Using Sentiment Analysis to See How the Negativity of Twitter Tweets in U.S. Concerning COVID-19 Changes During the Course of April 2020

Problem:
This capstone project aims to use sentiment analysis on coronavirus disease 2019 (COVID-19) Tweets in the U.S. to see how negativity towards the pandemic changes during the course of April 2020. The sentiment in this project will be categorized as positive, neutral, and negative.

Target Clients:
The primary clients this project targets are those that work in or deal with mental health and social analysis. Particularly, these clients would want to see how the social circumstances surrounding the COVID-19 pandemic and its lockdowns affect the social-welfare and mental health of the populace. COVID-19 is an infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that primarily affects the lungs. However, it was discovered that it can also negatively affect other organs like the heart, kidneys, and brain, causing a multitude of other health complications that can leave a lasting impact even after recovery from the initial infection. In the U.S., cases have risen more than 2.3 million with more than 121 K confirmed deaths as of now. It was first identified in Wuhan, China on December 2019 and has since spread worldwide, causing long-term lockdowns in many parts of the world. Besides taking a toll on medical health, it is indicated that the COVID-19 pandemic is also taking a toll on mental and social health, especially due to the stay-at-home orders.

Hence, this project could give mental healthcare workers and sociologists insights on how the COVID-19 pandemic situation is affecting the mental health of the populace who are heavily limited in their movements and social interactions to prevent the spread of the virus. Also, the results of this project could indicate how resources of psychological care could be distributed and utilized to mitigate any negative effects long-term pandemics and lockdowns may have on emotional health.

Data:
The datasets that will be used for this project were created by Shane Smith (https://www.kaggle.com/smid80) and posted on Kaggle.com under the CC0: Public Domain. They are available for download via the following links:

- Main Page: https://www.kaggle.com/smid80/coronavirus-covid19-tweets
- Early April: https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april
- Late April: https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april

Proposed Approach:
Multiple steps will be taken to create the sentiment analysis model for this project as well as to analyze the resulting predictions.

1. The multiple csv files consisting of the data will be imported, merged, and cleaned via Python 3. The variables that can be used for the sentiment analysis model and its predictions as well as give any additional statistical insights will be manually selected.

2. The merged and cleaned dataset will be explored visually via graphs created from Python 3.
3. As the original data does not have the sentiment labels (positive, neutral, and negative), the sentiment analysis model will be trained with the pre-labeled corpus of Twitter samples from the Natural Language Toolkit (NLTK) Python library to predict and create sentiment labels for the original data.
4. Then, the counts of the sentiment labels (positive, neutral, and negative) will be analyzed to see how they change during the course of April 2020.

Deliverables:

The final draft of the project will be presented via PowerPoint slides and delivered via Colab Notebook detailing each step taken and code written for the analysis of the project. A Github repository for the project will be created as well.