

HRVWCC Project Presentation
Jason Driscoll
ECON 8320

Objective

The initial objective of the project was to investigate whether the collection and analysis of data from the internet and social networking are effective in augmenting traditional data sources in evaluating the probability PMC Wagner is operating in a specific geographic area.

Methods

The project is an academic exercise that starts with the following assumptions:


- PMC Wagner operates in Africa and in at least some of the operations PMC Wagner's participation is either covert or not actively broadcast.
- The Human Rights Violators and War Crimes Center has an interest in identifying PMC Wagner's participation in operations in Africa but may not identify all operations through traditional means.
- Local people who are negatively affected by PMC Wagner operations may share information about their experiences on social media.

Given these assumptions, it is hypothesized that the negative impacts of PMC Wagner operations, particularly those involving suppression or political violence, may increase the amount of social media material mentioning PMC Wagner, or influence the polarity or subjectivity of social media material mentioning PMC Wagner.

The proposed first step for exploring this hypothesis is related to the collection and analysis of Twitter data from defined geographic regions. The intent is to achieve this goal by collecting Tweets mentioning PMC Wagner from set areas defined by the Twitter API's geo.fields parameters, therefore restricting the Tweets collected to those coming from countries in which PMC Wagner is known to operate. Tweets will then be subjected to some analysis, including analysis of the number of tweets from each geographic region during a specific time frame, as well as using the natural language processing tools available in Python to assess the polarity and subjectivity of the tweets. After analysis, visualizations of the results will assist in searching for patterns or insight for use in assessing the hypothesis.

Obstacles

Obstacles began to present themselves at the start of data collection. The most fatal of these obstacles involved difficulty determining how to access geographic information when using the Twitter API to send a search request. API requests including geographic parameters, which appeared to be correctly formatted, returned 403 errors, indicating the request was understood, but refused or forbidden:

403	Forbidden		The request is understood, but it has been refused or access is not allowed. An accompanying error message will explain why.	Check that your developer account includes access to the endpoint you're trying to use. You may also need to get your App allowlisted (e.g. Engagement API or Ads API) or sign up for access.
-----	-----------	---	--	---

Twitter API Response Codes and Errors, 2023/05/09

<https://developer.twitter.com/en/support/twitter-api/error-troubleshooting>

Investigation revealed the problem was related to the use of a free developer account offered by Twitter for academic experimentation. While these accounts are free and require very little in the way of application or approval, Twitter has recently implemented changes to the level of access these accounts provide, and the ability to use parameters to search for tweets using specific parameters is very limited. Since the geographic fields are among those the free developer account does not have access to, and the use of geographic information was a vital part of the proposed data collection and analysis, this obstacle immediately threatened to derail the entire project.

At this point the decision was made to switch to the Bing API in hopes that it would prove more feasible to use the Bing API to collect search results from specific geographic areas. While it did prove possible to collect results using the Bing API, it was discovered that Bing also placed limitations on the parameters a person with a basic account can use in the collection of news data, as well limiting the end points the API call can use to the 'Recent' endpoint, returning results for only the past seven days.

This discovery prompted a return to the use of the Twitter API with the thought that upgrading the access level of the developer account would provide access to the previously forbidden geographic parameters. It was discovered, however, that the account upgrade still did not result in access to the

geographic field information collected by Twitter for all tweets. It did provide access to the place.fields parameters, and these were implemented in the API calls immediately. And while these calls did return results, data related to the place.fields parameters were not present. After much investigation, it is likely the cause of the lack of data is that, unlike the geo.fields parameters, which are collected for each Tweet, the place.fields parameter is optional, and will only appear if designated by the user. Within more than two thousand Tweets collected using the API there were no instances of the place.fields parameter returning data.

Revised Methods

At this point the likelihood of making use of the geographic tools offered by the Twitter API or the Bing API appeared to be very low and the necessity of geographic data to the project indicated the need for an alternative method of collecting geographic data. As a result, the method of data collection was revised. The intent was to continue using the Twitter API, but instead of searching for Tweets mentioning Wagner from a specific geographic region, tweets in which PMC Wagner was mentioned alongside the keyword Africa were sought instead.

Using methods obtained from an article posted by Andrew Edward in *Toward Data Science* as a guide and template, it was possible to collect a significant number of Tweets fitting the criteria. Although another minor obstacle presented itself when it became apparent the time spent troubleshooting the geo.fields issues meant there was no way to get around the restriction to the <https://api.twitter.com/2/tweets/search/recent> in time to complete the project, this was resolved using a method suggested by Dr. White in which the API call was initially set to search the entirety of the seven-day period available, and was then modified and run again once daily for a few consecutive days. As a result, it was possible to collect 2,607 results from the period April 26, 2023, through May 06, 2023.

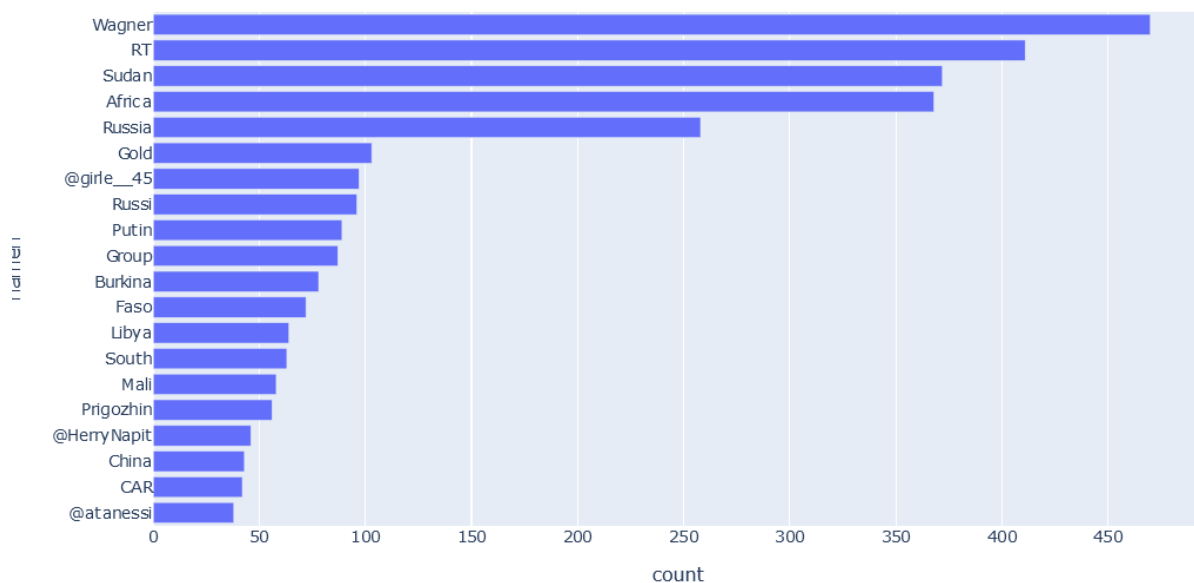
Data Cleaning

The majority of the 2,607 results collected were not relevant because, while they mentioned PMC Wagner and Africa, they did not mention a specific country that would indicate the geographic region to which the result should be tied. Additionally, calling Spacy's NLP method on the text of the full set of Tweets caused the Jupyter kernel to fail at a rate of about three times out of every four tried. This seemed like it was related to the workload, or the processing time required. I wasn't able to pinpoint which was the culprit, but since one of the evaluation methods for the project is that it is able to run on

other machines without modification, it seemed prudent to winnow out the non-relevant results. To accomplish this, and to facilitate the mapping of the results, the `natural_earth` data set from `geopandas` was used. By taking results restricted to the continent of Africa a list of African nations along with the mapping information was obtained. For some reason the numbers associated with each countries' place in the data set came with the names of the countries. Once these numbers were removed the data from the tweets was then iterated over to identify the Tweets mentioning a country in the list of African countries obtained from the `geopandas` data set. The result was a set of 563 Tweets fitting the criteria of mentioning PMC Wagner, Africa, and an African country name during the time period under review.

Results

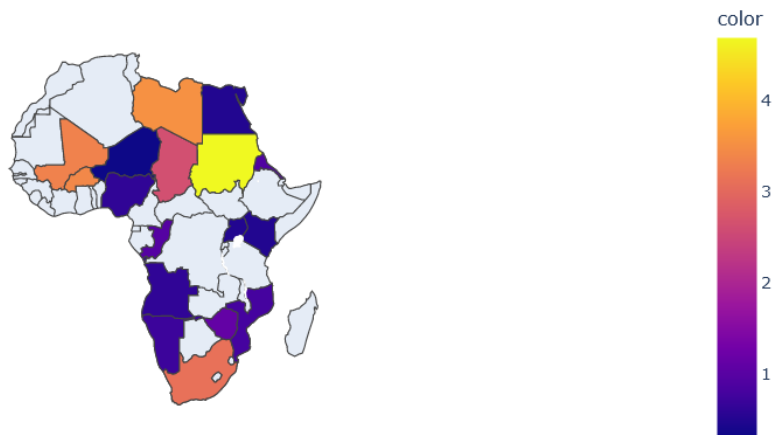
When the NLP method was called on this much smaller set of Tweets it was able to complete much more successfully. `Spacy` and `SpacyTextBlob` were used to enable processing of the data. The first attempt to sum the mentions of each country started with creating a list of proper nouns in the text. This produced an interesting bar chart, but of the top 20 proper nouns, a small minority were relevant country names:



The next attempt was a simple loop through the list of proper nouns directly tallying the countries mentioned and resulted in the following visual of the cumulative number of mentions of each countries' name within the collected Tweets. I'd like to mention that the significance of an outlier resulted in each of the other represented countries being the same color or close to the same color, so the log transformation was used to bring the differences between each country into clearer view. Also note that for the purpose of visualizing the number of mentions of each country, the likes, retweets, quotes and

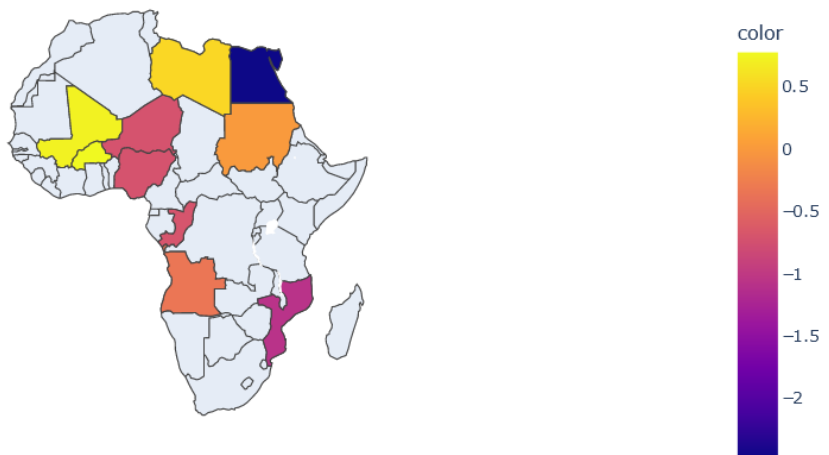
replies were tallied and added to the number of direct mentions to create a more comprehensive score reflective of the 'Impact' of tweets for each country.

Cumulative mentions of country name alongside Wagner



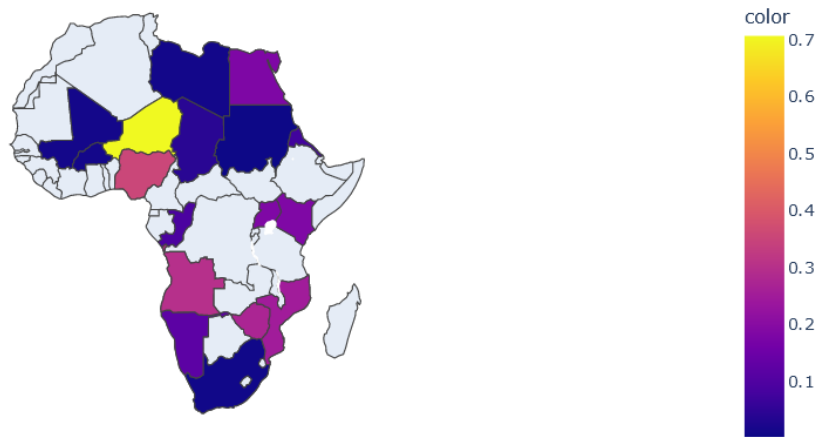
Also of interest were the polarity and subjectivity scores of the Tweets. These were assessed using Spacy and SpacyTextBlob and added to the data sets, producing the following visualizations:

Polarity of Tweets Related to Country and Wagner



And:

Mean Subjectivity of Tweets Related to Country and Wagner



There is some debate to be had about the best method of visualizing polarity and subjectivity. Both visuals originally showed the mean of the polarity/tweets. I then realized the large number of tweets, likes, retweets and replies coming from Sudan meant taking the mean minimized the differences between Sudan and the other countries. The final decision was to represent polarity with the cumulative numbers and subjectivity with the mean. The decision was primarily based on the scale used. The polarity uses a polar scale in which the distance from the midpoint provides relevant information, and therefore I think the cumulative total is an appropriate method of visualization. On the subjectivity visualization I decided to use the mean, and I am less sure that is appropriate. In fact, I am sure that it is *inappropriate*; to use the mean I suspect I would need to strip the number of retweets, likes, and replies out of the mentions. Unfortunately, this realization comes at the literal 11th hour, and my factorization of certain aspects of the project was disorganized enough that I feel tracking down the proper data set and making the changes to my code is, at this point, too risky.

Continuation

The project did not complete the objective, largely due to difficulties accessing the relevant information via the API. I feel, however, the results do indicate that there are differences in cumulative social media activity, polarity and subjectivity among countries and in areas in which PMC Wagner operates. If I were to continue work on this project there are some obvious avenues of continuation. First is obtaining access to the geo.fields parameters via the Twitter API. I believe this could be accomplished by requesting academic access, which would provide access to the desired geographic information, as well

as allowing the user to search the entire archive of tweets rather than being limited to the last seven days.

This would allow the use of the geographic information, which the documentation indicates can be as granular as using a set of 4 latitude and longitude coordinates to define specific geographic locations as small as a box of 25 miles on each side, as well as using tweets from a broader time frame to increase the relevance of the analysis by taking change in volume, polarity and subjectivity of tweets over time into consideration.

Work Cited

#Edward, Andrew (2021, Jun 13) An Extensive Guide to collecting tweets from Twitter API v2 for
#academic research using Python 3. Towards Data Science.
#<https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a>