# Finding Tweets About Real Events with NLP

Aidan Coco, Michael Wirtz

# Twitter Classifier

- The *New York Times* (NYT) would like us to build a model that can differentiate tweets as events or non events to help them find new leads for stories
- Keywords are not enough; without a model *fire* will return tweets about a real fire and tweets about someone’s fire party last night
- Can we use NLP and make a classification model which will help them differentiate?

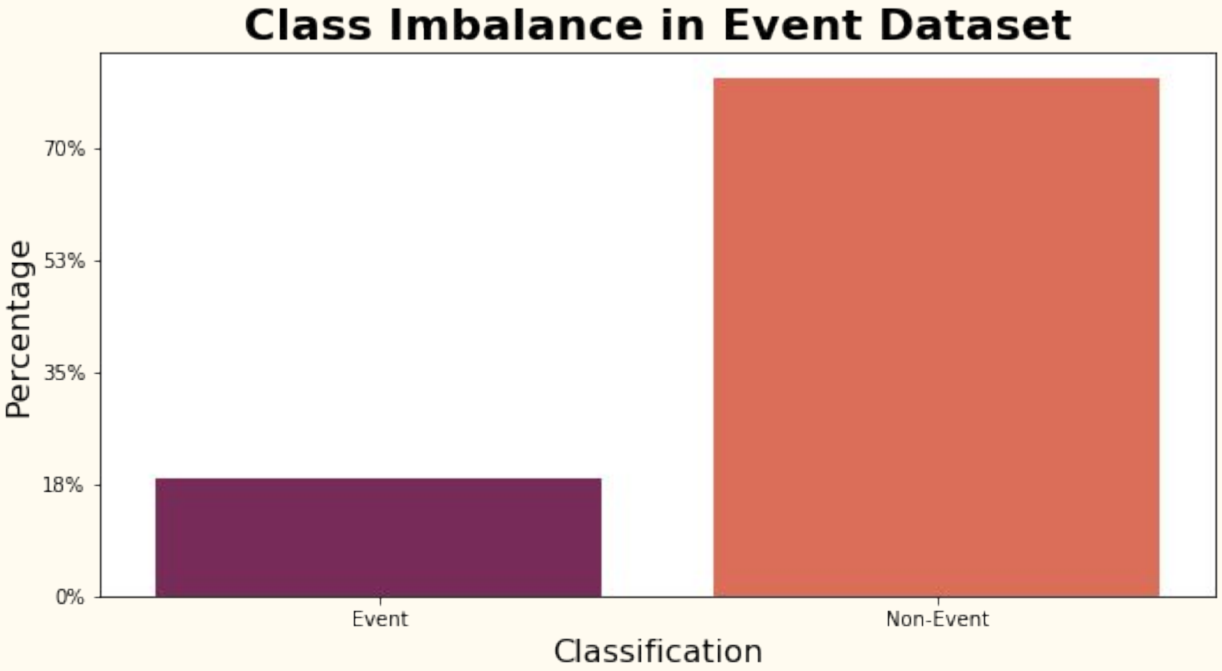
# Process

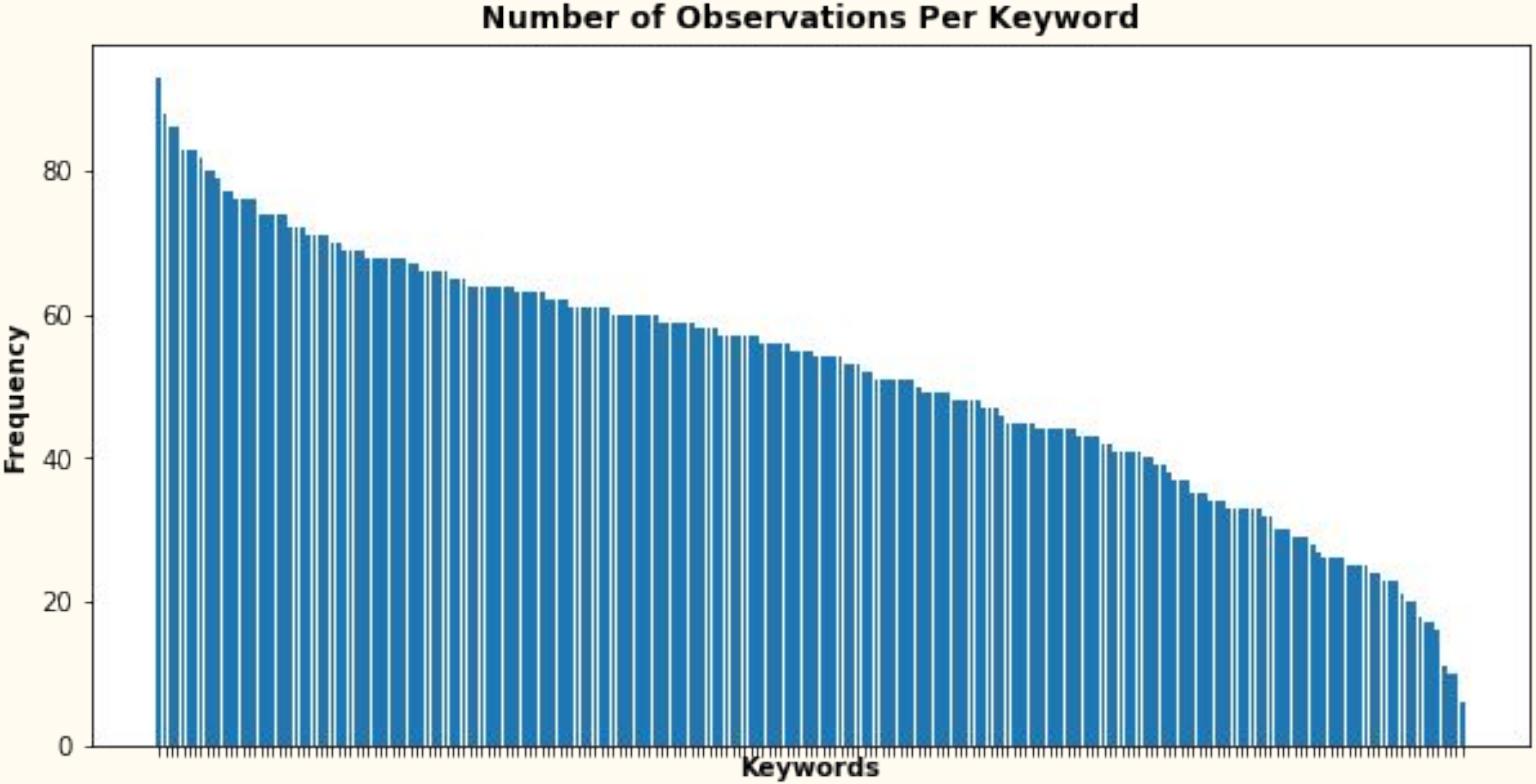
- Extensive EDA to find the best keywords to search for and how to differentiate between our two classes
- Features engineering: lemmatization, vectorization by TFIDF score
- Baseline KNN model
- Final Model: Naive Bayes (.89 accuracy, .68 F1 score)
- Applied to an entirely new holdout set
- Recommendations

# Data

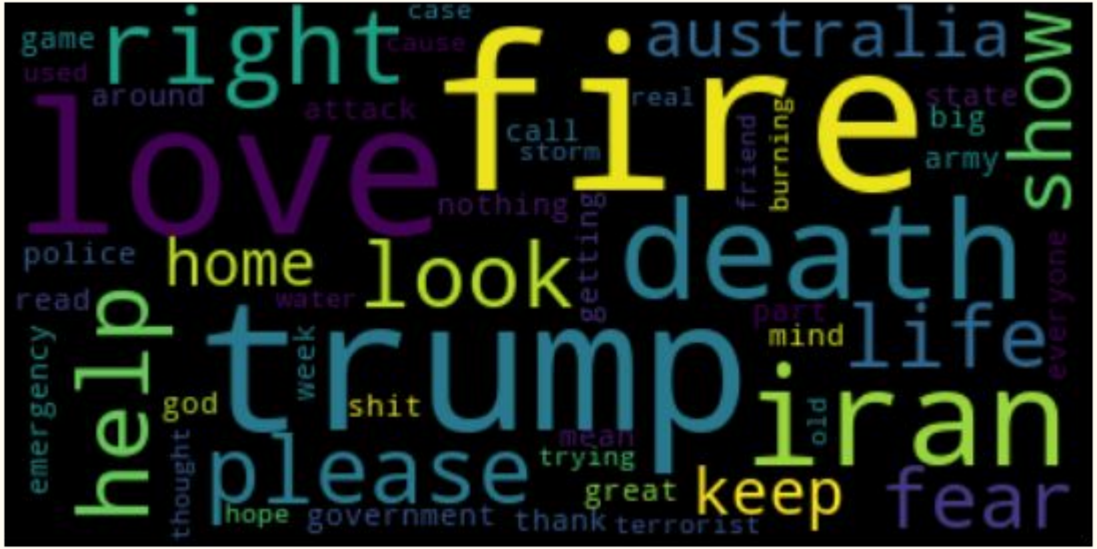
- Kaggle
- 11,000 Tweets found by searching over 200 disaster keywords
- Manually tagged as real disaster or not disaster (our target)
- Through our EDA we changed the classes to event and non event

# Class Imbalance

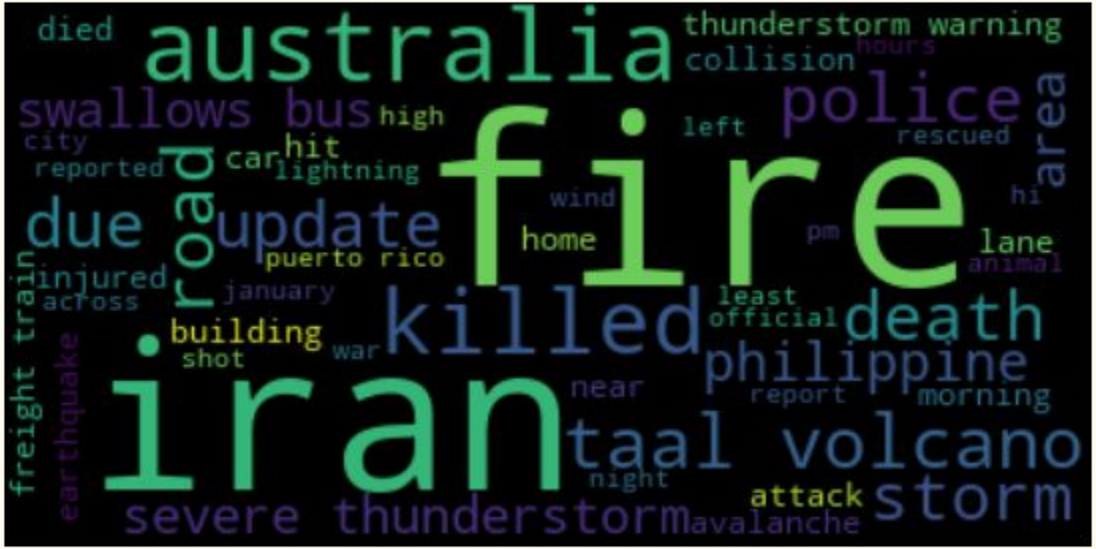



Number of Observations Per Keyword
Frequency
0
20
40
60
80
Keywords

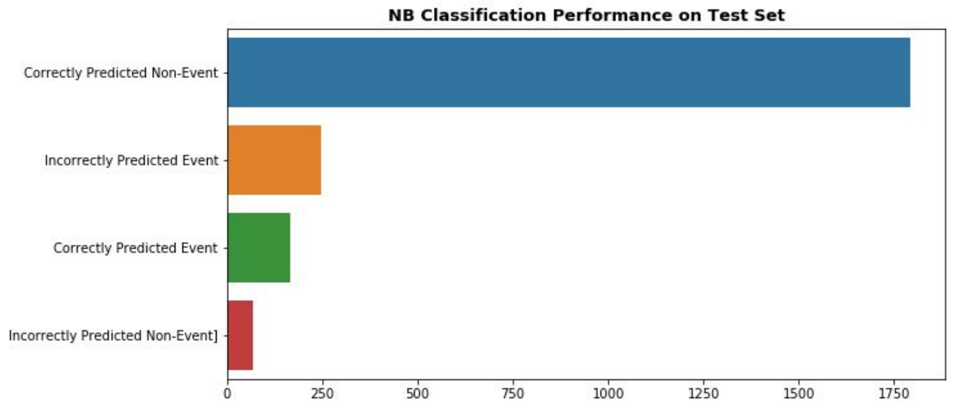
# Most Common Words: Non Events



# Most Common Words: Events



# Model Performance



- ➢ Best Model: Naive Bayes with an Alpha of 0.01
- ➢ .89 Accuracy
- ➢ .68 F1 Score

# Applied to a Holdout Set

- We used the Twitter API to search some of the most common keywords than manually tagged 100 of these tweets and ran our model on them
- .91 Accuracy
- .75 F1 Score
- Small sample size, but impressive performance for previously unseen data

# Recommendation

- The NYT should use the Keywords that were most likely to turn up real events when looking for stories: **violent storm, derailed, chemical emergency, hazardous, buildings on fire, body bag, sinkhole, derailment, collision and thunderstorm**
- With our model they can then filter these tweets further and only have to manually search through a tiny fraction of the words returned by the keyword search

# Future Steps

- More data
- Use pretrained vectors
- Neural Net