Project 2

Jigsaw Unintended Bias in Toxicity Classification

Abstract:

This is a Kaggle competition hosted by Jigsaw and google to identify toxic comment in online conversation. The dataset contains comments and relevant target in range 0 to 1. And several other features for further research. To build the classifier, the dataset is cleaned, feature engineered, built the model, train the model, and finally evaluated.

Methodology:

1. Data Cleaning:
   - For the training “comment_text” and “target” features were relevant. So, other columns are dropped.
   - Contraction is applied, punctuation, special sign, emoji’s are removed to improve the data quality.
2. Tokenizing and Embedding:
   - The training and testing data is tokenized and padded to have same length.
   - FastText and QuoraText embedding are combined, and fitted on tokenized dataset to create the embedding of the words.
3. Splitting the dataset:
   - The tokenized input variable and target variable is splitted for training and testing.
4. Model Building:
   - The model is built using Embedding layer, Bidirectional LSTM layer and Dense layer.
5. Model training:
   - The model is trained for 20 epoch with batch size of 248.
   - Early stopping and Reduce Learning rate is used to stop overfitting.
   - Model Checkpoint is used to save the best model.

Result Analysis:

- The training accuracy is: 0.9638 and loss is 0.0955
- The validation accuracy is: 0.96380 and loss is 0.095541
- The model seems converged well on the first epoch, the started overfitting. Reducing learning rate haven't helped much.

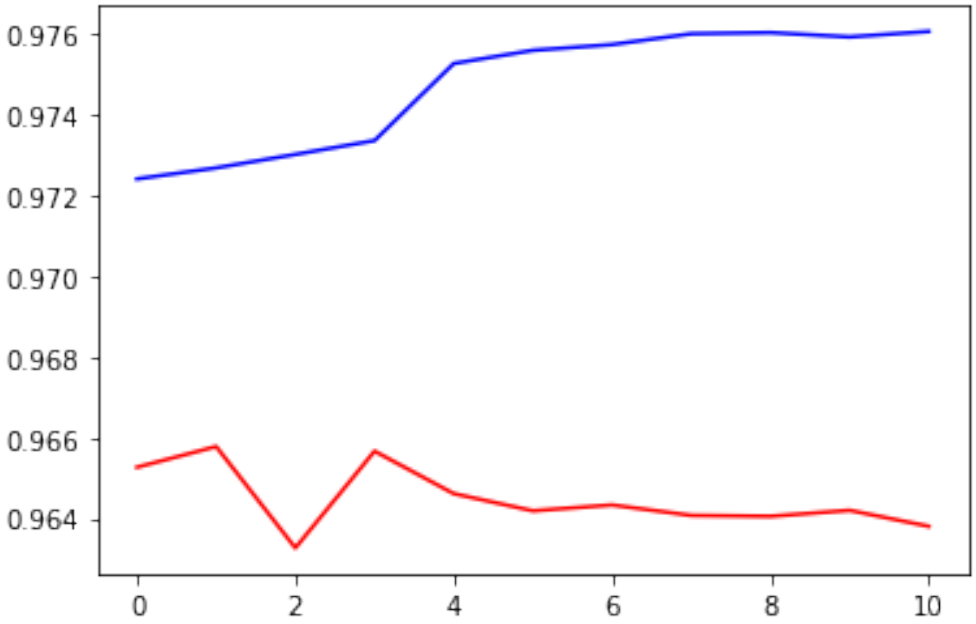

Fig 1: Training vs validation accuracy

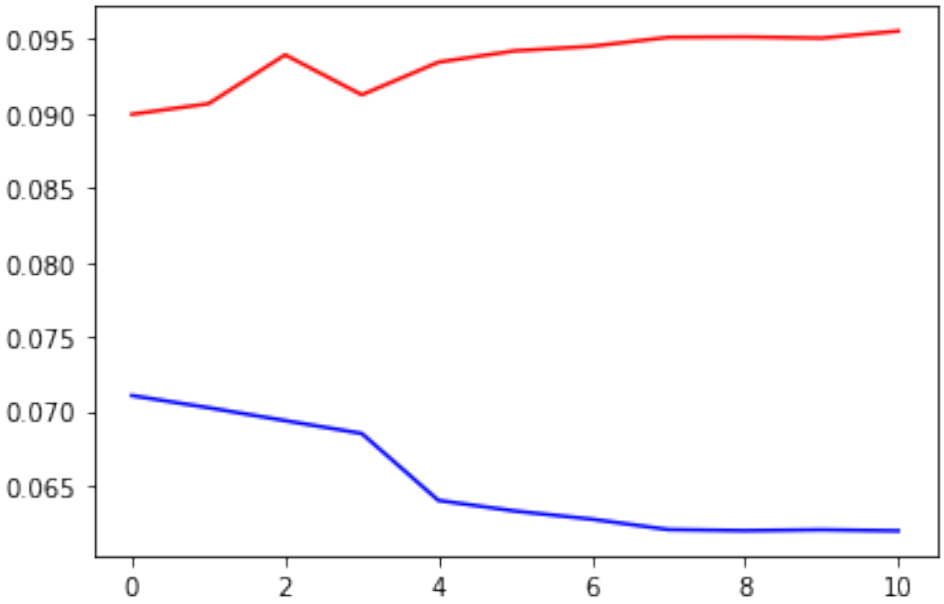

Fig 2: Training vs validation loss

Conclusion:

From the result analysis, it is clear that the model haven't overfitted. Early stopping and model checkpoint helped in the issue. The private score on Kaggle was 0.92284, which is pretty decent.