(1) To predict, given a tweet in English, its most likely associated emoji (**20** points)

For this part, I used bags of 2-gram vectors as train data and put it into the SVM model. As a result, the model has an F-Score of 32.567 (approximately 10% lower than the state-of-art).

(2) To predict, given a tweet in Spanish, its most likely associated emoji (**20** points)

Since we have fewer data in Spanish and the previous SVM model seems not to work well in this case, I tried to combine multiple models (SVM, Neural Network, Logistic Regression, and Random Forest) to improve the prediction accuracy. However, after using the multi-model, the performance dropped down 3% than the SVM. Finally, I picked the LR model which achieves the best performance among all and get an F-Score of 18.467.

(3) To use *any* type of multilingual transfer learning to see if you can use English data to improve Spanish emoji prediction or vice versa (**20** points)

In this part, I translated the English training data to Spanish and used a hashmap to match the outputs into Spanish emoji. After merging the translated data into the existing data, I put it into the LR model. However, the additional training data seems to have a bad influence on the model performance since the F-Score drops by 3% compared to the original model. One possible reason is that, after translating into Spanish, the relations among words become ambiguous, which may affect the model to mislearn the meaning of the sentences (hard to converge). The other explanation could be the way used to represent English sentences (bags of n-gram) may not be able to represent the meaning in Spanish, which makes it hard for models to fetch useful information from the data.