Capstone Project 3: Project Proposal

**DataSet** - Sentiment Analysis - Sentiment140 dataset with 1.6 million tweets

([click here](https://www.kaggle.com/kazanova/sentiment140))

**About the Data:** This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api. The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment.

### **Content**

It contains the following 6 fields:

1. **target**: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
2. **ids**: The id of the tweet (2087)
3. **date**: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. **flag**: The query (lyx). If there is no query, then this value is NO\_QUERY.
5. **user**: the user that tweeted (robotickilldozr)
6. **text**: the text of the tweet (Lyx is cool)

**Where will it be used?**

Through the use of sentiment analysis and AI, businesses are able to learn the posts that get higher positive engagements. On the other side, sentiment analysis highlights the social media marketing posts that attract negative responses from customers

**Problem Statement:**

Classification of people’s opinion or expressions into different sentiments. Sentiments include *Positive, Neutral*, and *Negative*, *Review Ratings* and *Happy, Sad*. Sentiment Analysis can be done on different consumer centered industries to analyze people's opinion on a particular product or subject.

**Approach:**

* + We will preprocess the tweets to remove unwanted characters
  + Use approaches such as tokenization, padding and word embedding
  + Build neural Network layers and predict the labels