Classifying a Tweet's Sentiment Based on its Content

Sara Robinson

Phase 4 Project

Flatiron School

Introduction

- Purpose
- Implications
- Natural Language Processing

The Data Science Process: OSEMN

Obtain

- CrowdFlower
- Brands and Product Emotions

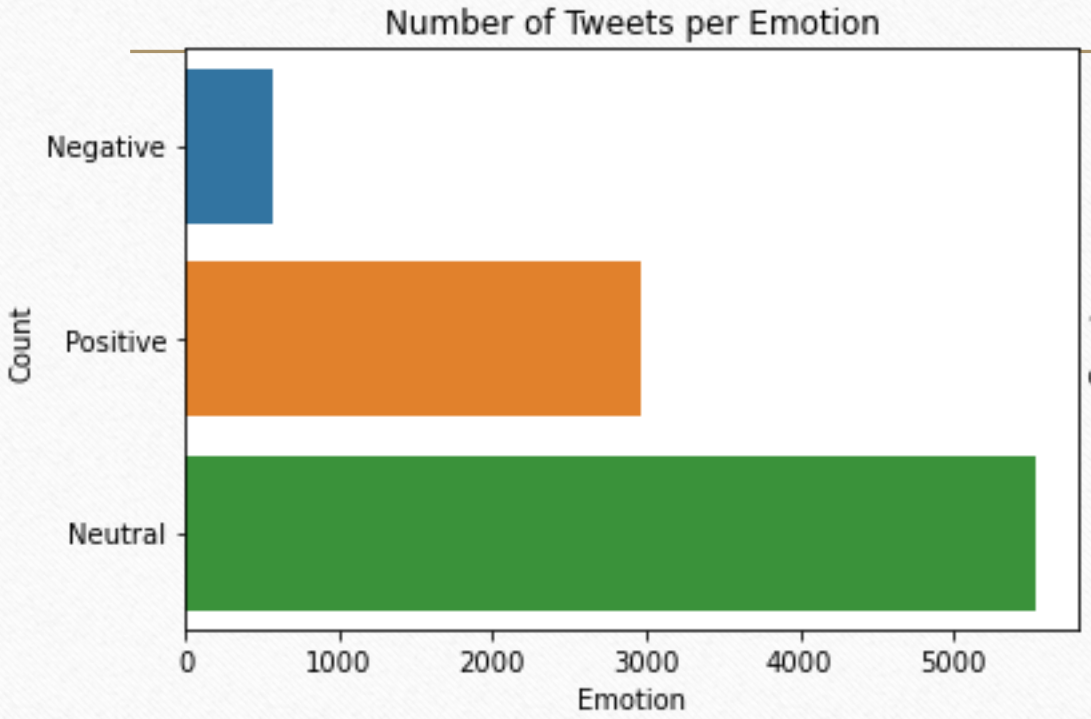
Scrub

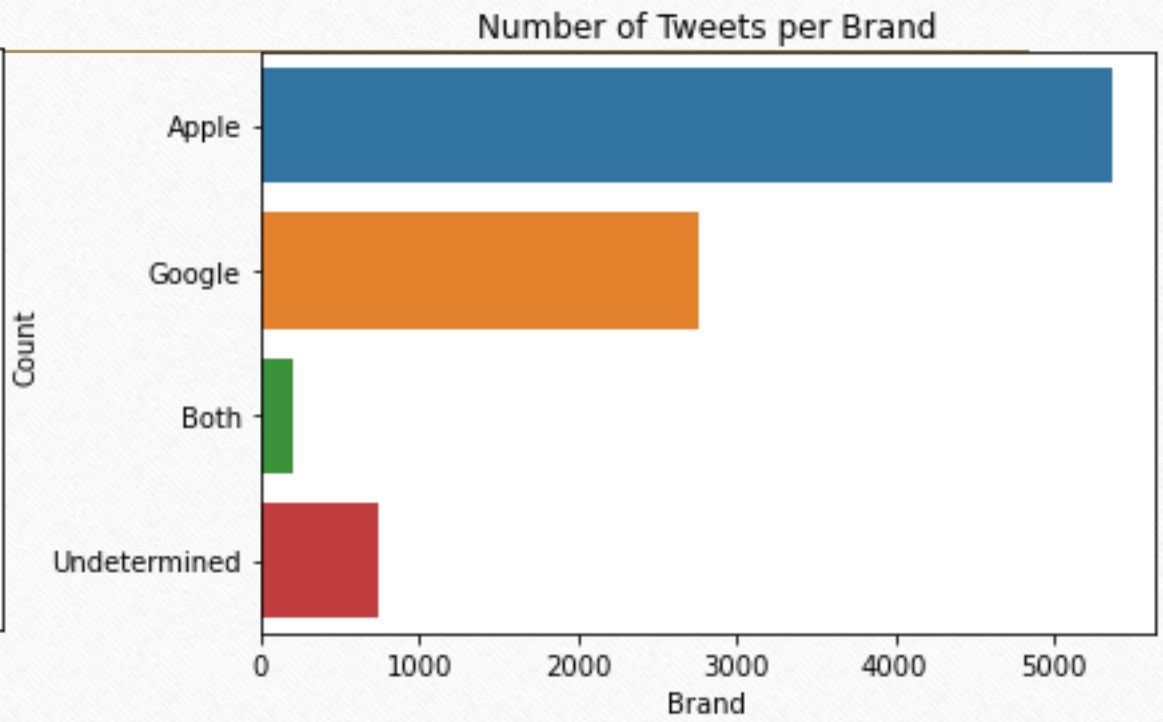
- Nomenclature
- Text Preprocessing

Explore

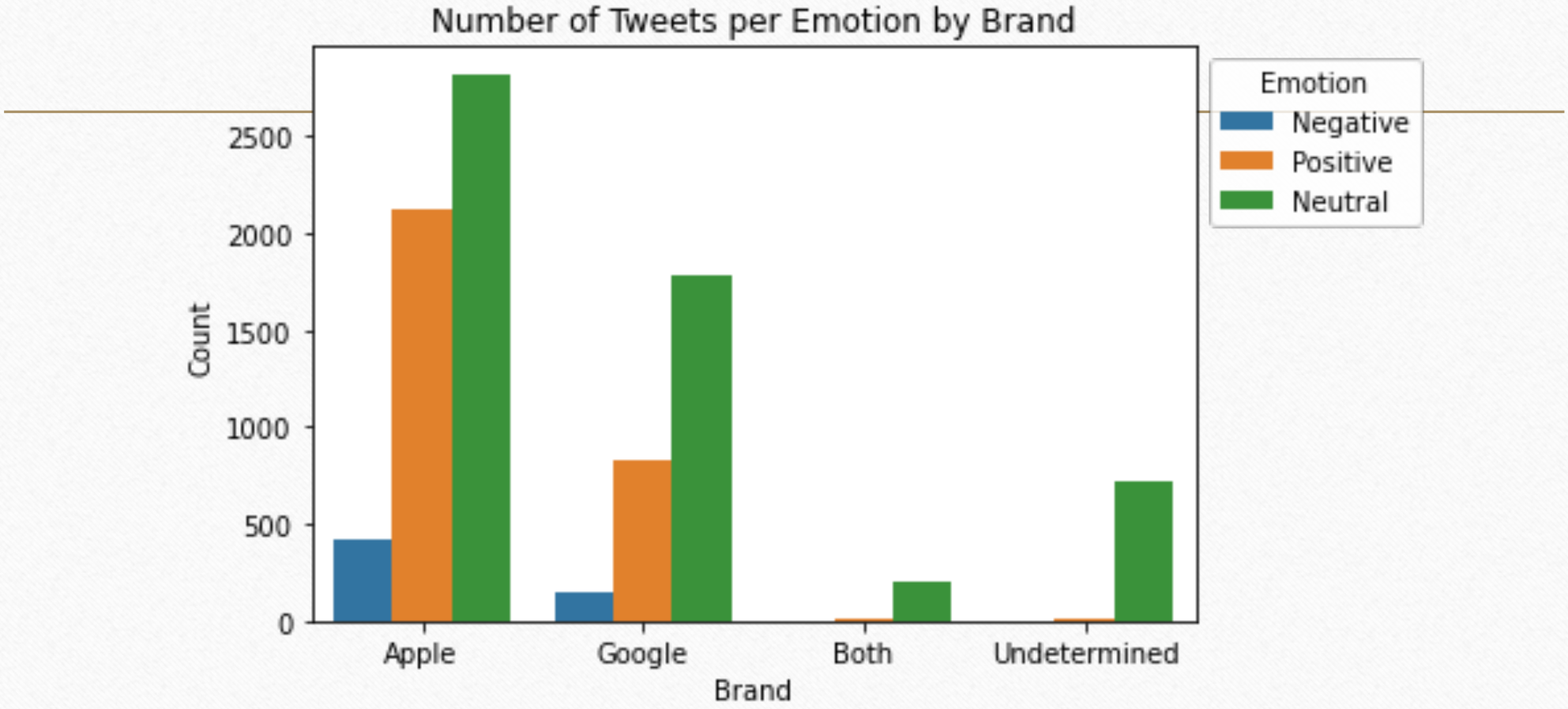
- Spread of Tweets
- Tweet Length
- Sentiment among Brands
- Most Used Hashtags
- Most Used Terms
- Common Phrases

How are the tweets spread?

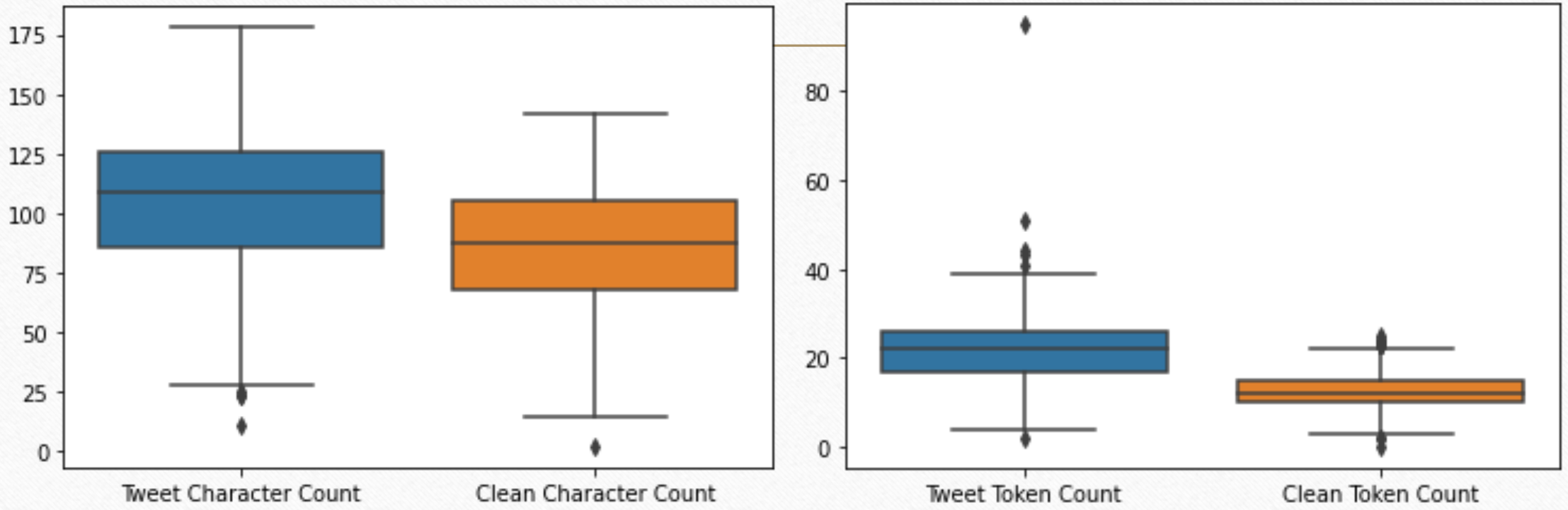




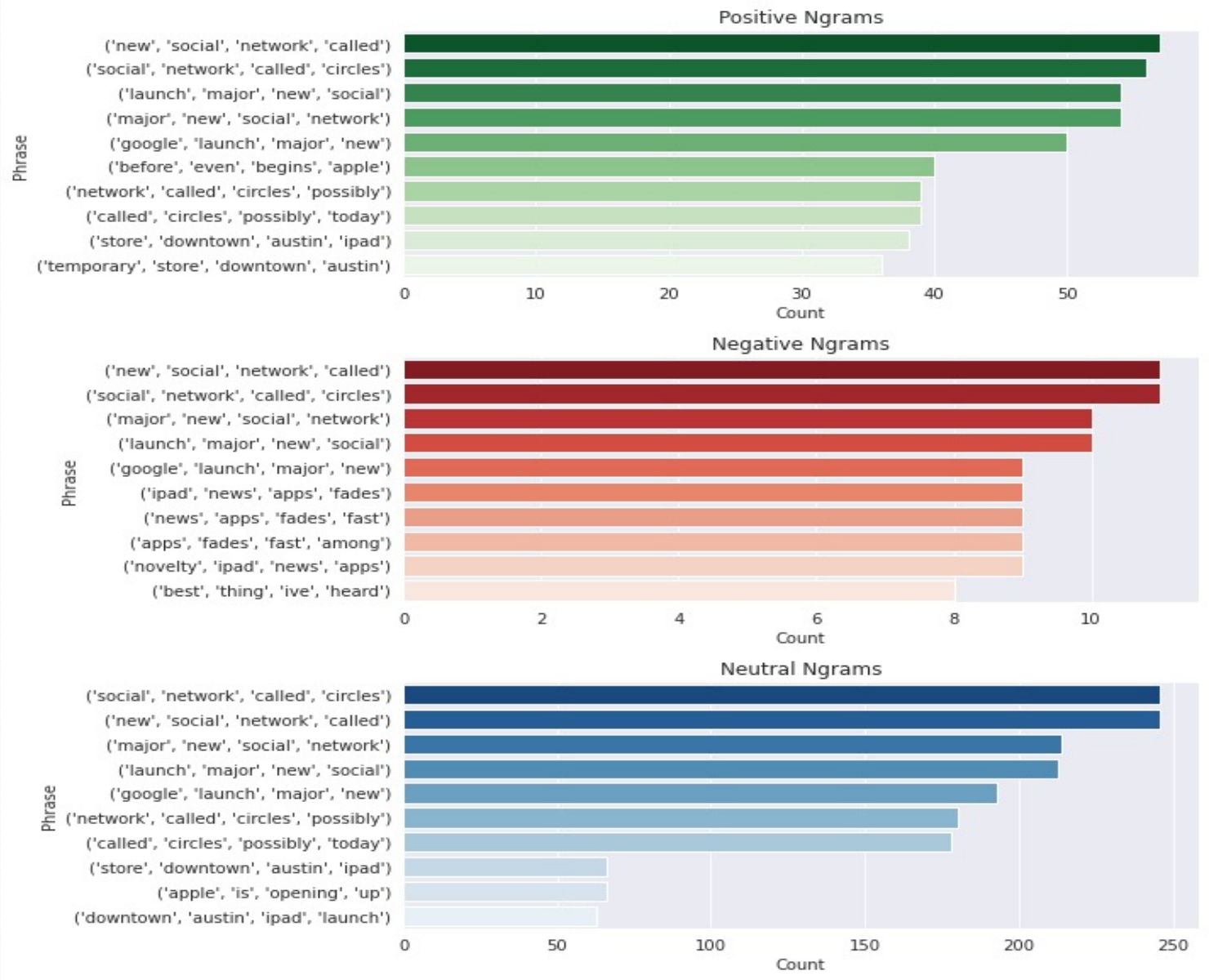
What is the spread of tweet sentiment among brands?



How much did preprocessing change the length of our tweets?

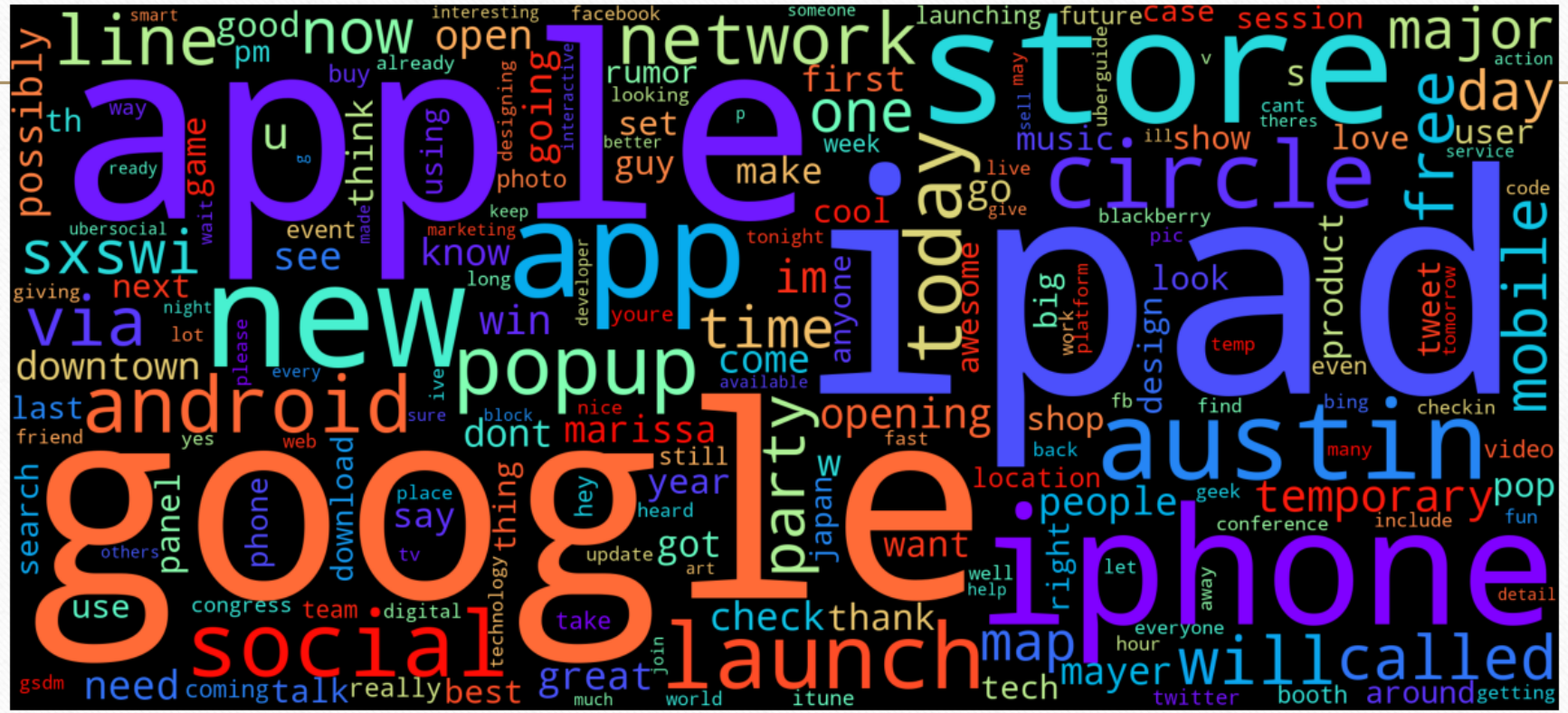


Do four-word phrases (ngrams) help distinguish between sentiments?



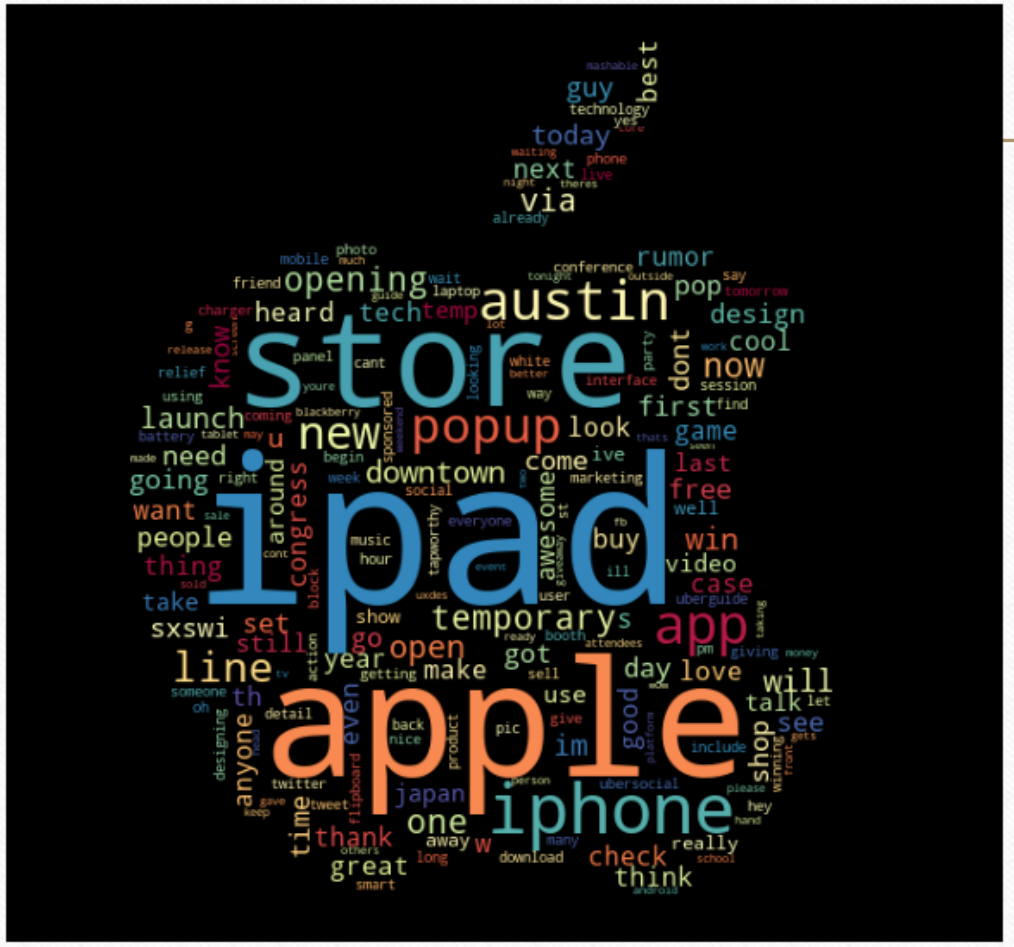
All Hashtags Word Cloud

All Tokens Word Cloud



Sentiment Tokens Word Clouds

Brand Tokens Word Clouds



Model

- 6 Models
- 2 Vectorizers

Comparing Baseline Model Scores

Model Name	Accuracy Score	Precision Score	Recall Score	F1 Score	Cohen's Kappa Score	Matthew's Correlation Coefficient	Neg PRC AUC Score	Neu PRC AUC Score	Pos PRC AUC Score	Neg ROC AUC Score	Neu ROC AUC Score	Pos ROC AUC Score
LogRegCV	0.69	0.63	0.55	0.58	0.373	0.377	0.34	0.81	0.63	0.82	0.75	0.77
MNBCV	0.68	0.66	0.51	0.54	0.342	0.345	0.31	0.8	0.62	0.78	0.74	0.75
LogRegTF	0.7	0.64	0.47	0.49	0.341	0.359	0.33	0.82	0.66	0.83	0.76	0.77
RFTV	0.68	0.65	0.49	0.52	0.304	0.324	0.35	0.79	0.59	0.8	0.73	0.74
RFCV	0.68	0.66	0.5	0.54	0.302	0.321	0.38	0.8	0.6	0.83	0.74	0.75

Model

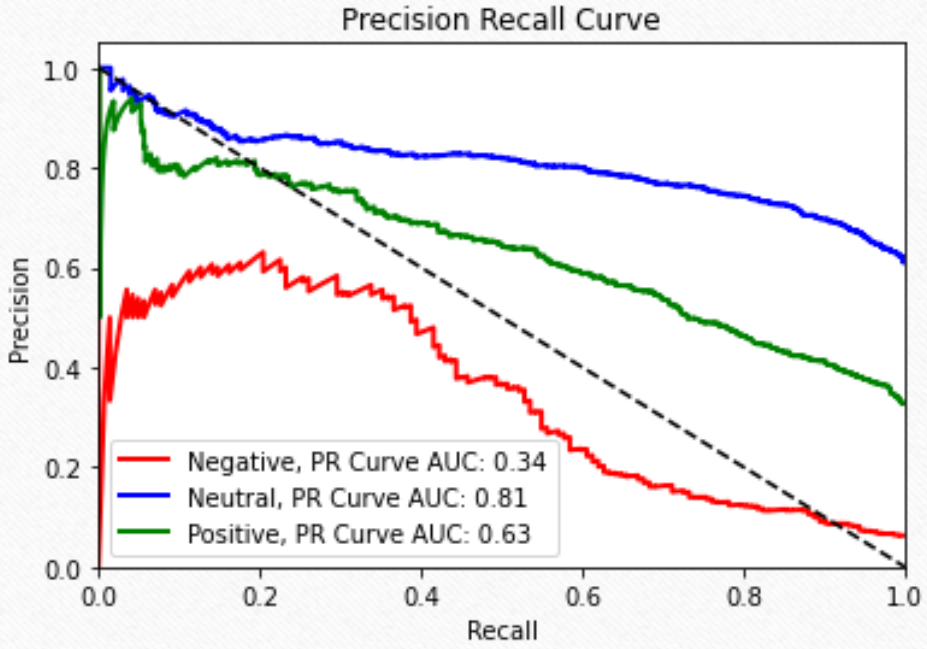
- Tuned Top Performing Models

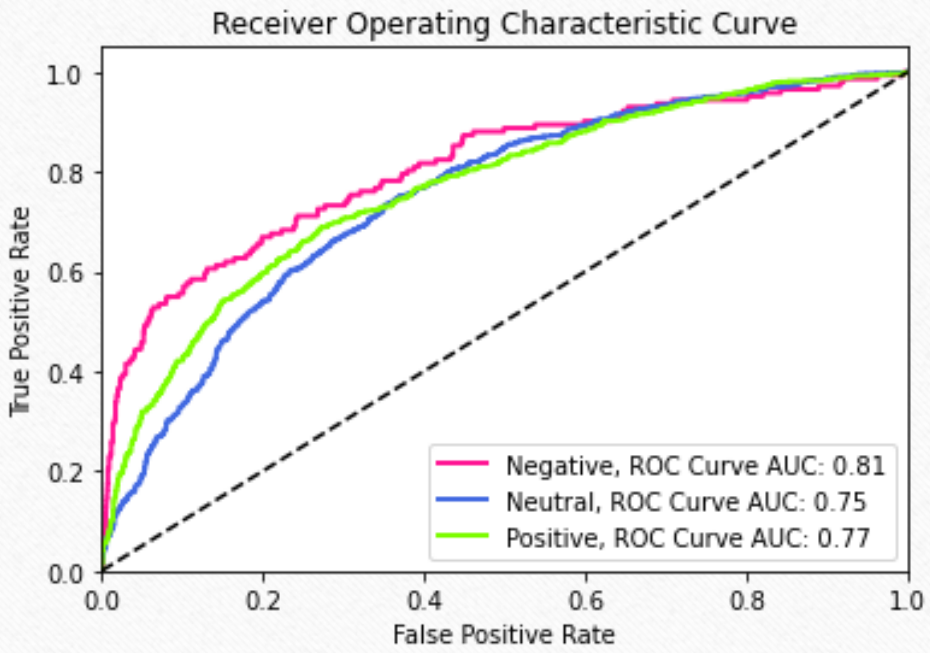
Comparing Tuned Model Scores

Model Name	Accuracy Score	Precision Score	Recall Score	F1 Score	Cohen's Kappa Score	Matthew's Correlation Coefficient	Neg PRC AUC Score	Neu PRC AUC Score	Pos PRC AUC Score	Neg ROC AUC Score	Neu ROC AUC Score	Pos ROC AUC Score
T1LogRegCV	0.67	0.58	0.61	0.59	0.382	0.382	0.34	0.81	0.63	0.81	0.75	0.77
T5LogRegCV	0.67	0.57	0.61	0.59	0.382	0.383	0.34	0.81	0.63	0.81	0.75	0.77
LogRegCV	0.69	0.63	0.55	0.58	0.373	0.377	0.34	0.81	0.63	0.82	0.75	0.77
T3LogRegCV	0.7	0.63	0.53	0.56	0.373	0.379	0.33	0.81	0.64	0.82	0.75	0.77
T6LogRegTF	0.65	0.55	0.62	0.57	0.369	0.373	0.36	0.82	0.64	0.84	0.75	0.77

iNterpret

- Final Model





Conclusions

- Sentiments
- Brands

Further Exploration

- Stemming/Lemmatization
- Class Imbalances
- SXSW ”Frequent Flyers”

Thank You