Phase4 Project - Natural Language Processing

By Rahul Krishnan

Natural language processing

What is Natural Language Processing?

- NLP refers to the branch of computer science concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

- NLP combines computational linguistics—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment.

source: [NLP definition]

Sentiment Analysis – Acme Online

Business Objective

1. Analyze tweets to check what customers are talking about.

2. Analyze tweets to identify the most popular product - pts 1&2 can be used to tweak Acme Online's inventory accordingly.

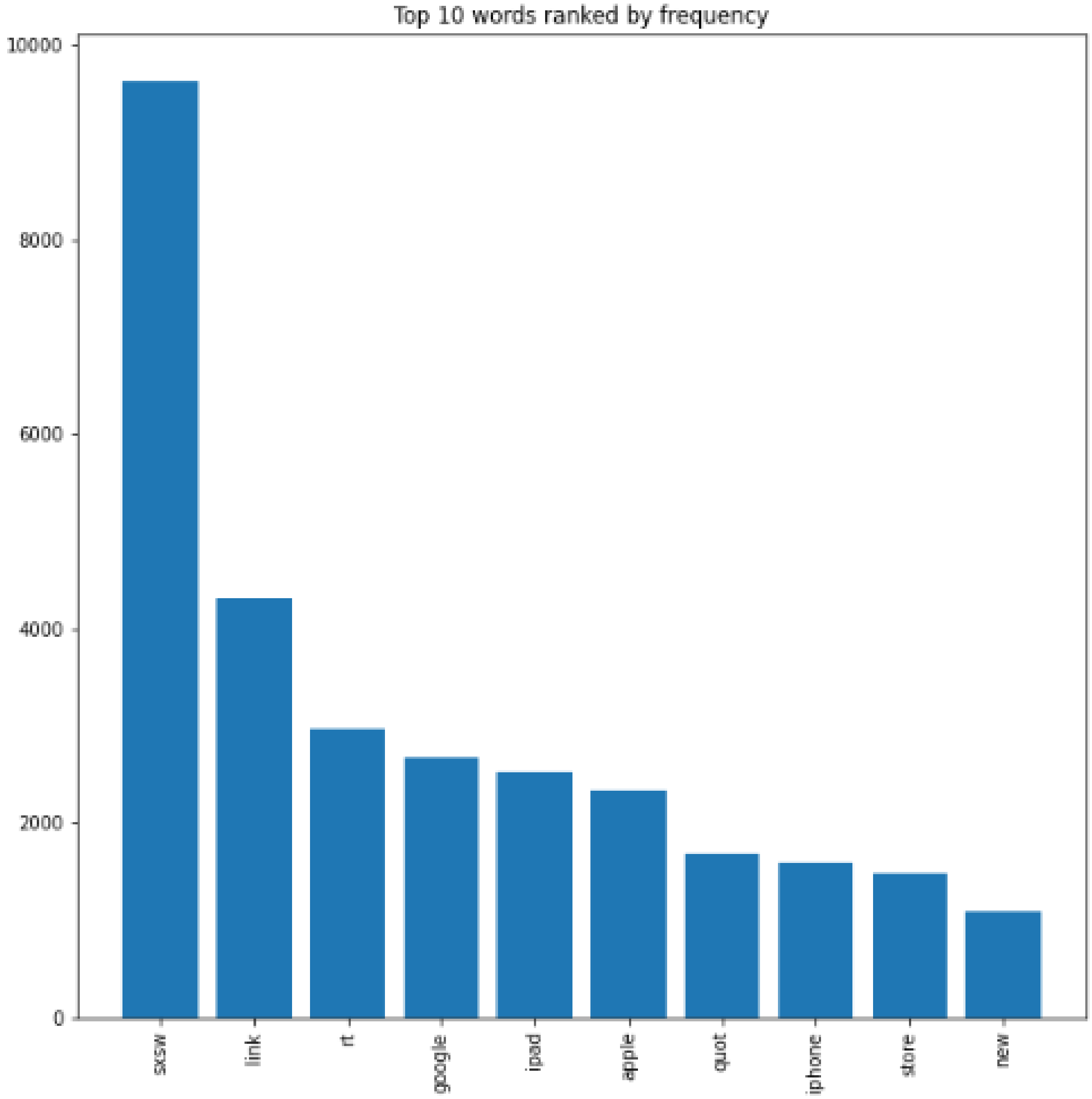
3. For each product, we will look to see what customers like/dislike to identify opportunities for improvement, if applicable.

4. Since human intervention was used identify products based on tweets, we will attempt to build a model using **NLP (Natural Language Processing)** to automate this. We will use the f1-score for model evaluations since minimizing False Positive and False Negatives is desirable.

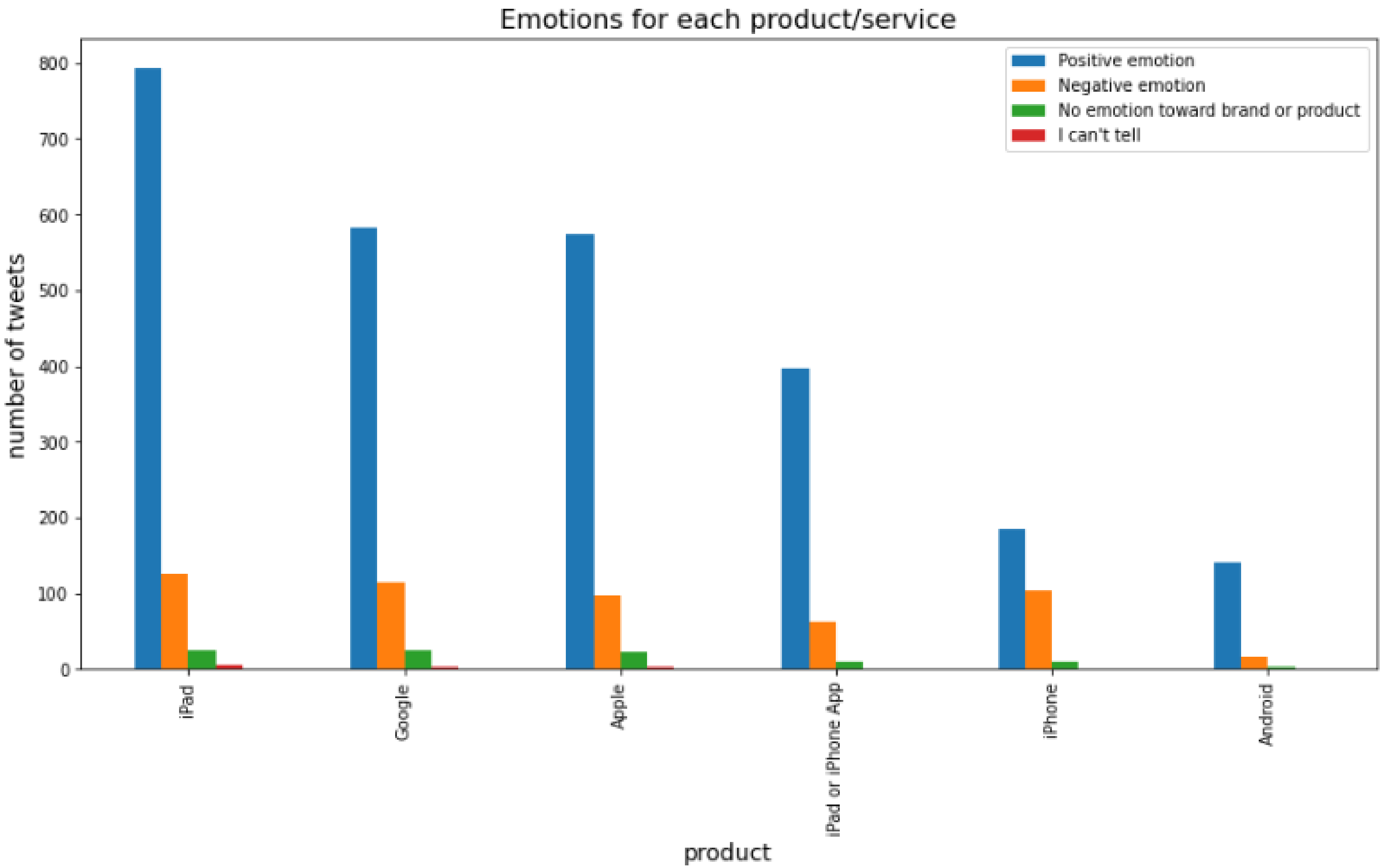
Dataset

Dataset sourced from CrowdFlower via data.world: https://data.world/crowdflower/brands-and-product-emotions

Trending Topic

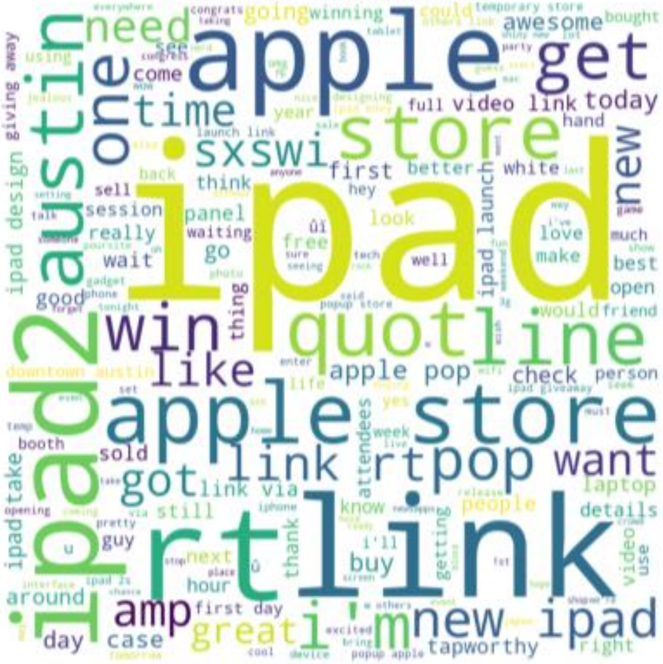


Most Popular Product

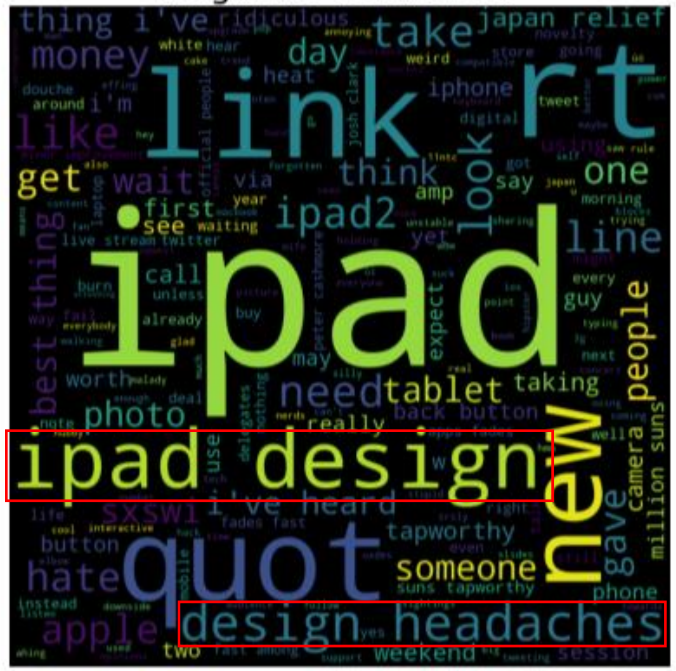


Likes/Dislikes

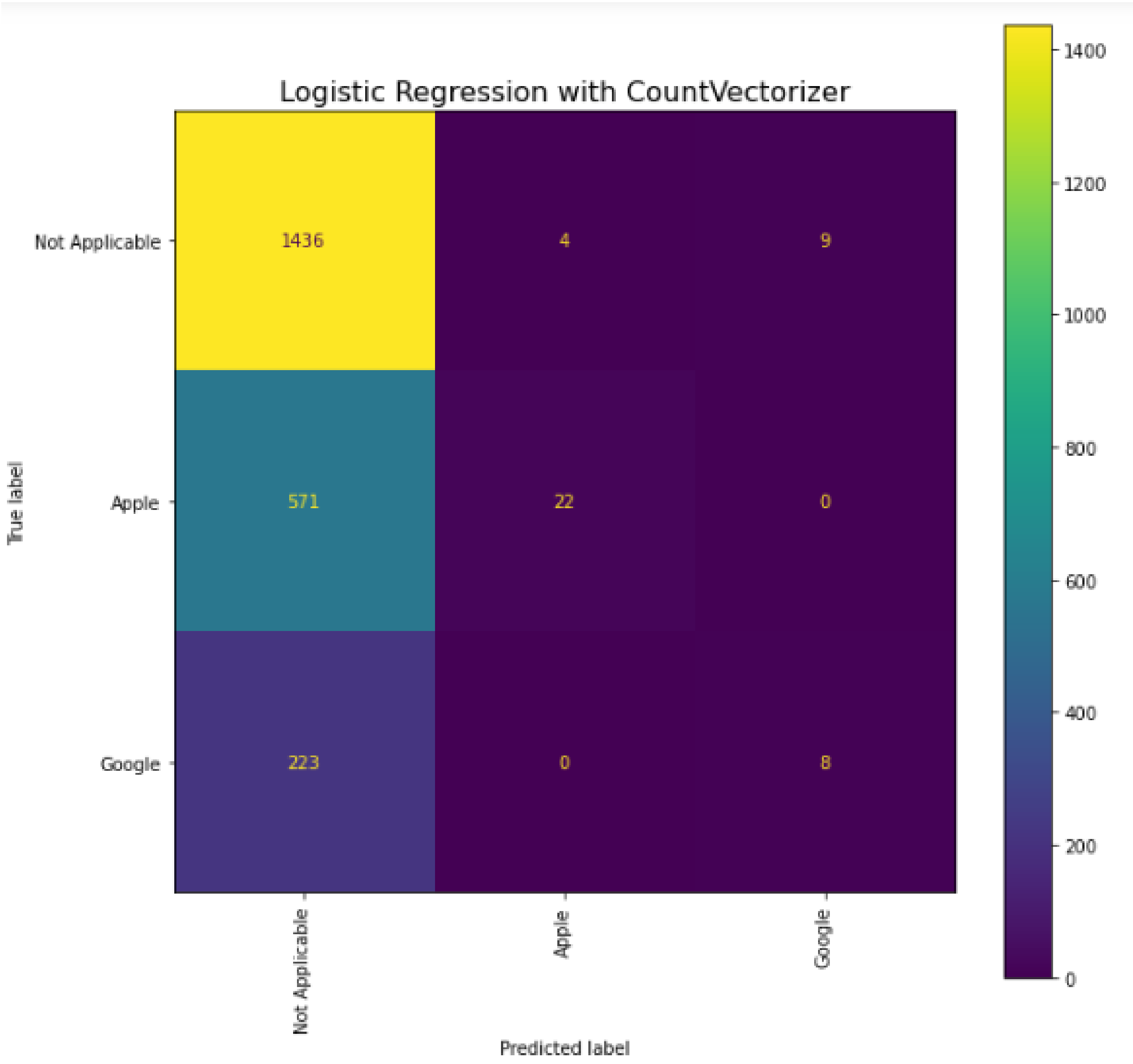
Positive words:iPad



Negative words:iPad



Best Model


Logistic Regression with CountVectorizer
True label
Not Applicable
Apple
Google
1436
4
9
571
22
0
223
0
8
Not Applicable
Apple
Google
Predicted label
1400
1200
1000
800
600
400
200
0

Conclusions

Recommendation:

- Since the iPad is the most popular product, Acme Online could look for opportunities to boost sales. Acme Online could also look to expand their portfolio by offering tablets from other manufacturers to see if this will bring more revenue.

Models:

- Part-of-Speech tagging can be used to create more features.
- Ensemble methods like XGBoost and Adaboost can also be trialed for modelling along with other word embedding techniques like fastText and Glove.

Limitations:

- More *varied* data is desirable. Current data is very imbalanced and localized to a time and place thus possibly skewing forecasts for the future. As next steps, that would be the starting point. We can then use it to improve model efficiency.