Twitter NLP Project

Meredith Newhouse

Purpose

The purpose of this project is to build a machine learning algorithm to predict the emotion of a tweet based on its content. To train the algorithm, the project will use tweets focused on a SXSW event that were manually categorized into positive and negative emotions.

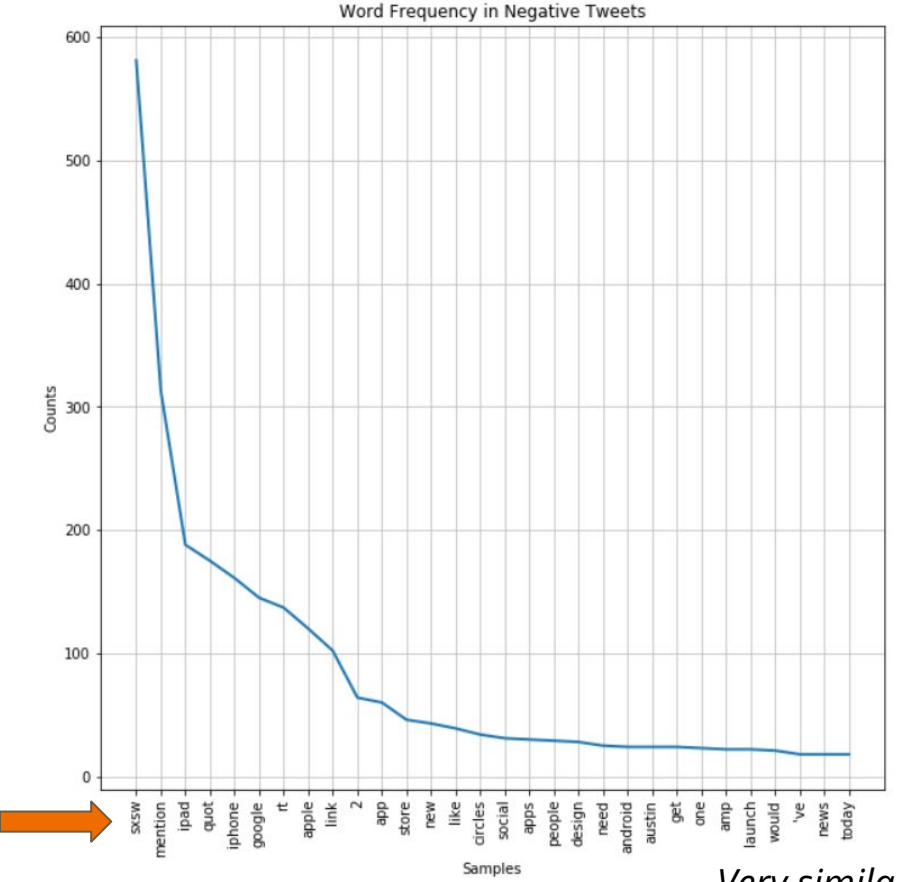
Data Used

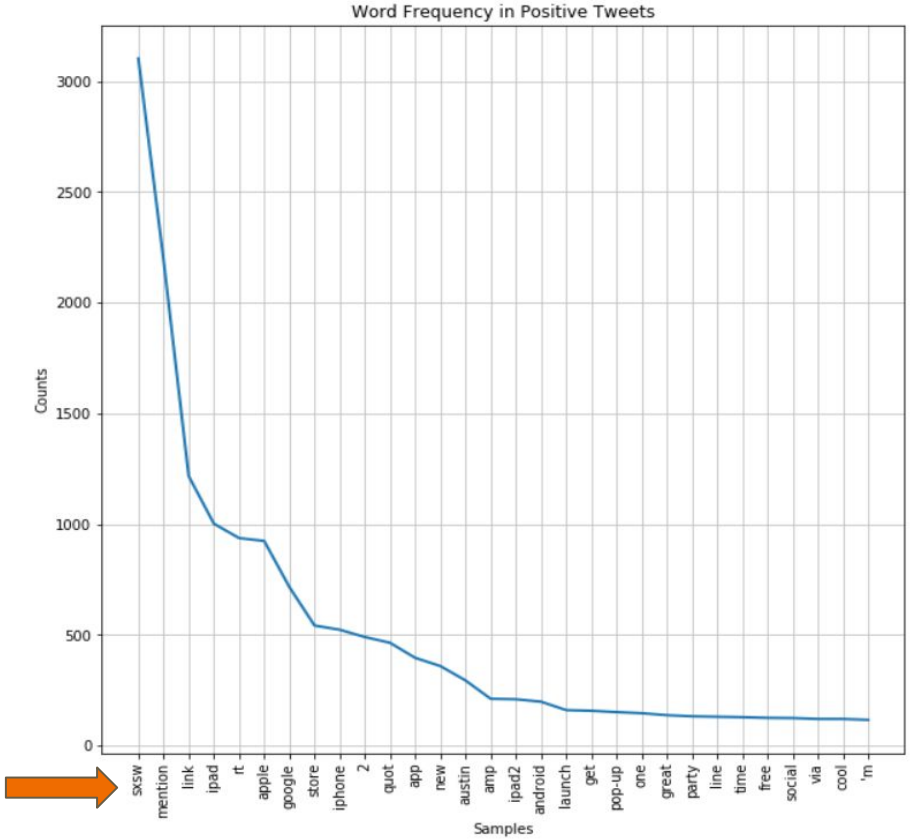
The data set was found on data.world

https://data.world/crowdflower/brands-and-product-emotions

In the data set, 2978 tweets were categorized as expressing a positive emotion. 570 tweets were categorized as expressing a negative emotion.

Exploring the Data



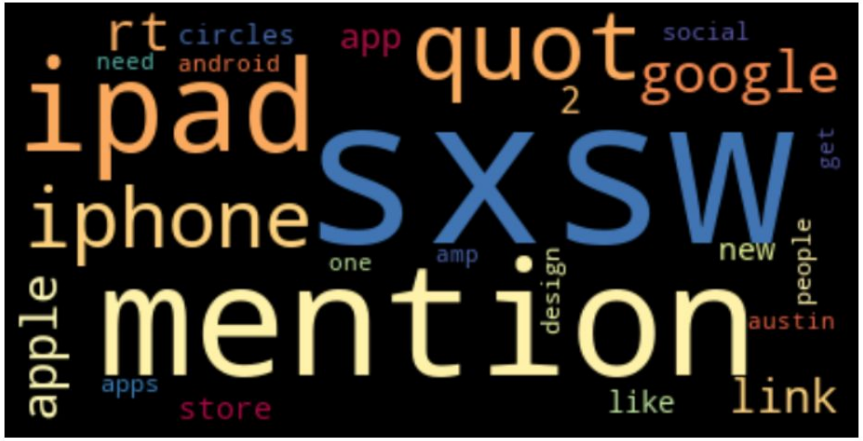


Very similar most common words

Word Clouds

Positive Tweets

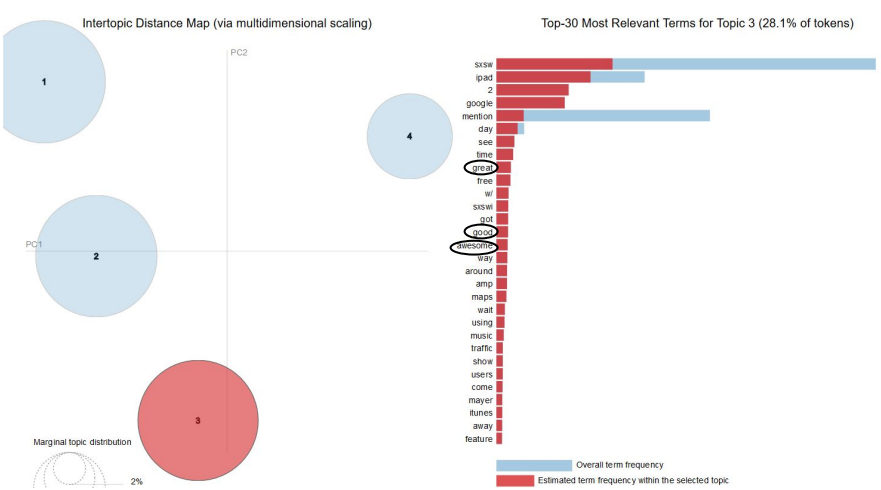
Negative Tweets



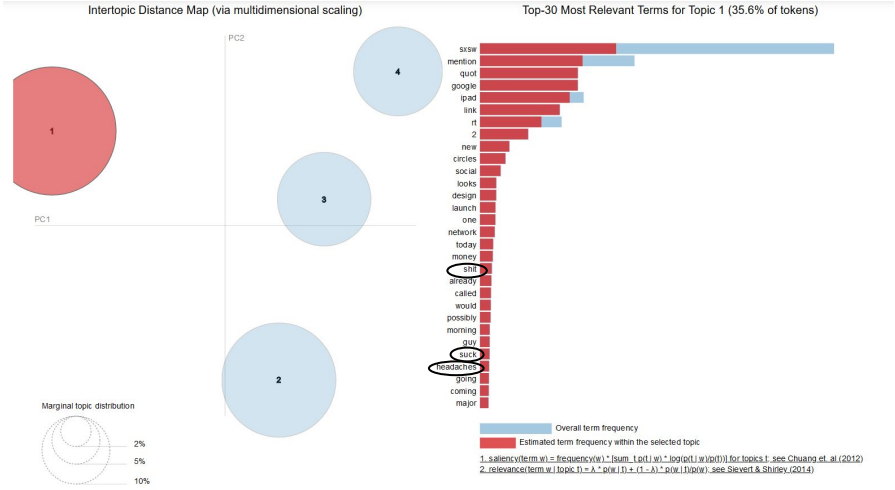
The similarity between words in the tweets may make it difficult for the model to differentiate between the two emotions.

Word Exploration

Positive Emotion Words



Negative Emotions Words



To explore what words may be used in the models for each emotion, I ran an LDA model to look at the types of words within each emotional category.

Modeling

Baseline Model:

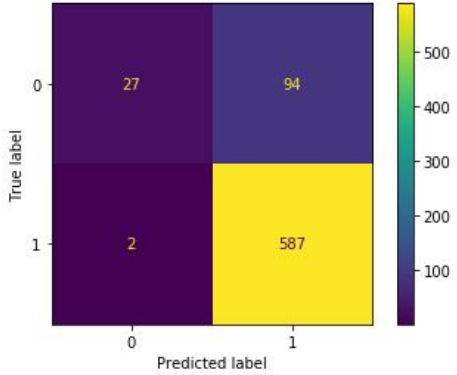
- The baseline model was a random forest classifier model with tfidf as a vectorizer.
- The model has clear overfitting, and is better at predicting the positive emotion than the negative emotion

Model Time: 0.8967099189758301

Training:	precision	recall	f1-score	support
0	1.00	1.00	1.00	449
1	1.00	1.00	1.00	2389
accuracy			1.00	2838
macro avg	1.00	1.00	1.00	2838
weighted avg	1.00	1.00	1.00	2838

Testing:	precision	recall	f1-score	support
0	0.93	0.22	0.36	121
1	0.86	1.00	0.92	589
accuracy			0.86	710
macro avg	0.90	0.61	0.64	710
weighted avg	0.87	0.86	0.83	710

Mean Absolute Error: 0.1352112676056338
Mean Squared Error: 0.1352112676056338
Root Mean Squared Error: 0.36771084782153735



Final Models

Best model:

- Multinomial Naive Bayes model performed the best
- Best ratio between correct negative predictions and correct positive predictions
- However, there is overfitting

Model Time: 0.17304468154907227

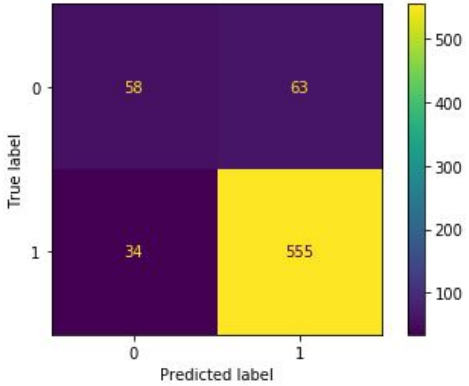
Training:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	449
1	1.00	0.99	1.00	2389
accuracy			1.00	2838
macro avg	0.99	1.00	0.99	2838
weighted avg	1.00	1.00	1.00	2838

Testing:

	precision	recall	f1-score	support
0	0.63	0.48	0.54	121
1	0.90	0.94	0.92	589
accuracy			0.86	710
macro avg	0.76	0.71	0.73	710
weighted avg	0.85	0.86	0.86	710

Mean Absolute Error: 0.13661971830985917
Mean Squared Error: 0.13661971830985917
Root Mean Squared Error: 0.3696210468978453



Final Models Cont.

- Tuning the models required balancing between positive emotion recall and negative emotion recall
- Some models could predict all positive emotion tweets correctly, but were not able to predict any negative emotion tweets.
- This Random Forest model has less overfitting than the Naive Bayes model
- However, its scores are lower overall.

Model Time: 0.3720052242279053

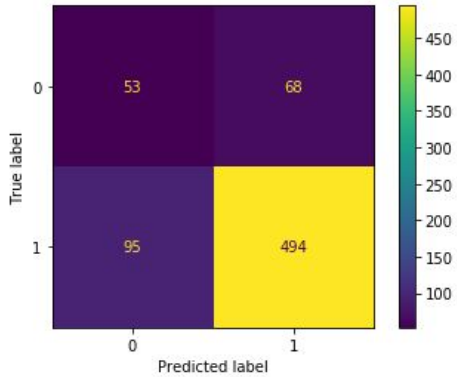
Training:

	precision	recall	f1-score	support
0	0.45	0.61	0.52	449
1	0.92	0.86	0.89	2389
accuracy			0.82	2838
macro avg	0.69	0.73	0.71	2838
weighted avg	0.85	0.82	0.83	2838

Testing:

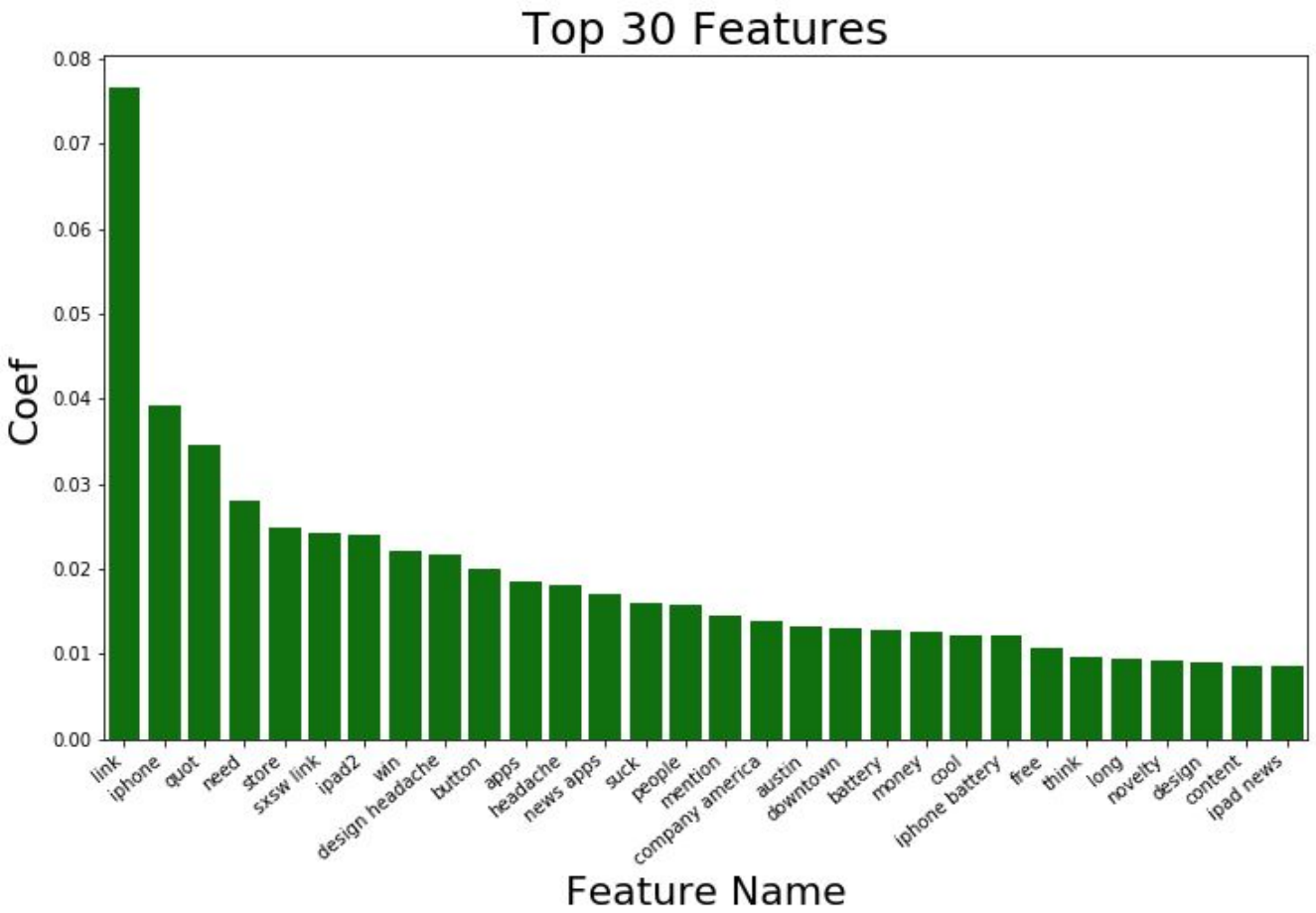
	precision	recall	f1-score	support
0	0.36	0.44	0.39	121
1	0.88	0.84	0.86	589
accuracy			0.77	710
macro avg	0.62	0.64	0.63	710
weighted avg	0.79	0.77	0.78	710

Mean Absolute Error: 0.2295774647887324
Mean Squared Error: 0.2295774647887324
Root Mean Squared Error: 0.47914242641278637



Interpretation

- Bigrams were used in the model to help contextualize words
- A max features of 1500 was used.
- Some notable words used in the Random Forest model:
  - ‘Design headache’
  - ‘Suck’
  - ‘Cool’
  - ‘Win’
  - ‘need’



Conclusion

- A model can be developed to categorize the emotion of a tweet.
- Some words used to categorize the tweet were
  - ‘Design headache’
  - ‘Suck’
  - ‘Cool’
  - ‘Win’
  - ‘Need’
- More data is needed to better categorize the emotions, especially in the negative emotion category.
- A trade off between positive emotion recall and negative emotion recall was required.
- Using bigrams, pairs of words, may help to provide context to improve the model.

Next Steps

- Collect more data, especially tweets with the negative emotion
- Integrate more tools to deal with overfitting like ensemble modeling methods and training more data.
- Try other methods like PCA or undersampling to deal with the class imbalance.

Thank You

Check out my github repo here: https://github.com/newhousem/NLPProject

Contact me: meredithnewhouse@gmail.com

Thank you to data.world for providing the data sets used in this analysis and Yish for helping to answer all of my questions.