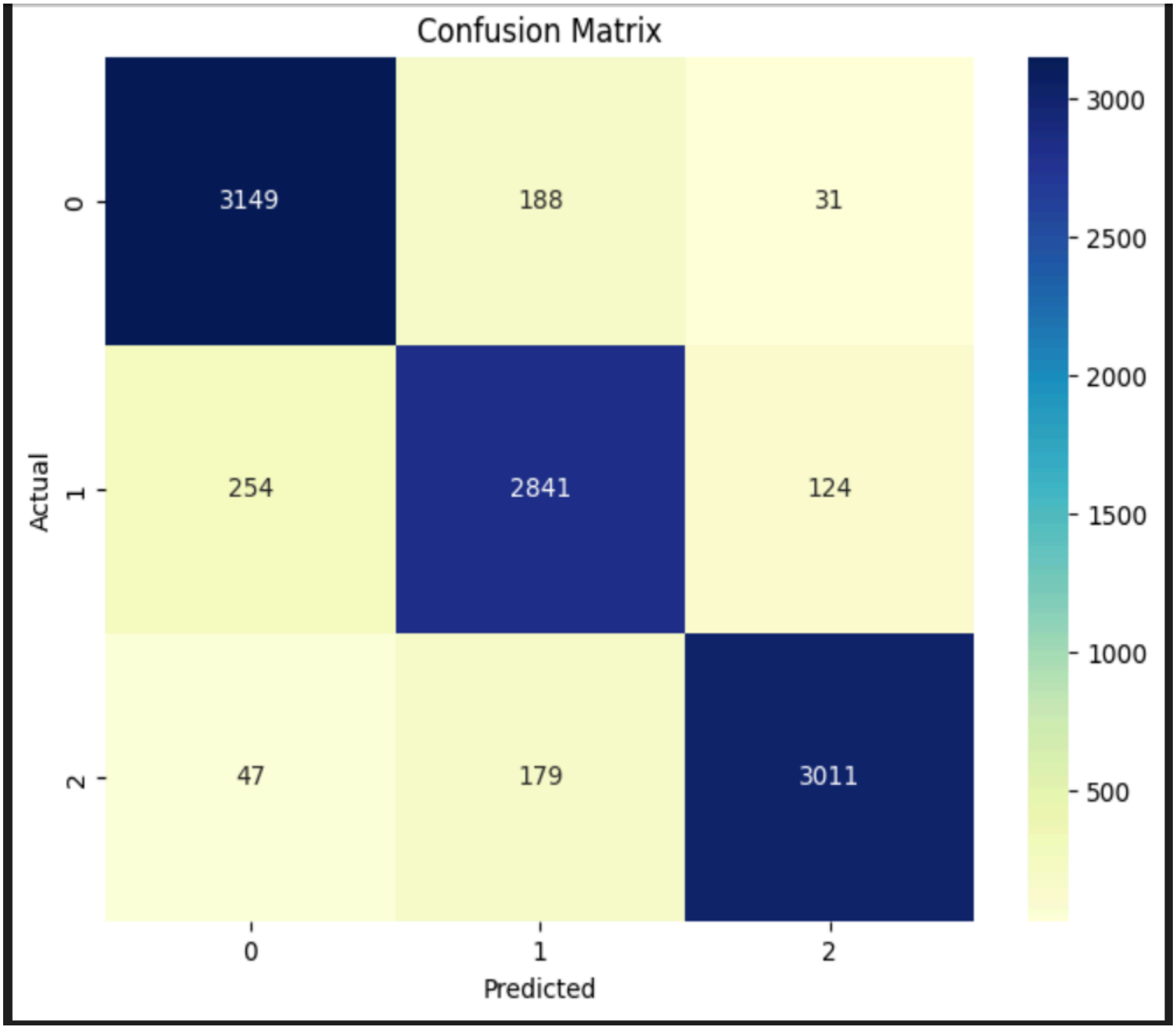
Class 2 (Contradiction): Most effective in contradiction detection with 3011 out of 3237 accurately predicted, indicating a strong capability in distinguishing contradictions.

**Conclusion:**

The use of an ensemble approach combining BERT and ALBERT models has proven to be highly effective, as reflected in the high accuracy and balanced performance across all categories. This strategy capitalizes on the individual strengths of each model, leading to improved overall classification accuracy.



```
Roberta + Alberta
Ensemble + Average Test Accuracy : 91.62%
Ensemble + Average F1 Score : 91.63%
Ensemble + Average Recall Score : 91.62%
```



|  | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 3149 | 188 | 31 |
| 1 | 254 | 2841 | 124 |
| 2 | 47 | 179 | 3011 |

### 2. MNLI Dataset

**Overview:**

- Model Performance: The ensemble achieves an overall test accuracy of 85.41%, with F1 and recall scores also around 85.41%, indicating a consistent and effective model across the task.

**Confusion Matrix Analysis:**

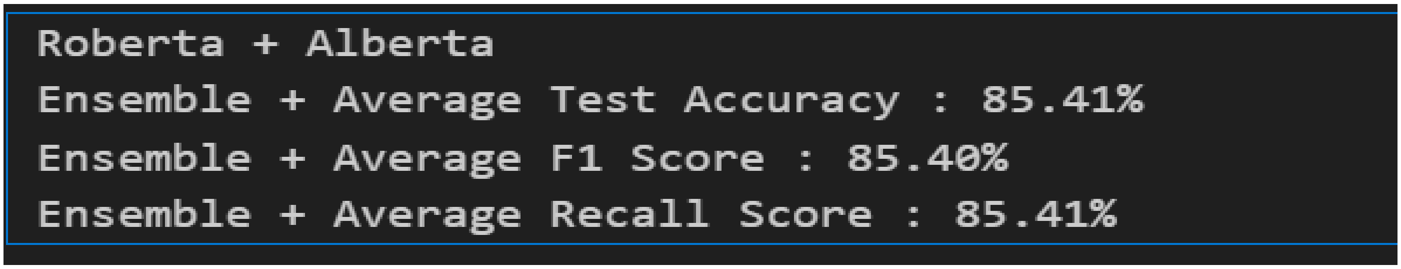
- Class 0 (Entailment): The model correctly predicted 3090 out of 3463 cases, showcasing robust accuracy in identifying entailment with minimal misclassification.
- Class 1 (Neutral): Out of 3129 cases, 2535 were correctly classified as neutral. This reflects the model's good performance, though with some room for improvement in reducing misclassifications with entailment and contradiction.
- Class 2 (Contradiction): Demonstrated strong performance in contradiction detection with 2773 out of 3240 accurately predicted, showing the model's capability in distinguishing this category effectively.

**Performance Metrics:**

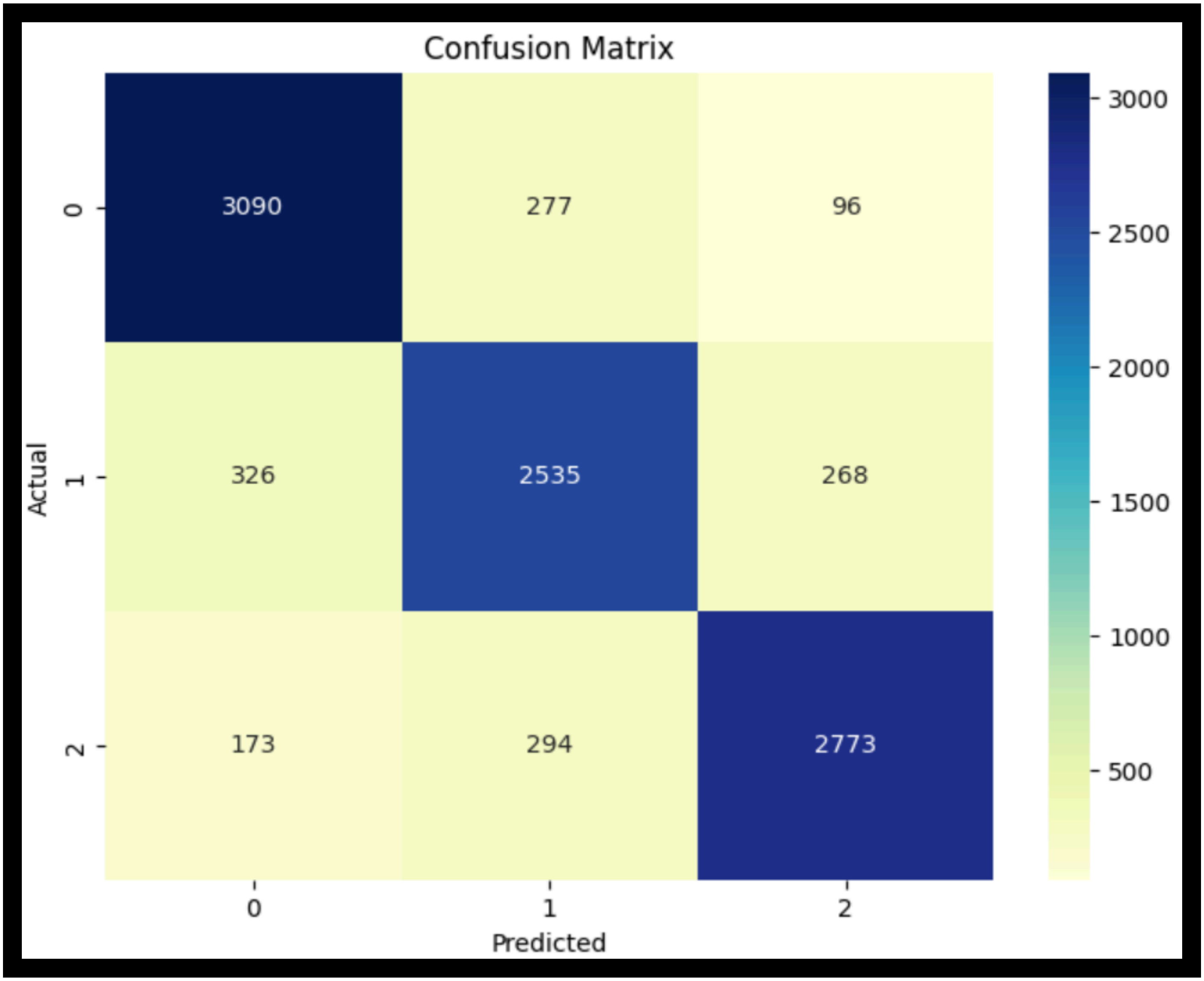
- Precision Average: Approx. 86%
- Recall Average: 85.41%
- F1 Score: 85.40%

**Conclusion:**

The ALBERT and RoBERTa ensemble is highly effective in handling the complexities of the MNLI dataset, showing particular strength in contradiction identification. The balanced metrics indicate that the model performs consistently across different types of sentence relationships.



```
Roberta + Alberta
Ensemble + Average Test Accuracy : 85.41%
Ensemble + Average F1 Score : 85.40%
Ensemble + Average Recall Score : 85.41%
```



### 3. SICK Dataset

**Overview:**

- Model Performance: The ensemble achieves an overall test accuracy of 89.56%, with an F1 score of 89.58% and recall of 89.56%, demonstrating a robust performance across all classes.

**Confusion Matrix Analysis:**

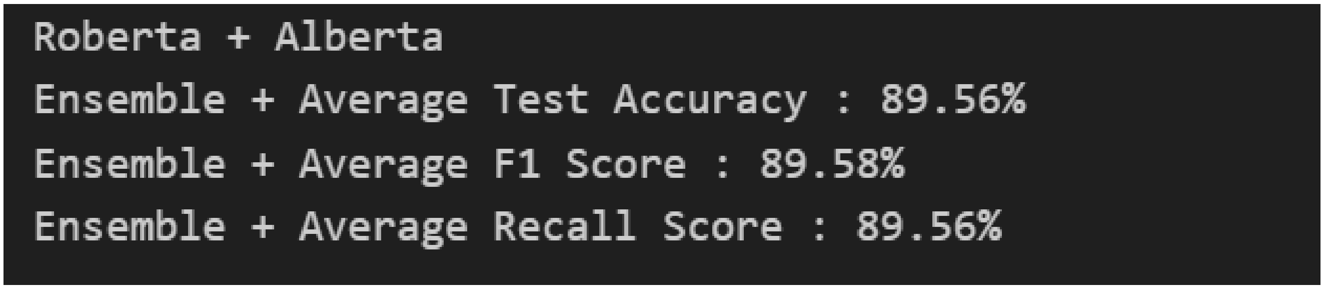
- Class 0 (Entailment): Shows strong performance with 1261 correct predictions out of 1404 cases, indicating a precise handling of entailment with very few misclassifications to other classes.
- Class 1 (Neutral): This category also shows solid accuracy with 2535 correct identifications out of 2790 cases. While there are some misclassifications as entailment and contradiction, the model mostly maintains high precision.
- Class 2 (Contradiction): Displays the most distinct classification with 598 correct out of 712 cases, signifying the model's effective differentiation of contradictions from other categories.

**Performance Metrics:**

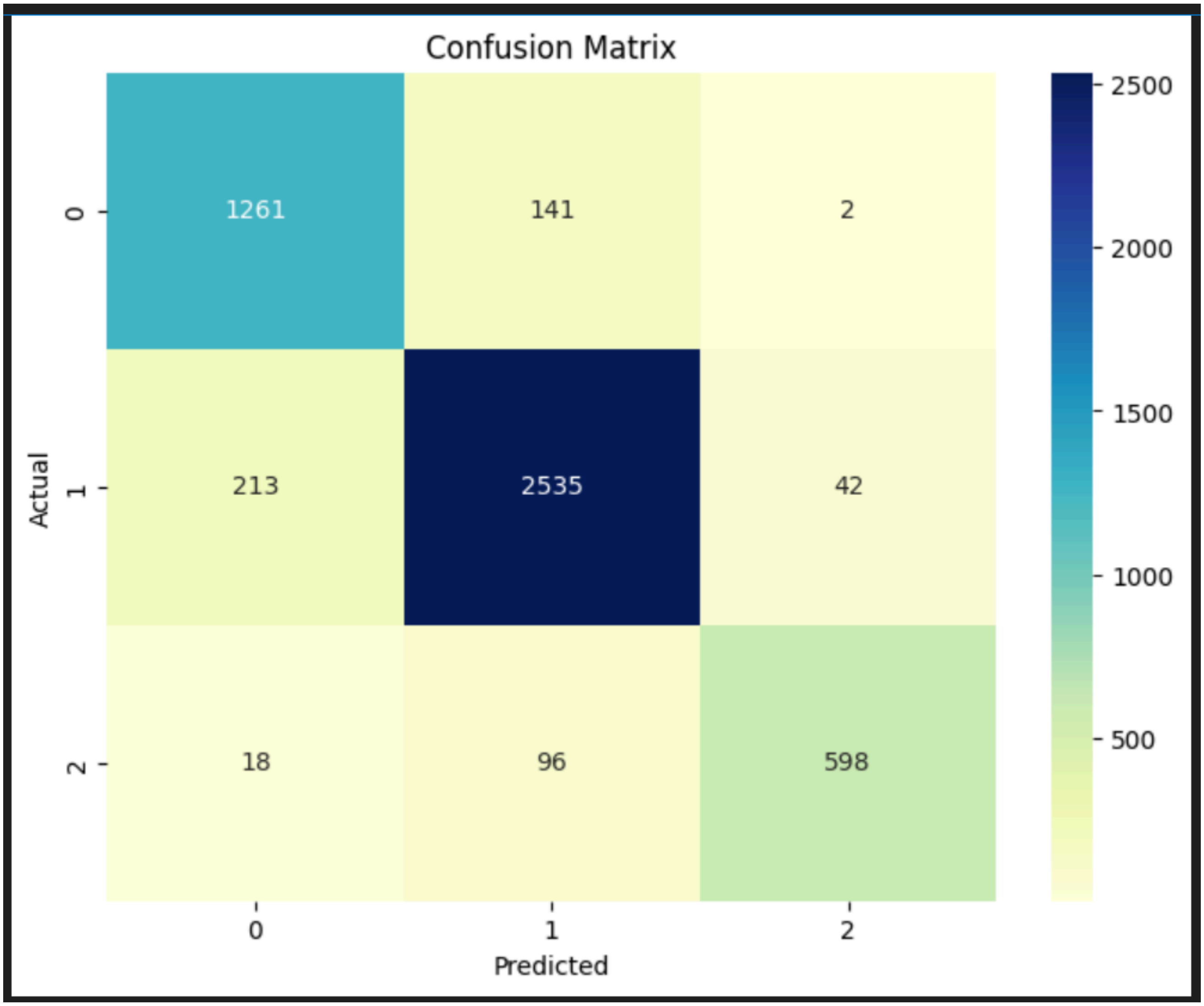
- Precision: Consistently high across all classes, supporting the model's ability to accurately classify each type.
- Recall: Nearly uniform across categories, indicating balanced sensitivity in detecting all labels.

**Conclusion:**

The ALBERT and RoBERTa ensemble effectively processes and understands complex sentence relationships within the SICK dataset, achieving commendable accuracy and balance across entailment, neutral, and contradiction classifications. This model setup highlights the strengths of combining these two powerful transformers to enhance NLI task performance.



```
Roberta + Alberta
Ensemble + Average Test Accuracy : 89.56%
Ensemble + Average F1 Score : 89.58%
Ensemble + Average Recall Score : 89.56%
```



## 4. PEFT

We  configured and trained a modified version of a Roberta model for sequence classification, specifically tailored for NLI (natural language inference) tasks, using the PEFT (Parameter Efficient Fine Tuning) approach to enhance model adaptability with reduced parameter updates.

### Configuration and Setup

**RobertaTokenizer:** Preprocesses text into format suitable for Roberta model input.
**RobertaForSequenceClassification:** Base model, fine-tuned here for a three-class sequence classification task.
**PEFT Configuration:** Adjusts the base Roberta model using the LoRA (Low-Rank Adaptation) technique which targets specific model components ('query', 'key', 'value') to enhance training efficiency without significant expansion in model size.
Device Setup: Assigns the model to run on a GPU if available, enhancing training efficiency.
**LoRA:** LoRA, or Low Rank adaptation, is a technique used in machine learning to enhance model performance in scenarios with limited labeled data. By leveraging a combination of labeled and unlabeled data, LoRA fine-tunes models efficiently, adjusting their parameters in a low-dimensional subspace. This approach is particularly valuable in domains like natural language processing and computer vision, where obtaining large labeled datasets can be challenging. LoRA optimizes model adaptability, making it a powerful tool for tasks where data scarcity is a hurdle to traditional learning approaches.

## Results and Analysis

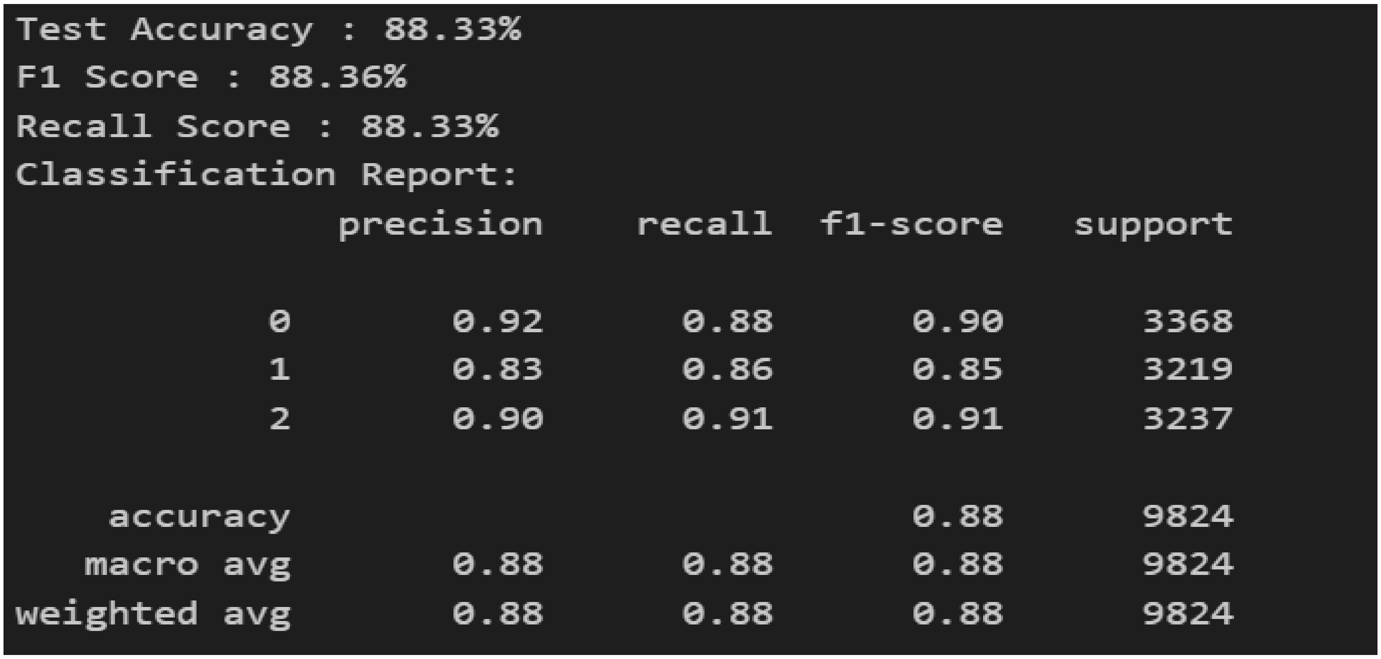
### 1. SNLI Dataset

**Overview:**
Model Performance: Achieved a high overall test accuracy of 88.33%, with precision, recall, and F1 scores all closely aligned, indicating a balanced and effective model across all classes.

**Confusion Matrix Analysis:**
Class 0 (Entailment): The model predicted 2949 true positives out of 3368 actual entailments, showcasing a precision of 0.92 and a recall of 0.88.
Class 1 (Neutral): Out of 3219 true neutrals, 2779 were correctly identified, resulting in precision of 0.83 and recall of 0.86.
Class 2 (Contradiction): Exhibited the best balance between precision and recall with 2950 out of 3237 contradictions correctly predicted, yielding a precision and recall of around 0.90 and 0.91 respectively.

**Performance Metrics:**
The model's weighted average scores across metrics suggest uniform efficiency in classifying all categories of the SNLI dataset. High scores in F1 and recall for all categories confirm the model's robustness in dealing with varied linguistic structures and logical relationships.

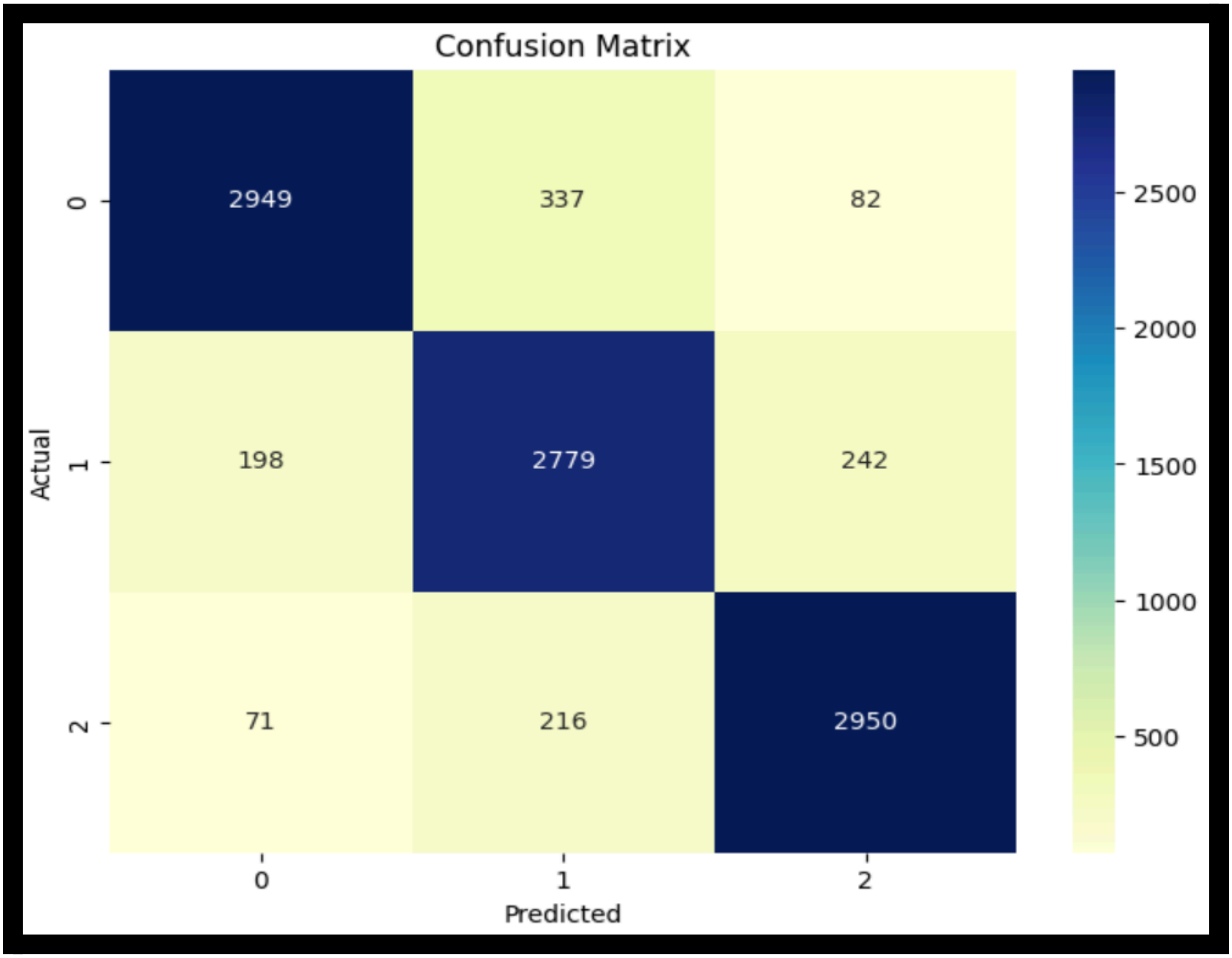
**Conclusion:**
The PEFT with LoRA approach demonstrates significant effectiveness in natural language inference tasks, managing to maintain a high standard of prediction accuracy across different types of sentence relationships.



```
Test Accuracy : 88.33%
F1 Score : 88.36%
Recall Score : 88.33%
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.88      0.90      3368
           1       0.83      0.86      0.85      3219
           2       0.90      0.91      0.91      3237

    accuracy                           0.88      9824
   macro avg       0.88      0.88      0.88      9824
weighted avg       0.88      0.88      0.88      9824
```



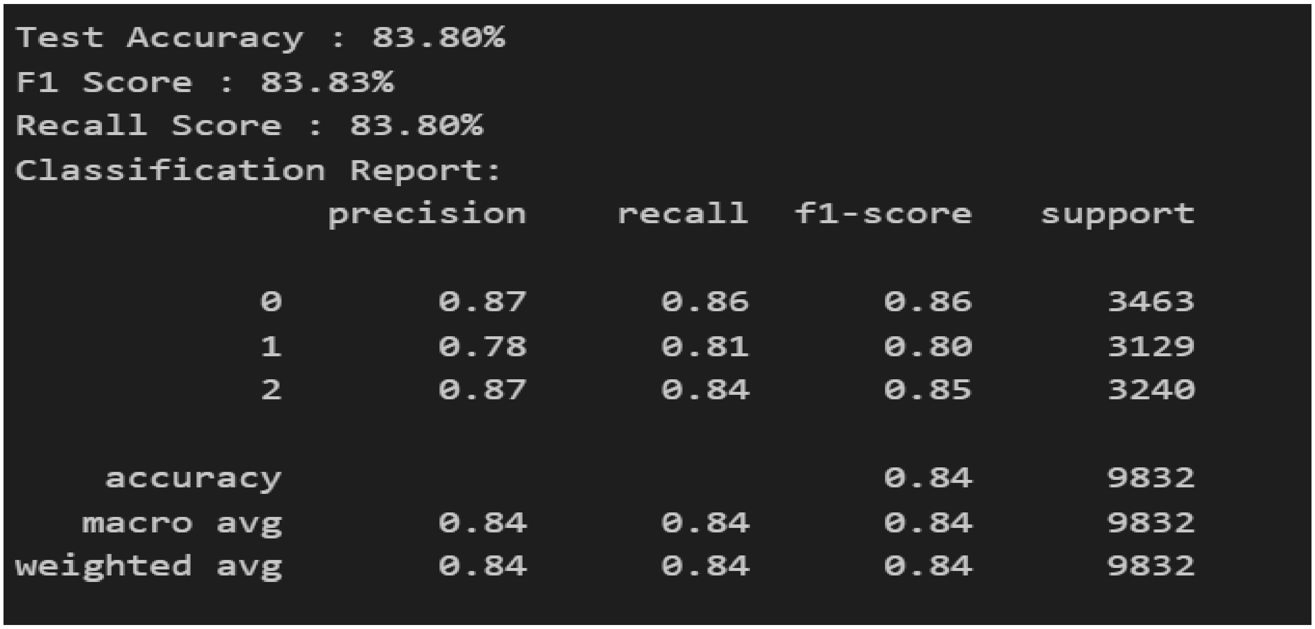
## 2. MNLI Dataset

**Overview:**
Accuracy: The model achieved an overall test accuracy of 83.80%, with closely matching precision, recall, and F1 scores across the three NLI classifications: entailment, neutral, and contradiction.

**Confusion Matrix Analysis:**
Entailment (0): The model correctly predicted 2990 out of 3463 cases, demonstrating a precision of 0.87 and a recall of 0.86. Misclassifications were mostly with the neutral class.
Neutral (1): Out of 3129 neutral cases, 2539 were accurately identified, yielding a precision of 0.78 and a recall of 0.81. Misclassifications were spread between entailment and contradiction.
Contradiction (2): Showed a strong performance with 2710 out of 3240 correctly predicted, resulting in both precision and recall around 0.87 and 0.84, respectively.

**Performance Metrics:**
The model shows balanced effectiveness across categories with F1 scores reflecting consistent performance in classifying complex sentence relationships. The slightly lower precision in the neutral category suggests potential areas for tuning to reduce false positives.

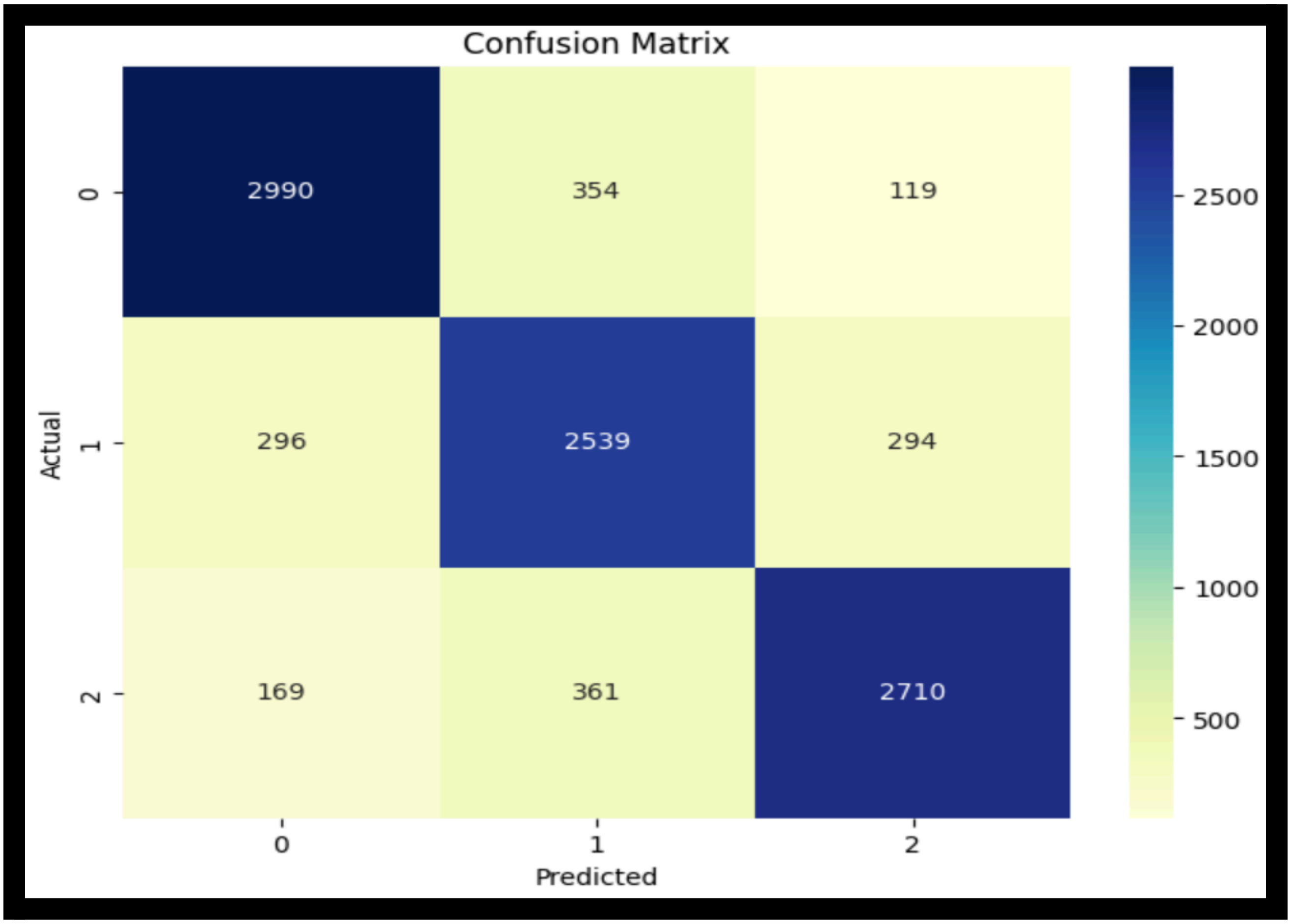
**Conclusion:**
The application of PEFT with LoRA on the MNLI dataset underscores the model's capability in handling diverse linguistic contexts and effectively distinguishing between varied logical relations in sentences.



```
Test Accuracy : 83.80%
F1 Score : 83.83%
Recall Score : 83.80%
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.86      0.86      3463
           1       0.78      0.81      0.80      3129
           2       0.87      0.84      0.85      3240

    accuracy                           0.84      9832
   macro avg       0.84      0.84      0.84      9832
weighted avg       0.84      0.84      0.84      9832
```



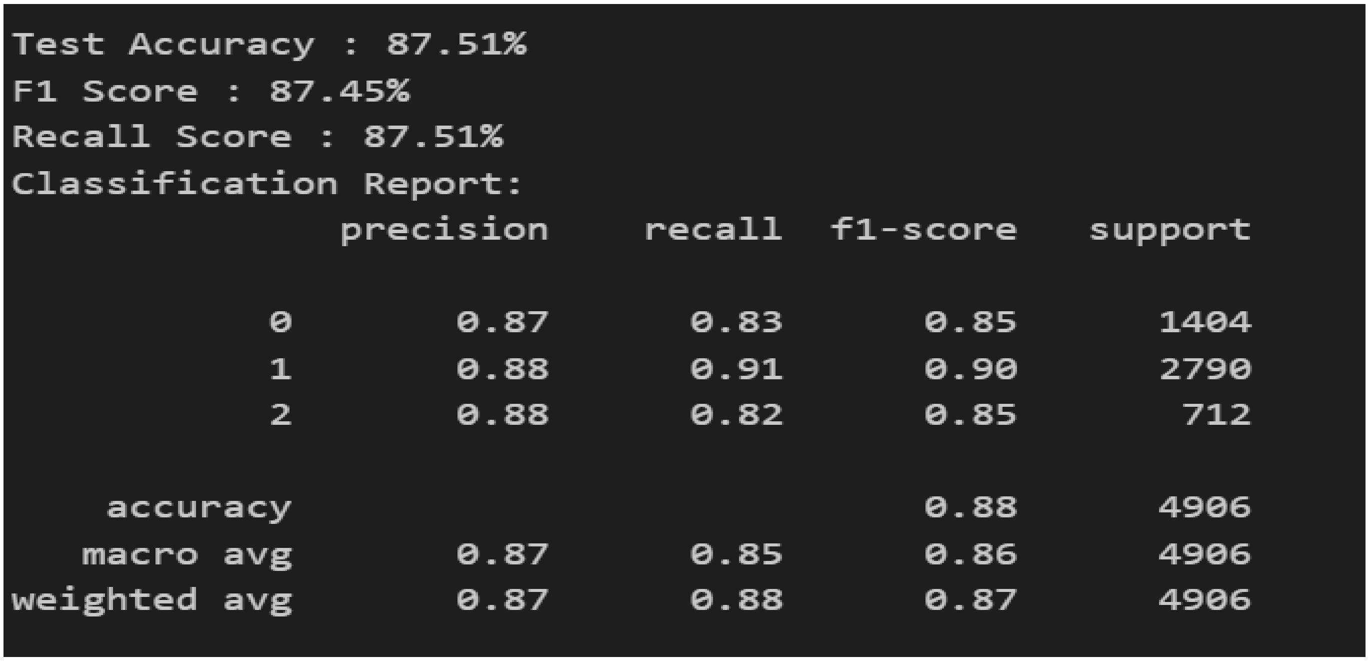
### 3. SICK Dataset

**Overview:**
Model Performance: The model achieves an excellent overall accuracy of 87.51%, with high precision, recall, and F1 scores across three classes: entailment (0), neutral (1), and contradiction (2).

**Confusion Matrix Analysis:**
Class 0 (Entailment): The model correctly identified 1161 out of 1404 entailment cases, demonstrating a precision of 0.87 and a recall of 0.83.
Class 1 (Neutral): Out of 2790 neutral cases, 2547 were accurately predicted. This results in a precision of 0.88 and the highest recall of 0.91.
Class 2 (Contradiction): For contradiction, the model correctly predicted 585 out of 712 cases, showing a precision of 0.88 and a recall of 0.82.

**Performance Metrics:**
Precision Average: 0.87
Recall Average: 0.85
F1 Score: 0.86

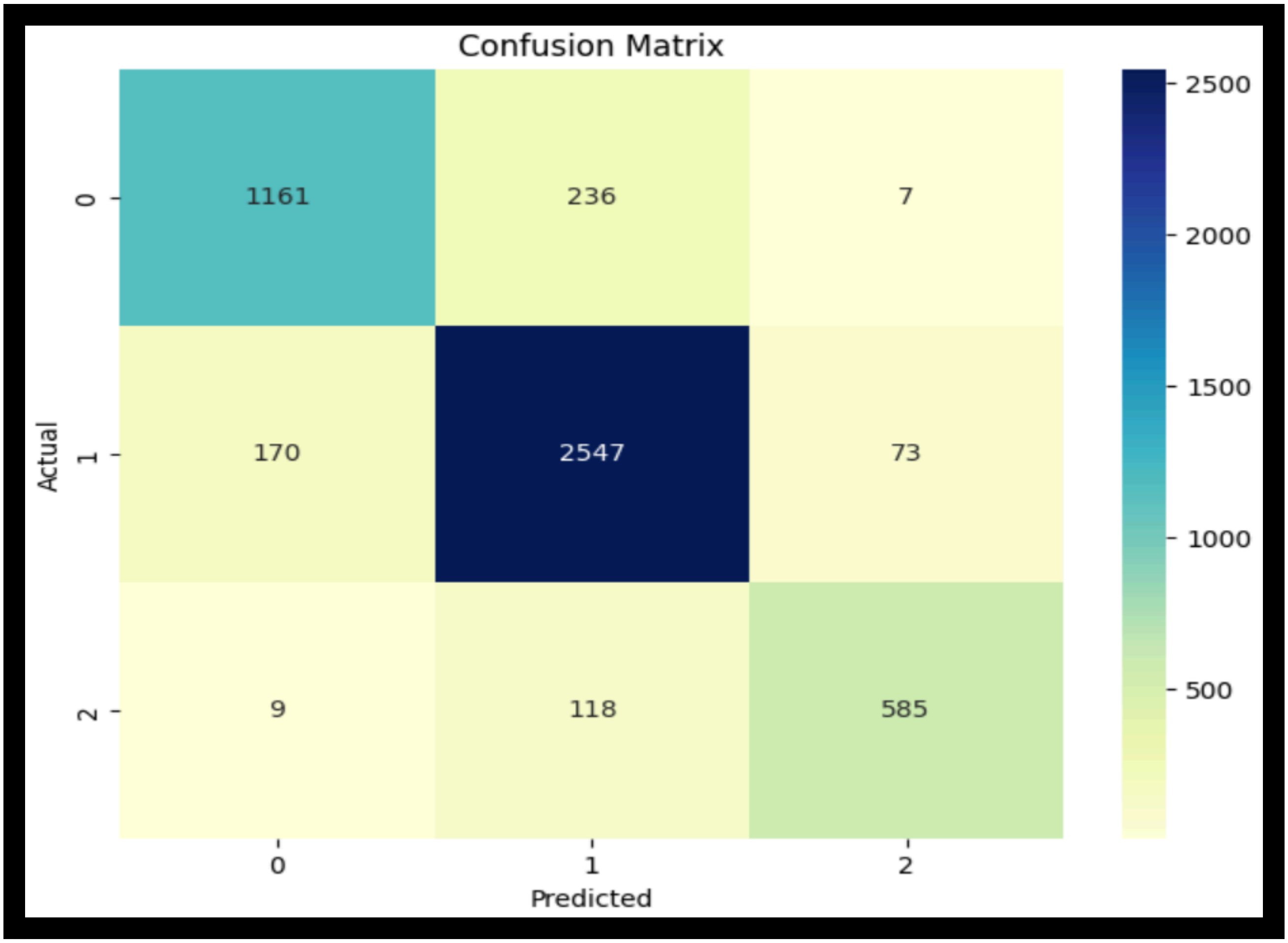
**Conclusion:**
The PEFT with LoRA model effectively distinguishes between different types of sentence relationships within the SICK dataset, showing a particularly strong ability in identifying neutral statements. The high performance across metrics indicates a well-tuned model capable of handling the complexities of NLI tasks.



```
Test Accuracy : 87.51%
F1 Score : 87.45%
Recall Score : 87.51%
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.83      0.85      1404
           1       0.88      0.91      0.90      2790
           2       0.88      0.82      0.85       712

    accuracy                           0.88      4906
   macro avg       0.87      0.85      0.86      4906
weighted avg       0.87      0.88      0.87      4906
```



## Overall Observations:

- The overall analysis shows performance metrics of various NLI models across three datasets: SNLI, Multi-NLI, and SICK.
- Traditional Models:
    - Logistic Regression and BiLSTMs show lower performance compared to more advanced transformer models, illustrating the advantage of deeper, contextual models for NLI tasks.
- Transformer Models:
    - BERT, RoBERTa, and ALBERT show significantly higher performance, with RoBERTa generally leading.
    - Enhanced strategies such as Max Epoch Selection for BERT, RoBERTa, and ALBERT show incremental improvements in accuracy.
    - RoBERTa with PEFT (Parameter Efficient Fine Tuning) slightly underperforms compared to its standard model, suggesting that for some datasets, straightforward implementations might be more effective.
- Ensemble Models:
    - Ensembles, particularly those using Max Voting (EWMV) and Epoch Max and Ensemble Aggregation (EMEA), significantly outperform individual models, with combinations involving RoBERTa generally showing the best results.
    - The EMEA strategy with RoBERTa + ALBERT shows the highest performance on the all dataset .

## Conclusion:

The use of ensemble techniques and advanced transformer models significantly enhances NLI task performance, confirming the value of combining multiple models' strengths. The top-performing ensemble configurations, particularly those involving RoBERTa and ALBERT, are highly effective across all datasets, making them preferable choices for robust NLI applications.