Classifying Sentiment of Tweets at SXSW Austin

Ryan McArthur, Flatiron School, DS

Introduction

Our Task:

Classify tweets by user sentiment, based upon the content of the tweet.

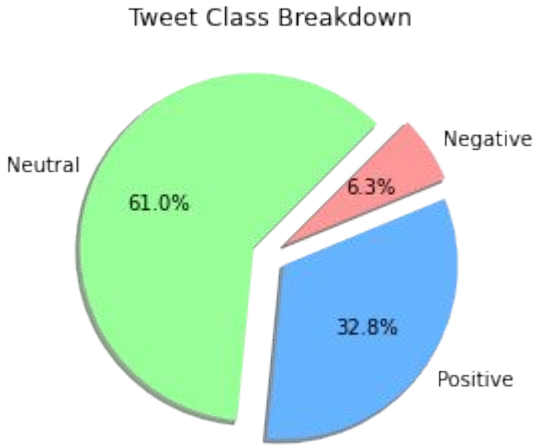
Convey insights into the dataset through visualizations

Data

We have 9093 tweets.

5802 of these tweets have no product in mention.

5545 are neutral. 2978 are positive. 570 are negative



Most used Words - Apple

Most Used Words - Google

Positive V Negative Words

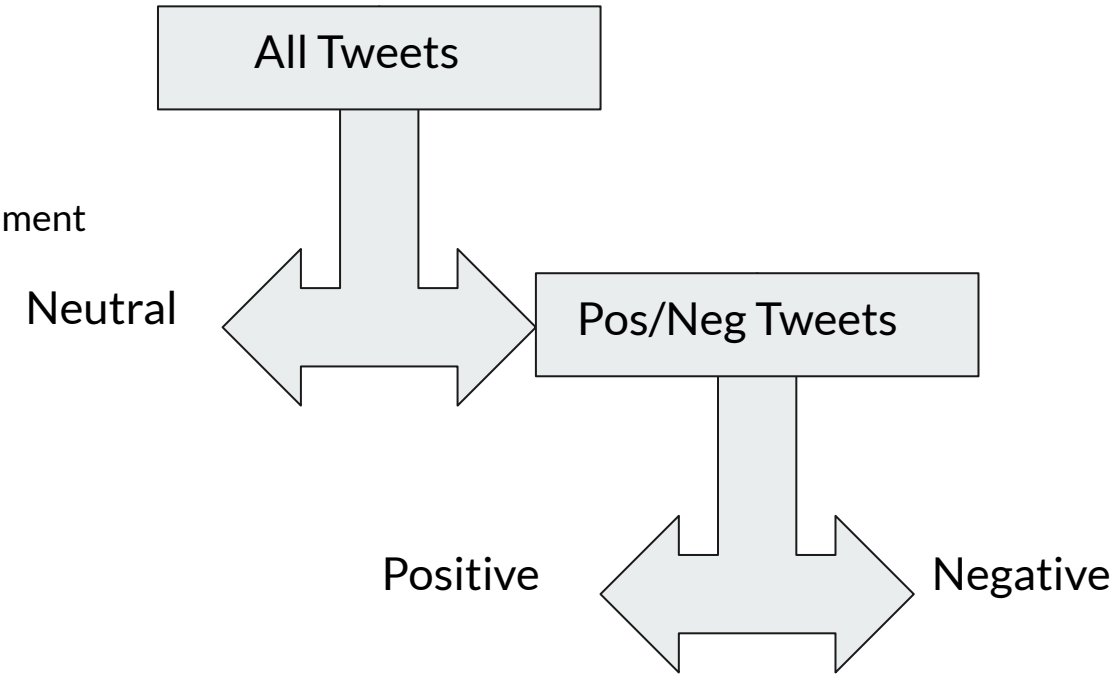
Positive

Negative

Modelling Flow

Phase 1: Classifying on Existence of Sentiment

Phase 2: Classifying on Negative V Positive Sentiment



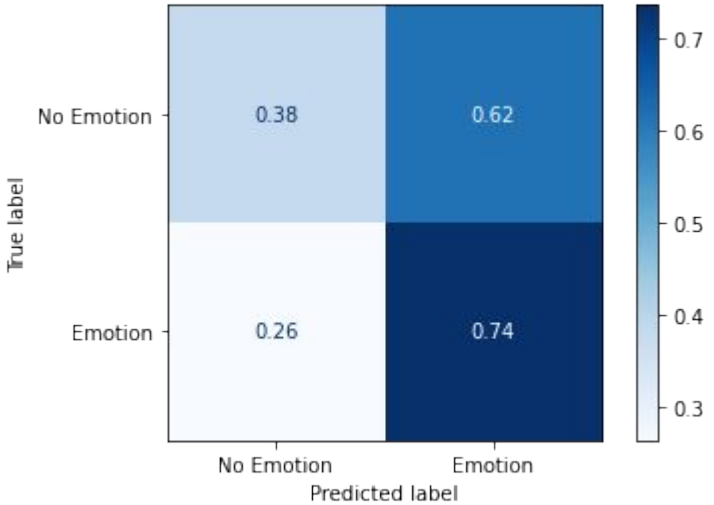
Modelling - Phase 1

Gaussian Naive Bayes

Recall: 89 %

F1 : 65%

Struggles with tweets that contain no sentiment, but excels with sentimental tweets.



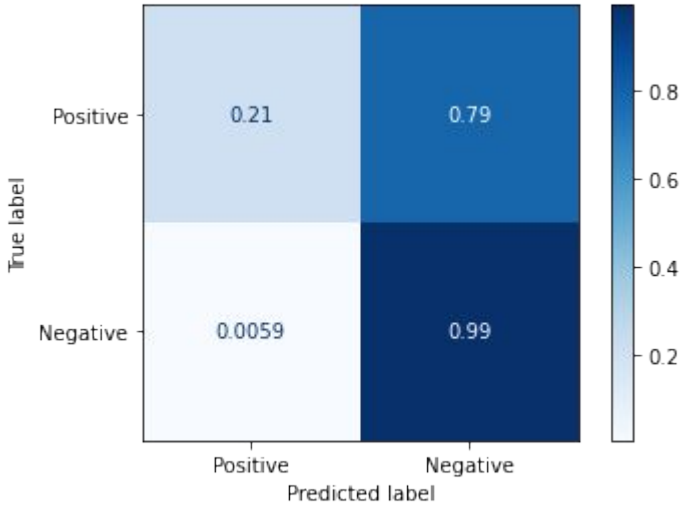
Modelling - Phase 2

Random Forest Classifier

Accuracy: 87%

F1 : 84%

This model performs well, especially with negative tweets.



Conclusions

Our models perform best when classifying tweets with negative sentiment, which are arguably the most important tweets to classify correctly.

Separating our classification into two models allowed us to both simplify our workflow, as well as address class imbalance.

Next Steps

Our dataset had some limitations that could be improved upon

Data was collected from a very small sample space

Data was heavily imbalanced, and not fully labelled.

Also, we could have analyzed our tweets to construct a list of most common phrases, much like we did with most common words. This would give better context to what twitter users average sentiments are.

Thank you!

Ryan McArthur

Rmcar522@gmail.com

Flatiron School, DS 2020