Using Social Media for Disaster Prediction and Warnings

By Benjamin Haile

March 13, 2020

Problem Statement

- While traditional methods for alerting on events such as hurricanes and tornadoes rely on information derived from official sources (e.g. USGS), this project aims to utilize Twitter activity to identify such an event. In practice, once the event is predicted, an alert can then be sent out across social media. **The outcome of this project will be a binary classification model that can analyze tweets and use them to predict whether a disaster is present and a warning must be sent.** As a proof of concept, this project will use archived tweets collected during the most dangerous days of Hurricane Sandy in 2012. The project's terminology will center around that of hurricanes specifically. In this situation, predicting no emergency while a hurricane approaches (false negative) is a much more dangerous outcome than predicting a hurricane when there is none (false positive). **Models will therefore be evaluated on recall as well as accuracy.**

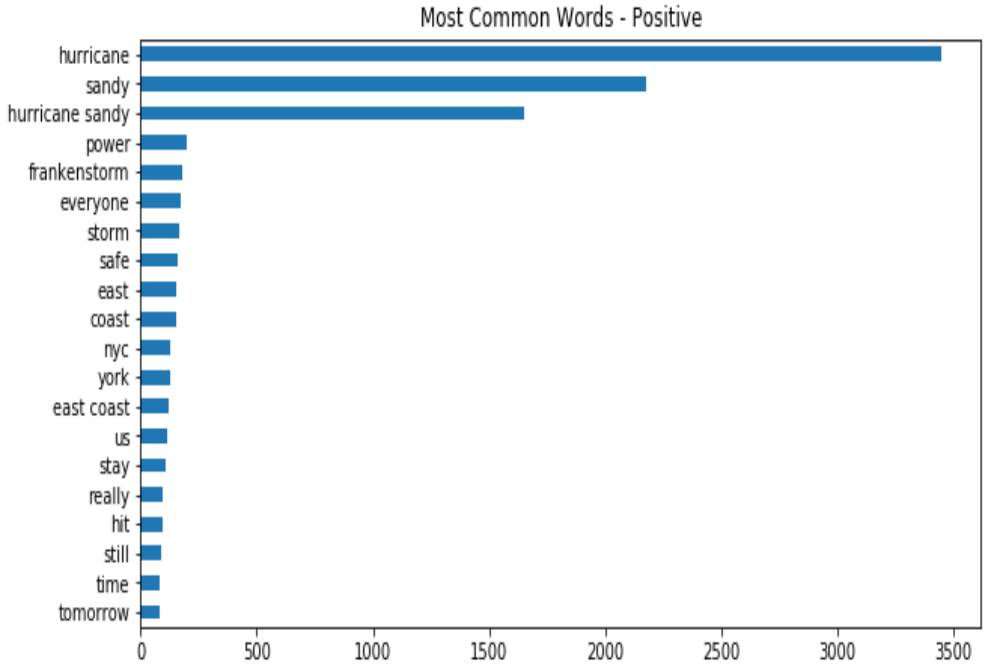
Data Source

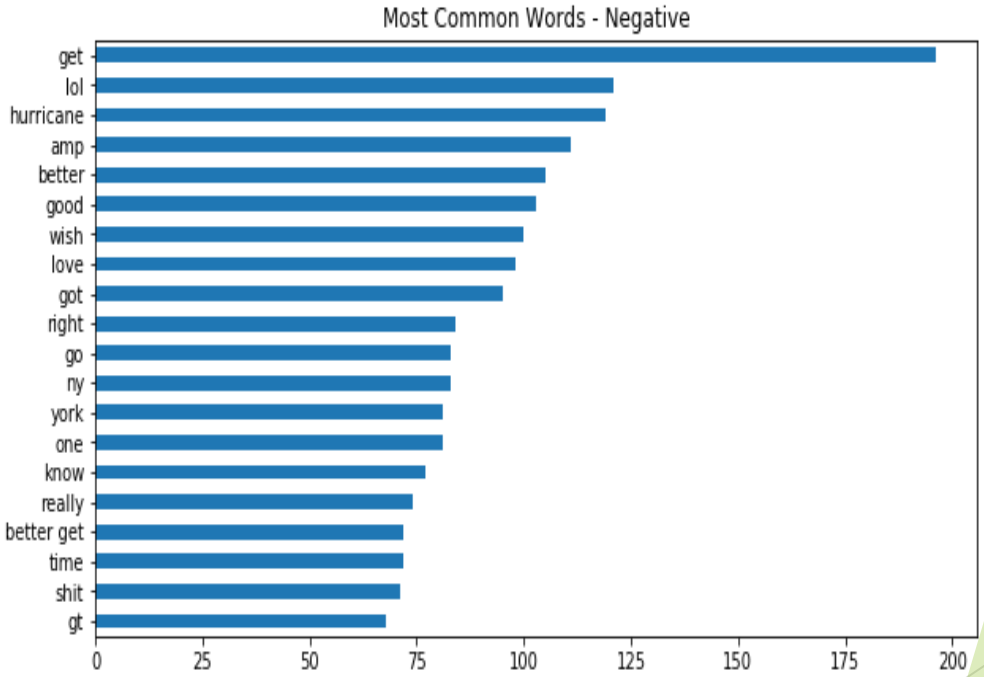
- CrisisLex
- Over 10,000 rows
- October 28-30, 2012
- Coastal New York and New Jersey
- Next: using Twitter API to collect live data or location and time

Methodology

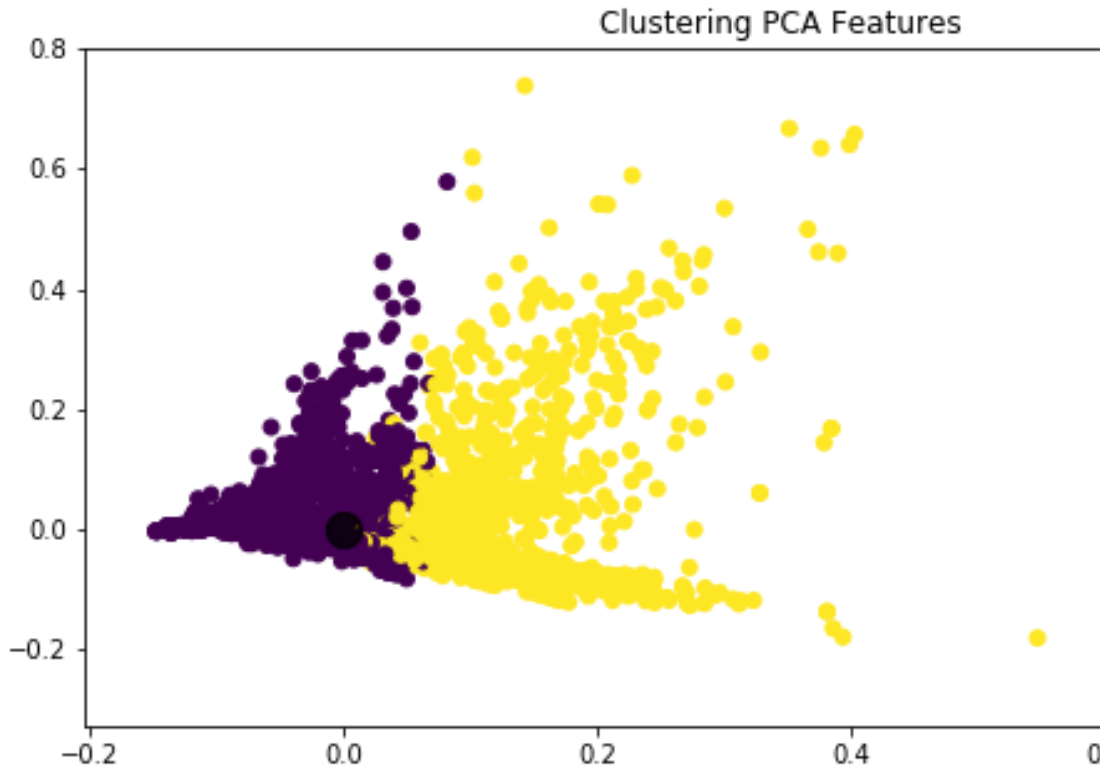
- Off-topic = 0, on-topic = 1
- Observe most common words, add to stop words if necessary
- Cluster the classes
- Run models with different versions of the stop words lists
- Experiment with two vectorizers

Comparing Common Words





Kmeans Clustering



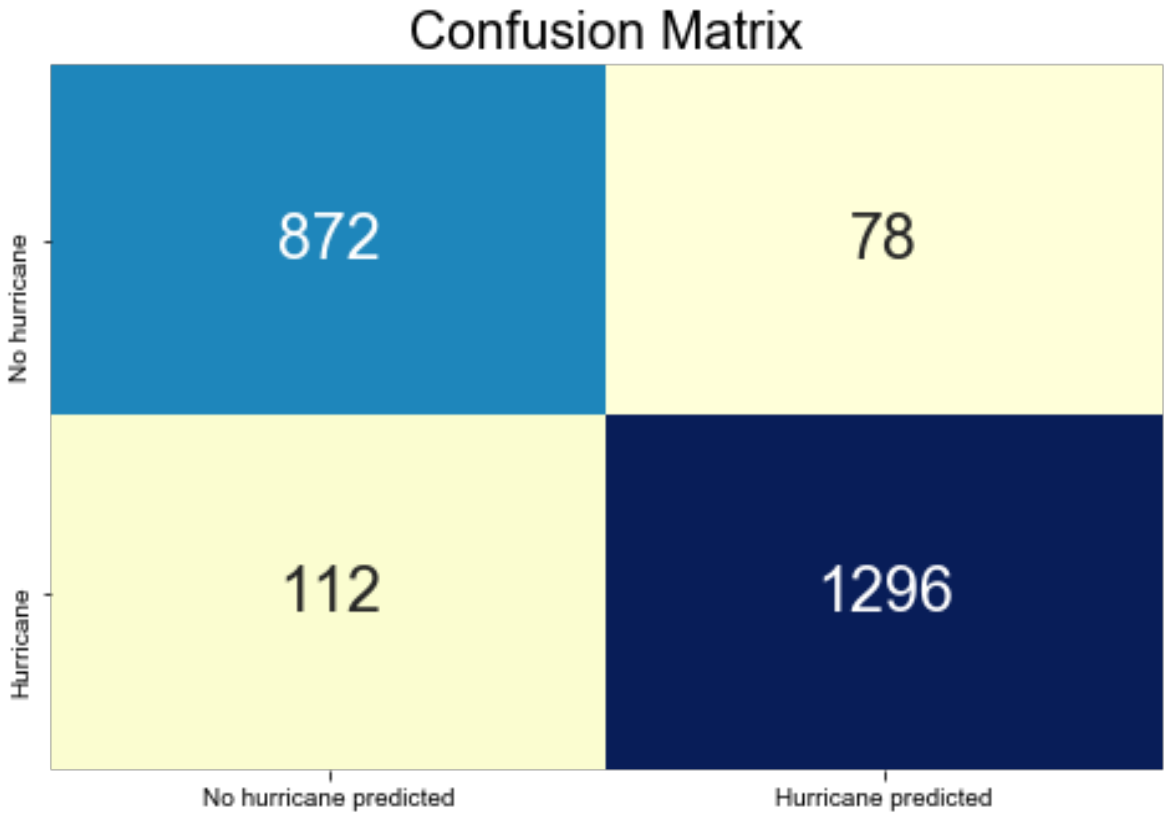
- Silhouette score = 0.0026
- Corpus transformed with TFIDFVectorizer
- PCA to view in 2D

Modeling

- Logistic Regression with CountVectorizer
- Logistic Regression with TFIDFVectorizer
- Random Forest with CountVectorizer
- Random Forest with TFIDFVectorizer

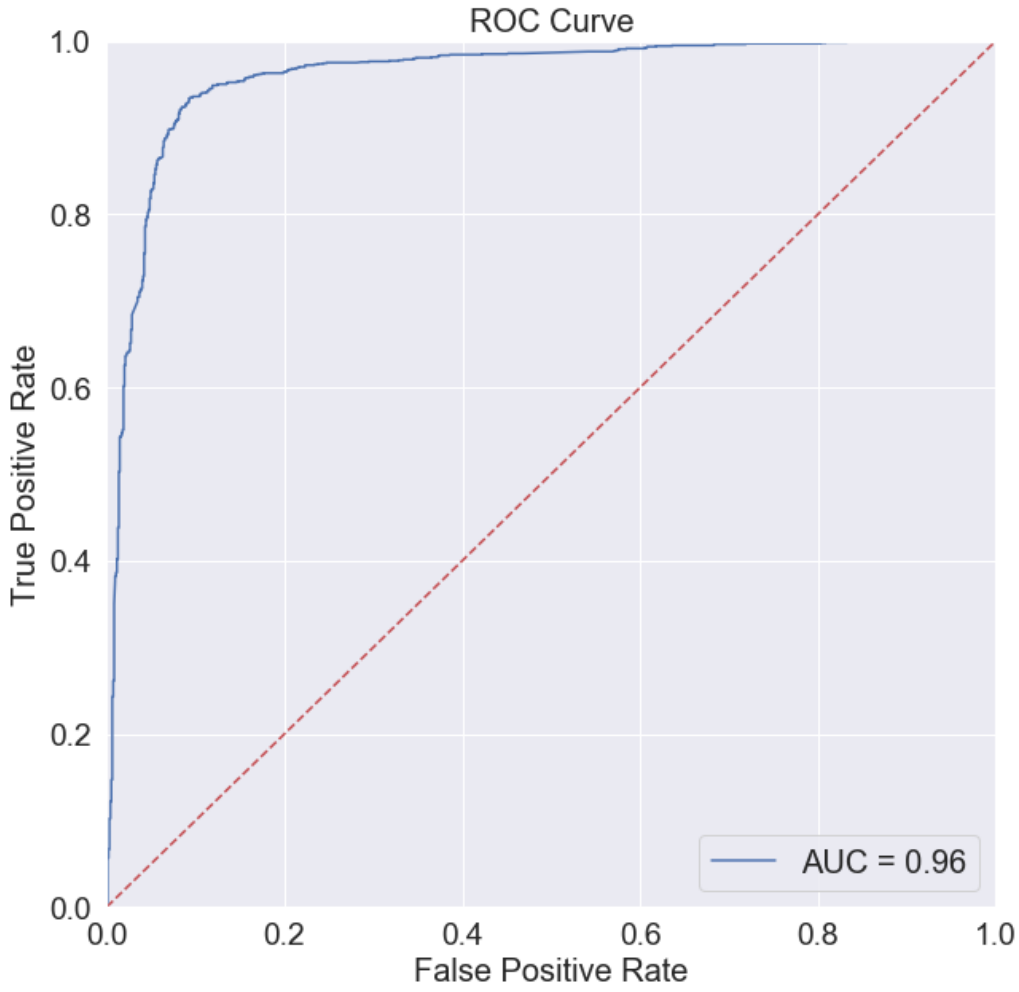
Model Selection

- Logistic Regression with CountVectorizer
- Recall = 92%
- Accuracy = 92%
- Baseline accuracy = 60% for positive
- Increasing stop words had a lesser effect on model performance
- Positive class was more numerous



ROC Curve

AUC = 0.96



Feature Coefficients

The most relevant words have strong coefficients

	word	coefficient
192	hurricane	5.113822
200	hurricanesandy	3.887277
135	frankenstorm	3.555969
353	sandy	3.515541
394	storm	3.431930
199	hurricanes	2.917694
324	power	2.503436
463	water	2.104734
83	damage	1.869608
51	building	1.821299
72	closed	1.767375
157	gone	1.669604
472	wind	1.366690
110	emergency	1.353689
131	flooding	1.353279
92	destroyed	1.278121

Conclusion and Next Steps

- This model successfully predicts the presence of a hurricane
- Use as the catalyst for an alert system
- Use Twitter API for live data
- Experiment with weights for retweets, handles, hashtags
- Generalize to different types of disaster
- Start predicting based on time, NHC: mid-August – late October