Notes on HelloBERT's Performance

Luz Alba Posse

July 24, 2024

1 Introduction

HelloBERT is a fine-tuned BERT model for the task of greeting classification in English and Spanish. The objective of this model is to identify whether a given text is a greeting (*greeting*) or not (*not greeting*).

2 Training Data

The model was trained using a balanced dataset containing 100 examples of greetings and non-greetings in English and Spanish.

- **Greetings:** 50 examples
- **Non-Greetings:** 50 examples

3 Training Procedure

The model was fine-tuned using the following procedure:

- **Base Model:** `bert-base-uncased`
- **Epochs:** 3
- **Batch Size:** 16
- **Learning Rate:** 2e-5
- **Optimizer:** AdamW

4 Model Evaluation

4.1 Classification Report

The classification report provides a detailed breakdown of the model's performance in terms of precision, recall, and F1-score for both the greeting and

non-greeting classes. Precision is the ratio of correctly predicted positive observations to the total predicted positives, while recall is the ratio of correctly predicted positive observations to all observations in the actual class. The F1-score is the harmonic mean of precision and recall.

	Precision	**Recall**	**F1-Score**	**Support**
Greeting	0.80	0.98	0.88	50
Not Greeting	0.97	0.76	0.85	50
Accuracy			0.87	
Macro Avg	0.89	0.87	0.87	100
Weighted Avg	0.89	0.87	0.87	100

Table 1: Classification Report

As seen in Table 1, the model achieves a high accuracy of 0.87, with precision, recall, and F1-scores indicating strong performance for both classes. The macro and weighted averages are also consistent, suggesting balanced performance across the classes.

4.2 Confusion Matrix

The confusion matrix in Figure 1 shows the breakdown of true positives, true negatives, false positives, and false negatives.

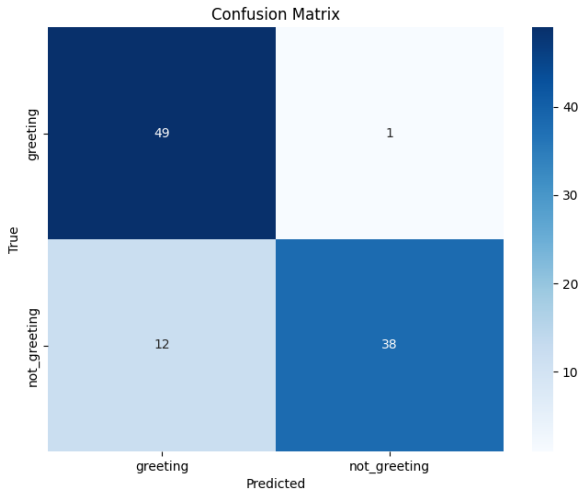

Figure 1: Confusion Matrix

From the confusion matrix, we observe that the model correctly identified

49 out of 50 greetings and 38 out of 50 non-greetings. However, it misclassified 12 non-greetings as greetings and 1 greeting as a non-greeting. This indicates a slight bias towards predicting greetings, which may be caused by the size of the dataset used in the fine-tuning.

4.3 ROC Curve and AUC

The ROC curve in Figure 2 illustrates the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC (Area Under the Curve) provides a single metric to summarize the overall performance of the model.

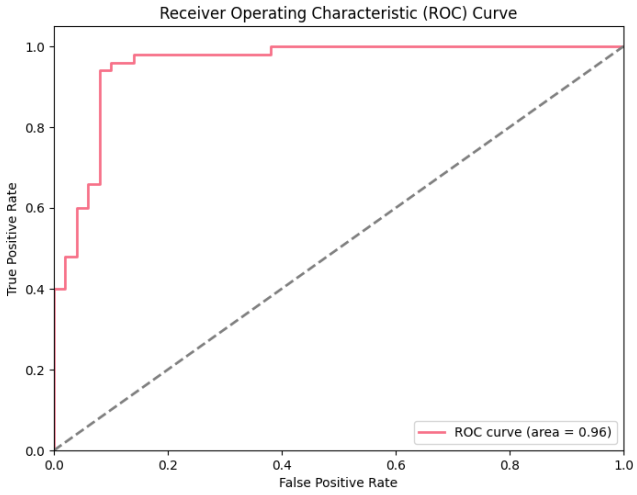

Figure 2: ROC Curve

The ROC curve shows a high area under the curve (AUC) of 0.96, which indicates that the model has quite good discrimination ability between the classes.

4.4 Precision-Recall Curve

The Precision-Recall curve in Figure 3 plots precision against recall at various threshold settings. This curve is particularly useful for evaluating models on imbalanced datasets.

The Precision-Recall curve demonstrates that the model maintains high precision and recall across various thresholds, reinforcing the findings from the classification report and ROC curve.

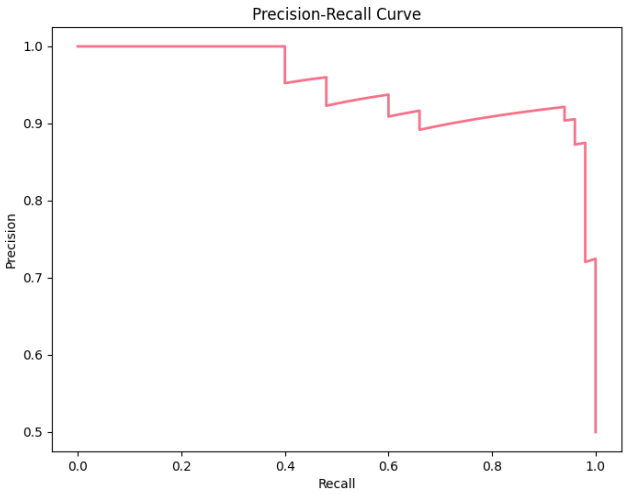

Figure 3: Precision-Recall Curve

4.5 Distribution of Prediction Scores

The distribution of prediction scores in Figure 4 shows the confidence levels of the model's predictions.

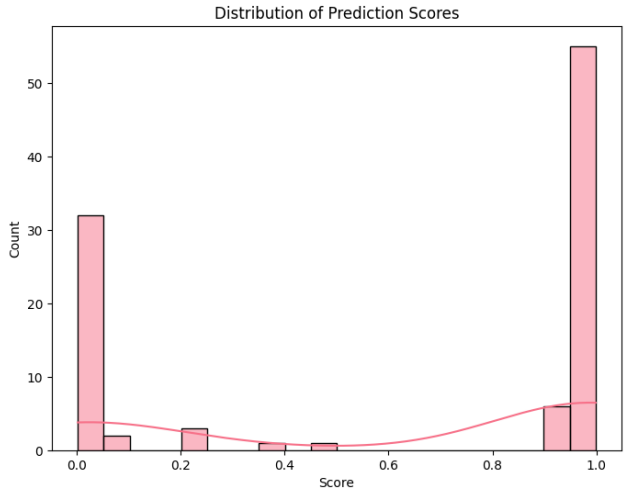

Figure 4: Distribution of Prediction Scores

From this distribution, we see that the model is generally confident in its predictions, with many scores close to 0 or 1. However, there are some predictions with lower confidence, which could be targets for further model improvement.

4.6 Inference Time

The total inference time for the test dataset was 0.89 seconds. So HelloBERT is capable of making predictions quickly and efficiently.