PROBLEM: MODEL OVERFITS; YIELDS LOW VALIDATION METRICS

When overfitting was presented, model experimentation and tuning hyper parameters were all tried. None of this helped as the validation accuracy didn't go above 80%. A hypothesis was that the training data was too small. The training data has a size of 2,885. A dataset on Kaggle was found to have satisfactory financial news headlines. This dataset was combined with the already present training data. This proved to no avail with the current model. After consulting with Sammy "Pepsipu" Hajhamid, it was suggested to use a pre-trained model to test if the problem routes within the code or the training data. With the BERT model, 3 tests were completed. Test I contains the BERT model with the training data plus the Kaggle dataset. Test II contains the BERT model with only the training data. Test III contains the BERT model with only the Kaggle dataset

Test I

With the combined dataset of the harvested training data and the Kaggle dataset, the best validation accuracy was about 81%. This is a little above average from what was seen with the trained model that is under development.

```
Epoch 1/5
266/266 [==============================] - 131s 438ms/step - loss: 0.5050 - accuracy: 0.7436 - val_loss: 0.4839 - val_accuracy: 0.7665
Epoch 2/5
266/266 [==============================] - 115s 431ms/step - loss: 0.1830 - accuracy: 0.9259 - val_loss: 0.6329 - val_accuracy: 0.7844
Epoch 3/5
266/266 [==============================] - 113s 426ms/step - loss: 0.0552 - accuracy: 0.9797 - val_loss: 0.7758 - val_accuracy: 0.8051
Epoch 4/5
266/266 [==============================] - 115s 434ms/step - loss: 0.0302 - accuracy: 0.9894 - val_loss: 0.9752 - val_accuracy: 0.8126
Epoch 5/5
266/266 [==============================] - 126s 472ms/step - loss: 0.0236 - accuracy: 0.9921 - val_loss: 0.9935 - val_accuracy: 0.7900
```

Test II

This test removed the Kaggle dataset and only contains the harvested training data. This proved very similar to the current training model. This is a lower accuracy than Test I suggesting that the training data may be the problem.

```
...  Epoch 1/5
     146/146 [==============================] - 75s 412ms/step - loss: 0.6065 - accuracy: 0.6811 - val_loss: 0.5785 - val_accuracy: 0.7123
     Epoch 2/5
     146/146 [==============================] - 61s 420ms/step - loss: 0.2718 - accuracy: 0.8932 - val_loss: 0.7007 - val_accuracy: 0.7990
     Epoch 3/5
     146/146 [==============================] - 59s 406ms/step - loss: 0.0879 - accuracy: 0.9701 - val_loss: 0.9281 - val_accuracy: 0.7834
     Epoch 4/5
     146/146 [==============================] - 59s 404ms/step - loss: 0.0399 - accuracy: 0.9864 - val_loss: 1.0883 - val_accuracy: 0.7799
     Epoch 5/5
     146/146 [==============================] - 63s 433ms/step - loss: 0.0224 - accuracy: 0.9926 - val_loss: 1.2778 - val_accuracy: 0.7764
```

Test III

This test involved only the Kaggle dataset. This produced the best results with about 90% validation accuracy. This is very acceptable for production and suggests that the Kaggle dataset is the best dataset to use.

```
Epoch 1/5
122/122 [==============================] - 71s 466ms/step - loss: 0.3789 - accuracy: 0.8284 - val_loss: 0.2745 - val_accuracy: 0.8937
Epoch 2/5
122/122 [==============================] - 55s 451ms/step - loss: 0.0817 - accuracy: 0.9685 - val_loss: 0.5123 - val_accuracy: 0.8773
Epoch 3/5
122/122 [==============================] - 47s 389ms/step - loss: 0.0371 - accuracy: 0.9898 - val_loss: 0.3710 - val_accuracy: 0.9039
Epoch 4/5
122/122 [==============================] - 49s 402ms/step - loss: 0.0140 - accuracy: 0.9951 - val_loss: 0.5112 - val_accuracy: 0.9080
Epoch 5/5
122/122 [==============================] - 48s 397ms/step - loss: 0.0115 - accuracy: 0.9969 - val_loss: 0.5618 - val_accuracy: 0.8937
```

Results

The results of these tests prove that the best results were produced by Test III, which contains only the Kaggle dataset. This suggests that the root of the issue is with the training data.

Conclusion

From these tests, it was deduced that the training data was “broken”. After a careful look, the training data appeared to contain blogs and “junk” headlines. This is the reason for the poor validation metrics. A new harvester system will have to be devised to only take useful headlines such as only press releases and form filings.