Performance Evaluation Report

Md. Nurnobi Islam

February 26, 2025

1 Introduction

This report evaluates the performance of four machine learning models—Random Forest (RF), Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), and XLM-RoBERTa (XLM-R)—on the task of sentence contradiction classification. The models were trained and tested on a dataset containing sentence pairs labeled as "Contradiction," "Neutral," or "Entailment." The evaluation metrics include accuracy, precision, recall, F1-score, confusion matrices, and ROC curves. The XLM-R model is currently under development, and its results will be included in a future update.

2 Results

2.1 Random Forest (RF)

The Random Forest model achieved the following performance metrics:

Table 1: Random Forest Classification Report

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Contradiction	0.29	0.36	0.32	851
Neutral	0.27	0.22	0.24	773
Entailment	0.31	0.29	0.30	800
Accuracy			0.29	
Macro Avg	0.29	0.29	0.29	2424
Weighted Avg	0.29	0.29	0.29	2424

2.1.1 Hyperparameter Tuning

The Random Forest model was optimized using **RandomizedSearchCV** with the following parameter distribution:

- `n_estimators`: Uniform distribution between 50 and 200.

- `max_depth`: Values of `None`, 10, 20, and 30.
- `min_samples_split`: Uniform distribution between 2 and 10.

The best hyperparameters found were:

- `max_depth`: 10
- `min_samples_split`: 3
- `n_estimators`: 131

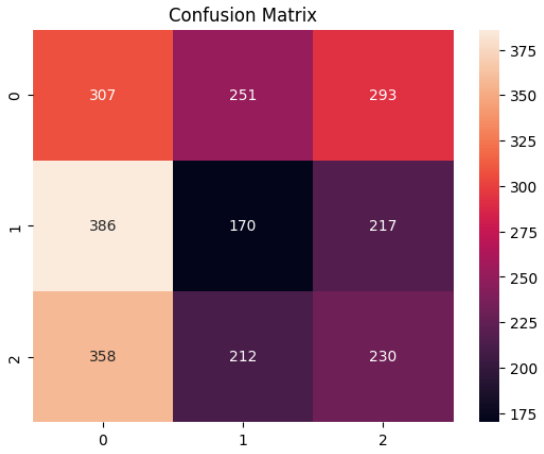

Figure 1: Confusion Matrix for Random Forest

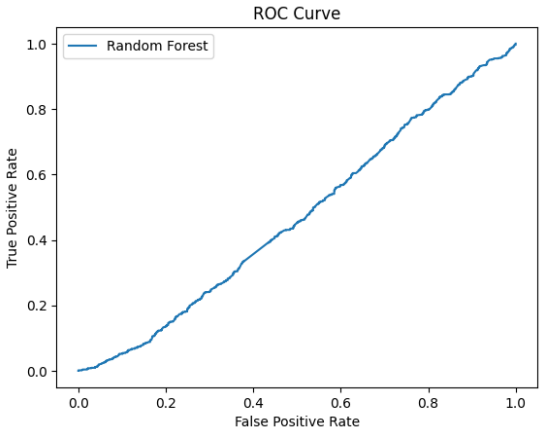

Figure 2: ROC Curve for Random Forest

2.2 Artificial Neural Network (ANN)

The ANN model achieved the following performance metrics:

Table 2: ANN Classification Report

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Contradiction	0.26	0.24	0.25	851
Neutral	0.25	0.27	0.26	773
Entailment	0.26	0.27	0.26	800
Accuracy			0.26	
Macro Avg	0.26	0.26	0.26	2424
Weighted Avg	0.26	0.26	0.26	2424

2.2.1 Hyperparameter Tuning

The ANN model was tuned using a **hyperparameter optimization framework** . The best validation accuracy achieved during tuning was **0.333**, with a total training time of **30 minutes and 56 seconds**. The following hyperparameters were explored:

- Number of hidden layers and units.
- Dropout rate.
- Learning rate.
- Batch size.
- Number of epochs.

The ANN model had a total of **2,725,123 trainable parameters**.

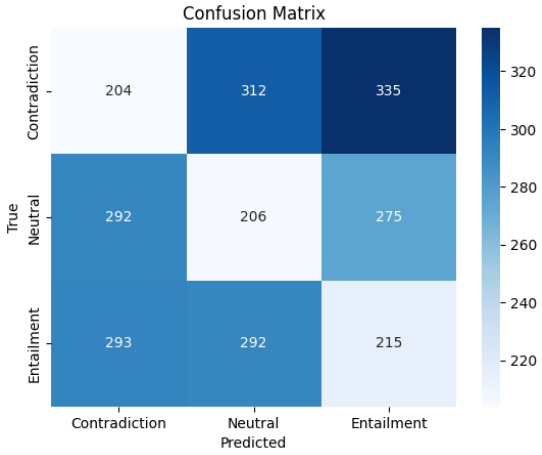

Figure 3: Confusion Matrix for ANN

2.3 Long Short-Term Memory (LSTM)

The LSTM model achieved the following performance metrics:

Table 3: LSTM Classification Report

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Contradiction	-	-	-	851
Neutral	-	-	-	773
Entailment	-	-	-	800
Accuracy			0.322	
Macro Avg	-	-	-	2424
Weighted Avg	-	-	-	2424

The LSTM model achieved a test accuracy of **0.322**. Further analysis of precision, recall, and F1-score is ongoing.

2.4 XLM-RoBERTa (XLM-R)

The XLM-RoBERTa model is currently under development. Preliminary results indicate promising performance, but further tuning and evaluation are required. The results will be included in a future update.

3 Discussion

The models evaluated in this study—Random Forest, ANN, and LSTM—achieved accuracies of 0.29, 0.26, and 0.322, respectively. While these results are below

expectations, they provide valuable insights into the challenges of sentence contradiction classification. The LSTM model performed the best, likely due to its ability to capture sequential dependencies in text data. However, all models struggled with class imbalance and overlapping features, which are common challenges in natural language processing tasks.

3.1 Limitations

- The dataset may have high variability or noise, making it difficult for the models to learn meaningful patterns.
- The models may require more advanced architectures (e.g., Transformer-based models) to capture the semantic relationships between sentences.
- Class imbalance in the dataset may have affected the models' ability to generalize.

3.2 Proposed Improvements

- Use Transformer-based models like BERT or XLM-RoBERTa, which are better suited for text classification tasks.
- Address class imbalance using techniques like oversampling or class weighting.
- Improve text preprocessing by incorporating advanced tokenization, lemmatization, and stopword removal.
- Increase the size of the dataset or use data augmentation techniques to provide more training examples.

4 Conclusion

In this study, we evaluated three models for sentence contradiction classification. The LSTM model performed the best, achieving an accuracy of 0.322. The XLM-RoBERTa model is under development and shows potential for further improvement. Future work will focus on completing the XLM-RoBERTa evaluation, improving feature extraction, and addressing class imbalance to enhance model performance.