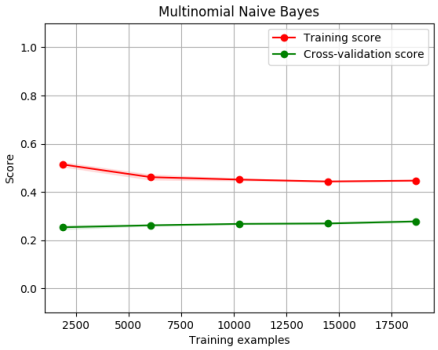
Samiha Samrose
samiha.mumu@gmail.com

Approach-1: Classification

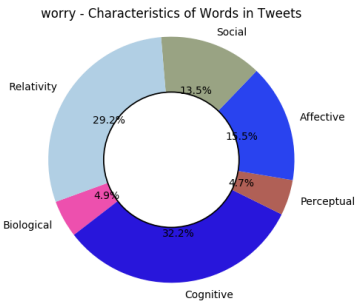
Task: Prediction of sentiment labels from tweet contents
Source code file: text_mining_classifier.py
Feature column: content
Class label: sentiment

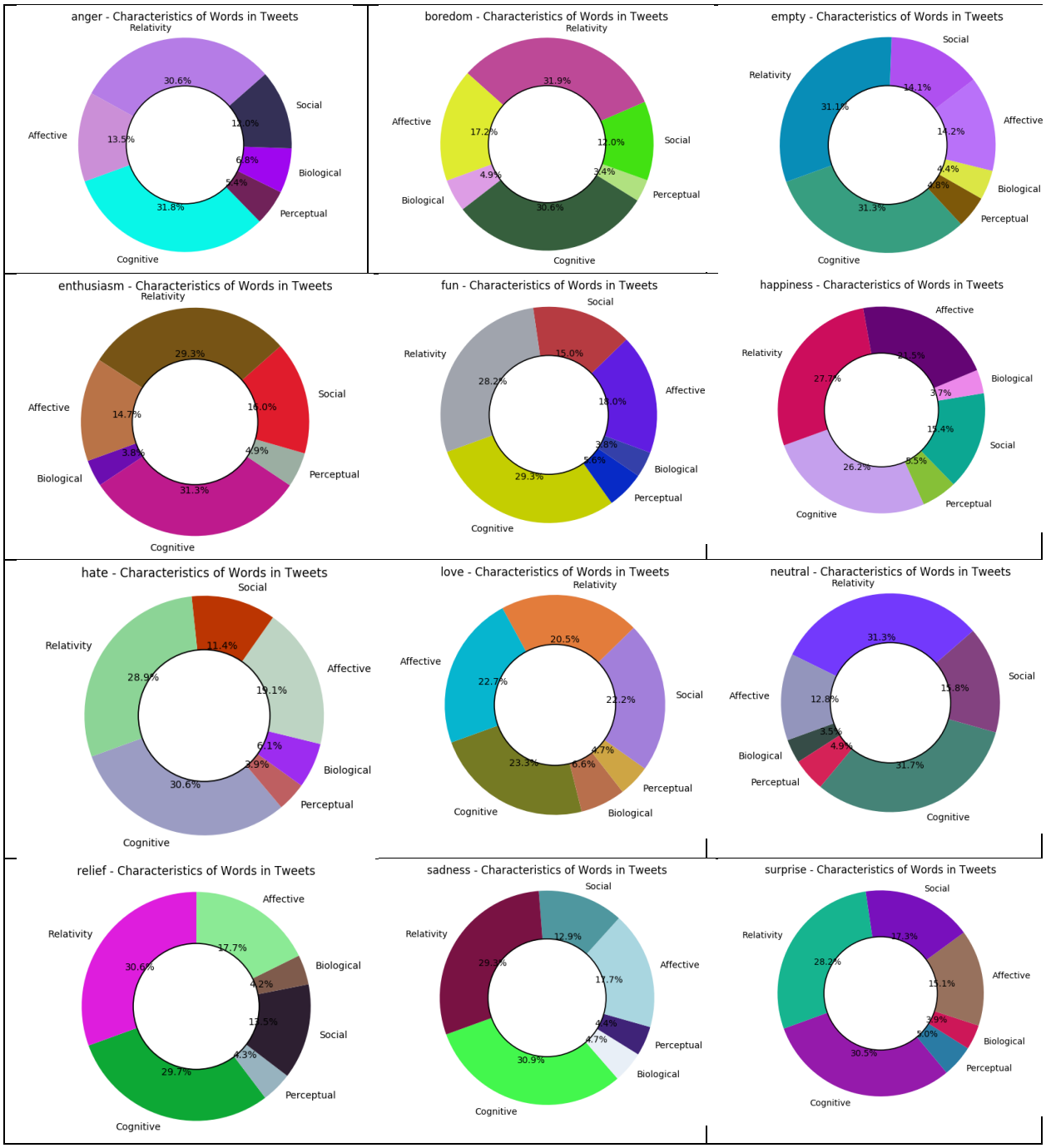
Prediction on text_emotion.csv:
70.00% in training set, 30.00% in test set
Technique-1:
Multinomial Naive Bayes
Train Accuracy: 43.7357142857 %
Test Accuracy: 28.2166666667 %
Technique-2:
SVM – SGD Classifier
Train Accuracy: 82.9821428571 %
Test Accuracy: 32.2166666667 %
Discussion:
Both the models are overfitting. (a) The cross-validation needs to be tweaked more, (b) needs more data, (c) other techniques can be applied



Approach-2: Language Model Analysis (using LIWC)

Task: Analyzing the characteristics of the words in tweet contents for each sentiment category
Source code file: liwc_category.py
Support code file: liwc_example.py
Feature column: content





Discussion:

Interesting pattern can be observed, such as Category: “Social” can have a boundary for positive-neutral-negative sentiments, and thus a classifier can be made to predict the sentiment using these categories.