**#Vaccine Sentiment Analysis 💉**

**PROJECT OVERVIEW**

The emergence of the Coronavirus pandemic has normalized vaccines to become a part of everyday discourse, as global efforts to increase vaccination rates continue. Considered to have an integral role in managing the pandemic, governments, employers, and other organizations have either implemented or considered vaccine mandates, and this has possibly become the most contentious topic in the last couple of years.

As expected, there are varying opinions about vaccines, which range from their safety or effectiveness, to whether the mandates infringe on human rights. Anti-vex sentiments well-predate COVID, however given how pervasive vaccine discourse has become, they seem to be on the rise. Understanding the causes and levels of vaccine hesitance is an important first step in addressing the factors that prevent people from getting vaccinated.

For this project, we took to Twitter to gather vaccine sentiments from the beginning of the pandemic to present, in an effort to address the following questions:

* What are the sentiments towards vaccines?
  + Positive, negative, or neutral
* How have these sentiments changed throughout the course of the pandemic, if at all?
  + Do pandemic milestones (e.g. waves, mandates, laws, etc.) impact opinions?
* What word associations are found with the word vaccine?
* Are we really as divided on vaccines as it seems?

Data extracted from Twitter will be stored in a Postgres database, analyzed using Natural Language Processing (NLP), and the findings will be visualized on a custom dashboard.

**THE TEAM**

Divyanshu Bisht | Farwa Ali | Sangeetha Praveen Kumar

The team members have not been assigned specific 'circle', 'square', or 'triangle' roles. All members are working collaboratively on each task, and each member has their own branch. The team communicates primarily through Slack, but also meet at least once a week via Zoom.

Next, implementing a natural language processing algorithm allows us to gather our sentimental analysis

* Machine Learning Libraries: nltk, sklearn
* Description of preliminary data preprocessing

**TOOLS & RESOURCES**

SOFTWARE: Python 3.9.3

Libraries & packages: Tweepy, SQLAlchemy, NLTK

Jupyter Notebook or Google Colab

Database: PostgreSQL

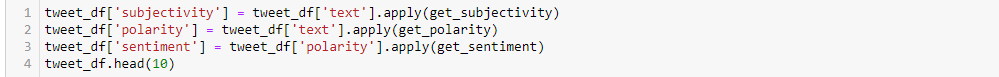
Dashboard software:

**THE DATA**

Data will be extracted using the Twitter API, which pulls "Twitter Objects" containing the text of a tweet. We will be focusing on tweets containing #vaccine from the start of the pandemic (around March 2020) to present.

**Data Processing**

1. Load historical twitter covid vaccine data from Kaggle
2. Clean tweets with clean\_tweet function(regex), tokenize and get ready for text classification. Also, clean up function for removing hashtags, URL's, mentions, and retweets
3. 2.Apply Textblob.sentiment. polarity and Textblob.sentiment.subjectivity, ready for sentiment analysis.



1. Apply analyze\_sentiment function on tweet texts to label texts with sentiment range from negative or positve or neutral.

**Description of preliminary feature engineering and preliminary feature selection, including their decision-making process**

1.Import CountVectorizerfrom sklearn.feature\_extraction.text. CountVectorizer is a tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. The value of each cell is nothing but the count of the word in that particular text sample.

2.Fit sentiment texts features with vectorizer, and target sentiment column.

Description of how data was split into training and testing sets Splitting into training and testing set so as to evaluate the classifier. The aim is to get an industry standard sample split of 80% train and 20% test.