MSBA Program

Practicum Project

Instructor: Brian Murrow

Final Report

by

Melis Diken
Matias Roca-Guifo
Zachary Vila

Navanti Group

COVID-19 Sentiment Analysis & Visualization

Executive Summary

For our Practicum Project at The George Washington University, we were sponsored by our client, Navanti Group, to produce a real-time sentiment analysis and tracking dashboard, which will aggregate and display COVID-19 misinformation transfer into the "mainstream". To carry out the analysis, we elected to use social media posts from the Twitter and Gettr platforms as a proxy for public opinion on the subject. Throughout the course of the semester, we met regularly with our client to refine the project scope, provide updates on the progress of the project, and incorporate feedback on our efforts. The final dashboard product is a result of those efforts and the analysis therein, which provides actionable insights and recommendations based on the reality that COVID-19 misinformation is in fact having an impact on mainstream media content.

To conduct our analysis, we first developed a scope for the information to collect. Overall, content remained limited to users located within the United States. First, definitions were drafted and approved by Navanti to compartmentalize the types of information to be harvested. Second, specific user accounts were researched and selected to represent the primary definition categories we had drafted. Third, ancillary requirements were developed to eliminate redundancies in the information, such as controlling for employer accounts where a user worked, which could lead to duplication of content. The project scope also required a fair representation of accounts across the political spectrum, along with a category to designate specific politicians currently holding political office.

Following completion of the dashboard, which includes five unique visualizations and automated interval-based updating, preliminary results show a marked association between mainstream sources and "fringe-like" post content. By viewing the K-means algorithm results in particular, one finds a clear clustering of all users along the upper regions of the y-axis, which represents a low "mainstream-like", high "fringe-like" volume of content. While the project scope is limited, these findings encourage further analysis and a broader look into both the causal factors of these phenomena and mitigation strategies for COVID-19 misinformation. In addition, the project authors recommend further extrapolation of the methodology to analyze misinformation on social media as a broad phenomenon.

Problem Statement

COVID-19 misinformation has spread widely in the mainstream public sphere. Navanti seeks to monitor "fringe" COVID-19 misinformation transformation into "mainstream" assertions across mainstream and non-mainstream sources. This monitoring will assess the inflection point of fringe information absorption into the mainstream. Navanti desires to undertake an intellectual excercise through the use of a living dashboard with COVID-19 sentiment analysis. This will facilitate the monitoring of key misinformation life cycles across mainstream and non-mainstream sources, with scientific sources serving as the anchoring "truth" by which relative levels of misinformation can be measured.

Hypothesis

The prevalence of COVID-19 misinformation is influencing mainstream public opinion in unprecedented ways. Fringe information that might previously have been rejected on scientific grounds is finding acceptance among large populations of people, and even among subject matter experts. To prove or disprove this hypothesis, we conduct analysis of COVID-19 themes using ongoing data streams. We then test our hypothesis through juxtaposition of fringe and mainstream media sources using scientific sources as a baseline. Finally, aim to view the analytics results through a data visualization pipeline, draw conclusions, and make actionable recommendations.

Scope & Methodology

Conceptual Definitions

- Misinformation - Information largely unsupported by or fundamentally contradicting facts derived from scientific research or empirical evidence.
- Mainstream sources[1] - institutional and news organizations that shape broad public opinion. These sources employ reasonable amounts of editorializing and opinion catering.
- Fringe sources - Persons or news outlets primarily sharing information derived from personal opinion not grounded in fact. These sources regularly employ ideological rhetoric or persuasion aimed at converting or entrenching individuals' personally held beliefs.
- Scientific sources - United States federal government agencies with national public health as their primary mission. They strive to provide the most accurate information derived from empirical studies that employ the scientific method (i.e. the Centers for Disease Control and the National Institutes of Health).

[1] Derived from the Harvard media study dated May 11, 2021.

Media Segmentation

- Defined source* sectors (see Appendix B for list of monitored accounts):
    - Scientific (anchor[2])
    - Mainstream
    - Fringe
    - Politician

Information Bounding

- Flag inherently political sources (i.e. holding political office)
- Exclude redundancies between persons and their associated organizations/affiliates to control for magnification effects
- Evaluate verified and unverified accounts to establish user scope
- Media source definitions determine categorization
- Overtly political[3] sources are prioritized for fringe content harvesting
- Exclusion of policymakers without a prominent, established following
- Flexibility to add source accounts over time, based on relevance and source definitions ("living" source list)

Twitter & Gettr Post Data

- Twitter: April 2022 onward
- Gettr: May 2021 onward
- Post attributes: Username, Date of Post/Tweet, Post/Tweet raw text

Data Analytics

Tools & Methods

- Twitter (Advanced Tier API) derived custom data extracts
- Gettr open source API (GitHub) derived custom data extracts
- AWS EC2 cloud computing for analytic processing power
- SQLite Database framework for data storage
- Data filtering & wrangling (Python)
- Browser-hosted dashboard visualization server (Bokeh module)
- Automated API querying capability for data ingest at regular intervals (via schedule module)

[2] To measure deviation from "truth", the anchor source is established.

[3] Defined as sources that either cater to a political ideology, or attempt to denigrate an opposing ideology.

Initial[4] Tracked Themes

- Covid-19
    - Words used for Covid-19 theme: covid, c@vid, pandemic, virus, disease, omicron, delta, SARS-CoV-2, variant, outbreak
- Vaccines
    - Words used for Vaccines theme: vaccin, vax, jab, mrna, pfizer, biontech, moderna, J&J, Johnson & Johnson
- Mandates
    - Words used for Mandates theme: mandate, mask mandate, vaccine mandate, vaccine card, passport, lockdown, quarantine, restriction
- Alternative Treatments
    - Words used for Alternative treatment theme: vitamin, zinc, ivermectin
- Health Organizations
    - Words used for Health Organizations theme: CDC, NIH, FDA, WHO
- Conspiracies
    - Words used for Conspiracies theme: bioweapon, lab
- Origin
    - Words used for Origin theme: Wuhan

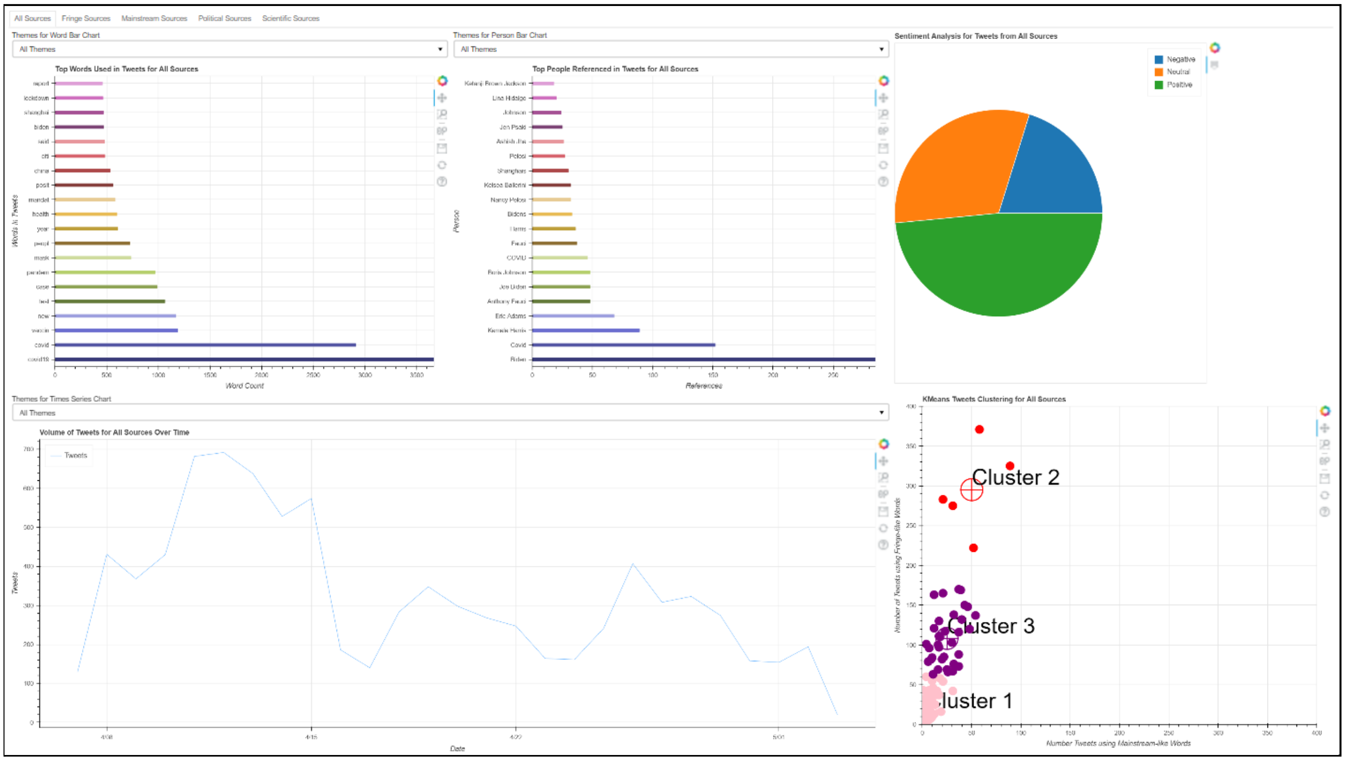
Algorithms

- K-means[5] Clustering
    - Unsupervised machine learning outlier detection and contextualization of user posts
- Sentiment Analysis
    - Pre-defined algorithm to classify user posts by positive, negative, or neutral overall sentiment

[4] Themes may be adjusted over time based on current trends.

[5] "K-means is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping sub groups (clusters) where each data point belongs to only one group" (Towards Data Science).

Dashboard & Visualizations


Themes for Word Bar Chart
All Themes

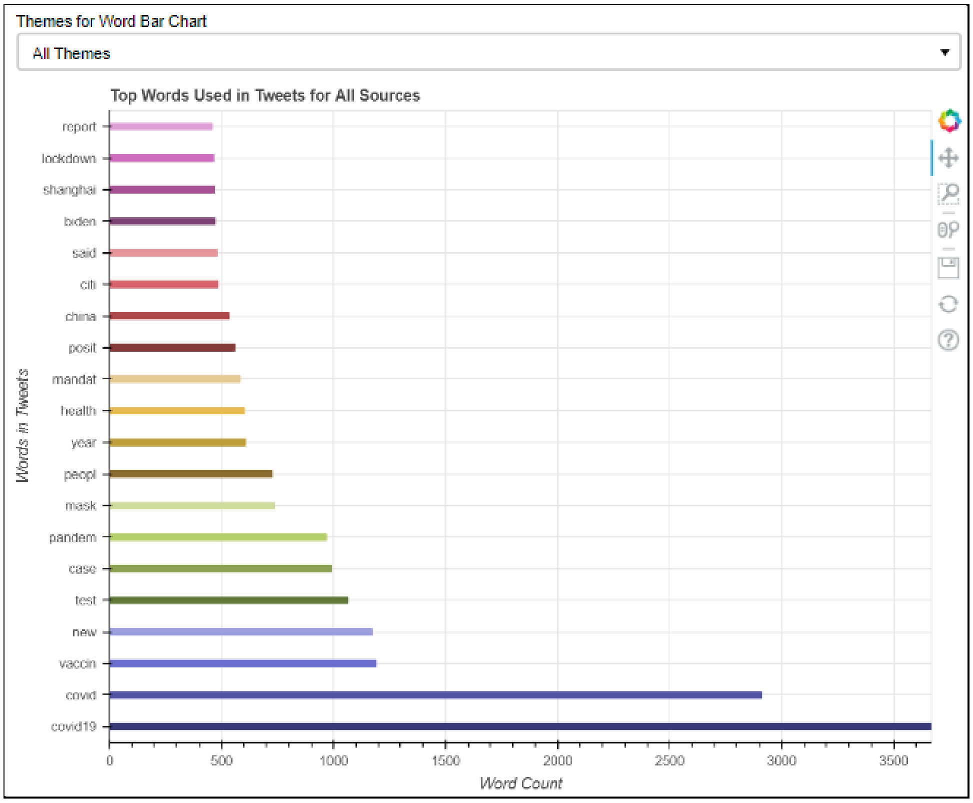
Themes for Person Bar Chart
All Themes

Themes for Times Series Chart
All Themes

Our dashboard consists of 5 visualizations, providing on-demand analysis of harvested social media content from Twitter and Gettr. Users are able to select from 5 distinct tabs, which segment the data by source type. There are also filters available to target specific themes in the data.

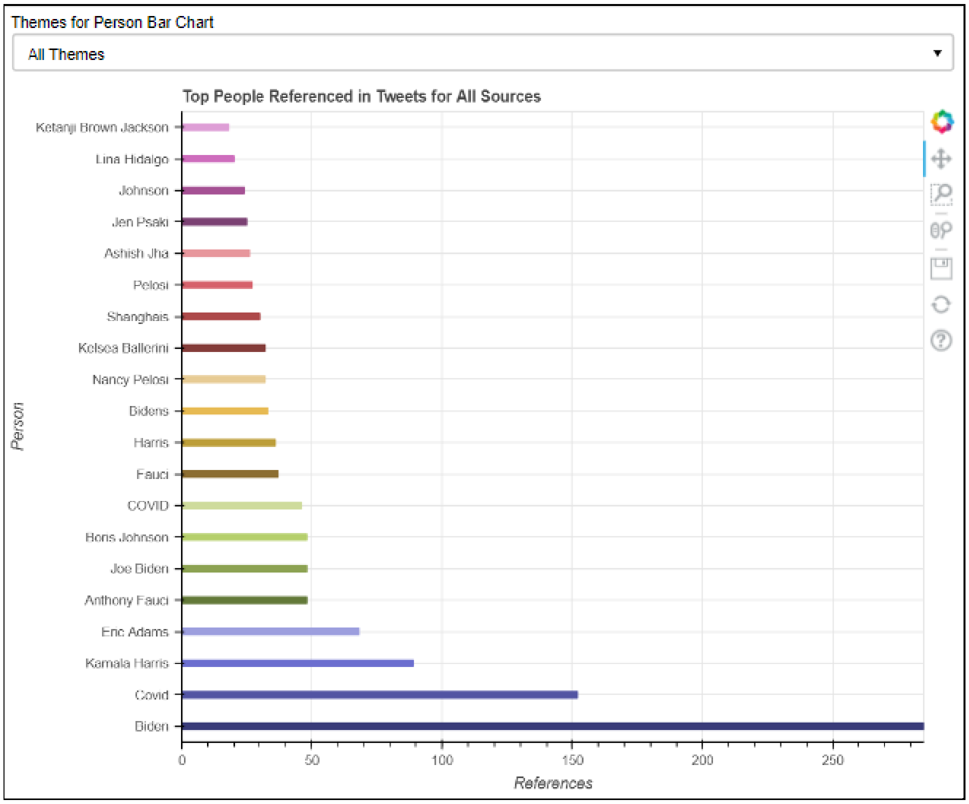
5/9/2022

Bar Charts - Trend Analysis



Themes for Word Bar Chart

All Themes ▾

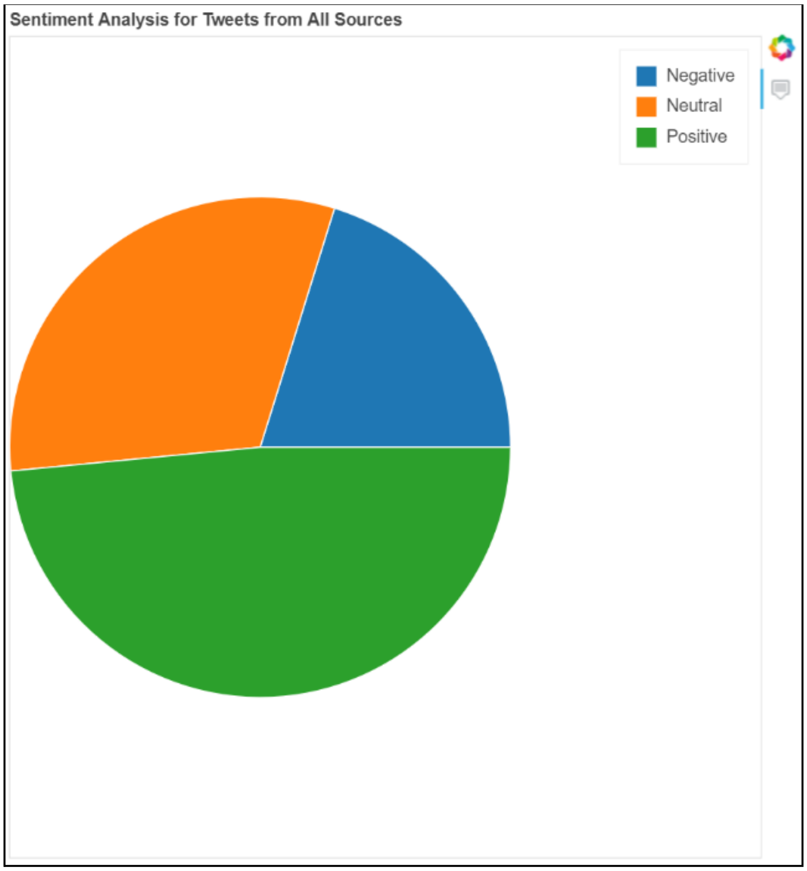


Themes for Person Bar Chart

All Themes ▾

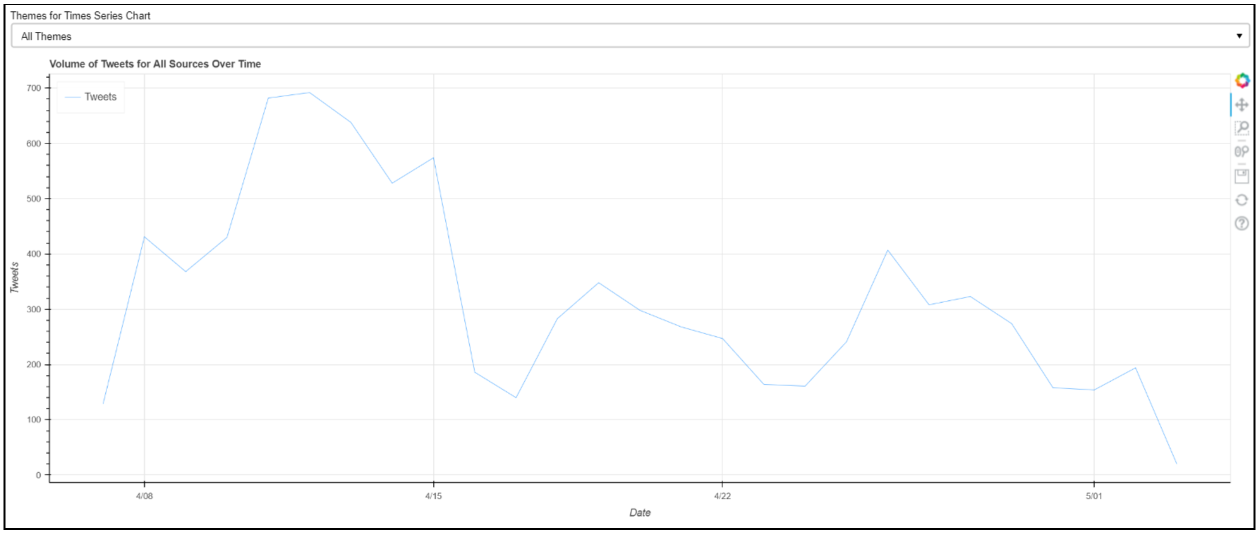
The Word and Person frequency charts provide a window into the top twenty entities occurring for each category, and can be filtered by theme.

Pie Chart - Sentiment Analysis Classification



The pie chart provides an overview of the general sentiment users are contributing at a given moment in time, categorizing the harvested posts as positive, negative, or neutral.

Time Series - Frequency Analysis

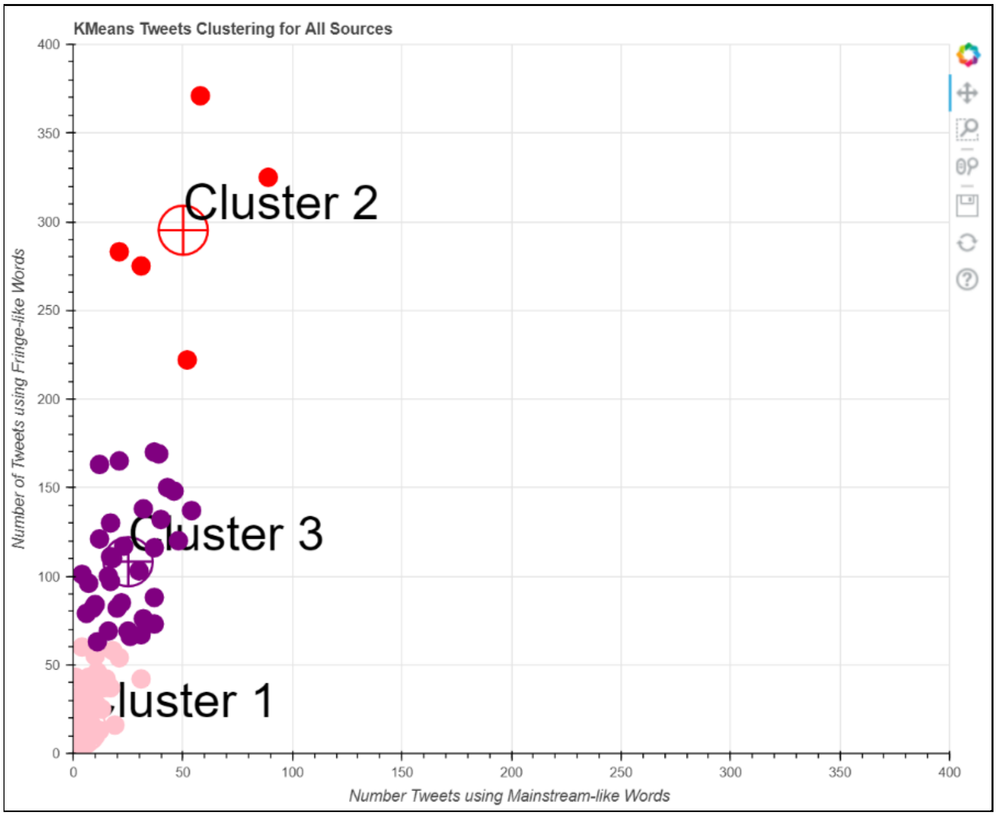


Themes for Times Series Chart

All Themes ▾

The time series chart provides insight into the change over time in the data, and can be used to compare volume of posts in a given period of time. The chart can be filtered by theme.

K-means Clustering Analysis - Unsupervised Machine Learning



The K-means algorithm provides a clustering visualization one can examine to better understand the nature of information being shared on Twitter and Gettr. Clusters with high numbers of Fringe-like tweets will fall higher on the y-axis, while posts with mainstream-like content will fall further right on the x-axis.

Conclusions

The project analysis shows a marked shift in COVID-19 sentiment towards fringe-like posts among the user accounts examined. This is particularly apparent when examining the K-means algorithm outcome using Twitter data. These data have been bound to provide the majority of mainstream content, though analysis shows the sentiment of the information falls far from a reasonable expectation of minimal fringe-like content. On Twitter, this shift towards fringe narratives and content surrounding COVID-19 is not isolated among particular groups or users; rather, it is a full-scope embracing of fringe-like content as defined. Moreover, our analysis confirms our hypothesis that within the project scope, fringe misinformation is affecting the expected types of content circulating among mainstream users. However, we cannot confidently assert a causal factor or factors for the genesis of the ready acceptance of misinformation among the users studied; we only endeavor to articulate the nature of the phenomenon itself.

Project Recommendations

- Increase the size of fringe, mainstream, and scientific sources so that they can be more reflective of each group
- Identify additional keywords associated with Covid-19 that can be used to filter posts on Twitter and Gettr
- Do not utilize an automated refresh system through AWS due to incurring significant costs with little added benefit. Gettr data can only be refreshed daily due to a lack of timestamps in the data, and Twitter would not see significant benefit from refresh more frequently than daily.
- Expand on filters for added targeting of analysis to key information areas
- Ensure maintenance and expansion of project data capture, as analysis and insights will improve as dataset grows
    - Additional data attributes (Retweets and Likes) could capture dissemination potential
- Creating a logistic model to predict a user’s information category
    - Using keywords we find in the Dashboard from each category as dummy variables to train the model

Further Study

Developing an understanding of this phenomenon has application beyond COVID-19 misinformation and could be applied to a number of public forums where misinformation is prevalent. We provide several potential areas of exploration here:

- The effects of social media on alternative personality prominence
- The effects of misinformation surrounding the following events:
    - Ukraine War
    - 2020 U.S. Presidential Election
    - Major U.S. National Security events

Appendices

Appendix A: References

Python

- Recommended AWS VMI Environment:
    - Python version 3.9,
    - Jupyter Notebook
    - Command Line Interface (e.g. Terminal or other program)

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Bokeh documentation — Bokeh 2.4.2 Documentation

bokeh.server — Bokeh 2.4.2 Documentation

schedule · PyPI

Gogettr Github: GitHub - stanfordio/gogettr: Public API client for GETTR, a "non-bias [sic] social network," designed for data archival and analysis.

Tweepy Github: GitHub - tweepy/tweepy: Twitter for Python!

Media Sources

https://en.wikipedia.org/wiki/News_media_in_the_United_States

https://adfontesmedia.com/

misinformation-and-disinformation-4-16-2020.pdf (cmu.edu)

COVID: Top 10 current conspiracy theories - Alliance for Science (cornell.edu)

Myths and conspiracy theories on vaccines and COVID-19: Potential effect on global vaccine refusals (nih.gov)

Understanding COVID-19 Conspiracy Theories Listicle - President's Writing Awards (boisestate.edu)

Eight Persistent COVID-19 Myths and Why People Believe Them - Scientific American

Tracking Viral Misinformation: Latest Updates - The New York Times (nytimes.com)

IFCN Covid-19 Misinformation - Poynter

- Mainstream

https://projects.iq.harvard.edu/futureofmedia/index-us-mainstream-media-ownership

- Fringe - Right-Wing

https://www.brookings.edu/techstream/prominent-political-podcasters-played-key-role-in-spreading-the-big-lie/

https://en.wikipedia.org/wiki/Alternative_media_(U.S._political_right)#:~:text=The%20 term%20right%2Dwing%20alternative,liberal%20bias%20of%20mainstream%20media

- Fringe - Left-Wing

https://en.wikipedia.org/wiki/Alternative_media_(U.S._political_left)
https://www.reddit.com/r/BreadTube/comments/8nxf9b/a_big_huge_list_of_channels_discriptions/
https://www.reddit.com/r/LefTwitch/comments/bwzef8/a_short_list_of_leftist_twitch_streamers/
https://breadtube.tv/channels/
https://www.reddit.com/r/BreadTube/comments/9kwkxa/master_list_of_leftwing_youtube_and_podcast/

- YouTube Political Channels

https://www.pewresearch.org/journalism/2020/09/28/a-closer-look-at-the-channels-producing-news-on-youtube-and-the-videos-themselves/
https://intpolicydigest.org/the-right-wing-vs-the-left-wing-on-youtube

Sentiment Analysis
https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

Appendix B: Information Sources (Gettr & Twitter Accounts)

- See spreadsheet list attached

Appendix C: Detailed Work Plan (Jira)

- Link: Navanti gwu - Roadmap - Jira (atlassian.net) and image attached

Appendix D: Python Code & SQLite Databases

- See Jupyter notebooks and database files attached

Appendix E: Project Deliverables

Stakeholders

- Navanti, plus potential clients based on historical relationships

CUSTOMERS / PARTNERS / TARGET MARKETS

Chemonics
Development works here.
fhi360
THE SCIENCE OF IMPROVING LIVES
USAID
SAFERWORLD
INTERNATIONAL ORGANIZATION FOR MIGRATION
IOM·OIM
THE UN MIGRATION AGENCY
THE WORLD BANK
IBRD · IDA | WORLD BANK GROUP
giz
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH
MSI
A TETRA TECH COMPANY
VOA
Voice of America
U.S. AGENCY FOR GLOBAL MEDIA
UNDP
OTI
OFFICE OF TRANSITION INITIATIVES
MBN
unicef

Deliverables

- Two real-time dashboards segmented by source types (applies to Twitter only) and filterable by theme categories with five distinct visualizations:
  - Time series of post counts
  - Sentiment Analysis (pie chart)
  - Top 20 filterable words by frequency (bar chart)
  - Top 20 persons mentioned by frequency (bar chart)
  - K-means Clustering
- EC2 cloud-based processing service (AWS account)
  - x1e.2xlarge instance (8 Cores)
- Two SQLite databases (segregated Twitter and Gettr post data)
- Bokeh Server application
- Development code (Python)
- PowerPoint Presentation
- Work Papers