**2021 Undergraduate Big Data Challenge: Data Portal**

[Instructions for Accessing Datasets](https://docs.google.com/document/d/1n1p4PEjIM2EXFQECC_bhn7JzNmxYBKJuXKJRsVB2EZU/edit?usp=sharing)

**Main Datasets**:

1. [CoAID: COVID-19 Healthcare Misinformation Dataset](https://ui.adsabs.harvard.edu/abs/2020arXiv200600885C/abstract)
   1. Dataset includes diverse COVID-19 healthcare misinformation, including fake news on websites and social platforms, along with users' social engagement about such news. CoAID includes 4,251 news, 296,000 related user engagements, 926 social platform posts about COVID-19, and ground truth labels.
   2. <https://github.com/cuilimeng/CoAID>
2. [FakeHealth](https://github.com/EnyanDai/FakeHealth)
   1. Dataset addresses challenges in Fake Health News detection, which includes news contents, news reviews, social engagements and user network
3. [COVID-19 Fake News](https://github.com/diptamath/covid_fake_news/tree/main/data)
   1. Collection of data from various social-media platforms (Twitter, Facebook, Instagram etc.) regarding Covid-19, of which some are classified as fake or real.
4. [COVID-19 Tweets Dataset](https://github.com/lopezbec/COVID19_Tweets_Dataset)
   1. An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic

*Note: Given the Tweet and Reply ID, you can trace back to the post with this:* [*twitter.com/anyuser/status/tweetIDnumbergoeshere*](http://twitter.com/anyuser/status/tweetIDnumbergoeshere)

**Supplementary Datasets:**

1. [Covid-19 Vaccine Tweets](https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets)
   1. Tracking hashtags relating to vaccine
2. [COVID Fake News Dataset](https://zenodo.org/record/4282522#.YKXcGZNKhhE)
   1. Dataset containing list of COVID Fake News/Claims which are shared all over the internet
3. [Meta-Analysis of Wikipedia Coronavirus Citation Sources](https://zenodo.org/record/3901741)
   1. Cited sources on wikipedia pages

**General Data Portal - Use these portals to find/search for data/datasets:**

1. [Pushshift.io](https://pushshift.io/)
   1. Web Application that ingests data from various social media companies such as Twitter and Reddit and makes that data available for researchers and academic institutions.
2. [Google Dataset Search](https://toolbox.google.com/datasetsearch)
3. [Kaggle](https://www.kaggle.com)
   1. Kaggle is a subsidiary of Google LLC, an online community of data scientists and machine learning practitioners.
4. [Harvard Dataverse](https://dataverse.harvard.edu/)
   1. Repository for research data
5. [Youtube Data API](https://developers.google.com/youtube/v3)
   1. Use the Youtube Data API to search for videos matching specific search terms, topics, locations, publication dates, and much more.
6. [Twitter API](https://developer.twitter.com/en/docs)
   1. The Twitter API provides the tools you need to analyze, contribute, and engage with the conversations happening on Twitter.
7. [Google Trends](https://trends.google.com/trends/)
   1. Google Trends is a website by Google that analyzes the popularity of top search queries in Google Search across various regions and languages.