Executive Summary

This project was conducted to create a library of emotive words that can be later used for sentiment analysis. Current high profile events show the power of online platforms and media and their ability to influence individuals and communities.

We found that the Voting Classifier best fit the data and our end goals. The Voting Classifier ensemble model scored ~0.06 better on mean accuracy than the closest model, random forest classifier. The training score was below the cross val score, which may mean the model is not overfit, but the fact that the training score is consistently lower than the cross val score is odd. There may be unaccounted for leakage occurring. Additionally, the Voting Classifier has multiple models from which to pull feature names (words) and feature importance. This allowed for us to pull many words from the classification model.

Overall there were words that were gathered that intuitively make sense, however, it is recommended that classification models are not used to create a library of emotive words. Although, results include words that make sense intuitively this process does not scale well due to the conflict between limiting features for modeling while simultaneously wanting a large library.