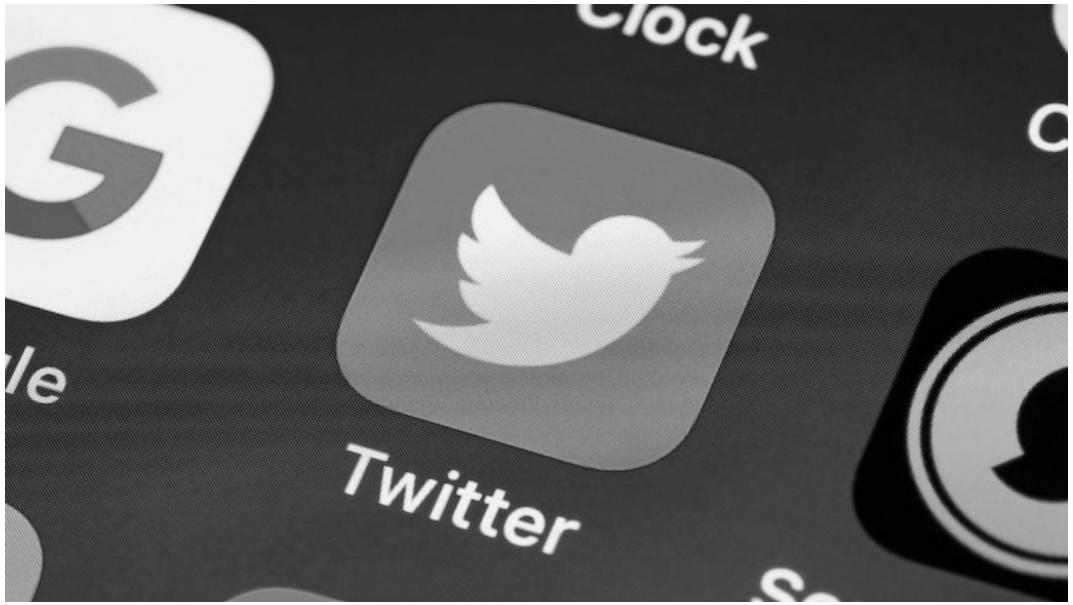
NATURAL LANGUAGE PROCESSING – TWITTER SENTIMENT CLASSIFICATION

CATHERINE FRITZ – PROJECT 4

7/15/2021

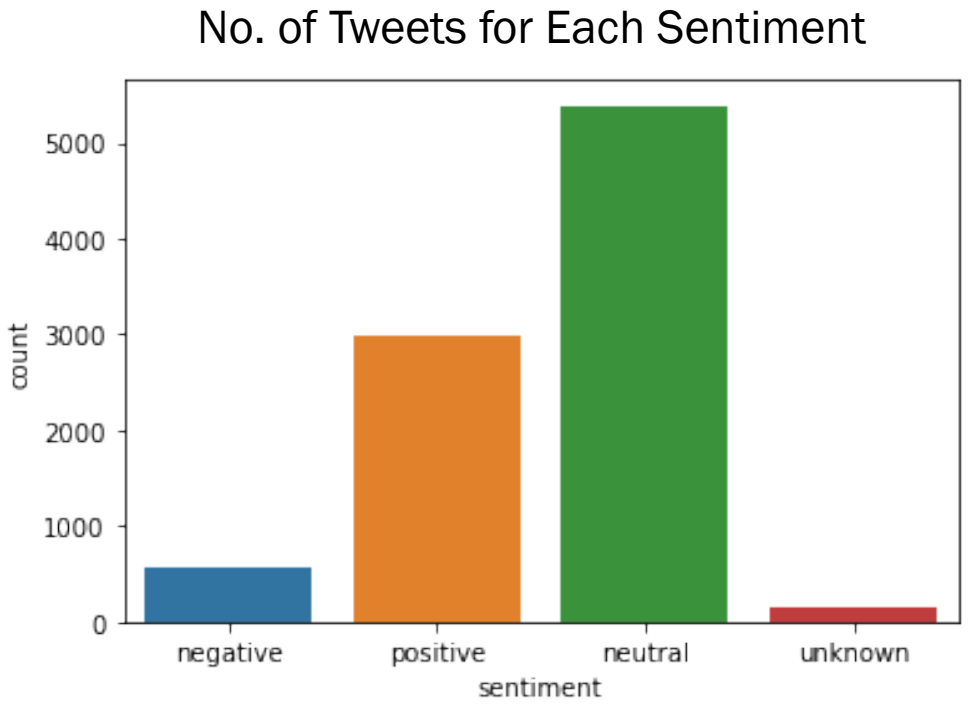


BUSINESS UNDERSTANDING

- Businesses need to get feedback on their products
- Product reviews only one source
- Informal reviews on social media
- Predict sentiments based on social media posts

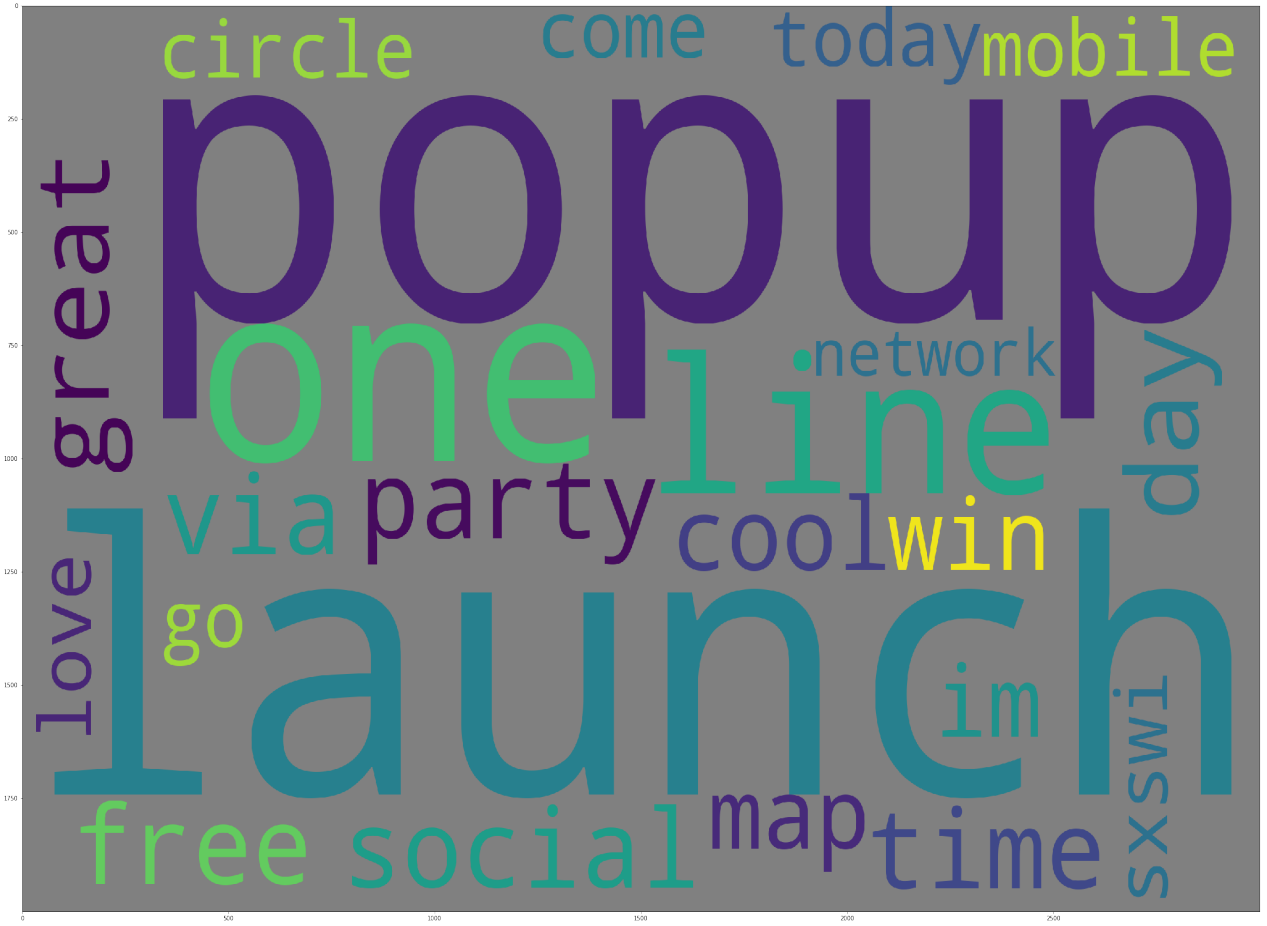
DATA UNDERSTANDING

- Data classified by a human as positive, negative, neutral, or if the sentiment is unknown
- Subject of the tweets center around Apple or Android products.
- Data imbalance



DATA PREPARATION

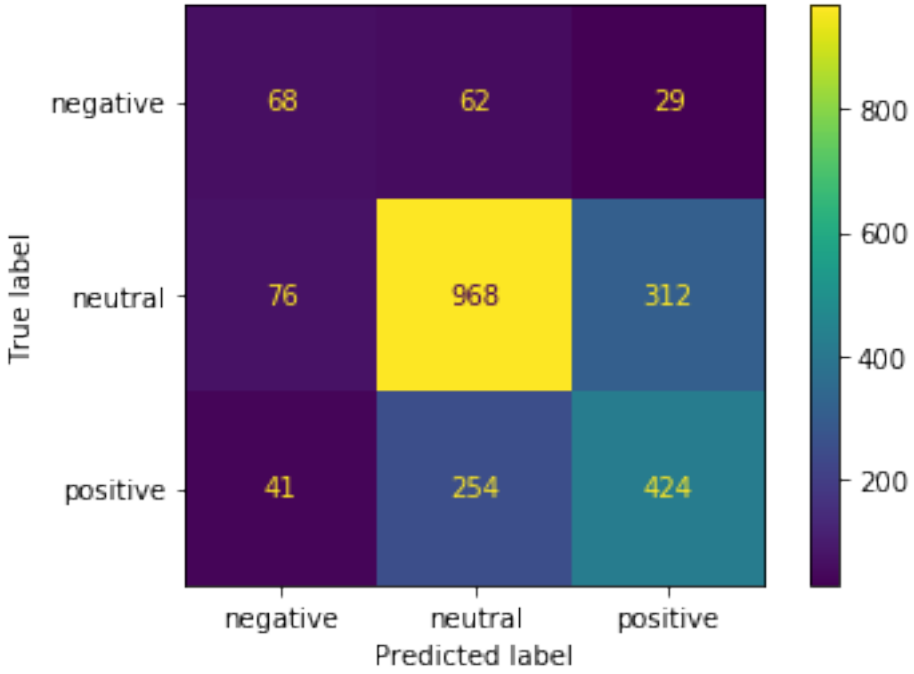
Top Words for Positive Sentiment



- To make the data suitable for modeling, the following steps were taken:
- Remove undesirable characters
- Remove Twitter specific text like @ tags
- Tokenize the text
- Remove common stop words for the English language and also specific to Twitter (like "RT" or "link").
- Lemmatize the stop words to consolidate similar words

MODEL & EVALUATION – LOGISTIC REGRESSION

- Winning Method: Logistic Regression
- Captures negative and positive data
- Training Accuracy is: 87.0%
- Validation Accuracy is: 65.4%



CONCLUSION

- Based on the current model and data, can predict if a tweet is positive or neutral.
- Not good at picking up on negative sentiment, which is probably the most useful to know.

FUTURE WORK

1. Solve class imbalance problem
2. Get more data

THANK YOU

NAME: CATHERINE FRITZ

EMAIL: CMFRITZ0@GMAIL.COM

GITHUB: @CMFRITZ

LINKEDIN: LINKEDIN.COM/IN/CATFRITZ

ADDITIONAL INFORMATION CAN BE FOUND AT

HTTPS://GITHUB.COM/CMFRITZ/PROJECT_4_NLP