The files in this folder are separated based on the part of the project they were focused on. Below you can find a brief description of each file’s purpose:

* **Data\_Loading**
  + Data is scraped using Tweet ID’s using the Twitter API and saved to a csv file
* **Data\_Cleaning**
  + Data is loaded and text preprocessing is performed
    - Lower case, removal of punctuation, etc.
    - Create a version of the text data without stop words to use for Bag-of-Words
* **Bag-of-Words**
  + Tokenizes tweets
  + Converts to Bag-Of-Words representation
  + Logistic Regression classifier on Bag-Of-Words tweet dataset
  + Evaluate on various training set sizes
* **LSTM**
  + Train LSTM for classification
  + Evaluate on various training set size
* **BERT\_Supervised**
  + A fully supervised BERT model for classification is trained
  + Run analysis on various training sizes to explore performance
* **Unsupervised Clustering**
  + Dimensionality reduction of BERT embeddings (PCA + t-SNE)
  + Various Clustering algorithms
    - K-Means
    - Agglomerative/Hierarchical Clustering
    - Spectral Clustering
    - GMM
    - OPTICS
    - DBSCAN
* **Deep Clustering**
  + BERT is first pre-trained using Deep Clustering as an auxiliary trask on most of the data
  + Model is fine-tuned using a set of labeled examples
  + Evaluate on test set