## Twitter Analysis

Objectives: Build a program to measure the following on a regular basis, and output results:

1. Timing and “reaction speed” to BPCI Advanced-related announcements from CMMI
2. Level and variety of engagement with CMMI tweets and announcements:
3. Geographic analysis in heat-maps or other visual representations:
4. Content analysis (positive/negative, keywords, word cloud)
   * Many Twitter mentions are sales pitches from vendors or conveners such as Archway Health and NaviHealth. May be useful to know who is in the space, and what problems they’re highlighting as something they can solve.
   * These Tweets also point to other activity and discussion (e.g. conference presentations, webinars, appearances on other blogs and podcasts).

**Basic-level solution, Advanced-level solution**, **Complete-level solution**

1. **Twitter data extraction:**
   1. **Identify BPCI Announcements.** Identify historical BPCI-Advanced related announcements to use as test cases for building Twitter extraction tool; ask Health folks for sources.
   2. **Identify BPCI Related Tweets.** From announcements, identify tweets related to each announcement. The identification process will involve iterating on possible Twitter searches to identify the set or sets of query parameters that will minimize false positives (inclusion of irrelevant tweets) and false negatives (exclusion of relevant tweets). Conduct manual review and curation of these tweets, announcements, and their relationships. Conduct a more extensive manual review
   3. **Identify CMMI tweets and announcements.** Extract any other related tweets or announcements using same process as 1a/1b. We need to clarify with Health the distinction (if any) between the relevant announcements/tweets between objective A and B.
2. **Clean and prepare tweet data:**
   1. **Clean up json tweet data.** Some tweet jsons have further jsons nested into their structure. Will need to parse and flatten out json data so we have a normalized data set of relevant tweets.
   2. **Extract geolocation of all relevant tweets.** Geolocation, if present, is embedded in the tweet data json files but it may be a tweet-level or a user-level characteristic. It might also be worth seeing if additional geospatial data can be obtained from separate Twitter API requests for specific users. In the past I’ve had trouble generating location for a significant portion of tweets.
3. **Analyze data for each announcement/objective:**
   1. **Objective A:** Iterate on more specific definitions of “reaction speed“ and timing. Can recycle some concepts from the SIE work to for this, like time elapsed to maximum tweet impressions.
   2. **Objective B:** Similar iteration and recycling concepts as objective B, like number of favorites/retweets generated, and number of impressions. There are some more literature on measuring engagement and influence we could investigate into using to measure these concepts in a more sophisticated fashion.
   3. **Objective C:** Doesn’t have code to recycle from SIE, will need to explore this from scratch. Need to work with Health to better understand what kinds of geospatial relationships between tweets are meaningful. Would need to read more about what kinds of geospatial analyses exist. More in-depth exploration of literature.Even more in-depth exploration of literature.
   4. **Objective D:** As part of the brainstorming we did in June, I already have written code to do this. I have set up some basic classification of sentiment for tweets using the nltk’s VADER scoring method. I also generated some bag of words visualizations like word clouds using python packages like wordcloud. We could port over most of what I did to produce similar output for each announcement. We would then tinker with using Twitter queries to separately identify conversations from vendors. We could explore additional and more sophisticated kinds of visualizations and NLP analyses that can be done with tweets. Would need to read more about what is out there to recommend additional analyses. We could also explore substituting the VADER scoring method with alternate classification, applying topic models to extract common topics, and do more in-depth reading into how else we could group and classify concepts identified in the tweet corpus
4. **Export output into an accessible format**

Export data into raw csv/xlsx files. Hand off to Health analysts to manually format and clean up presentation of information.

Build a basic offline GUI with buttons to open up graphs, charts, and data on command. Health then curates the output of interest and manually transmits to client.

Fully integrate the output into the existing Tableau dashboard or other web-based tool for seemless content sharing with client.

1. **Running steps 1-4 to run regularly**

Set up Python files on shared drive, and instruct health staff how to run from command line when ready for updated results.

Set up an offline GUI, and train health staff how to use it to rerun code and obtain updated output. Build in “change” exhibits that compare current and previous exhibits if desired.

As part of integration with a web-based tool, have the code automatically run at regular intervals. Will also automatically send update reports to health staff, and alerts if the automatic run raises an error.

1. **Ongoing maintenance of monitoring tool**

External changes may cause the existing version of the monitoring to raise errors when rerun. This will require DMG staff to spend time troubleshooting the issue, and fixing the error. DMG staff could train health staff to do basic adjustments such as altering search queries or other simple changes to how the tool works. DMG staff could be held on retainer to make modifications as client priorities shift and change.

**LOE Estimates**

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Basic | Advanced | Complete |
| 1. Twitter data extraction | 15 | 20 | 25 |
| 2. Clean and prepare tweet data | 15 | 15 | 15 |
| 3. Analyze data for each announcement/objective | 45 | 65 | 85 |
| 4. Export output into an accessible format | 5 | 20 | 60 |
| Total | 80 | 120 | 185 |
| 5. Running steps 1-4 to run regularly | 5/month | 10-15/month | 10-30/month |
| 6. Ongoing maintenance of monitoring tool |

**Delivery Dates**

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Basic | Advanced | Complete |
| 1. Twitter data extraction | 23-Aug | 23-Aug | 23-Aug |
| 2. Clean and prepare tweet data | 30-Aug | 30-Aug | 30-Aug |
| 3. Analyze data for each announcement/objective | 13-Sep | 20-Sep | 20-Sep |
| 4. Export output into an accessible format | 20-Sep | 27-Sep | 27-Sep |
| 5. Running steps 1-4 to run regularly | - | - | - |
| 6. Ongoing maintenance of monitoring tool | - | - | - |