1. **Data**

```
count: 31962 row, 3 col
id : int -> id of a post

label : int ->
                0 (no offense) # count -> 18217
                1 (offense)    # count -> 13745

tweet  : str ->  the post from a user
```

2. **Estimator**

```
Logistic Regression
    maxIter: int = 1000
    regParam: float = 0.2
    elasticNetParam: float = 0.0
    tol: float = 1e-6
    fitIntercept: bool = True
    threshold: float = 0.8
```

3. **Evaluation Metrics' Results**

```
  TP : 10766
  FP : 3962
  TN : 14255
  FN : 2979

  precision: 0.73
  recall:    0.78
  accuracy:  0.77
  f1:        0.75
```

It was preferable for us to minimize FP than FN. (model shouldn't incorrectly predict the positive class). Thus precision has more value for us than recall. For our case to condemn someone that he/she used hate speech (but actually didn't) is more dangerous, than to skip someone who did, but we predicted that didn't. (Like in court cases.) That is why the threshold is above 0.5 in Estimator's hyperparameters.

4. **Device & Computation Time**

```
CPU: 12 Intel Core i7-8750H 2.20 GHz
RAM: 15.5 GiB
GPU: NVIDIA GeForce GTX 2080 8 GB
```

```
Spark configs:
    max_workers = 12
    partition_count= 128
    max_memory_used=4 gb
```

```
Train (tT)  ->  5.2 sec
Predict (pT)  -> 0.01 sec
```