* **Over All Observations:**
  + For some of the bigrams, the MLE score was zero, which in turn gave a zero score for some of the tweets. To overcome this issue, I used smoothening to substitute one for zero.
  + Without substituting the zero, all three models’ scores were zero for every tweet.
  + Among all the three LM models, LM\_off seems to perform well on the test data with the highest average score of 0.09145 for the offensive subset.
  + The training and test set both have around 70 percent instances of non-offensive as compared to the offensive.

Below are the individual observations of all three LM models for each test set:

* **LM FULL** with Test data:

**Average score - full\_TestData:** 0.017621986976653608

**Average score - not\_off\_TestData:** 0.020469582983241183

**Average score - off\_TestData:** 0.010265697292969044

**Observations:**

* LM model is trained with a full training dataset and tested over test data. By observing the average score of each test dataset, non-offensive test data has the highest score. It means the model analyzed the language correctly because the full training dataset has around 8840 non-offensive instances out of 13240.
* **LM Non-offensive** with Test data

**Average score - full\_TestData:** 0.021027589036362072

**Average score - not\_off\_TestData:** 0.02161883956946718

**Average score - off\_TestData:** 0.019500191825840556

**Observations:**

* By observing the average score of each test data against the non-offensive LM model, it’s evident that the LM model captures well over non-offensive test data with an average score of 0.0216 as compared to other test data.
* Full and non-offensive test datasets’ average scores do not have a big difference, which is somewhat acceptable because as we know the trained model is non-offensive, and in the test dataset there is around 70 percent of instances belong to the non-offensive language.
* **LM Offensive** with Test data

**Average score - full\_TestData:** 0.07068087385766639

**Average score - not\_off\_TestData:** 0.06263796733626759

**Average score - off\_TestData:** 0.09145838237127994

**Observations:**

* The offensive LM is trained by passing the offensive dataset only and tested over the test dataset. It turned out that the offensive test dataset has the highest average score among the non-offensive and full datasets. This shows the language model is decently capturing the offensive language.
* Second highest average score belongs to the full test set, which means the trained LM\_off model captured the language correctly. Since the test data has more instances of non-offensive language as compared to offensive, still it doesn’t affect the trained model pattern i.e., to capture offensive language.