Data is manually collected from https://twitter.com/search?q=genetic&src=typd&lang=en

LogisticRegression is used

a term will have a high TF-IDF score if it appears frequently in a comment but not very often throughout the corpus. TF-IDF vectors can be generated at different levels of input tokens, such as word level, n-gram level, and character level. In this project, n-gram is chosen. Using TF-IDF, m comments string can be translated into a m*n matrix, where m is the size of our training set and n is the number of features. The matrix value is the TF-IDF score and the matrix is the input of the model. Logistic regression (LR) is a widely used model to do classification.