* Any surprises from your domain from these data?

I have to say, I was surprised at how generally positive the sentiment was for Biden tweets. Depending on your news source, it can seem like nothing but vitriol directed at one side or the other but when viewed as an aggregate they are generally positive! It was interesting to see that the Trump dataset was far more volatile than the Biden dataset, with occasional hours that were a net negative sentiment. Biden’s set, however, only had a handful of hours that were negative, and only just so.

While the Trump data was much more volatile, its high point tops out at 853 versus 3512 for Biden. I have not had a chance to match the peaks to notable events on the election timeline – that is a phase I am excited for. I hope to be able to pinpoint the event that triggers the tweetstorms!

* The dataset is what you thought it was?

Yes, it was surprisingly well done! I did have a few rows I had to prune because they didn’t import correctly. We are talking about something on the scale of 40-50 rows out of 867,000 and 693,00 rows respectively that’s pretty good! I cleaned the tweets up and removed special characters. Since the collection of tweets is all-inclusive, they were not filtered for English only. About 32% of Tweets sent worldwide are in English. Given that we are looking in to the US election, it’d be a safe assumption to say most of the tweets are in English but it is something that I plan on looking in to.

* Have you had to adjust your approach or research questions?

Not much of my approach has had to change, fortunately! The dataset I picked was fairly complete, well documented, and fairly large as well. I was worried that I would not have enough data to do any meaningful analysis, or worse, too much data! I have definitely had projects that the data crunch brought my poor desktop to its knees, so it was a valid concern.

The first research question I had survived my first look at the data, which is a good sign. Moving forward, I will continue to flesh things out and try to suss out the thread of the storylines I am hoping to answer with this data.

* Is your method working?

Knock on wood, my approach seems to be working so far. I had not worked with sentiment analysis in a while so it took me a little bit longer to get that working correctly but the benefit of having 9 classes worth of coding projects is there’s a chance I have a snippet of code I can reuse.

Once I got the sentiment (both in a polarity score version, and one that classifies sentiment as positive, negative, or neutral), it was time for the moment of truth. I threw the data in Tableau and… it was what I was hoping to see! I ended up setting the time period to one hour. With so much data, I could have gone down to the minute but I felt like the visuals would not work as well. There were some \*very\* clear spikes as well that I hope to investigate. Overall, I am happy with my first stab at the data and I am excited to dive deeper into the analysis portion.

* What challenges are you having?

Reigning in the scope of the project most definitely. There are so many questions I could answer but instead of trying to answer everything I had to refine my questions, realizing that some of the questions were two sides of the same coin. I had some difficulty importing the data actually into Tableau – originally, the CSV file came through with a lot of columns with null values even though they weren’t in the original dataset. I ended up exporting just the columns I was using in an Excel file and that cut the file size down to a much more manageable 50 MB or so.