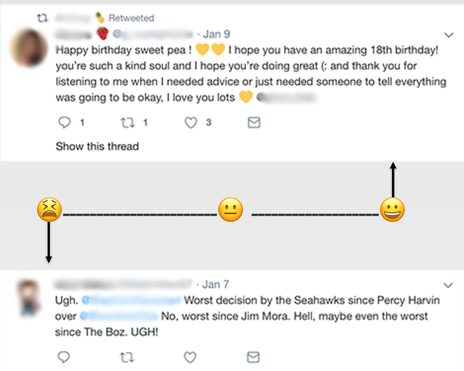
Twitter Sentiment in Under 140 Characters

Tyler Atkinson

Introduction

What is sentiment analysis?

Sentiment analysis is simply working out if a piece of text is positive, neutral, or negative depending on the type of wording they use, any emoji's etc. Examples of positive and negative tweets would look like:



Process

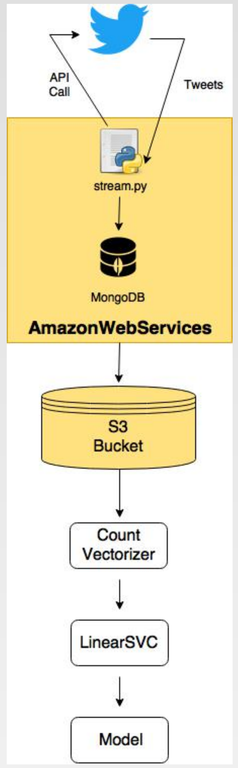
- *I collected about 800k tweets using twitter's streaming API that were sent from inside the United States.*
- *To clean the tweets, I used regular expressions to remove any links, hashtags, or usernames.*
- *To get sentiment, I used a package built into the python package nltk called Vader that returns a value from -1 to 1, -1 being a negative sentiment and 1 being positive.*
- *I then decided to create a range for positive, neutral, and negative tweets and chose the following scale*

  Compound score >= .1 ➔ Positive 😀

  -.1 < Compound score < .1 ➔ Neutral 😐

  Compound score <= -.1 ➔ Negative 😫
- *After choosing these boundaries, I ended up with 334115 positive, 302507 neutral, and 172794 negative tweets.*

Modeling



After getting my positive, neutral, and negative classes, I created combinations of the following models paired with both CountVectorizer and TfidfVectorizer:

- *Naive Bayes*
- *LinearSVC*
- *Random Forest Classifier.*

Ultimately, the LinearSVC along with CountVectorizer performed the best.

Analysis

I broke my analysis into two parts:

Sentiment and location

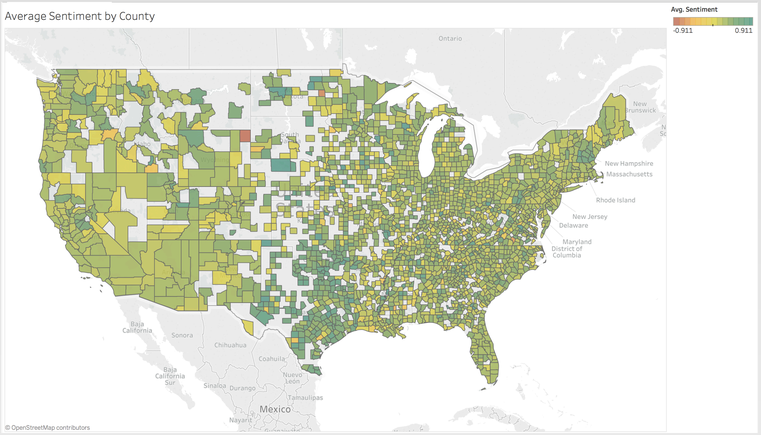
- *I was curious to see if there was any relationship between how "happy" someone was and their location. As you can see to the right, I chose to look at the average sentiment for both football teams playing in the super bowl. I found there was not much of a relationship, but cleaning my data up further could help.*

Sentiment and hashtags / keywords

- *I grouped tweets by similar hashtags or keywords and was able to look at characteristics such as people tweeting more from a certain area about a given topic and a higher average sentiment for an area. I found that overall, I collected a lot more tweets from densely populated areas such as Los Angeles, Seattle, and Miami.*

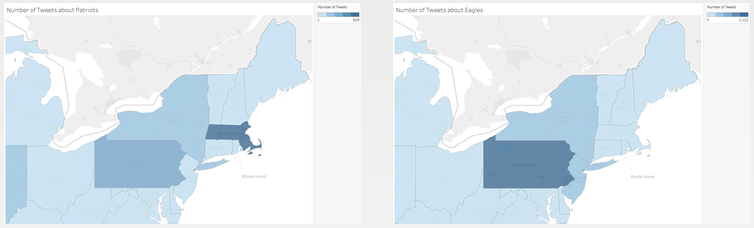
Sentiment and Location

To see if there was a relationship between someone's location and the happiness of their tweet, I graphed the average sentiment of every county I had in my dataset. I found there were counties with only one tweet, which obviously does not accurately represent that county.

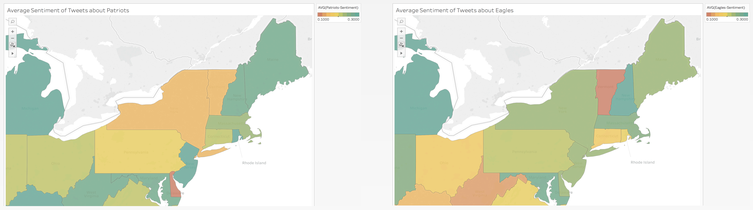


Applications

Count of Tweets about Patriots or Eagles



Average Sentiment of Tweets about Patriots or Eagles



To apply this to a real topic, I filtered tweets containing words and hashtags that related to the two NFL teams playing in Super Bowl LII. I created two groups, one about the Patriots ('pats, brady, patriots' etc.) and one about the Eagles('eagles', 'Philadelphia', 'nick foles', etc.). Not surprisingly, most people from Massachusetts were tweeting about the Patriots and same for Pennsylvania and the Eagles.

After filtering the tweets above, I graphed the average sentiment for each state. I did not find much of relationship but cleaning my data up further by removing tweets that don't accurately reflect sentiment (spam for example) could help.

Other Potential Applications

- *Filter tweets about a given company to get overall sentiment*
- *Get overall sentiment of given user based off their old tweets*
- *Word clouds for positive and negative sentiment for a given topic*
- *Get sentiment of a newly released product*
- *See if a certain region has a more positive sentiment for a topic*
- *Identify potentially untapped markets or product features*

Tech Stack

mongoDB, pandas, Flask, scikit learn, python, tableau, amazon web services

Contact

LinkedIn.com/tyleratk
Github.com/tyleratk
tyler.atk@hotmail.com