Capstone 2 - Final Report: Valery Lynn

Title: National patterns in race tweets

Problem: Racial bias is difficult to measure due to its sometimes ambiguous and subtle presentation. Most empirical measures of racial bias or discrimination are conducted using individual self-reporting or experimental testing that attempts to obtain results using indirect or subconscious actions. A research study being conducted at UCSF, Department of Epidemiology & Biostatistics, is using Twitter data to create place-level measurement of sentiment towards racial/ethnic minorities. The study analyzes more than one million U.S. tweets related to race/ethnicity for sentiment. A sample of the tweets (6,600) have been human-labeled with 90% agreement between the researchers. The researchers have made the data available to me to help them classify the tweets as positive, negative, or neutral. A previous study was conducted using sentiment analysis, but they were unable to classify the tweets into more than two categories. They got around this by collapsing the neutral category into the negative category. This project will be my attempt to use machine learning NLP methods to classify the tweets into the three categories originally proposed by the researchers.

Data: A stata file was obtained from Thu Nguyen, a post-doctoral fellow (Associate Specialist) at UCSF, Department of Epidemiology & Biostatistics. It contained the original tweets, along with results from a previous sentiment analysis, names of the people who hand coded the data (conducted after the unsupervised sentiment analysis), the labels, and the state, county, and zip code for the location of each tweet. Each tweet has an identification number attached as well as a location FIPS code that was determined from the latitude and longitude locations from the tweets. The FIPS code will be used to create the U.S. heatmap by county after the final prediction. The training set has 6,481 rows, each with one tweet. The unlabeled set to be predicted is 1.25 million tweets, of which the training set was extracted for manual labeling last summer.

The tweets were obtained from March 2015– April 2016 using Twitter's Streaming Application Programming Interface (API) to continuously collect a random sample of publicly available tweets. This API gives access to a random 1% sample of tweets. The data was restricted to only tweets with latitude and longitude coordinates for the contiguous United States and the District of Columbia. In total, there are 79,848,992 million general topic tweets from 603,363 unique Twitter users in the dataset. A keyword list of 398 racial/ethnic group terms and slurs was derived from racial and ethnic categories used by the U.S. Census and the online database of racial slurs. These keywords were used to identify tweets that contained at least one race-related word. This resulted in 1.25 million tweets.

Out of this, the training set was created by randomly selecting a subset of 6,481 tweets for manual labeling. Three coauthors labeled three possible responses: negative or positive. Inter-rater reliability among the coders was 90% for sentiment rating. In addition, the labelers indicated whether the tweet used discriminatory or stereotyping language about a racial group, whether the tweet mentioned a non-food location (e.g., Jewish Center), and whether the tweet should be excluded because it was not race-related (relating to individuals, people, or culture).

Approach: A previous sentiment analysis was conducted using the Stochastic Gradient Descent Classifier in Python software version 2.7. This is a linear classifier that has the value “0” or “1” and is appropriate for classifying into only two groups. To accommodate these restrictions, the researchers joined the neutral category with the negative category as this gave more accurate results in preliminary analyses. The purpose of my study is to develop machine learning classification algorithms that will classify tweets into three categories: positive, negative, or neutral.

I tested five classification algorithms: Multinomial Naive Bayes, Support Vector Classification (SVC), Logistic Regression, Random Forest and Gradient Boosting Classifier. I created a sparse matrix for each of the three categories. Labelers scored a tweet as a “1” in the “happy_manual” column if the tweet was positive and a “1” in the “sad_manual” column if the tweet was negative. I created a third column for the neutral tweets giving a score of “1” if both of the positive and negative columns were scored as “0”. I trained the algorithms as binary outcome variables in each column (columns are each label) against the preprocessed text of the tweets.

Tweets were preprocessed using the “re” package in python for regular expressions. Words with contractions were replaced with whole words, url strings were removed along with punctuation. The text was featurized using both CountVectorizer and TfidfTransformer from SKlearn. The featurized data were first visualized using an ecdf (empirical cumulative distribution function) to determine the hyperparameters df_min and df_max which will be used to tune the featurizer during classification. It also gives a good overview of the variation in the data. Most of the variation is contained in the first 25 tweets, showing that there is not a great deal of overall variation in the tweets.

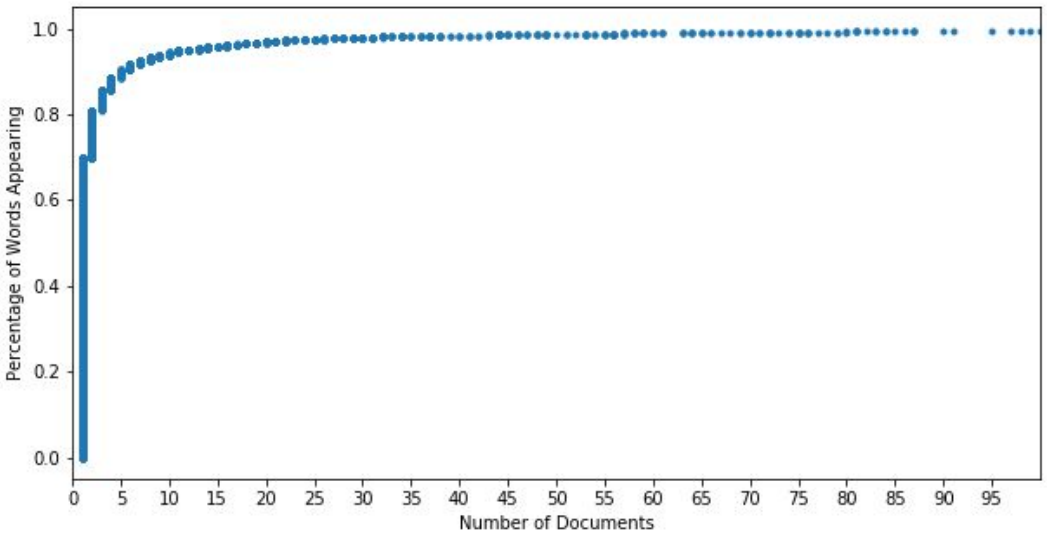

Figure 1. ECDF for tweets in the corpus.¶

Next, the featurized data was used to generate word cloud visualizations. There was a qualitative difference between the word clouds generated for the overall corpora, positive tweets, negative tweets, and neutral tweets. This suggests that a classifier is a reasonable method to use with this data.

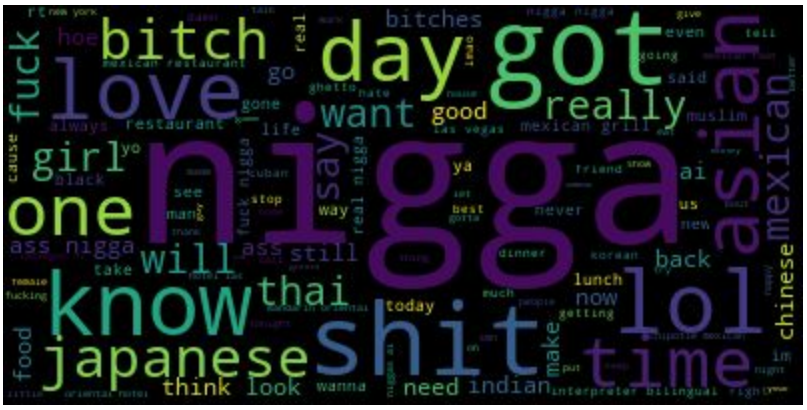

Fig 2. Word cloud for all the tweets.

There is a mix of positive and negative words. The highest used word is “nigga.” Looking at the overall corpora alone it is difficult to give meaning or explanation. The following three images show some striking differences.

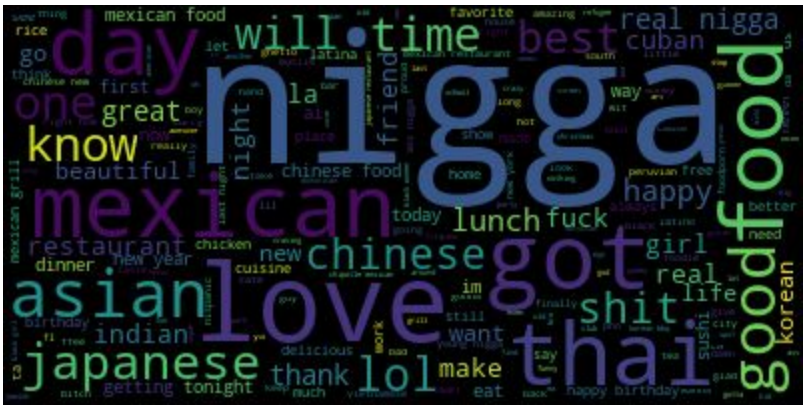

Fig 3. Word cloud for positive tweets.

In the positive tweets many of the swear words are missing that were seen in the overall word cloud. Positive words such as "beautiful", "happy", "thank" and "good" are present.

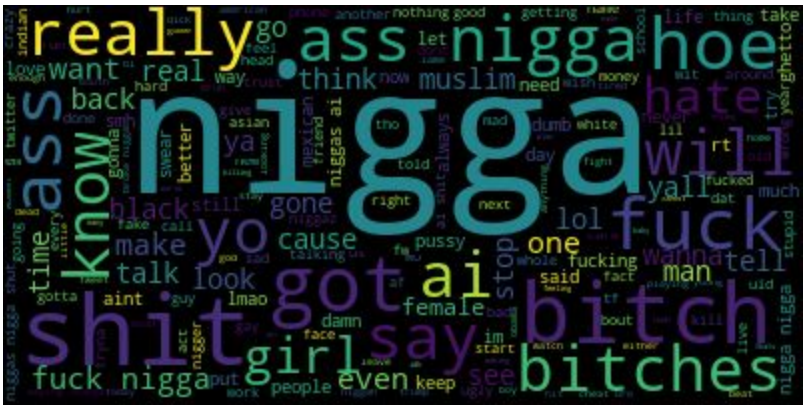

Fig 4. Word cloud for negative tweets.¶

This word cloud contains many swear words! Note that 'girl' is larger here, indicating that it is present more often. Here also are disparaging words used against females. Negative words such as "stop" and "hate" are seen.

Figure 5. Word cloud for neutral tweets.¶

There are fewer swear words. Many more words referring to ethnic food and "las vegas" as well as “hotel”.

Conclusion: These word clouds indicate that there is a qualitative difference in the labeled tweets for positive, negative, and neutral categories. All of the categories have the word "nigga" used predominantly, giving this word an ambiguous meaning. This illustrates the difficulty that an unsupervised classifier or cluster algorithm would have in determining sentiment.

The next section describes tested supervised classifiers and how models were developed to predict the labels "positive", "negative", and "neutral" on unlabeled tweets.

Model Selection

I am interested in classifying the highest number of positive labels for each category (note that here we are using 'positive' to mean a classification of '1' versus '0', not the 'positive' label for race sentiment). For this, I maximized recall, sometimes called sensitivity, which is the true positive rate. Recall is the ratio of positive instances that are correctly detected by the classification algorithm. The numerator is the number of true positives (TP) - tweets classified belonging to a category when they actually belong to the category determined by the labeled data. The denominator is the sum of the true positives plus the number of false negatives (FN) - tweets classified as not belonging to a category when they actually do belong to the category. False negatives should have been classified as positive (belonging to a category). Recall tells us how well the classifier detects positive instances.

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

This is contrasted with precision, which is the accuracy of the positive predictions. The numerator is the same, but the denominator is the sum of true positives and false positives - tweets that should have been classified as positive for a category, but were not. High precision will lower the number of positively classified instances, but they will be more certain in the classification. High precision is good for scenarios where you must be as correct in a prediction as possible. Examples are recommendation systems where you want the recommended content to be correct and can sacrifice having larger numbers of recommended content.

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

If we were concerned only with getting the highest accuracy possible, we would tune on precision, but the cost would be that we would predict much fewer (albeit correct) positive instances. Tuning for recall will lower the accuracy, but we will lower the risk of leaving some positive instances out. The difficulty (and interesting challenge) is that there is a tradeoff between recall and precision. Increasing one will reduce the other. The F1 score gives a picture of how well the classifier performs for each. It is the harmonic mean between precision and recall. This score is high for classifiers that have similar precision and recall.

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

Finally, 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'). A perfect classifier will have an AUC = 1, and one that is purely random will equal 0.5. So, we are looking for the highest AUC score above 0.5.

Since this data is highly unbalanced for each category, there may be more negative instances than positive instances. The ROC curve isn't the best metric to use. Instead, it is better to use a Precision-Recall Curve. However, this is not easily done in Sklearn with every classifier. It depends on calculating the decision_function, which isn't available for every classifier (e.g., MultinomialNB or RandomForestClassifier).

To evaluate these classifiers I need a metric that is shared among them all. Therefore I used Recall, F1 and the AUC scores. “Best overall” score is given to the best out of 3. In the case of a tie, it will go to best recall.

Table 1. Top performing classification models for each category label.

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

We want to predict the highest number of tweets in each category and are willing to accept a few incorrect classifications in order to catch all the correct classifications. These metrics attend to both precision and recall so that the overall choice of classifiers for each category will minimize false negative claims.

Conclusion: The positive label is best predicted using a tuned LinearSVC classifier. The negative label is best predicted using a tuned MultinomialNB classifier, and the neutral label is best predicted using a GradientBoosting Classifier.

Prediction and Results

Results were obtained by using the selected models to predict negative and positive tweets. Tweets that were not classified as negative or positive were considered to be neutral. To obtain these results I used the classifier with the highest AUC score (MultinomialNB classifier for the negative label - AUC: 0.7393) , then removed the negative tweets from the dataset. Next, I used the MultinomialNB classifier for the positive label to predict the positive tweets, and removed these from the dataset. Those remaining were labeled neutral.

The final results report the number of positive, negative, and neutral tweets in the U.S., in each state and each county. Differences between states and again between counties show that there are marked differences within each geographical area. County maps indicate stark differences between counties within the same state.

Each tweet is labeled with a Federal Information Processing Standards (FIPS) code. This code uniquely identifies counties in the United States. The data are aggregated by state and county and the sum of tweets for each county makes up a new column of tweet labels: positive, negative and neutral. This is a sample of tweets so the best way to represent the results is with percentages of each label for each state and county present in the dataset. I divided the number

of tweets for each label (positive, negative and neutral) by the total number of tweets within that state or county to obtain percentages. Data on the names of counties and states by FIPS code were obtained from “Population Estimates” found on the US Census Bureau website (www.data.gov).

Results for the U.S.

Predicted tweets for each label were aggregated and divided by the total tweets in this dataset. Nearly half of all tweets were neutral. There was a high percentage of positive tweets, 31.94% than negative tweets, 20.72%.

U.S. Positive Tweets: **31.94%**

U.S. Negative Tweets: **20.72%**

U.S. Neutral Tweets: **47.34%**

Results by State Level

Next, I aggregated by state level and the following , Tables 2 - 4, give the top 10 states for each category.

Table 2. Top 10 states with highest percent positive tweets.

Rank	State	%_Positive
1	NV	66.69
2	OR	46.08
3	ME	44.67
4	NH	41.71
5	CA	41.51
6	WA	40.36
7	NY	39.22
8	WY	38.58
9	CO	37.48
10	MT	36.84

Table 3. Top 10 states with highest percent negative tweets.

Rank	State	%_Negative
1	MS	29.65
2	LA	29.01
3	MD	27.82
4	AL	26.80
5	DE	26.71
6	SC	26.38
7	MI	25.43
8	AR	25.31
9	GA	25.17
10	OH	25.15

Table 4. Top 10 states with highest percent neutral tweets.

Rank	State	%_Neutral
1	LA	57.55
2	OH	55.72
3	ID	55.19
4	OK	54.84
5	AL	52.99
6	DE	52.77
7	MS	52.53
8	SD	52.38
9	GA	52.19
10	CT	51.86

Aggregating at the state level can show good overall trends, but it doesn't answer the question regarding place-level measurement of sentiment towards racial/ethnic minorities. This data is contains FIPS codes so that I was able to agreggate by county to reveal a far more complex spacial distribution of the types of tweets.

Results by County Level

There were 2,615 counties represented in this data, making it too cumbersome to display in table form. A better way to view the data is with chloropleth maps, sometimes referred to as heatmaps. These show the percentage of each type of tweet by county using color shading. Figures 5, 7, and 8 show results for positive, negative, and neutral tweets. Counties in the maps that are white are either counties in the dataset that did not have tweets with those labels, or were counties that were not represented in the dataset (529 counties). In general, states with the most missing counties were in the western portion of the U.S. (Nebraska, Kansas, Texas, Missouri, and South Dakota).

Table 5. Top 10 counties with highest percent positive tweets.

Rank	County	State	% Positive
1	Waupaca County	WI	86.05
2	Pamlico County	NC	85.27
3	Custer County	NE	84.91
4	Waseca County	MN	83.33
5	Santa Cruz County	CA	81.54
6	Lewis and Clark County Inyo County	MT CA	80.00
7	Nez Perce County	ID	75.00
8	Drew County	AR	74.00
9	Douglas County	WI	73.91
10	Boone County	AR	73.68

Table 6. Top 10 counties with highest percent negative tweets.

Rank	Area_Name	State	%_Negative
1	Brown County	IN	81.92771084
2	Hampshire County Madison County	WV FL	75
3	Spencer County Johnson County Dickens County Newton County Dickey County Parmer County Marion County Harrison County	IN GA TX IN ND TX GA MO	66.66666667
4	Appomattox County Scott County Poquoson City	VA IN VA	62.5
5	Emmet County Coosa County Monroe County Miller County Winnebago County	IA AL WV MO IA	60
6	Accomack County	VA	59.18367347
7	Echols County Union County Newton County Estill County	GA IN MS KY	57.14285714
8	Barren County	KY	56.17977528
9	Marion County	AL	52.63157895
10	Morgan County	IN	52.51141553

Indiana tops the list of negative tweets by appearing 6 times, followed by Georgia and Virginia with 3 times each. Note however that neither Indiana or Virginia were in the top 10 rankings for states with negative tweets. This demonstrates how important it is to aggregate at the county level to capture geographic diversity.

374 counties had 100% neutral tweets so a top 10 ranking doesn't make sense. There were no counties with 100% of positive or negative tweets. There were more counties with higher percentages of positive tweets than there were negative tweets. The state with the most counties

with the highest negative tweets was Indiana. The states tying for the most counties with the highest positive tweets were Arizona, California, and Wisconsin.

The following maps (figures 6 - 8) display the results for each county. Note that there are 529 counties missing from this dataset (no racial tweets from that location). They are shaded white in the maps.

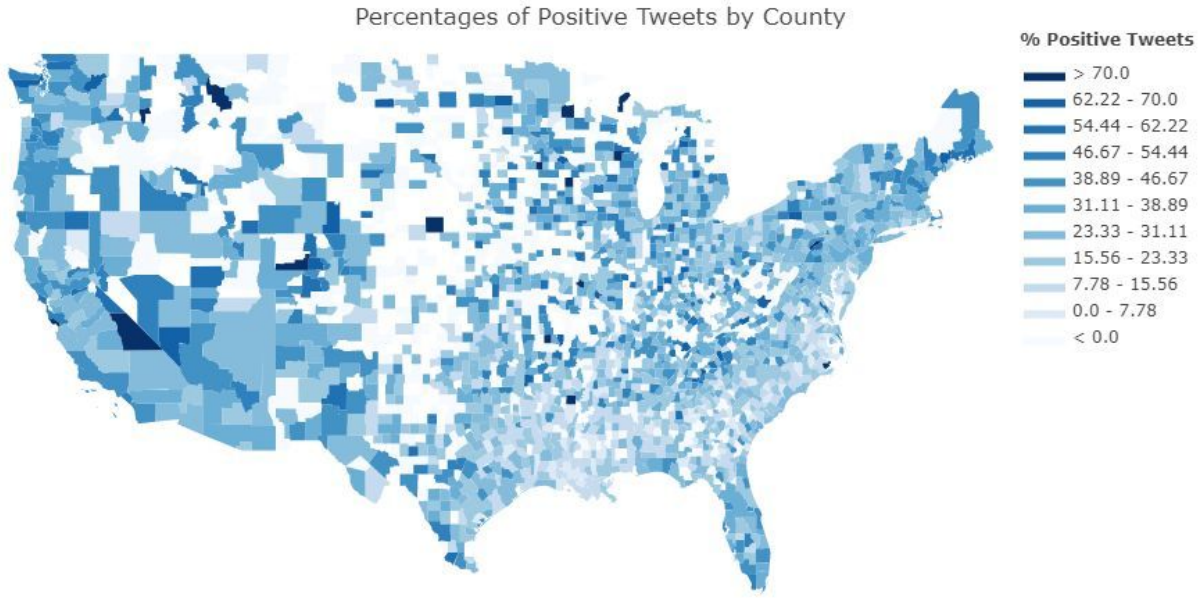

Figure 6. U.S. chloropleth map of percentages of positive tweets by county.¶

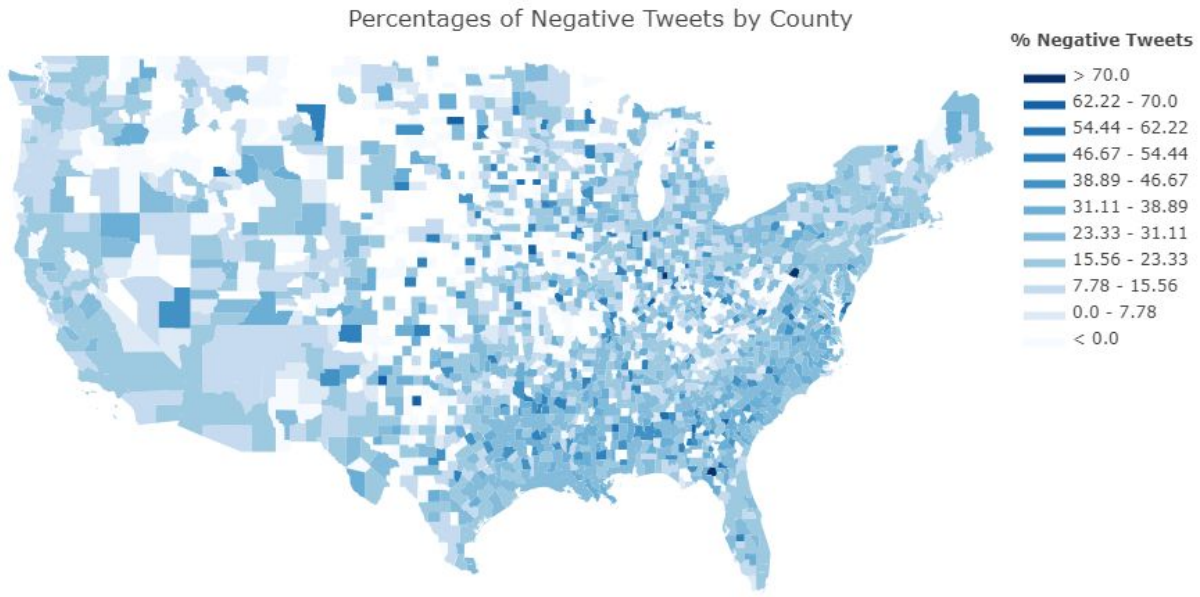

Figure 7. U.S. chloropleth map of percentages of negative tweets by county.¶

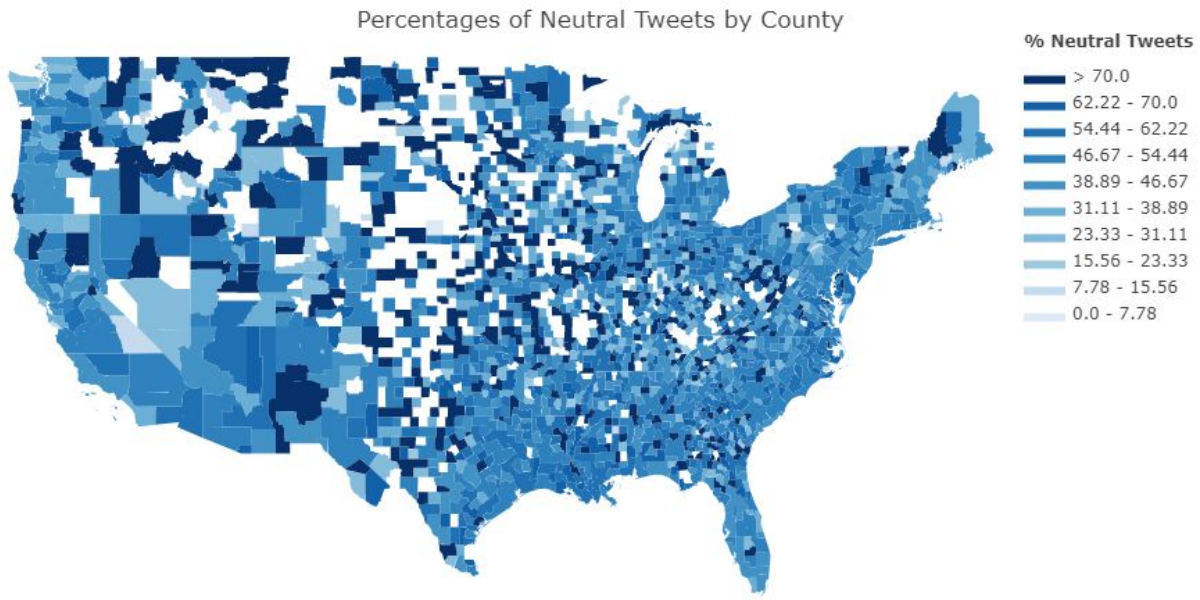

Figure 8. U.S. chloropleth map of percentages of neutral tweets by county.¶

Conclusion and Next Steps

Three machine learning classification algorithms performed adequately on this dataset. I chose to treat each label as a binary classification problem rather than use multilabel classification for three labels. Both MultinomialNB and RandomForestClassifier are appropriate for multilabel classification and could easily be performed as a follow-up with this cleaned data. In addition, I plan to create state cloropleth maps with these results to display differences between states. Finally, it would be interesting to do correlations with positive and negative tweets and population density, majority political party, and access to education. In fact, any county level measurement could be used by itself or in combination to predict racial sentiment in tweets using the training set for this data. There are limitless questions that could be asked and answered with this data in combination with other county or state level data.

References

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning - ICML '06. doi:10.1145/1143844.1143865

Géron, Aurélien. Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly, 2018.

“Population Estimates.” *Data.gov*, Publisher US Census Bureau, Department of Commerce, 2 Mar. 2018, https://catalog.data.gov/dataset/population-estimates.

Sentiment analysis. (2018, December 10). Retrieved from https://en.wikipedia.org/wiki/Sentiment_analysis

Smetanin, S. (2018, September 01). Sentiment Analysis of Tweets using Multinomial Naive Bayes. Retrieved from https://towardsdatascience.com/sentiment-analysis-of-tweets-using-multinomial-naive-bayes-1009ed24276b