Proposal for Capstone project 2: *Sentiment analysis of Twitter tweets on company Stock price*

1. **What is the problem you want to solve?**

As investors we are constantly following the behavior of stock markets. This affects our emotions and motivates us to buy or sell shares. The endeavor here is to understand the effect of social media reactions and emotions on the stock prices. We will be using sentiment analysis to analyse the impact of tweets written on certain stocks on their respective share price.

2. **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis?**

My client is a large hedge fund trying to understand the impact of micro blogging such as tweets from Twitter on the stock prices.
My client would use the outcome of this analysis in modeling their stock forecasts by including twitter feeds in their modeling.

3. **What data are you using? How will you acquire the data?**

This dataset I am using is a part of the paper published in 2020 IEEE International Conference on Big Data Special Session and is available in Kaggle. This dataset contains tweets that are written about Amazon, Apple, Google, Microsoft, and Tesla by using their appropriate share tickers. This dataset contains over 3 million unique tweets with their information such as tweet id, author of the tweet, post date, the text body of the tweet, and the number of comments, likes, and retweets of tweets matched with the related company.

Data source:
https://www.kaggle.com/omermetinn/tweets-about-the-top-companies-from-2015-to-2020

4. **Briefly outline how you'll solve this problem? Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

There are two aspects I would like to explore.

1. Is there any correlation between the volume of tweets with the volume of stock traded the next day.

Here a linear relationship between the variables is not assumed, although a monotonic relationship is possible.

We will start with the null hypothesis that two sets of data are uncorrelated.

We will calculate correlation coefficient value. A value between -1 and 1 that represents the limits of correlation from a full negative correlation to a full positive correlation. A value of 0 means no correlation. The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation.

We will also calculate the p-value for our hypothesis test.

2. Also, we would like to go beyond calculating the correlation between the tweet volume and stock price, by looking into how the sentiment of the tweets impact the stock price.We need to find a way to categorize the tweet sentiments as either positive or negative. We will build models to predict stock price based on these tweets.

The hypothesis here will be that the sentiment of the tweet has no correlation with the share price of the company.

5. **What are your deliverables?**

My deliverables will include:

- A GitHub repo containing the complete work for each step of the project.
- A slide deck
- A report