The programming language used in this project is python. Before the project starts, open all the data of develop, test, and train. Then print count, full, glove and tfidf respectively to visually see the structure of the data and prepare for subsequent data processing. It can be seen that the full data is a string of original tweets. Count stands for bag of words. After removing very frequent and uncommon words, the remaining word strings are mapped to their IDs according to the content of vocab.txt, and each tweet is represented by a list of (ID, word count) tuples. The data in tfidf is similar to the data in count, but the tfidf value is used to measure the importance of features. The data in Glove maps each word to a 100-dimensional Glove "embedding vector". These vectors are trained to capture the meaning of each word. Then add the vectors of words in the tweet to get a single 100-dimensional representation of the tweet.

The data selected for this project are counts, tfidf and gloves. Extract all the data in the tweet column of the data set. For the processing of the two data count and tfidf, a two-dimensional array of all zeros is first introduced. Each column corresponds to a word ID, and each row represents a tweet. Then, according to the number of words corresponding to the words in count, the 0 of the corresponding word ID is overwritten to form a brand new two-dimensional array. Data processing in tfidf is the same as count data processing. The tfidf value corresponding to the word is used to overwrite the 0 of the corresponding word ID to form a brand new two-dimensional array. The data in Glove is directly converted into a set of values, and each value represents the 100-dimensional value of the tweet.

After processing the data, call the machine learning model to use each data set to make predictions. Then compare with the correct Lable to calculate the accuracy of each model. The baseline is calculated using zero-r. Compare the baseline with the accuracy measured above.