Preprocessing:

- Stop_words and clearning.ipynb:

1. Find Stop words: stops.txt
2. Apply the cleaning pipeline to the raw data(speech.csv). The dataset after cleaning is clean_data.csv

BaseLine_Logistic_Model

- Logistic model.ipynb:

Train the logistic model with text in clean_data.csv. The top 100 importance words from positive group and negative group are in the folder top100_importance word

XLNet

- NewsXLnetTrained: Train the XLNet model and save the trained models.

1. Train with cleaned data:
https://drive.google.com/drive/folders/19MyoQCms9CNHCOdeBCwZD2uRT7rlipqh?usp=share_link
2. Change input by only leave nouns in text:
https://drive.google.com/drive/folders/1eFRgnsFDAspV76ORtM6KH0kFmszZneo6?usp=share_link
3. Change input by shuffling the sentence:
https://drive.google.com/drive/folders/1Gdu0EMayRnVHp6yCkKAH8ehYctNjI3j9?usp=share_link

- Xlnet_attention_score.ipynb: Generate top 100 importance words using attention-score method for the three XLNet models and the result are in the folder top100_importance
- Classification.ipynb : Choose to have ten clusters based on elbow method; Apply k-means clustering on documents based on the tfidf score with max feature 1800. Name each clusters based on top 100words. Classify the top words for positive and negative group based on type of document that it has the highest occurrence in.

Classification result

- Classification result and pie chart for top words of XLNet model, logistic model and Bert Model(Twitter)