# Capstone 2 - Project Proposal

## Binary Sentiment Classification with IMDB review dataset

## **Introduction**

Binary sentiment classification is analyzing customers opinion such as online reviews or survey responses as positive or negative. It helps to understand how well a product/service is doing in the market and owners can take swift action based on the analysis. In this project we will analyze movie reviews from IMDB and classify them into positive or negative.

## **Dataset**

[Large Movie Review Dataset](http://ai.stanford.edu/~amaas/data/sentiment/) by Stanford provides a set pre classified movie reviews for training and testing. Both training and testing dataset has 25000 reviews each with equally distributed positive and negative reviews. A negative review has a score <= 4 out of 10, and a positive review has a score >= 7 out of 10. Reviews with more neutral ratings are not included in the train/test sets.

## **Approach**

## We will apply the several machine learning – supervised learning approach to do this and compare the model accuracy.

1. Text Preprocessing
2. Train with Random Forest classifier
3. Train with SVM classifier
4. Train with Naïve Bayes classifier

We will also use different NLP packages to understand the ease of use of these packages.

## **Deliverables**

* Code
* Report
* Slide Deck

## **Citations**

* [Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher - Learning Word Vectors for Sentiment Analysis](https://www.aclweb.org/anthology/P11-1015/)