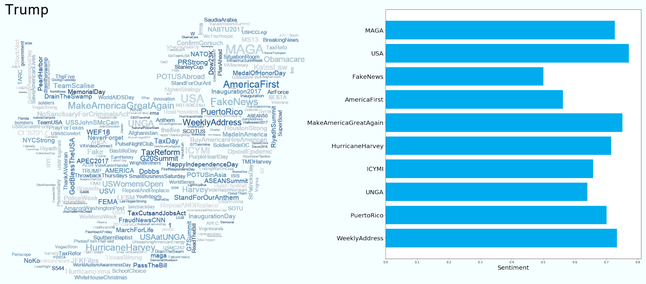
Statistical Urban Sentiment Analysis

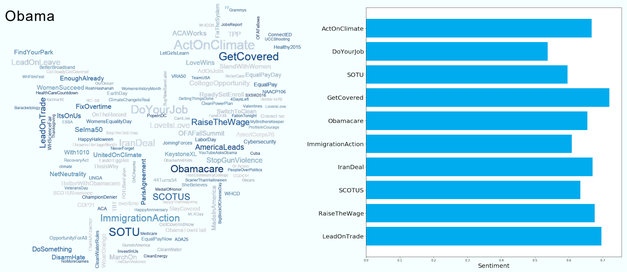
Kai Bernardini & Vasily Ilin

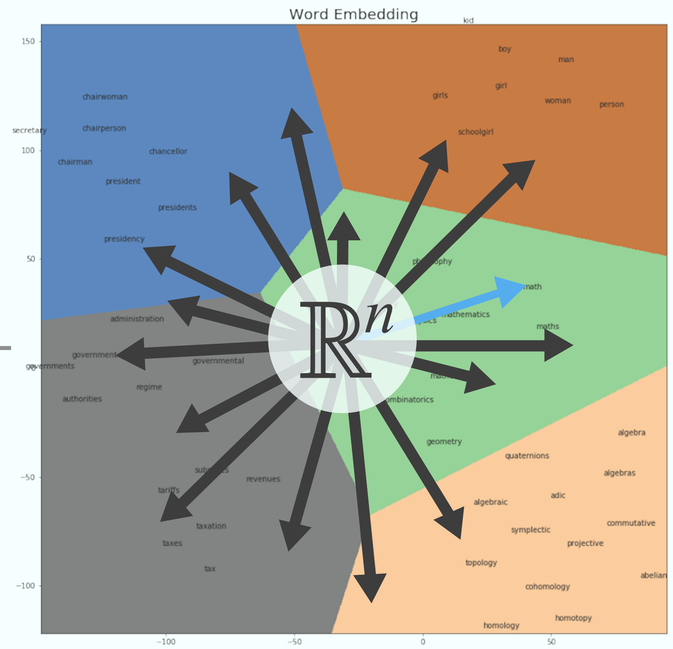
Trump

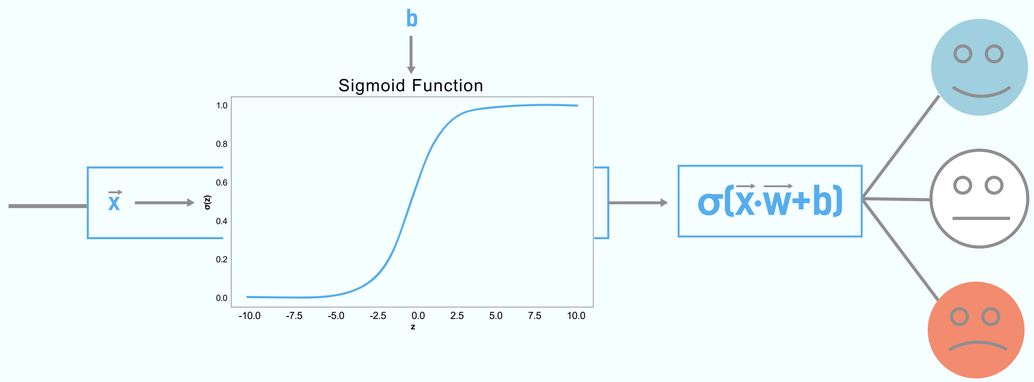


Term Frequency Inverse Document Frequency is a great vector space model.

Obama





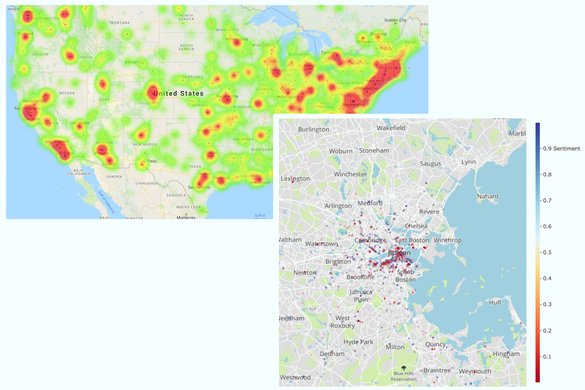


Vector Space Model of Text

- Each sentence is converted to an n-gram tuple array
- Each element in the tuple array containing n consecutive words is represented as a one-hot encoded vector in a Vector Space
  - The dimension of the vector space is the size of the vocabulary
  - Sentences are converted to vectors by aggregating the individual word vectors
  - Each vector is weighted according to Term Frequency, Inverse document frequency (TF-IDF)
- A transformed Corpus of tweets is now represented as a sparse data matrix X with ground truth labels y
- An L2 regularized Logistic Regression model is Built on the sparse Datamatrix which models Conditional Probability of y given X
- Weights bias are chosen to maximize the log-likelihood under a hypothesis function
  - Subject to the constraint that the L2 norm of the parameters does not exceed some value.
  - Bayesian interpretation is a Gaussian Prior over the parameters
- New tweets are Transformed and Classified

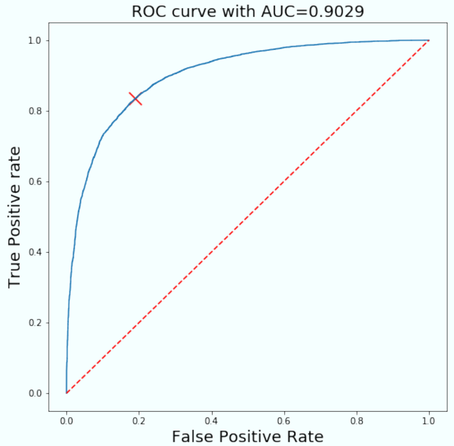
Generic Twitter Data

- Dataset containing over 1.5 million sentiment labeled tweets is used as a starting point for training and validating sentiment classification model.
- 500,000 tweets were retrieved from all over Massachusetts using Twitter API
  - Each tweet is vectorized and scored
  - Each hashtag associated to the tweet is stripped and stored along with class conditional probability.
- 1,000s more were retrieved from all over the US



Model Performance

- The training set has 150,000 features and 1.5 million examples.
- This matrix fits into memory, as the vast majority of the entries in the matrix will be fixed (hence, we can use sparse matrices) Note using the full data matrix assuming 8 bytes per entry would require 1800 Gigabytes.
- Cross Validation was used to determine the optimal value for penalization. On a holdout Testing set, the model achieved an accuracy of 82% with an AUC of approximately .90



TWITTER SENTIMENT ANALYSIS

Sentiment analysis is a form of natural language processing that attempts to isolate and extract subjective information encoded in text by a writer.

In our case, we seek only to determine if the subject matterof a tweet is generally positive, or generally negative. For example, a big collection of tweets can be classified as positive or negative to determine public opinion on a particular topic. For each tweet collected, its predicted sentiment is.recorded, If the tweet has any hashtags, they are recorded and associated to the tweets sentiment score.