



TWITTER SENTIMENT CLASSIFICATION

Nick Gigliotti

Business Problem



Create an accurate predictive model that can classify novel tweets as expressing positive, negative, or neutral sentiment.



Find out what people are saying about Apple and Apple products at South by Southwest, 2011.



Make some PR recommendations for the period immediately following the event.

Dataset

Tweets related to South by Southwest, 2011

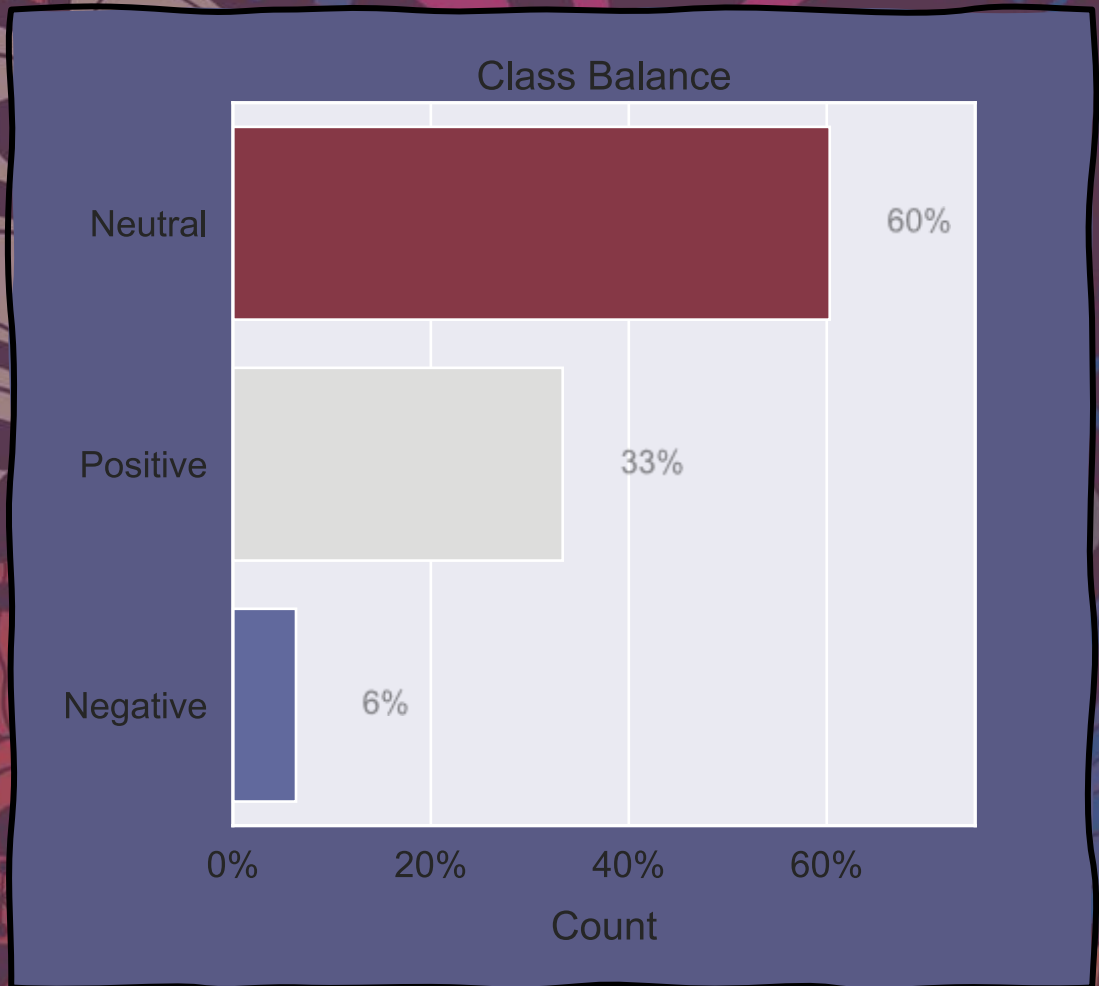
- Annual conference and arts festival
- Austin, TX
- Major site for iPad 2 Launch

Crowdsourced sentiment labels

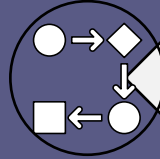
~9k tweets: just text and labels

- Tweets are labeled with *Positive*, *Negative*, *Neutral*, or *Uncertain*
- Tweets are labeled with brands and products related to the sentiment

Sparse 'Negative' class

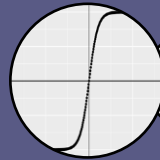


Methods



Iterative Process

- Many parameter sweeps
- Careful attention to detail



Logistic Regression

- Performed better than:
 - Naïve Bayes
 - Linear SVM



Preprocessing

- Lowercase
- Porter stem
- Mark negation



TF*IDF Vectorization

- Term Frequency * Inverse Document Frequency
- Balances term count with **corpus-wide rarity**
- A high score means high term importance



VADER Analysis

- Valence Aware Dictionary and Sentiment Reasoner
- Rule-based sentiment analysis tool
- Gives each tweet a Positive, Negative, Neutral, and Compound polarity score

Model

Balanced Accuracy

- **Score:** 0.66
- Equivalent to average recall
- Equivalent to average of diagonal

High Negative Recall

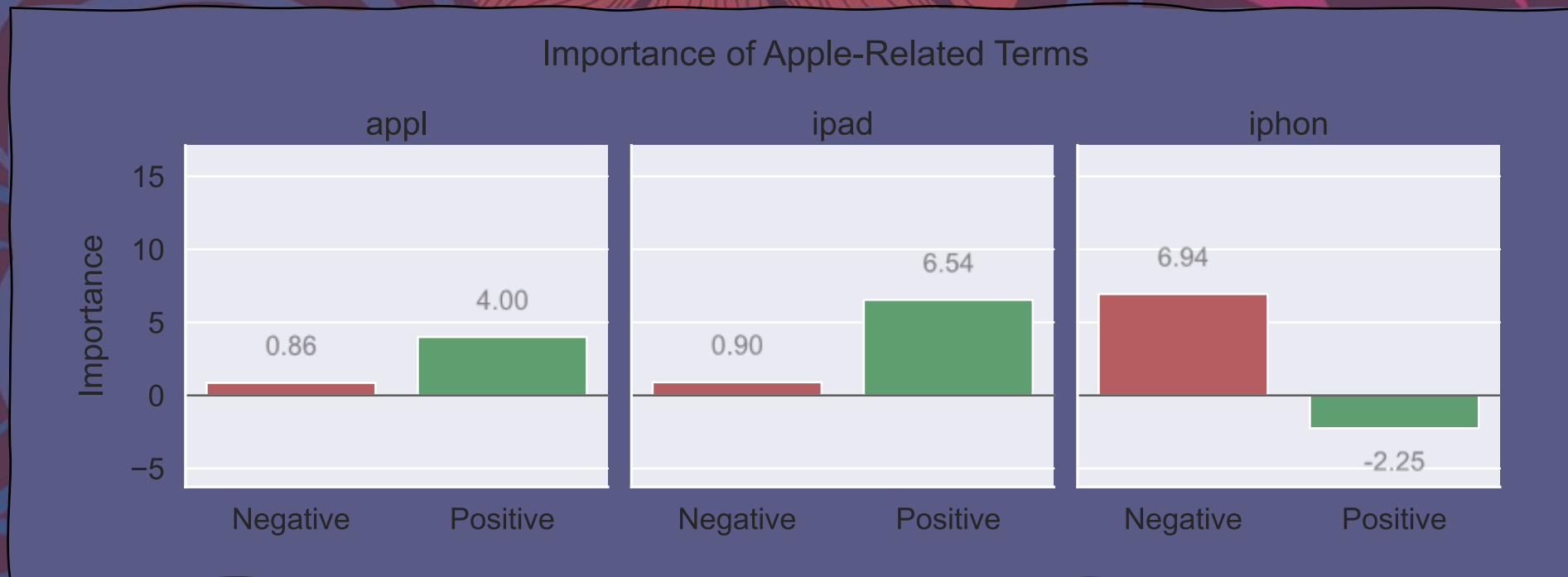
- Negative Support: 6% of samples
- Positive Support: 33% of samples
- **Equal Positive and Negative recall**

Sacrifices Neutral Recall

- Neutral Support: 60% of samples
- Most common and least important
- Not a subject of research

		Normalized Confusion Matrix		
True label	Negative	0.67	0.18	0.16
	Neutral	0.12	0.65	0.23
	Positive	0.087	0.24	0.67
		Negative	Neutral	Positive
		Predicted label		

- Top features for prediction are TF*IDF words.
- 'batteri' and 'iphon' are highly associated with Negative.
 - Complaints about iPhone battery life
- 'ipad' is highly associated with Positive.
- Punctuation and common terms like 'austin' and 'sxsw' are associated with Neutral.



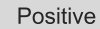
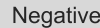
Apple-Related Coefficients

Observations:

- 'iphon' is associated with Negative
- 'ipad' is associated with Positive
- 'apple' is associated with Positive
- People are much more excited about iPads than iPhones.
 - Tweets from iPad 2 launch.

One of Apple's values?

Relinquish some control to send the message that people are free to use Apple's products however they wish.



Apple is known for wanting total control of its products, even well after they leave the shelf.

The original comment was in response to the suggestion that Apple might create its own proprietary newsstand for e-magazines.

Flipboard:
game-changer
or just a fad?
(2011)

- Many people are complaining about the iPhone's limited battery life.
- Others are rejoicing at having a charger (see the positive word cloud).
- Some people mocked the spotty AT&T service which many iPhone users have.
 - You already solved this problem by allowing Verizon to carry the iPhone.

- Improve the iPhone's battery life and turn it into a new marketing campaign. People will pay attention.

iPad 2 Popup Store was a Success

- People couldn't stop talking about the iPad 2 popup store.
- Terms like 'shiny new' and 'jealous' have high importance.
- On the negative side, there was viral tweeting about donating one's iPad 2 money to Japan relief.
- 'design headaches' had to do with a benign talk given about iPad app design.

Recommendation:

- Do another temporary popup store for your next big product launch.



Conclusion

- I built an accurate model (0.66 balanced accuracy) despite the small and noisy dataset.
- There is still room to improve the model.
- I arrived at three recommendations:
 - Send the message that Apple cares about individual freedom by relinquishing control in a public manner.
 - Improve the iPhone's battery life and use it as a rallying point for a marketing campaign.
 - Open another temporary popup store for your next big product launch.



		Normalized Confusion Matrix		
True label	Negative	0.67	0.18	0.16
	Neutral	0.12	0.65	0.23
	Positive	0.087	0.24	0.67
		Negative	Neutral	Positive
		Predicted label		

Future Work



- Stacking Classifiers
 - Combine the results of multiple classifiers.
 - Increases accuracy and both positive and negative recall.
- Sophisticated Vectorization
 - Use a Doc2Vec model trained on a large corpus of tweets.
 - Explore other sophisticated vectorization algorithms.



THANKS

Nick Gigliotti
ndgigliotti@gmail.com