**What is your baseline model's overall accuracy for the Test (PRIMARY) and Test (SECONDARY) data sets?**

**Our baseline model accuracies are following:**

* **Model Training Accuracy on Test Primary Data**
* **0.737**
* **Model Scores on Test Secondary Data**
* **0.606**

**What is your enhanced (with new engineered features) model's overall accuracy for the Test (PRIMARY) and Test (SECONDARY) data sets?**

**Our baseline model accuracies are following after feature engineering:**

* **Model Training Accuracy on Test Primary Data**
* **0.737**
* **Model Scores on Test Secondary Data**
* **0.65**

**What new features did you add to your model?**

We added the following feature to the data.

* Sentiment: 1 for positive and 0 for negative sentiment
* Cluster: We divided our text into two clusters and assigns a value of 1 if a specific text falls into the first cluster and assigns a value of 0 if it falls into the second cluster.
* Emotion: We apply emotion analysis to the text and divided it into two emotions joy and fear. There is 1 for joy and 0 for fear.

**Did your features improve model performance? Why do you think they did (or did not)?**

No, our model accuracy increased by just 0.5 times. This is a neglectable improvement. In my view, our new feature has no impact on the target feature. There are no direct or logical relations between dependent and independent features. So, this is why our model did not show noticeable improvement.

**Did your model have better accuracy on the Test (PRIMARY) or Test (SECONDARY) data set? Why?**

Our model has the best accuracy on primary data because our model is somehow aware of the trend of input features. On the opposite side, the secondary dataset is naïve to our model and even our model did not see any instance of secondary data before testing. Secondly, the secondary dataset is quite large than the primary data. Also, we did use only 70% of primary data for model training keeping in mind primary data is smaller than secondary data.  Therefore, there is almost a 13% gap between the testing accuracies of primary and secondary test data.