Jon Atkins

Intro to GIS 101

**A Statistical Analysis of the Correlation between Geographic Political Polarization and the Tone of Social Media Interaction**

**Intro/Research Question**

Twitter, like many other social media platforms, has become an increasingly important medium in which political discourse occurs. This is never more present than in the election cycle. In the United States election cycles have become points of social tension with social media as the arena in which many political debates rage, commentary arises and is shared.

Many studies have shown that politics and government in America have become increasingly polarized over the past few decades, but can the same be said for political discourse among citizens? States are large geographic entities, often with diverse political interests. Can these political differences be represented and this regional diversity categorized? With the rise of social media and natural language processing, we have a way to record, measure and quantify tone of conversation between people across the world. While the tone of social media interactions about politics can often take on a confrontational tone, can this tone be quantified and can it be related to this political and geographic diversity? Does the tone of political discourse change between those who come into regular contact with those who disagree with them and those that do not?

**Background**

In the last full national election in 2012, the neighboring Wisconsin and Minnesota both voted for victorious Democratic president Barack Obama and Democratic Senators (Tammy Baldwin and Amy Klobuchar) by significant majorities. However, both states hold significant conservative minorities and a closer look at the results reveals two very divided Midwestern states. There are huge clusters of highly Democratic regions and highly Republican regions. Many of the people in these regions can feel like they are living in different states, rarely coming into contact with the opposition in person. However, that geographic divide is in theory removed in social media. Thus the research question for this study concerns whether the tone of political conversation on social media changes for those that live on the borders of these partisan strongholds or in outlier areas.

**Data Sources and Tools Used**

For this project, a lot of outside resources were used, in the form of data sources, data management libraries, data analysis tools, data scraping modules and other important tools. First off, the precinct shapefiles and precinct-level data is surprisingly difficult to obtain for many states. After failures to obtain relevant or usable data for notorious battleground states like Ohio, Virginia or Colorado, Wisconsin (known for its particularly strong partisan divide) and Minnesota (known for its easy to access precinct shapefiles and proximity to Wisconsin) were chosen to ease the process. The shapefiles and precinct-level voter data and demographics were taken from the Office of the Secretary of State for both states.

Obviously, the tweets were scraped from Twitter.com, specifically from its Search API v1.1 using a custom-written python script. For the Python search, an open source library named TwitterSearch was used. To store the data, python module Pandas was employed and then VADER Sentiment Analysis\* was used to analyze the tweets and the Google geocoder API (via python geocoder library) was used to mass geocode the 100,000+ results of the queries.

*Libraries/APIs Used*: Pandas, TwitterSearch, VaderSentiment, Twitter Search API, Google Geocoder API, Python Geocoder, Openpyxl.

**Methodology**

Analysis for this study was based on precinct delineations and precinct-level data from the 2012 general election. Data was procured and analyzed according to a ‘Percent Democratic’ variable:

*Percent Democratic*is an attempted generalized, binary representation of a precinct’s partisan preferences. It is an attempt to both remove the impact of non-major party candidates as well as the complicating factors of specific candidates. The formula is as follows:

((DEM\_PRES\_VOTES/TOT\_PRES\_VOTES)

+ 1 -  (GOP\_PRES\_VOTES/TOT\_PRES\_VOTES)

+ (DEM\_SEN\_VOTES/TOT\_SEN\_VOTES)

+ 1 - (GOP\_SEN\_VOTES/TOT\_SEN\_VOTES)) / 4

From there, two different spatial analyses were performed:

* *Hotspot Analysis*to find the most strongly Democratic and Republican areas, as a preliminary measure and visual representation of divides. Next,
* *Local Moran’s I Cluster Analysis*to find outlier precincts, which were mapped and combined without regard for the partisan lean of the polarization.

To gather tweets a number of steps were taken. First a dictionary of election-related terms was created. These 57 keywords ranged from generic ‘*election*’ or ‘*politics*’ to key issues like ‘*immigration’* and “*social security*.” It also included each of the presidential candidates’ names, many variations on the party names and even common hashtags like ‘*feelthebern’* & ‘*nevertrump*.’

From there, a smallest possible circle was drawn around the two states (as seen above), and the origin and radius were used to establish a query using Twitter’s search\_by\_location feature.

As the tweets were collected for the week from May 3rd-10th 2016, *VADER Sentiment* Analysis was used to find the tone (on a scale from –1 - 1) of the tweet’s text and, where possible, the coordinate or place information were recorded. For the vast majority of tweets that included no specific location information, the ‘location’ given in the user profile was used and geocoded. In all, around 50,000 tweets were able to be geocoded, of which 26,538 were within the sample area. These tweets were mapped and coded.

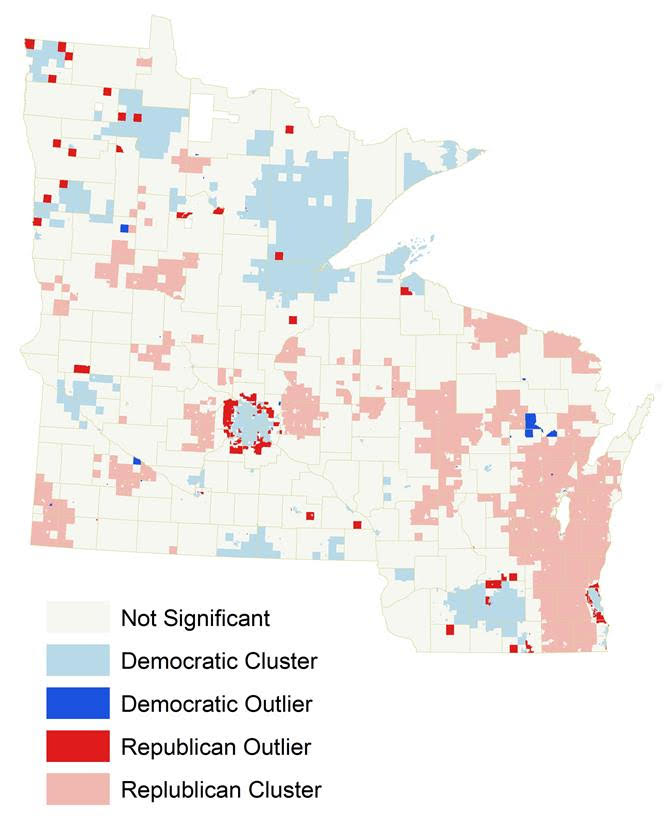
**Difficulties Encountered**

In completing this analysis a multitude of difficulties were encountered. And even more potential issues with the conclusions were discovered.

*Difficulties*:

* As mentioned above, the precinct and voter data was particularly difficult to obtain. I spent hours upon hours attempting to track down up-to-date (or at least matching) precinct shapefiles and precinct-level data. Most of the decision on Wisconsin and Minnesota was that they were the two best states that I found for documenting this data (MN in particular had useful compilation of data already in a shapefile, requiring no tricky joins or the validations that come with that).
* When dealing with Twitter Location data is important to realize that very few tweets come with the geotagged information available. The “coordinates” feature is opt-in and very few users opt to do that. And even so, it is probably a very specific subset of people who choose to share their location that way. It is possible, though, by combining data from the “Coordinates” variable, the “Place” variable, and, as a last resort, the “User[‘Location’]” variable (the location shared by a user on their profile) to come up with a reasonable estimate for a significant minority of the tweets. However this data is far more general and less accurate than the coordinate data. Furthermore, the serach\_by\_location feature doesn’t really work, and requires much validation afterwards to sort out the data you actually want.
* Also with regards to the tweets, it is impossible to scrape tweets from more than seven days ago except when searching for specific users’ public timelines. Thus, the data is probably heavily skewed to current events trends. Furthermore, it is very easy to run up against the API hit limit for the searches. It is technically only possible to search for 2500 Tweets every 15 minute (unless you want to pay a few thousand dollars for the entire archives – in which case there is no seven day limit either) so the script needed to be reran and tweaked to avoid hitting this limit repeatedly and messing up the data.
* This API limit was also an issue with the geocoding service. Google only offers 10,000 free API hits a day without payment, and after about 5 different partial run-thrus (of around 20,000 tweets each) that failed to save because of this and other issues, I elected to just pay for the 100,000 API hits a day and run a lot of pre-analysis filtering on the data to avoid reaching this limit with data that would not prove helpful (outside the geographic boundaries/non-useful location data).
* Another technological issue is the viability of the sentiment analyzer used. It was at least possible to download the source code directly to avoid paying for API hits with this but the overall effectiveness of Sentiment Analysis, no matter how rigorous, at this point in time is up for debate. After a few test run-thrus I attempted to correct for unintended consequences but it was very difficult to tinker with the 7,500-word dictionary to make a difference, as detailed above.
* Furthermore, there were many potential sources for error in the actual analysis because of the small sample size (only about 20 precincts out of 12,000 in total were considered Democratic Outliers) and the conflating factors when looking at only two states. Also the sample of tweets was small enough that it could have been influenced by a relatively (compared to the normal size of Twitter) small amount of re-tweets of a specifically polarized tweet.

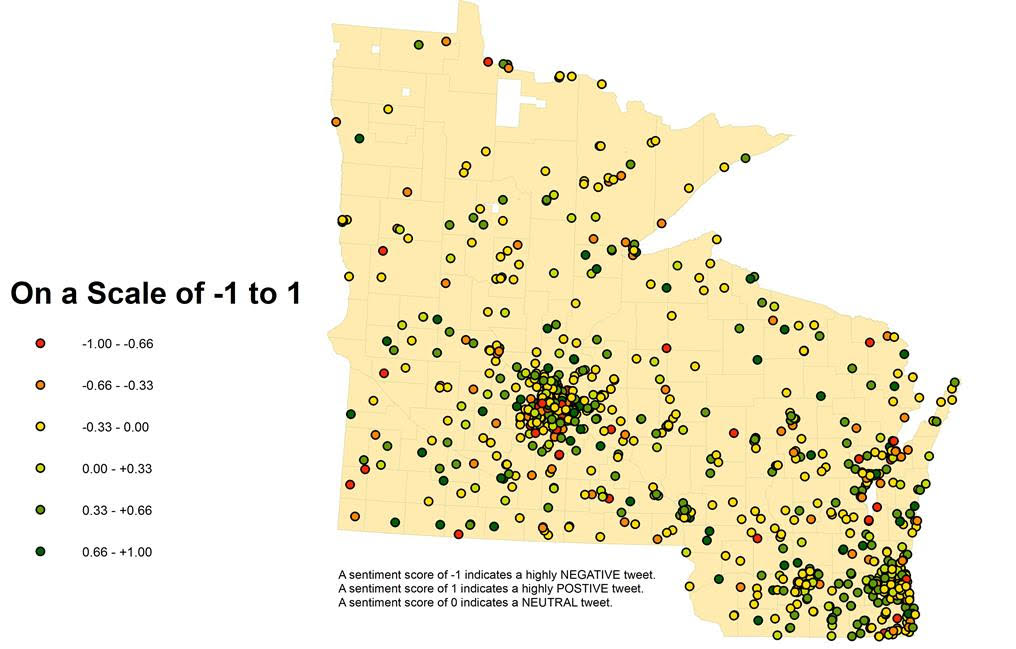
**Conclusions**

*Wisconsin*is a notoriously politically divided state. Simultaneously capable of electing Tammy Baldwin and voting for Bernie Sanders as re-electing Scott Walker and once supporting Joseph McCarthy. A precinct-level analysis of the state shows this divide starkest  between the extremely Democratic Milwaukee proper and the fiercely Conservative suburbs. This divide then extends again as you travel further west to the Democratic stronghold of Madison, near the University.

*Minnesota*shows a similar, if less intense trend. The Democratic stronghold of the Twin Cities quickly gives way to the most Republican of suburbs that surround the urban area in a ring. Furthermore, like a smaller Madison, remote Duluth merges the liberal tendencies of a college town and mid-size city to create another political pole in the state.

*Precincts Analysis*:By the precinct-level analysis, it is possible to see that the most politically polarized outlier areas exist mainly in suburbs of these major cities and remote locations away from population centers.

*Twitter*: Because most of the tweets were not geocoded beyond the Town level (mostly based on user-listed locations which rarely if ever specify the precinct), both data was dissolved into county and official town (where applicable) layers and then batched and analyzed accordingly.



The overall results showed a slight positive tone (.060 in the tweets, but drastic differences based on the search keyword). As for geographic results, the tweets were fairly spread across the country but unsurprisingly tended to cluster around the areas of high population.

For the geographic correlation, results, as can be seen in the graph, did not show a strong correlation (-.068) between the variables of geographic political polarization and positivity of tweets. Thus it appears, at least for these states and this week, there is little verifiable link between geographic political polarization and the tone of political discourse on social media.

**Annotated Citations**

**\*** Hutto, C.J. & Gilbert, E.E. (2014). Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eight International Conference on Weblogs and Social Media(ICWSM-14). Ann Arbor, MI, June 2014

*This version of sentiment analysis was particularly useful as it came with a built in dictionary and accounted for word order. It is intentionally built for use with social media and as such includes such things as abbreviations and emoticons. It would have been much more difficult to do this project without this as every other attempt I found involved creating my own sentiment analyzer. Nonetheless, I had a few issues with it, mainly that I couldn’t figure out how to add to the library. I attempted to add terms that come up in common political discussion nowadays such as “Nazi” and “tremendous,” as well as the urls of popular memes from the search window, in order to more accurately represent the twitter universe. However none of my additions would work. Nonetheless, this is an incredible library.*