**Tweet Analytics - Activities**

1. **Introduction Session ( w-1)**
2. **Problem statement and Business Benefit – detailed discussion**
3. **Understanding columns and rows of given input transaction data**
4. **Some basic Analysis of given input data using excel pivot**
5. **Review and clarification of questions**
6. **Python and Jupiter notebook installation and environment setup**
7. **Data Cleaning and Pre-Procession (EDA - Exploratory Data Analysis) (w1,2)**

* **Read base data excel – this data has no Label (class)**
* **Label column creation**
* **Check duplicate rows , check columns having constant value**
* **Check missing values in each column and imputation if required.**
* **Delete rows which are nor relevant**
* **Drop columns which are nor relevant**
* **Check for outlier ( Not required now)**
* **Add derived features**
* **Tweet\_source distribution**
* **Label id distribution**
* **Convert label to numeric**
* **Explore Negative tweet and its category**
* **Country wise positive, Negative and Neutral**
* **WordCloud of each positive, negative and neutral class**
* **Save first round cleaned data**

1. **Processing tweet text data (w-3)**

* **Removing user handles starting with @**
* **Removing numbers and special characters**
* **Removing urls**
* **Removing single characters**
* **Tokenizing**
* **Removing stopwords**
* **Expanding not words**
* **Lemmatizing the words**
* **Stemming the words**
* **WordCloud of each positive, negative and neutral class with clean tweet**
* **Save cleaned data**
* **Vectorization using TF-IDF - Term frequency and Inverse Document frequency**
* **Apply a machine learning model to check current accuracy**
* **First Review with BITS**

1. **Check the correctness of each class df using TextBlob (w-4)**

* **Incorporation - Feedback and suggestions of last review**
* **Compute sentiment scores (polarity) and labels**
* **Sentiment statistics per news category**
* **Mege positive\_correctness , negative\_correctness , neutral\_correctness**
* **Filtered out weak score , if any**
* **Save corrected class data**

1. **Supervised Learning ( w4 , w5 ,w6)**

* **Read pre-processed csv file**
* **Data preparation**
* **Train-test-valid data splitting**
* **Use important supervised machine learning algorithms from using sklearn package**
* **LogisticRegression**
* **Gradient Boosting**
* **Random Forest**
* **KNeighborsClassifier**
* **SVM**
* **LSTM**
* **Model accuracy of each model and comparison**
* **Find and understand all important matrix to get best accuracy and store Mode Report file (csv)**
* **Review and discussion**

1. **Create the models using H2O package (If time permits) ( w-6 , 7)**

* **Install h20**
* **Brief about H20 package**
* **Data preparation for H20**
* **Use important supervised machine learning models from H20 for smote data**
* **H2OStacked Ensemble**
* **H2ORandom Forest**
* **H2O Gradient Boosting**
* **H2O-Grid-search**
* **H2OAutoML**
* **Deep Learning**
* **Model accuracy of each model and comparison**
* **Find and understand all important matrix to get best accuracy and store Mode Report file**
* **Review and discussion**

1. **Final project report preparation and gap filling, if any ( w-8)**