“Tells” in Tweets: Developing a reliable Twitter sentiment analysis classifier for tech companies

Michael Burak

Problem statement:

-Classifying tweets about your brand *reliably* is key.

-Twitter data is full of noise but can be very valuable.

-Developing a way to sift through the noise is key.

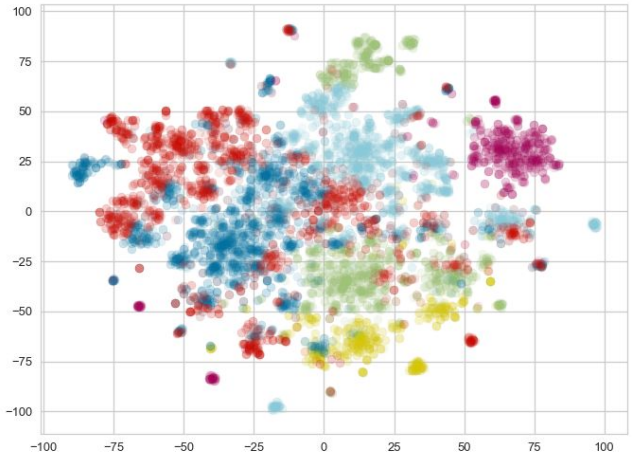
Sample Clusters:

Dark blue : ipad, sxsw, link, mention, apple, rt, the, design, new, line

Light green : network, social, circles, called, launch, major, new, today, possibly, google

Red : apple, store, popup, sxsw, link, austin, ipad2, mention, line, open

Clusters of Similar Words



Business value:

-Reliable classifiers: good classifiers for the modern world.

-Through studying Apple and Google’s wins and missteps, stand on the shoulders of giants.

-The data: Tweets about Apple and Google during SXSW 2011 and emotions expressed.

Methodology: an overview

-Sentiment analysis: The study of emotions in text.

-Cohen's Kappa: Focus on reliability, how much better categorizing with a model is compared to guessing, given the data categories.

Cohen's Kappa statistic (κ)	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.20–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

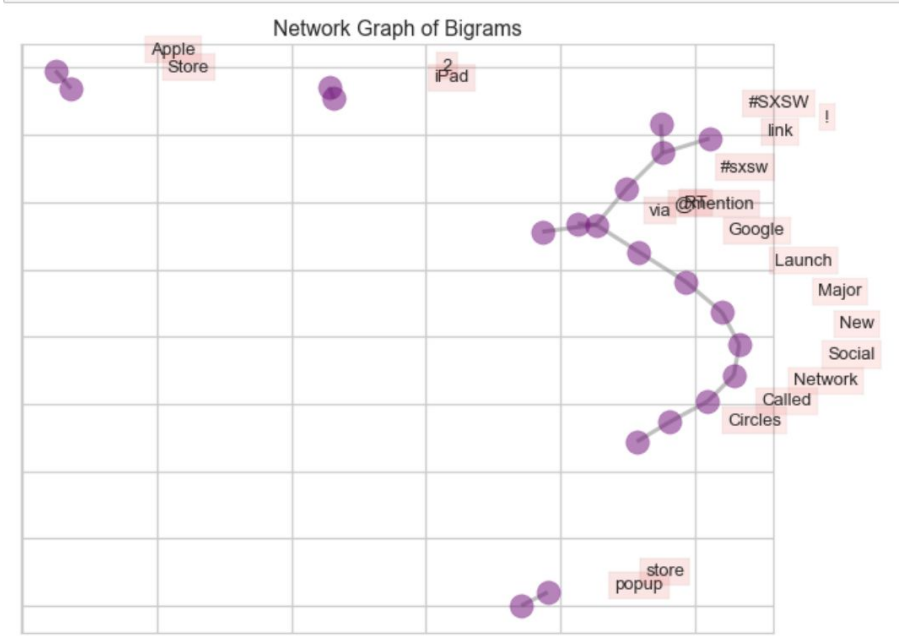
-Naïve Bayes: Estimates probabilities, classifies well(with some tweaks.)

Business recommendation 1:

-Go big: capitalize on big events.

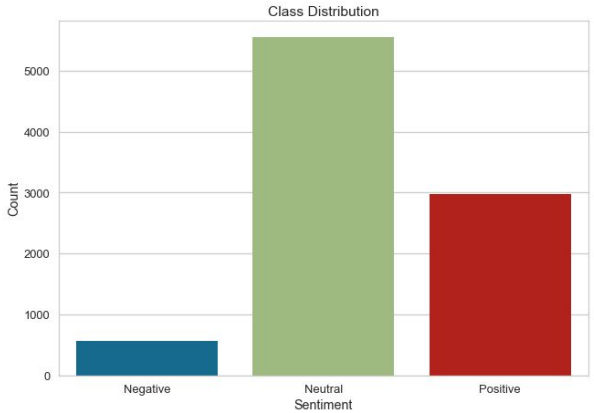
-Apple’s popup store and iPad 2

-Google’s social network “Circles”



Business recommendation 2:

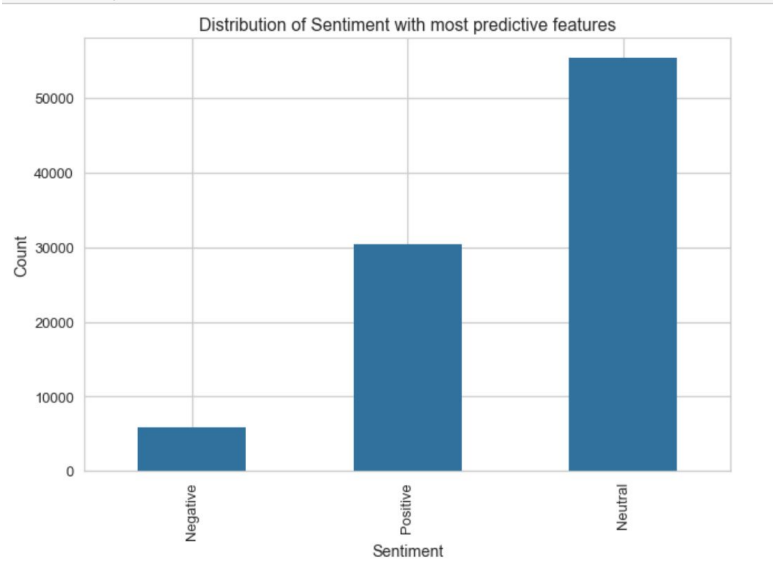
-Keep your data clean and representative: data integrity and imbalances' effects.



Business recommendation 3:

- Utilize reliable classifiers & metrics to get consistent results with noisy, changing data like tweets to keep up to date.

-Model results: Final model Cohen's Kappa: .5 / fair reliability.



Future Work:

-Anomaly detection

-Deep Learning

-More Machines(Support Vector that is.)

Thank you!