**Introdction:**

2020 US presidential election is one of the impactful events in the world. And whether the current president, Donald Trump, will win the election and still serve as US president is hard to predict. In this case it is important to know his attitude toward the 2020 election. I want to do some sentiment analysis on Donald Trump’s tweets during 2016 and 2020 election and see if there’s any changes.

**Description of dataset**  
For this project, I will use Donald Trump’s tweets in the year of 2020 and 2016 before election. I will only use two properties: content of the tweets for data mining and date for filter. I will probably use Twitter API to get more data if I need some other properties (e.g. who will be retweeted by Trump).

2016: [https://www.kaggle.com/benhamner/clinton-trump-tweets](https://www.kaggle.com/benhamner/clinton-trump-tweets/home)

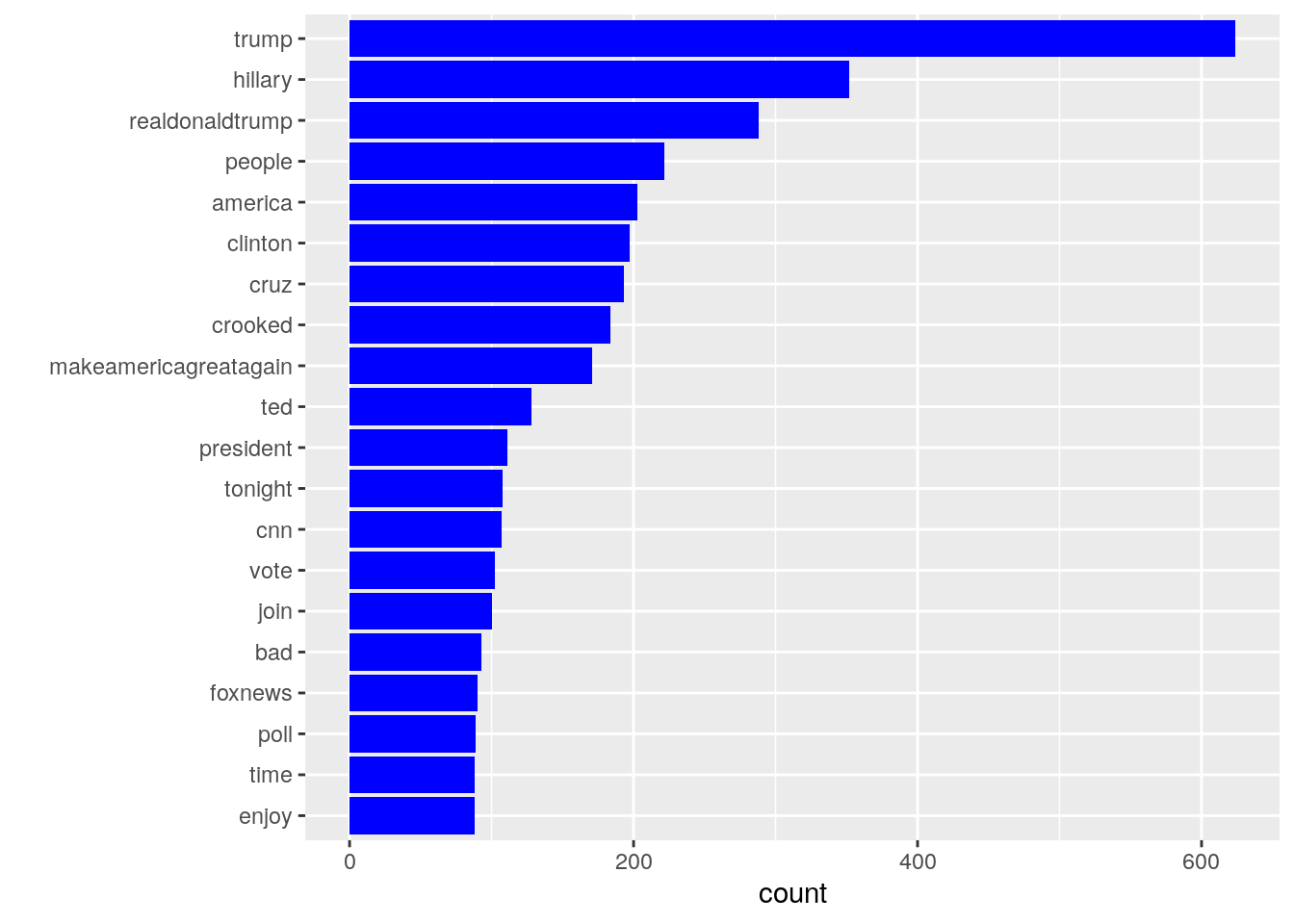
2020: <https://www.kaggle.com/austinreese/trump-tweets>

Please take notice that 2020 dataset is an example because there is still 2 months before the election, and maybe I will mine the data myself instead use one on Kaggle.

**Preliminary Figures**

I analyze Trump’s 2016 tweets (I’m not sure to use the 2020’s above) and there’s some results.

This is the top 20 of most used terms by Trump in 2016.



**Questions:**

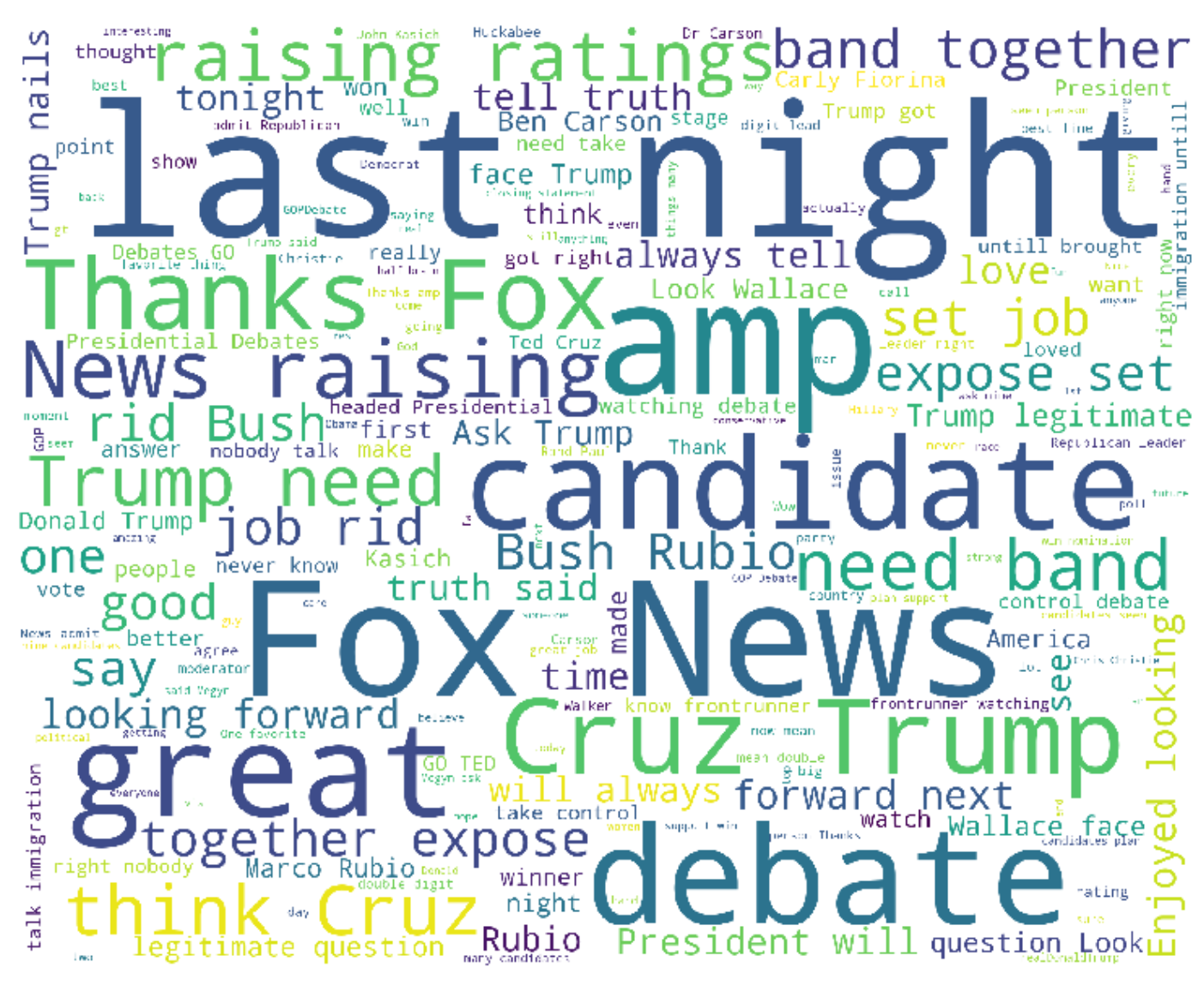
What is the difference of sentiment for Trump’s tweets in 2016 and 2020?

How did his sentiment change according to time series? And why? (Are there any representative words can reflect this changing?)

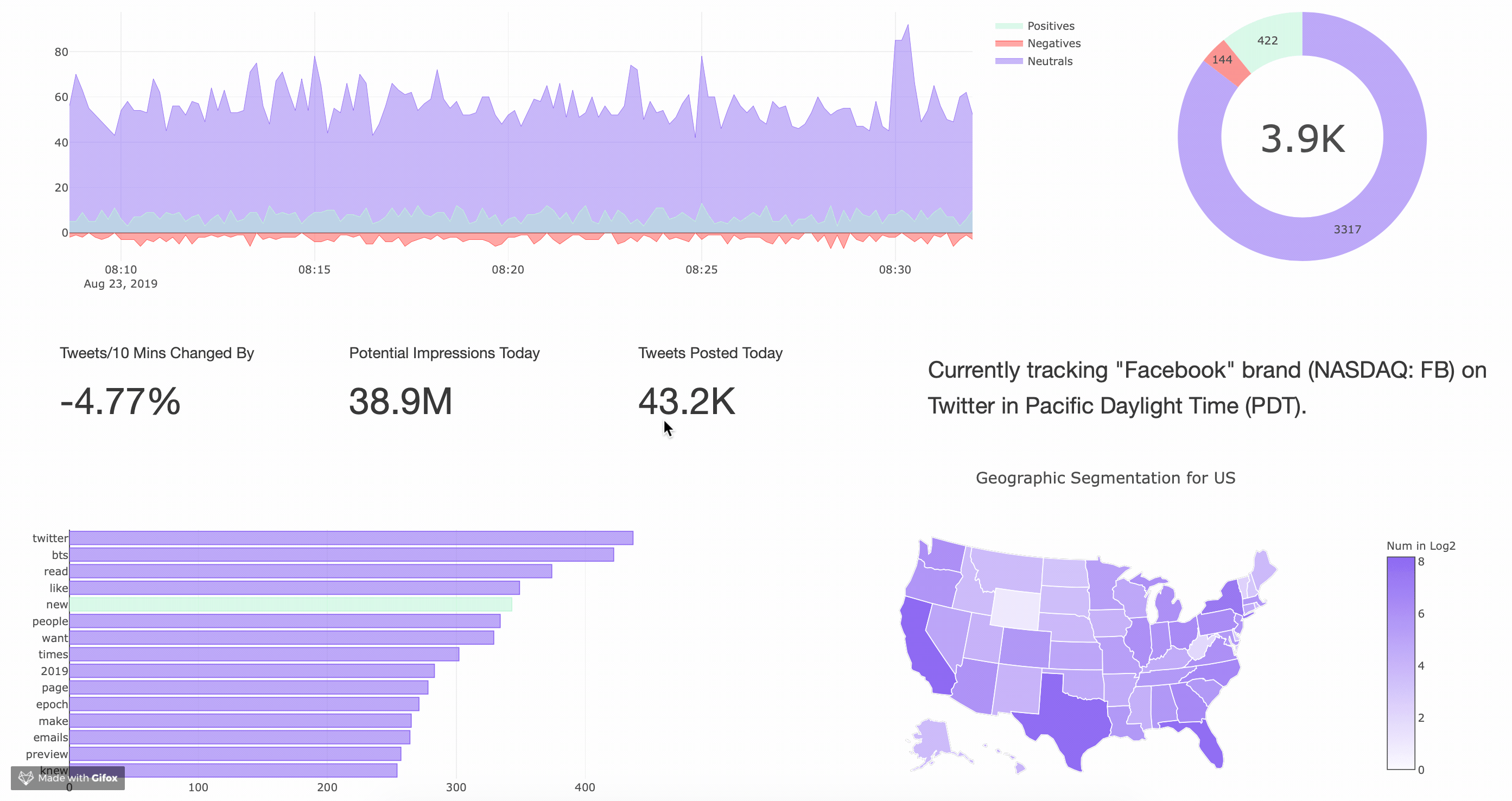
How was Trump’s attitude toward some of the trending news and celebrities? (e.g. Joe Biden, COVID-19 …Which words has the highest co-occurrence with these words? And what was the sentiment these words stand for?)

**Modeling**  
In my prior experience ridge regression works better in NLP tasks. But I will try naïve bayes, decision tree and neural network and compare their results.

A popular visualization in sentiment analysis (and other NLP tasks) is WordCloud. It will be a figure like the following one.



And maybe I can visualize the changing of sentiment like the following one. This one is real time. I think a real time one is too complex for me but a line chart and labels sounds like what I can do.



**Criteria:**

These is lots of things I can do with this dataset. At least I shall complete the first two questions, with a clear model use for sentiment analysis and some figures to present the results.

Then I can continue with the third one and perhaps other potential questions. (e.g. Who will likely be retweeted by Trump? What is the sentiment of these tweets?)