Progress Report

1) Which tasks have been completed?

We have gathered all the data required for our machine learning dataset. This involved building a bot which could scrape the top 50 comments, number of comments, and number of likes from 10 instagram celebrities' last 20 posts. We then took the top 50 comments and wrote a script that passes them through NLTK's sentiment analyzer, taking the average sentiment scores for each post and appending that to our dataset csv. We then dropped the comments themselves from the csv, because they won't mean anything to the ML algorithm. We also have split the dataset into thirds (67 entries each) to begin ranking posts as liked or disliked/mixed/neutral by hand.

2) Which tasks are pending?

All that is left is to rank the dataset and the main goal/outcome of our project, which is to pick a machine learning algorithm that fits our purposes best, and feed it the dataset we rank so it can hopefully predict the reception of an instagram post based off the four input variables (# of followers, # of likes, # of comments, average sentiment score).

3) Are you facing any challenges?

There are no challenges which we are **currently** facing. This project has not been a breeze however, specifically building the instagram bot forced us to perform a lot of research regarding the instagram API, as well as learning and using unofficial API's (which by extension required us to research topics like proxy networks, data privacy, and even terms of service contract legality & consequences).

One concern, which might become a problem later, is how the NLTK VADER sentiment analyzer handles things like emojis and gifs. Currently, it just reports them as truly neutral. We have decided to move ahead, because all it does is bring the sentiment analysis average score closer to 0, but won't outright flip the score from positive to negative or vice versa. We are, however, already planning for this to become a problem, and are searching for emoji sentiment analysis datasets to train the VADER analyzer to be more accurate with emojis. One possible solution is to use a Naive Bayes Classifier for the sentiment analysis as either positive or negative.