* Business Problem
  + Sentiment Classification
    - Domain
      * Movie Reviews
  + Manager:
    - Questions
      * What about the Dataset?
      * Language?
      * Amount of Data?
      * How much time do we have?
      * How are we gonna serve this?
      * What type of movies?
      * Class details
    - Risk:
      * Accuracy depends on data quality?
      * Accuracy depends on data distribution?
      * Accuracy depends a lot on difference/similarity b/w labels.
* Data Science Team (Internal):
  + Analyze the Data (EDA: Exploratory Data Analysis)
    - Pre-Process
    - Explore
    - Avisualize
  + Baseline Model
    - Partition
    - Modelling
    - Evaluation
  + Fine Tuning Process
    - Feature Engineering
    - Algorithmic Exploration
    - Parameter & Hyperparameter Tuning
    - Evaluation
  + Deployment
    - Library
    - API
* EDA
  + Class Distribution
  + Dataset Distribution
  + Noise
    - Missing Value
  + Report
    - File is Tab Separated
    - Number of Records
      * 156060 (Train)
    - Number of Columns
      * 4
      * Phrase
      * Sentiment
        + 0 - Negative
        + 1 - Somewhat Negative
        + 2 - Neutral
        + 3 - Somewhat Positive
        + 4 - Positive
    - Class Imbalance
    - Duplicate Records
    - Text
      * Noise
        + Case Normalization
        + Word Different forms
      * Cleaning
        + Stopwords
        + Lemmatization

Done

* + - * + Lowercase

Done

* Baseline
  + Vectorization
    - TF-IDF
      * Unigram + Bigram + Trigram
  + FE
    - N-Gram
      * Unigram + Bigram + Trigram
  + Model Selection
    - Classification
      * Multi Class
      * Logistic Regression
      * Random Forest
  + Partitioning (Train Data Set from the Kaggle - Entire DS)
    - 80 - Train
    - 20 - Test
  + Evaluation
    - Confusion Matrix
      * Need to work on precision
      * Will have to treat class imbalance
  + **Baseline Conclusion:**
    - Numbers are encouraging, we can aim for models with high accuracy.
    - Solve class imbalance.
      * Its impacting our accuracy
        + Under Sampling & Over Sampling

K-NN

* + - We should explore powerful models
      * XG-Boost
      * Deep Learning based
    - Seems we have good features in the data set
      * Enough variation for model classification
      * Improved feature engineering might help.
    - Param & Hyperparam
* Fine Tuning
  + RF
  + GRU & LSTM
* Final Model
  + LSTM - Final Model
* Deployment
  + Library
    - Read Me
      * Step 1:
        + from senti\_lib import PredictSentiment
      * Step 2:
        + Create Object of Type PredictSentiment

Ex

db = DashBoard()

* + - * Step 3:
        + Call the function input

Ex

output = pred.input(review\_text)

* + - * + Input to this function is “review\_text”

Ex

output = pred.input(review\_text)

* + API