Dnyanai Surkutwar
Santa Clara, CA

Folder details:

1) Part_I -> Making the Dataset (Part_I_Making_the_Dataset.ipynb) (2 files for Dec3_4 and Dec 10_11 updates)
2) Part_II -> Classifier (Part_II_Classifier.ipynb)

Requirements:

1) Making a tweets dataset.
2) Generating Labels for the dataset.
3) Scraping articles from the relevant links in the tweets.
4) Creating a classifier for classifying the tweets into 'Case Studies' and 'Press Releases' labels.

Approach:

1) I used Tweepy to scrape tweets from Salesforce twitter account - @SalesforceNews as I had to build a dataset.
2) We need data in two labels – 'case studies' and 'press releases '.
3) I, then used Selenium, requests, and Beautiful Soup to get the correct labels and scraping the body from the website mentioned in the respective tweets.
4) Once, I had a dataset I made a BERT classifier on a pre-trained BERT model for 2 output classes – 'case studies' and 'press releases'.

Future Steps:

1) The classifier is mostly overfitting as we do not have a lot of data, this can be rectified by making a pipeline to build the dataset continuously. So, appending the tweets to the old dataset as new tweets are scraped is a good next step to make sure enough data is gathered.
2) Various classifier's can be implemented and compared once the dataset is big enough.
3) We could also generate text summarization of the scraped articles to understand the two categories better possibly building a more generalized classifier.