# CS437 – Vaccine Classification by Tweet Content

# Code Demo Instructions

## Overview

These instructions are intended to be utilized alongside the Colab notebook linked in the team's GitHub readme, as well as in the Project Write-Up PDF. Additional in-line instructional comments can be found within this Colab notebook, and a description of models utilized can be found in the ml.md file within the repo. High-level project and process documentation can be found within the Project Write-Up PDF.

All aforementioned documents are stored within the team's GitHub repo, located below:

GitHub: https://github.com/reed271/437_Final_Project

## Step-by-Step Instructions

1. Navigate to the following Colab Demo:
   a. https://colab.research.google.com/drive/12dLOr1zcvVOal9gBJ4AK_drWn2EGjKl5?usp=sharing
2. Download the Twitter dataset from the following source:
   a. https://www.kaggle.com/datasets/prox37/twitter-multilabel-classification-dataset

3. Upload the file into Google Drive's local space using the upload() command, activated by running the first code block within the Colab notebook.
4. Run the second code block to import necessary libraries and initialize the Natural Language Toolkit.
5. Run the third code block to get and clean the raw source data, converting it to a form that is usable by our ML models.
6. Run the fourth code block to initialize the primary evaluate() function, which is used to evaluate the performance of any given model on any given dataset.
7. Run the fifth code block to initialize the secondary evaluate function, used for plotting results graphically.
8. Run the sixth code block to initialize the generate_data() function, which returns various permutations of train and test data splits, as well as min_df and ngram variations.

9. Run the seventh code block to initialize the LDA model.
10. Run the eighth code block to actually generate the data permutations that will be used in the fitting, training, and testing process.
11. Run the ninth code block to produce analytics on the class distribution of the source dataset.
12. Run the tenth and final code block to execute the driving code of the program. Here, we instantiate our seven ML models to be tested and add them to a dictionary. We iterate through this dictionary, training the models on the optimal datasets and measuring their accuracy and F1 scores.
    a. Output: Plots of each model's performance at classifying each given label, as well as tables showing overall accuracy and F1 Scores for each model, separated by count vectorized and TF-IDF vectorized datasets.