***Twitter Sentiment Analysis of Climate Change***

**Goal and Project Overview:**

Twitter is an extremely popular social media platform that many people across the world use to express their thoughts regarding various different topics. The goal of my capstone project is to analyze twitter data regarding topics relating climate change, and how people’s tweets have changed and evolved between 2015 and present day (5 years) using sentiment analysis.

Additionally, I conduct further analysis on my dataset to analyze and identify trends in negative or positive tweets and demonstrate these trends in a form of different visualizations such as graphs, word clouds, and maps.

**Source of Dataset:**

The primary dataset I will using in this project is a one that I have to capture. I will capture tweets relating to climate change for the past 5 years. The keywords that I will use include but not limited to climate change, global warming, fossil fuel, greenhouse gas, pollution etc. This dataset will be limited to English speaking tweets; therefore, my dataset is limited to people that can only voice their opinions in English. To capture these tweets I will use python packages such as Tweepy to access my Twitter API account and stream the tweets and write the data into a CSV file. The dataset will be in a CSV file, with columns including UserID, Tweet, Date Time of Tweet, Hashtag and Location of Tweet.

Currently my dataset contains 60k tweets, but I plan to run the twitter stream longer and expand on keywords, therefore it will grow in size.

The dataset is then classified into three categories: positive, negative and unidentifiable.

**Positive**: Tweet indicates that climate change is real.

**Negative**: Tweet indicates that climate change is not real

**Unidentifiable**: Tweet contains climate change terminology but does not indicate if its positive or negative. Examples include news or factual statements.

**Data Processing and Methodology:**

First, I will clean the raw data obtained from twitter by removing certain characteristics from the tweets such as removing emojis and whitespaces, which are not needed for my sentiment analysis. Then I will normalize the dataset, as the number of active twitter users vary every month. After cleaning the data and normalizing the dataset, I will be using machine learning algorithms such as Naïve Bayes or Support Vector machine to see which algorithm best classifies the tweets into the right category.