I started the Hackathon with a very simple approach of converting texts to numbers. Divided the complete hackathon in few simple stages:

1. Data Reading
2. Data PreProcessing
3. Text Analysis
4. Model Building

**Data Reading:**

Reading both test and train files to generate a combined dataframe.

**Data PreProcessing:**

Data Processing consists of following steps:

* Step A : Converting html entities
* Step B : Removing "@user" and "#words" from all the tweets
* Step C : Changing all the tweets into lowercase
* Step D : Apostrophe Lookup
* Step E : Short Word Lookup
* Step F : Emoticon Lookup
* Step G : Replacing Punctuations with space
* Step H : Removing extra white spaces
* Step I : Replacing Special Characters with space
* Step J :Removing all the words when http or pic comes in text
* Step K : Replacing Numbers (integers) with space
* Step L : Removing words whom length is 1
* Step M : Replacing mom alternatives with "mother"

**Text Analysis:**

This gives an in depth idea of what are the most frequently occurring words in each sentiment class. This also gives an idea what the most common words are all three sentiment classes are so that we can remove them.

**Model Building:**

Tried 3 different models :

1. XGBOOST
2. GBM
3. AdaBoostClassifier

And finally proceeded with GBM with some hyperparameter tuning.