**Polarity vs Subjectivity of State-Level Public Transit Related Tweets (Map w/polarity/subj data)**

Using Twitter's API, the Tweepy module, and the Natural Language ToolKit, we scraped all available tweets regarding public transit and mentioning each specific state (if available, 35/50 states had available tweets about public transit at the time of analysis). We performed sentiment analysis on these tweets and extracted a subjectivity and polarity score for each state’s tweets about public transit.

A subjectivity score closer to 1 (on a decimal scale from 0 to 1) means a given state tweeted more opinion-based content regarding their local transit, and a score closer to 0 means a state tweeted more fact-based content regarding their local transit.

A polarity score closer to +1 (on a decimal scale from -1 to +1) means a state tweeted more positive content regarding their local transit, whereas scores closer to 0 meant more neutral content, and scores closer to -1 meant more negative content.

The dots on the map are sized according to polarity—larger circles mean a state is tweeting more positively about their local transit, and vice versa.

Notable in this map are the small sizes of the New York and Massachusetts dots, indicating the somewhat negative polarity of their tweets overall. This could be an indicator of dissatisfaction with local transit, but whether it is a referendum on users not liking the transit system overall or taking issue with imperfections in the system is a topic for further study.

**Polarity vs Subjectivity of Recent Tweets from the US Overall about Public Transit**

Similar to the polarity and subjectivity data at a state level, we also scraped all available tweets regarding public transit that were geotagged in the United States overall rather than just at the state level. We performed sentiment analysis using Natural Language ToolKit on these national-level tweets and found that at this large scale, tweets about public transit are generally pretty neutral, with a polarity value of 0.0726. Tweets about public transit were somewhat subjective with an overall subjectivity score of 0.3659.

These national-level results make sense in the context of the medium, Twitter is generally a place for users to post opinionated content. Any positivity or negativity about public transit seems to disappear at the national scale, suggesting that people have stronger positive or negative feelings about their local transit than public transit as a whole.

**Scikit-Learn Machine Learning Linear Regression**

Using data from the Federal Transit Administration, we performed a linear regression as a supervised machine learning algorithm to understand how well funding and previous year’s ridership predicts the current year’s ridership. We collected the FTA data which provided urban transit funding from 1998-2019, and total ridership in the same time band. We then engineered two features, which was two lag columns that calculated the previous year’s ridership and the previous two year’s ridership.

The linear regression model used urban funding data, previous year’s ridership, and the ridership from two years ago as features to predict the current year’s ridership. We normalized the data using StandardScaler to improve the model, and ended up with a reasonably good fit as shown by the predicted values of the model in orange and the actual values in blue.

The evaluation metrics also make a moderately compelling case for the utility of this model. The RMSE metric tells us that the model can predict ridership within about 500,000 riders, an error rate of roughly 5% given an approximate annual ridership of around 10,000,000 people.

This model can provide some utility in predicting future ridership based on potential future funding, a use case that would behoove local lawmakers making a case for increasing funding to their local transit authorities. However, if this model were to be used in a pre-mass-COVID-vaccination landscape, it would require much more data and tuning given that public transit ridership has shifted seismically thanks to COVID restrictions.