***Performance Evaluation Report***

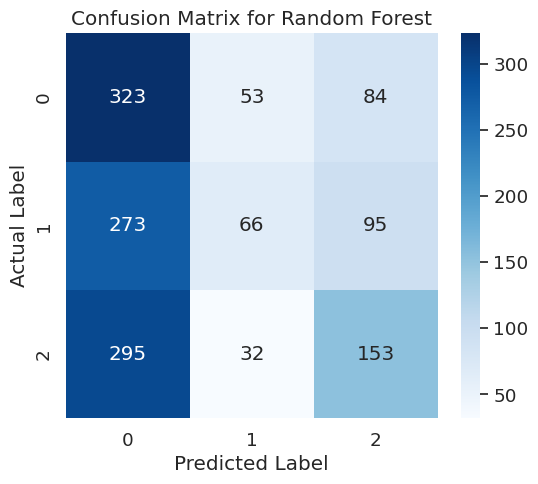
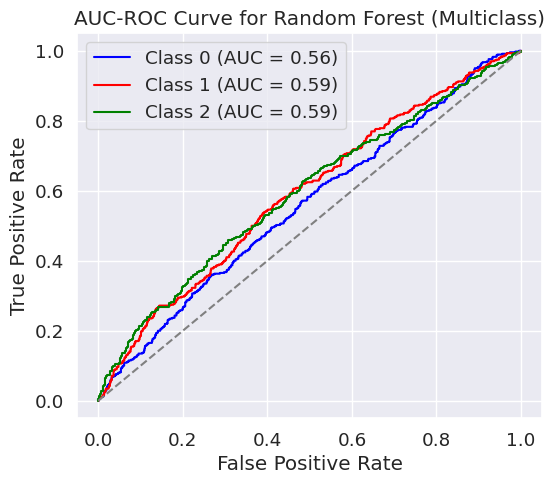
This report evaluates the performance of ML, DL and pre-defined models for classifying sentence contradiction across different languages. Here, 0 =

Each model's performance is assessed based on accuracy, precision, recall, and F1-score. The results are then analyzed to determine the most effective model for multilingual classification tasks. Moreover, confusion matrix and the AUC-ROC curve are also plotted for visualizing the model performance more transparently.

1. **Random Forest:**

Random Forest is an ensemble learning method that uses multiple decision trees to improve classification performance. Hyperparameter tuning was performed for each language, with varying results across datasets. The average accuracy is 34.13%.

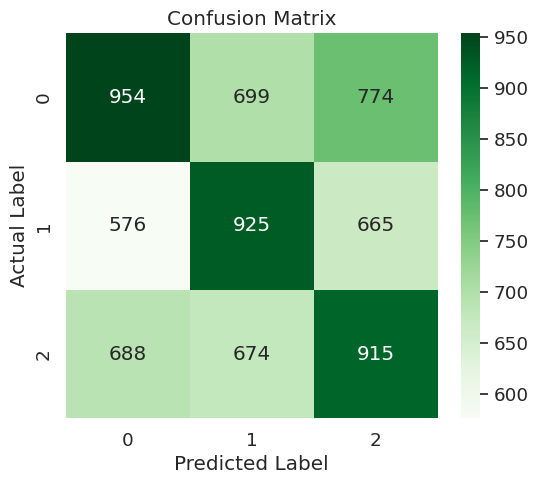
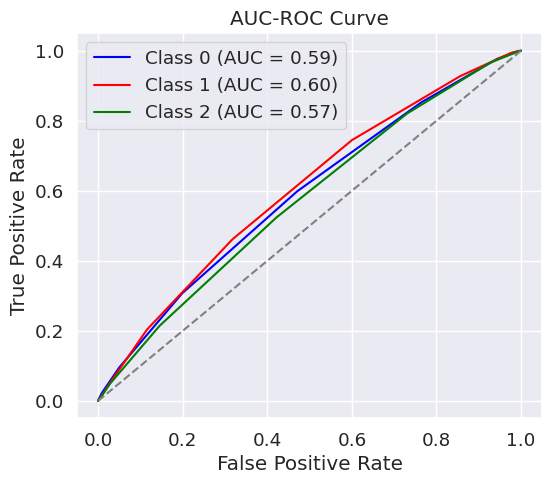
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-score | Support |
| 0 | 0.36 | 0.70 | 0.47 | 460 |
| 1 | 0.40 | 0.15 | 0.22 | 434 |
| 2 | 0.46 | 0.30 | 0.36 | 480 |
| Macro Avg | 0.41 | 0.38 | 0.35 | 1374 |
| Weighted Avg | 0.41 | 0.38 | 0.35 | 1374 |



2. **XGBoost:**

XGBoost is a gradient boosting framework known for its high efficiency and predictive accuracy. After hyperparameter tuning, XGBoost showed improved performance over Random Forest, with an average accuracy of 38.8% across languages.

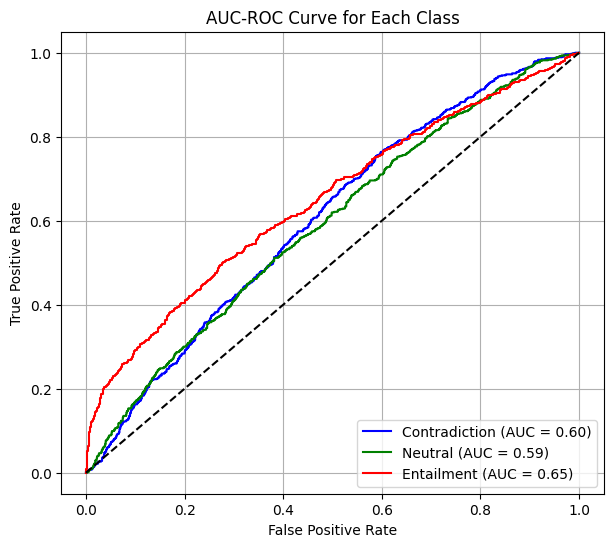
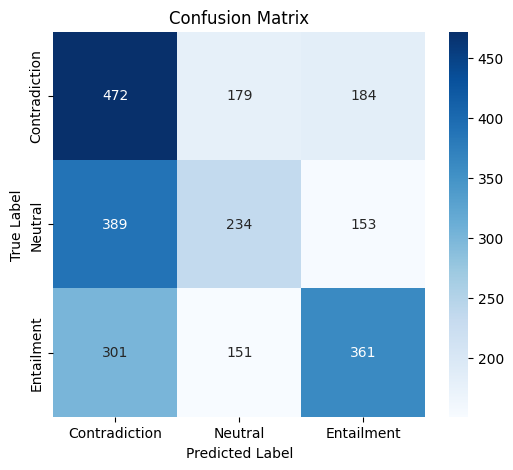
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-score | Support |
| 0 | 0.39 | 0.58 | 0.47 | 460 |
| 1 | 0.42 | 0.35 | 0.38 | 434 |
| 2 | 0.55 | 0.37 | 0.44 | 480 |
| Macro Avg | 0.45 | 0.43 | 0.43 | 1374 |
| Weighted Avg | 0.45 | 0.43 | 0.43 | 1374 |



3. **Bi-LSTM:**

Bi-LSTM (Bidirectional Long Short-Term Memory) is a recurrent neural network-based model that captures long-range dependencies in text data. Its accuracy for the multilingual dataset was 44%.

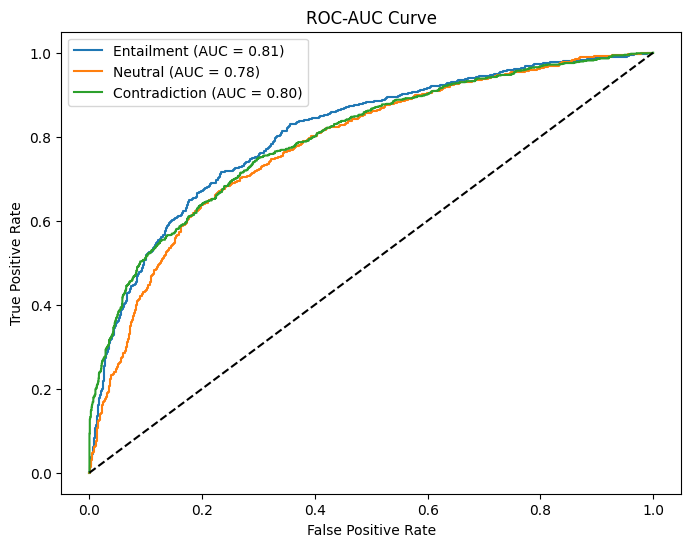
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-score | Support |
| 0 | 0.41 | 0.57 | 0.47 | 835 |
| 1 | 0.41 | 0.30 | 0.35 | 776 |
| 2 | 0.52 | 0.44 | 0.48 | 813 |
| Macro Avg | 0.45 | 0.44 | 0.43 | 2424 |
| Weighted Avg | 0.45 | 0.44 | 0.43 | 2424 |

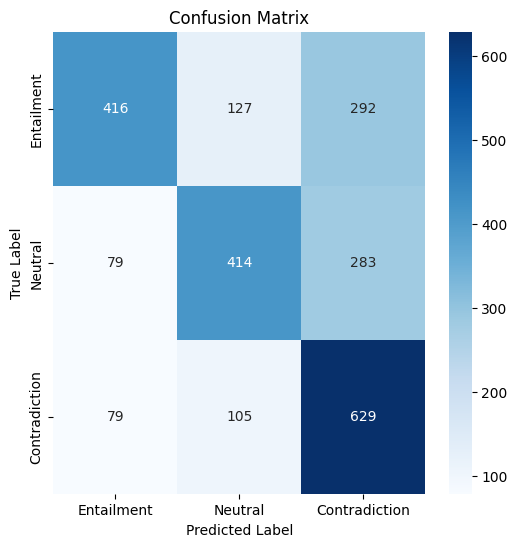


4. **m-BERT:**

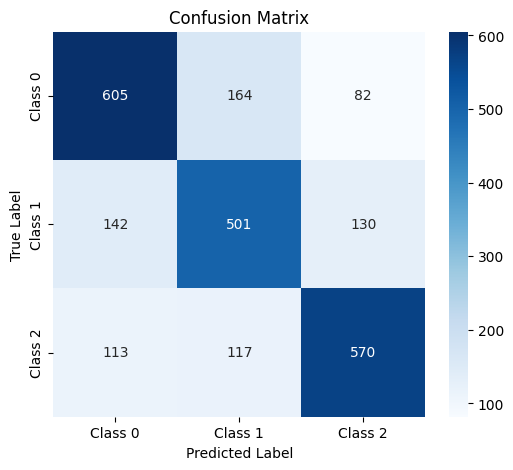
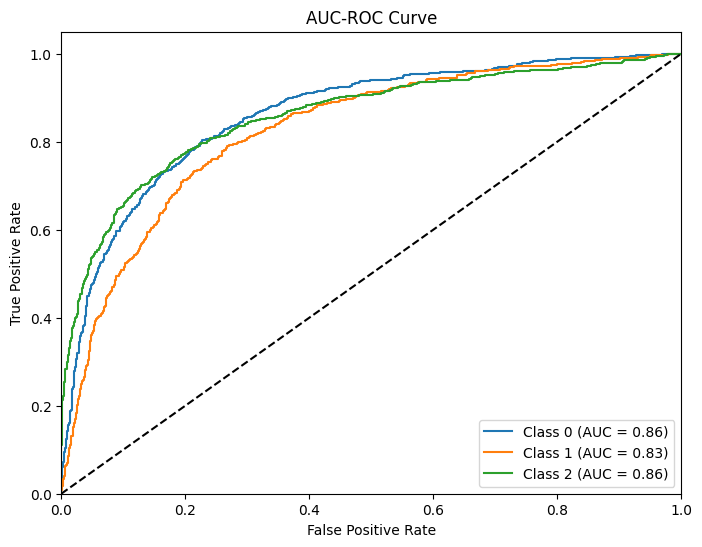
Multilingual BERT (m-BERT) is a transformer-based model designed for multilingual natural language processing (NLP) tasks. It demonstrated an accuracy of 62.54%, significantly outperforming tree-based models and Bi-LSTM.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-score | Support |
| 0 | 0.66 | 0.63 | 0.65 | 835 |
| 1 | 0.61 | 0.57 | 0.59 | 776 |
| 2 | 0.60 | 0.67 | 0.64 | 813 |
| Macro Avg | 0.63 | 0.62 | 0.62 | 2424 |
| Weighted Avg | 0.63 | 0.63 | 0.63 | 2424 |





**5. XLM-R:**

XLM-R (Cross-lingual Language Model - RoBERTa) is an advanced transformer model pre-trained on large-scale multilingual data. It achieved the highest accuracy of 69.14%, making it the most effective model for this task.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-score | Support |
| 0 | 0.70 | 0.71 | 0.71 | 851 |
| 1 | 0.64 | 0.65 | 0.64 | 773 |
| 2 | 0.73 | 0.71 | 0.72 | 800 |
| Macro Avg | 0.69 | 0.70 | 0.69 | 2424 |
| Weighted Avg | 0.69 | 0.69 | 0.69 | 2424 |

Overall, the result indicate that transformer-based models (m-BERT and XLM-R) significantly outperform traditional machine learning models (RF and XGBoost) as well as recurrent networks (Bi-LSTM). XLM-R achieves the best accuracy, making it the most suitable model for multilingual text classification tasks.