National patterns in race tweets

Valery Lynn, MS
Data Science Career Track

Problem:

- Racial bias is difficult to measure
  - Ambiguous and subtle presentation
- Most empirical measures are conducted using indirect or subconscious actions:
  - Individual self-reporting
  - Experimental testing

Current Study at UCSF:

Department of Epidemiology & Biostatistics

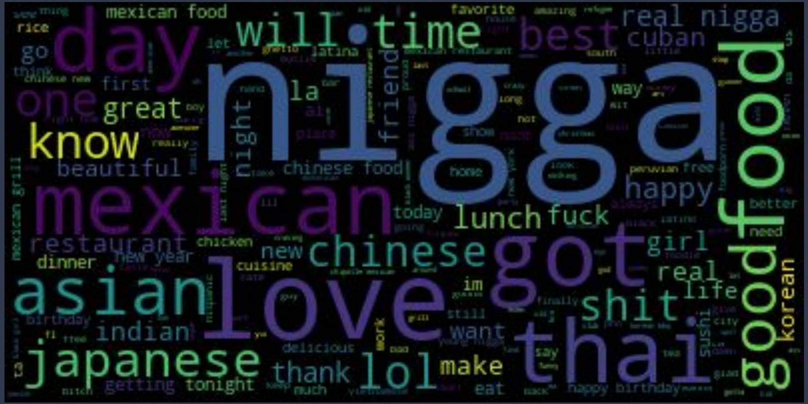
- Uses Twitter data to create place-level measurement of sentiment towards racial/ethnic minorities.
  - 1.25 million U.S. tweets related to race/ethnicity

The Data :

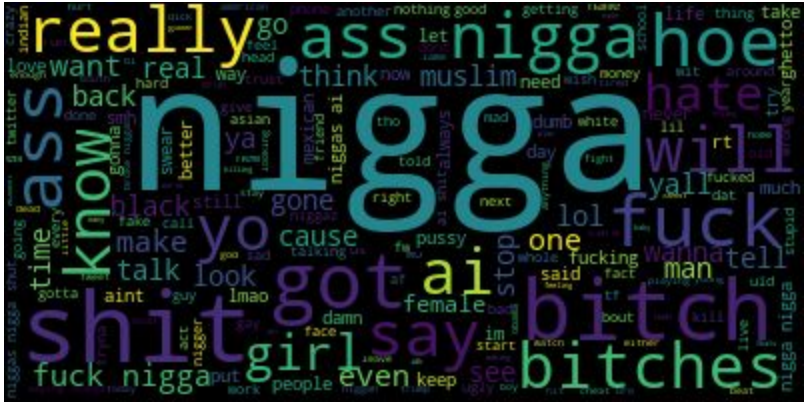
- A sample of the tweets (6,481) :
  - Hand-labeled for racial sentiment:
    - Positive, Negative
    - Neutral category imputed for neither of the above
  - 90% agreement between the researchers.
- Remaining data are unlabeled
- All data are stamped with geographical location:
  - Contain FIPS code
  - Can be used to create a U.S. cloropleth heatmap.

The Data Story:

- My first task was to visualize the three labeled categories:
  - Three word clouds show observable differences in the categories



Positive Tweets



Negative Tweets

Neutral Tweets

Model Selection Goal:

- Classify the highest number of positive (“1”) labels for each category.
- Optimize for recall

$$Recall = \frac{TP}{TP + FN}$$

- Use precision to calculate F1

$$Precision = \frac{TP}{TP + FP}$$

- F1, the harmonic mean between precision and recall

$$F1 = 2\frac{P \times R}{P + R}$$

Model Selection cont.:

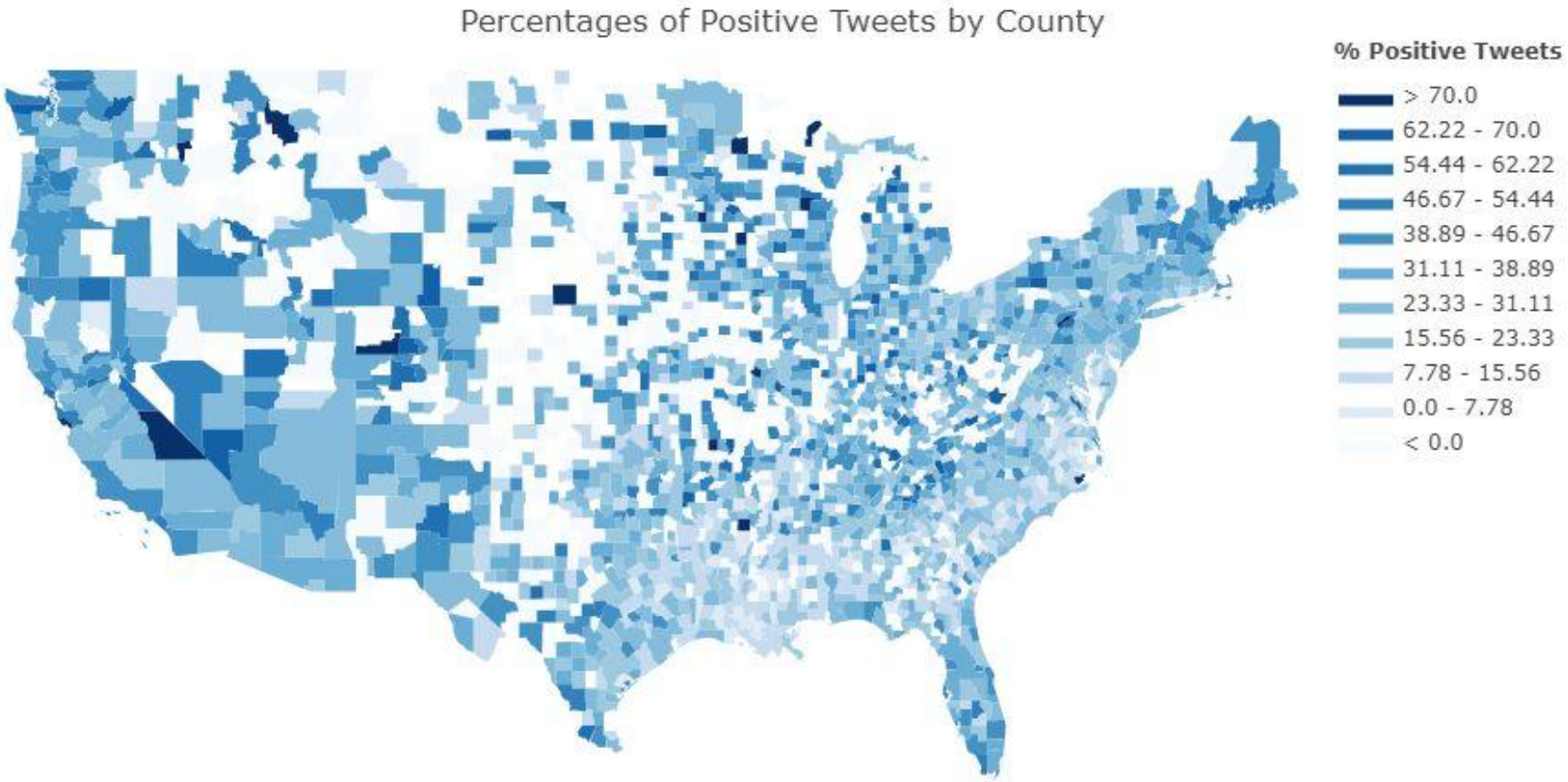
- The area under the ROC curve (AUC) is a way to compare classifiers using above metrics.
- It is the area under the curve plotted as the True Positive Rate (Recall) against the False Positive Rate
    - The ratio of the negative instances ('0') that are incorrectly classified as positive ('1')

Sklearn Models Tested:

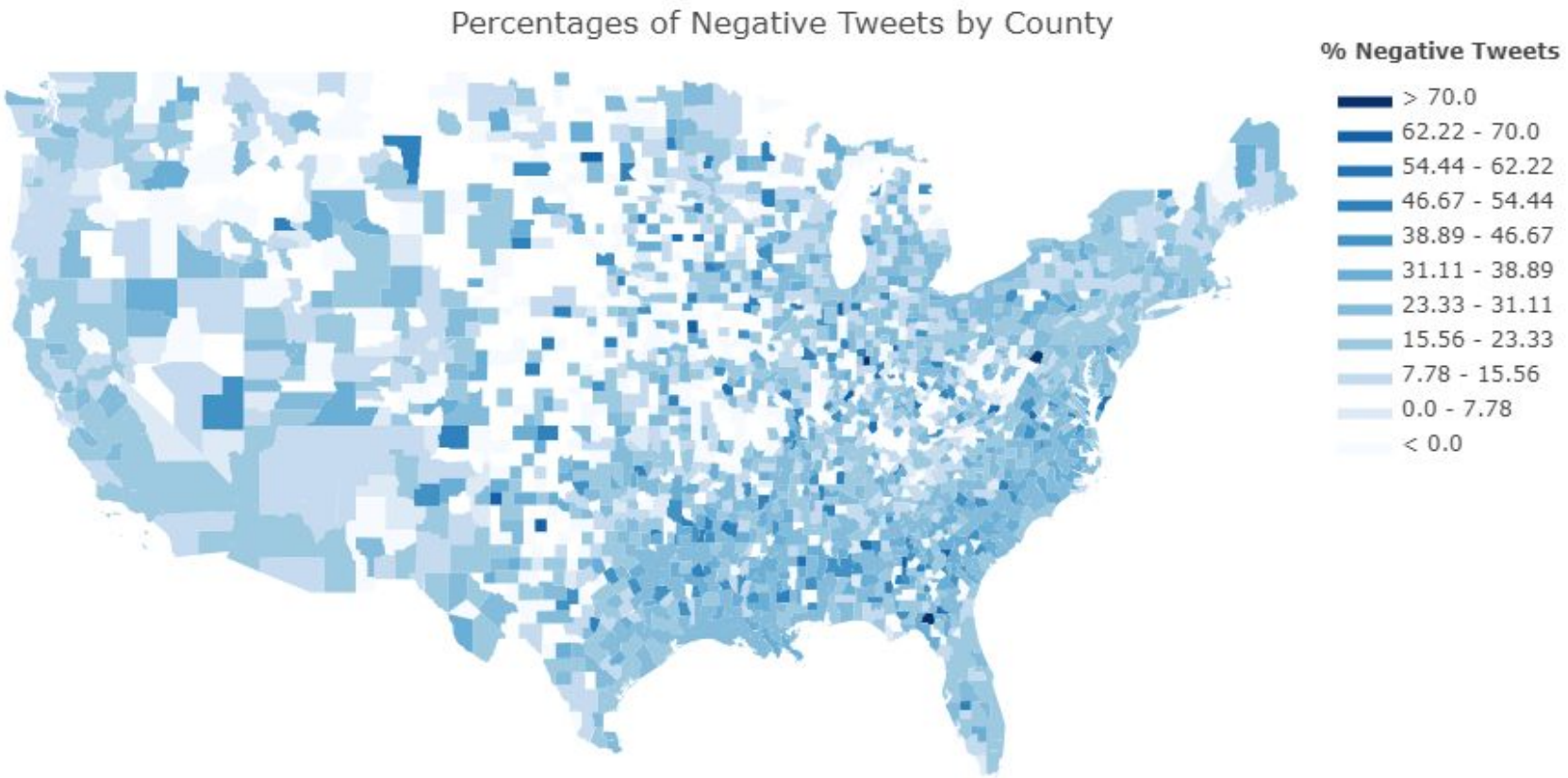
1) MultinomialNB Classifier
2) LinearSVC Classifier
3) LinearRegression Classifier
4) RandomForestClassifier
5) GradientBoostingClassifier

Models Selected:

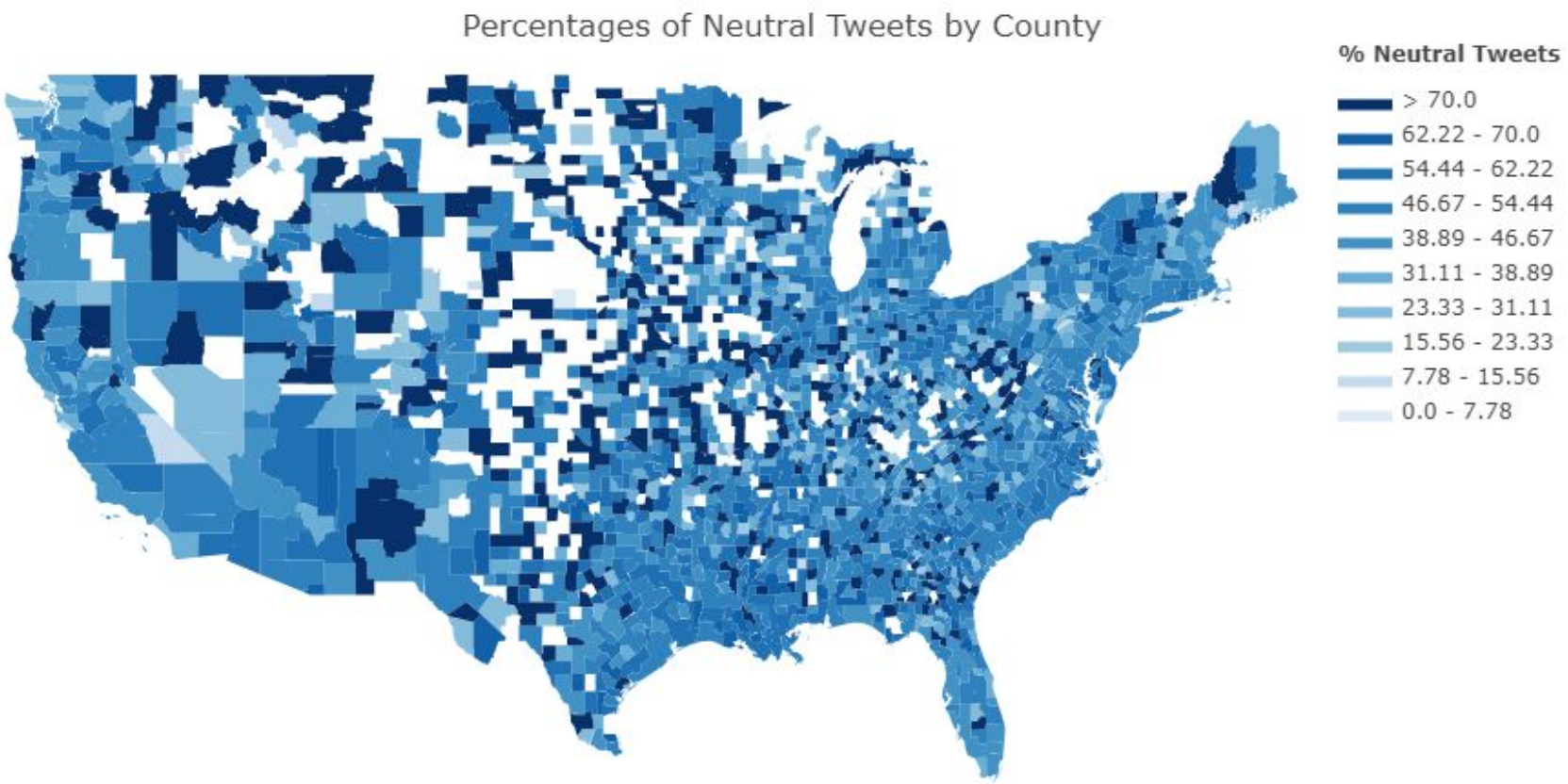
Category	Best Recall	Best F1	Best AUC	Best Overall
Positive	LinearSVC	LinearSVC	MultinomialNB	LinearSVC
Negative	MultinomialNB	MultinomialNB	MultinomialNB	MultinomialNB
Neutral	GradientBoosting Classifier	GradientBoosting Classifier	LogisticRegression	GradientBoosting Classifier



Positive Tweets: Percentages by County (U.S.)



Negative Tweets: Percentages by County (U.S.)



Neutral Tweets: Percentages by County (U.S.)