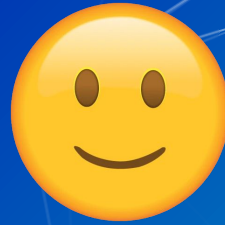


Tweet Sentiment Analysis



Flatiron School Module 4 Project

By Shawn Sobieski and William Newton

Data and Methodology

The image is a screenshot of the Twitter mobile application interface. The layout is divided into three main sections: a left-hand navigation sidebar, a central Home feed, and a right-hand recommendations panel.

Left-hand Navigation Sidebar:

- Twitter logo icon
- Home (with house icon)
- Explore (with hashtag icon)
- Notifications (with bell icon)
- Messages (with envelope icon)
- Bookmarks (with bookmark icon)
- Lists (with list icon)
- Profile (with person icon)
- More (with three dots icon)
- A large blue "Tweet" button at the bottom.

Central Home Feed:

- Header: "Home" with a star icon.
- Post 1: A tweet by "Kevin Roose" (@kevinroose) 3m ago, stating "Quite a Mod Monday." It includes a retweet icon and a "Retweeted" label by "danny". The tweet content is a video player showing the YouTube logo on a red background with many small red triangles.
- Post 2: A tweet by "Mozilla" (@mozilla) at the bottom, partially visible.
- Below the video player, there is a text snippet: "YouTube bans Stefan Molyneux, David Duke, Richard Spencer, and m... YouTube is taking more action against supremacy channels and hate speech theverge.com".
- Interaction icons at the bottom of the feed: reply, retweet (19), like (37), and share.

Right-hand Recommendations Panel:

- Search bar: "Search Twitter".
- Section: "What's happening".
- Item 1: "COVID-19 · This morning Gilead Sciences detailed its pricing plans for COVID-19 drug remdesivir" with a small image of a Gilead Sciences box.
- Item 2: "#ShopSmall Learn more at ShopSmall.com Promoted by American Express".
- Section: "Trending in United States".
- Item 3: "Twitch".
- Section: "World news · June 27, 2020".
- Item 4: "Russia denies allegation that it offered bounties to militants in Afghanistan for killing American troops" with a small image of a soldier.
- Section: "Trending in United States".
- Item 5: "JK Rowling".
- Link: "Show more".
- Section: "Who to follow".
- Item 6: "General Motors Fl..." with a "Follow" button.

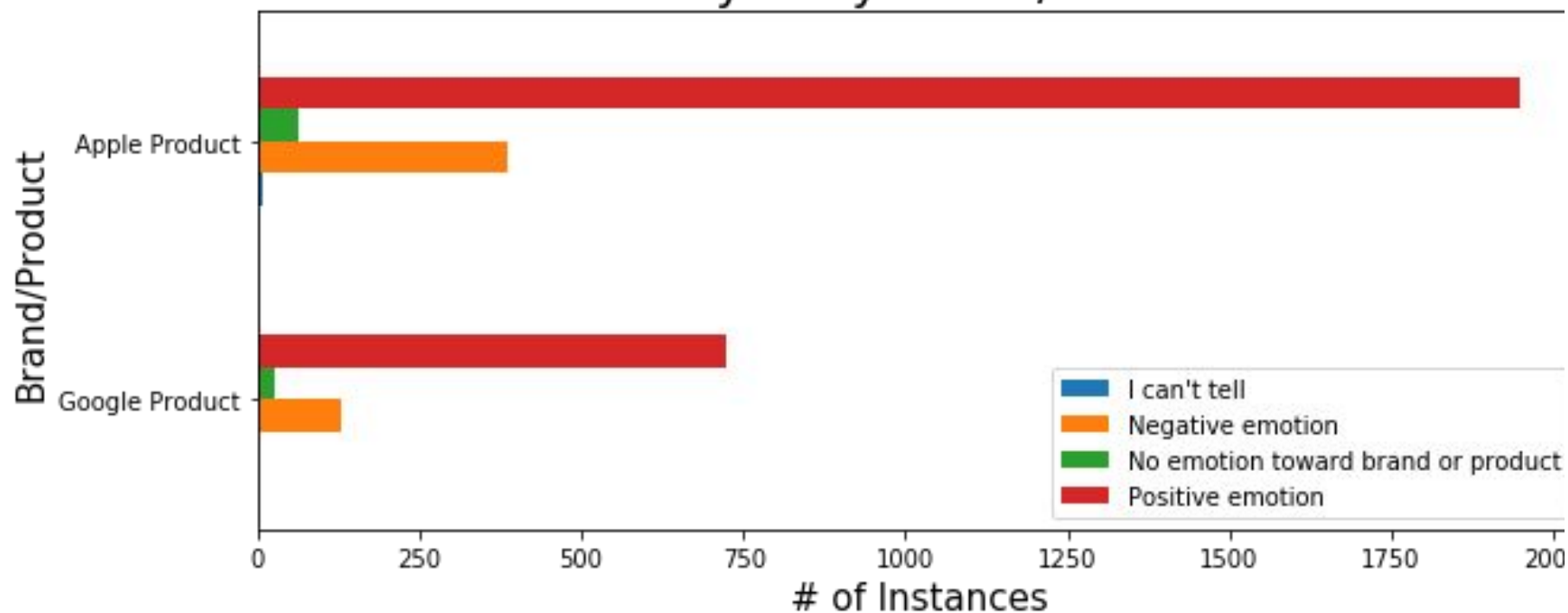
Data and Methodology

- Data used for this project was obtained from Data.world
 - <https://data.world/crowdflower/brands-and-product-emotions>
- Data consists of Tweets collected at a major-telecom provider sponsored tent at the SXSW music festival in Austin, Texas
- Data was labeled with either Positive, Negative, No Emotion, or Unclear sentiment
- If a brand or product was identified in the Tweet, it was labeled as either a Apple or Android Product

Goal:

Use the data set to
build a model to predict
sentiment in future
tweets

Sentiment Analysis by Brand/Product Combined

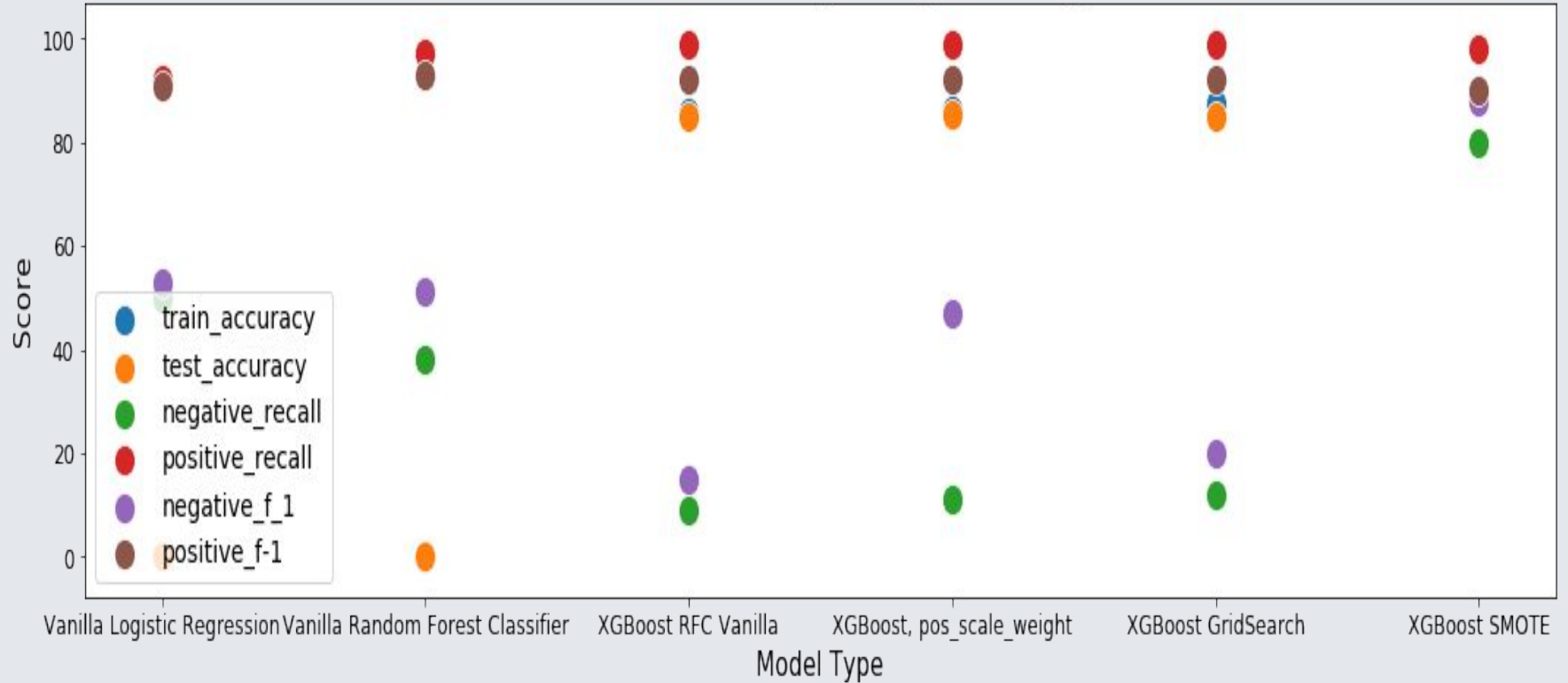


Findings

The background of the slide is a blue gradient, transitioning from a darker blue on the left to a lighter blue and greenish tint on the right. Overlaid on this background is a complex network of white lines and dots, forming various geometric shapes like triangles and polygons. Some of these shapes are more prominent and interconnected, while others are fainter and more isolated. The overall effect is a modern, tech-oriented aesthetic.

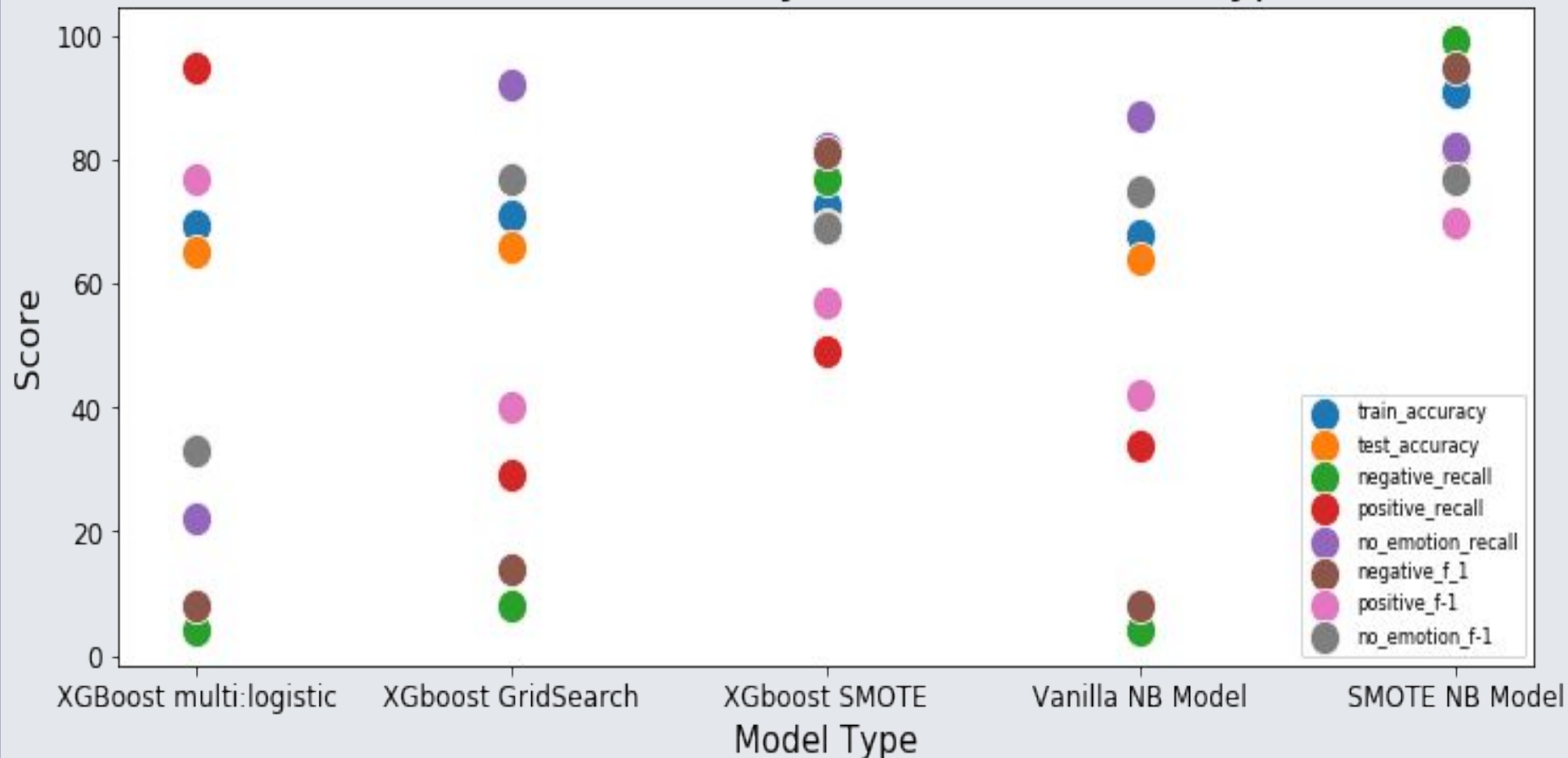
model_#		model_type	train_accuracy	test_accuracy	train_test_diff	negative_recall	positive_recall	negative_f_1	positive_f-1	notes
0	1	Vanilla Logistic Regression	0.0	0.0	0.0	50.0	92.0	53.0	91.0	Vanilla model very basic, needs tuning
0	2	Vanilla Random Forest Classifier	0.0	0.0	0.0	38.0	97.0	51.0	93.0	Vanilla model very basic, needs tuning
0	3	SVM: Balanced Class Weights, Linear Kernel	81.3	76.4	4.9	67.0	78.0	47.0	85.0	Model overfits and class imbalance still causing issues
0	4	SVM: Balanced Class Weights, Linear Kernel, GridSearch	84.5	75.6	8.9	64.0	78.0	45.0	84.0	Model still overfits and less predictive for balanced data
0	5	SVM: SMOTE	99.2	88.9	10.3	93.0	85.0	89.0	89.0	Model overfits but SMOTE has solved class imbalance
0	6	XGBoost RFC Vanilla	85.8	85.1	0.7	9.0	99.0	15.0	92.0	Model is basically guessing 1 every time
0	7	XGBoost, pos_scale_weight	86.1	85.4	0.7	11.0	99.0	47.0	92.0	Class imbalance still causing issues with model
0	8	XGBoost GridSearch	87.7	85.1	2.6	12.0	99.0	20.0	92.0	Class imbalance still causing issues with model
0	9	XGBoost SMOTE	89.8	89.2	0.6	80.0	98.0	88.0	90.0	SMOTE fixed class imbalance. Best performing baseline

Evaluation Metrics by Binary Model Type



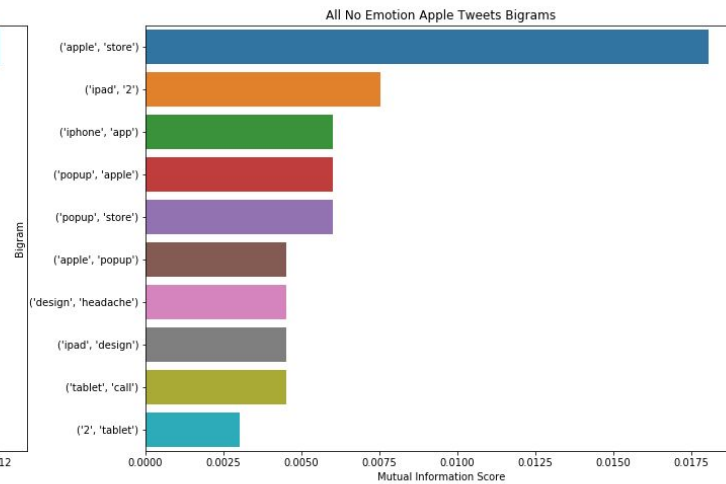
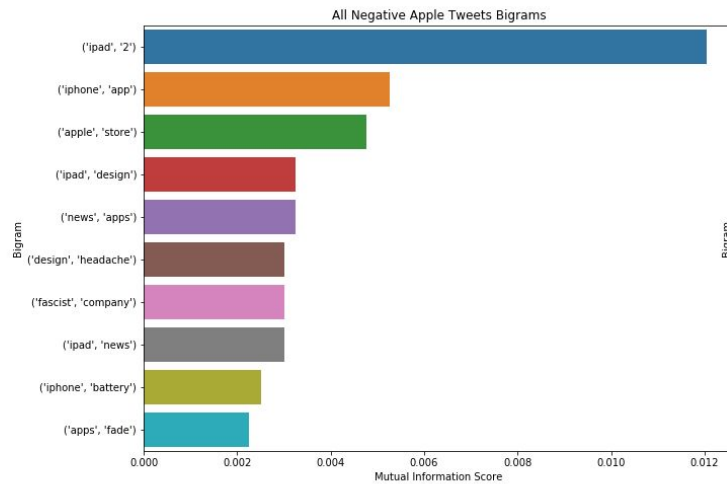
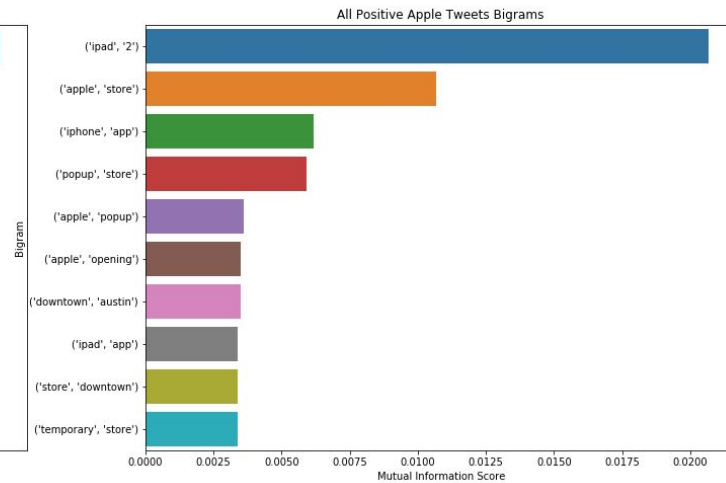
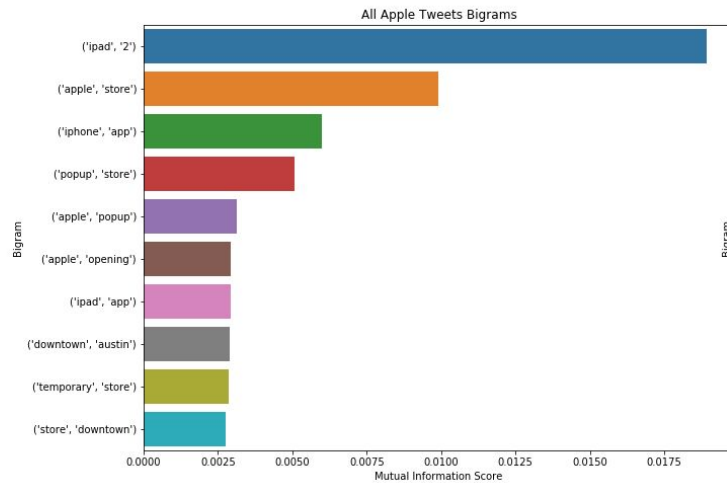
model_#	model_type	train_accuracy	test_accuracy	train_test_diff	negative_recall	positive_recall	no_emotion_recall	negative_f_1	positive_f-1	no_emotion_f-1
1	SVM: Vanilla Linear Kernel Class Weight Balanced	93.0	66.6	26.4	34.0	58.0	75.0	38.0	57.0	75.0
2	SVM: GridSearch Balanced Class Weights	75.0	61.8	13.2	49.0	55.0	67.0	38.0	53.0	71.0
3	SVM: SMOTE	94.7	83.6	11.1	100.0	82.0	70.0	96.0	75.0	79.0
4	XGBoost multi:logistic	69.4	65.2	4.2	4.0	95.0	22.0	8.0	77.0	33.0
5	XGboost GridSearch	71.1	65.9	5.2	8.0	29.0	92.0	14.0	40.0	77.0
6	XGboost SMOTE	72.7	69.4	3.3	77.0	49.0	82.0	81.0	57.0	69.0
7	Vanilla NB Model	67.9	63.9	4.0	4.0	34.0	87.0	8.0	42.0	75.0
8	SMOTE NB Model	91.1	81.3	9.8	99.0	82.0	82.0	95.0	70.0	77.0

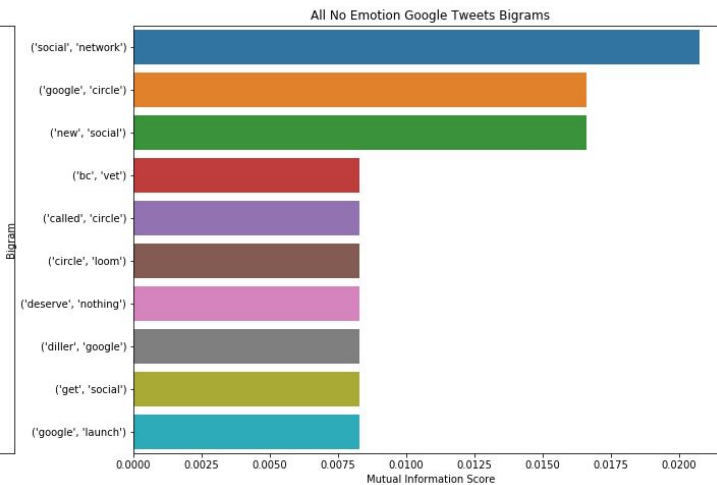
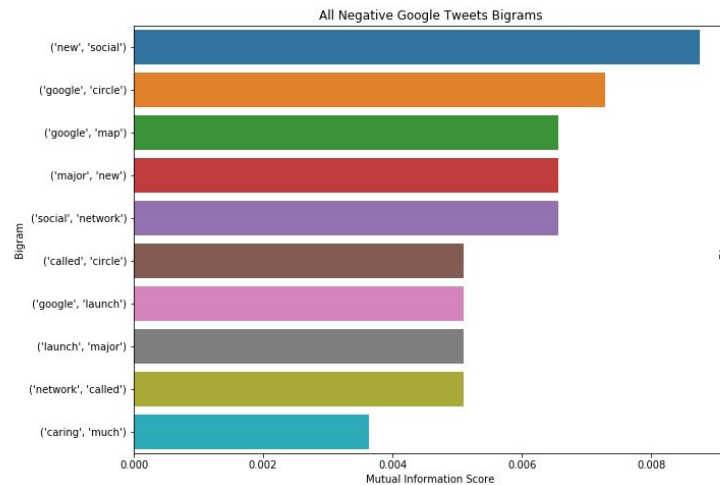
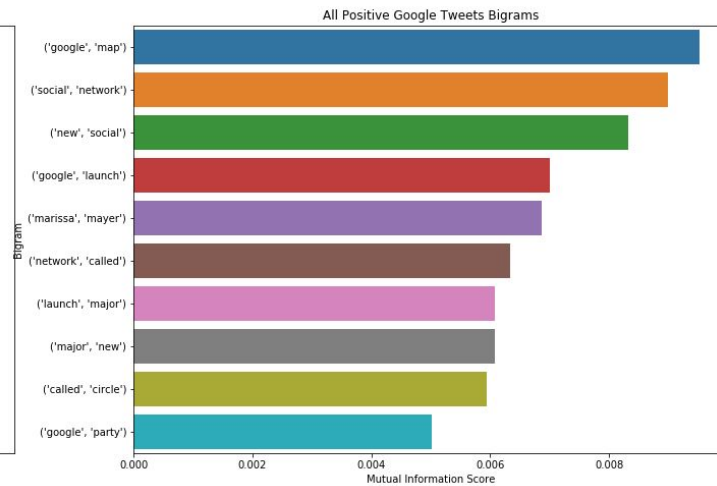
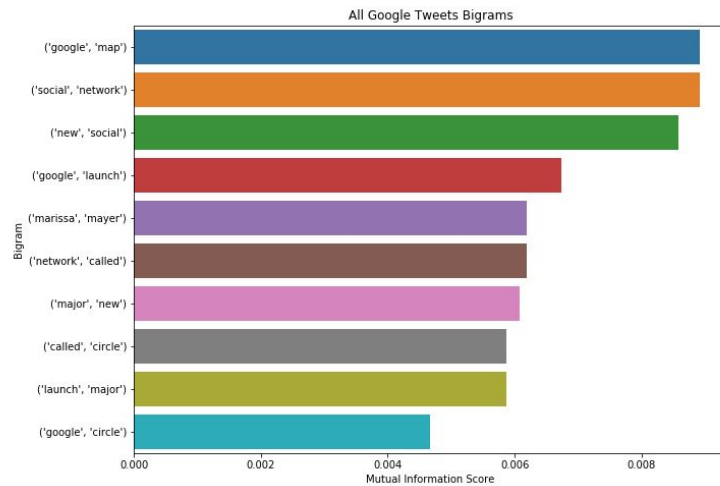
Evaluation Metrics by Multi-Class Model Type



Recommendations

The background of the slide is a blue gradient, transitioning from a darker blue on the left to a lighter blue on the right. Overlaid on this gradient is a complex network of white lines and dots, forming various geometric shapes like triangles and polygons. These shapes are interconnected, creating a web-like structure that suggests a network or a system of relationships. The lines are thin and white, while the dots are small and white, acting as nodes in the network.





Recommendations

- Utilize this tool to quickly analyze future tweets at nearly 90% accuracy
- Investigate positive sentiment for products that draw successful buzz
- Investigate negative sentiment tweets for areas of needed improvement

Future Work

- Breaking down positive and negative tweets by even more granular product details
- Establishing even more precise and generalizable data by collecting from different sources and over a longer time frame in order to eventually reduce or eliminate the need for human labeling



Thank you!

Any Questions? Feel free to email us

shsobieski@gmail will4856@gmail