Project Proposal

Team Members: Adarsh Balakrishnan, Daniel Gilroy, Philip Iannucci, Brandon Vidro

The objective of this collaborative project is to provide an analysis tool to our stakeholders that will allow them to gain insight into the overall reviews of applications. It will also allow them to predict the rating of an application based on textual reviews with a high level of accuracy. For example, they may use the rated reviews to determine how much they would pay for an application or which ones to publish their advertisements on or invest in.

Problem Statement: The team is provided with data consisting of application reviews from Amazon which contain the sentiments of what the reviewers thought about the products reviewed. The team is also presented with a multi-class neural network model which predicts the output reviews. The model was trained on 1555 records and tested on 667 records. After pre-processing and running some feature extraction operations, the performance of the model was obtained and proved to be bad with an overall accuracy of 60%. This was because the model failed to predict the right classes in the testing phase and seemed to predict only one class. The goal of the team is to propose a new model and train the model to improve its performance and accuracy.

As a user, I would like to take a set of data involving textual reviews for applications and predict the corresponding ratings, on a unit scale of 0-1, with at least an 80% accuracy rating.

Method:
The dataset was first reviewed, and it was found that it contained textual reviews grouped together by application identifier, date reviewed, and review rating (scale 0-1). After reviewing the initial neural network model, the model will be re-designed to predict reviews on textual classification and review rating so that a newer version of the model has better accuracy as compared to the previous version.

The methods used will involve the following:

1) Assumed formatted dataset containing textual reviews, grouped by application identifier, date reviewed and review rating (scale 0-1)
2) Perform an analysis as follows:
    a) Clean the input data
    b) Perform some form of textual classification to be determined
    c) Identify one or more suitable models to predict reviews based on textual classification and review rating
    d) Optimize any hyperparameters via cross validation of trained model(s)
3) Present a final analysis tool that takes the data in 2.a and produces a rating of application.

The model will be developed/re-designed using Microsoft Azure ML Studio.

The progress of the project will be managed using Trello.

Timeline: The span of work will cover roughly 10 weeks and the work will be broken into weekly sprints. The sprints will be partitioned into program increments as outlined below, after each program

1) Program increments 1: 1 sprint
    a) Identify project proposal and key requirements (Sept. 7-13)
2) Program increment 2: 4 sprints (Sept.14 - Oct.18)
    a) Sprint 1:
        i) Identify data cleaning and extraction of all columns
        ii) Identify textual classification techniques
        iii) Identify possible models for predicting rating on aforementioned classifications
        iv) Define model validation requirements
        v) Define base model training requirements
        vi) Define hyperparameter training requirements
        vii) Define final summary of model evaluation
    b) Sprint 2:
        i) Test proposed textual classification techniques
        ii) Test possible models for predicting rating on aforementioned classifications
        iii) Alter workflow for training models
        iv) Alter workflow for hyperparameter training requirements
        v) Alter workflow for final summary of model evaluation
    c) Sprint 3:
        i) Create final architecture diagrams and other ancillary documentation for final implementation plan.
    d) Sprint4:
        i) Buffer for overrun of preceding sprints
3) Program increment 3: 2 Sprints (Oct. 2 – Nov.8)
    a) Sprint 1:
        i) Construct first pass at final product
    b) Sprint 2:
        i) Revisit/review Sprint 1 results for optimizations
4) Program increment 4: 1 Sprint (Nov. 9-23)
    a) Prep for final presentation
        i) Retrospective of entire process
            (1) Identify what worked
            (2) Identify what didn’t work
            (3) Identify what could have been done better

Program Increment	Sprints	Goal	Timeline
1	1	Identify project proposal and key requirements	Sept. 7-13
2	1	Identify data cleaning and extraction of all columns	Sept. 14 - Oct. 18
		Identify textual classification techniques	
		Identify possible models for predicting rating on aforementioned classifications	
		Define model validation requirements	
		Define base model training requirements	
		Define hyperparameter training requirements	
		Define final summary of model evaluation	
	2	Test proposed textual classification techniques	
		Test possible models for predicting rating on aforementioned classifications	
		Alter workflow for training models	
		Alter workflow for hyperparameter training requirements	

		Alter workflow for final summary of model evaluation	
	3	Create final architecture diagrams and other ancillary documentation for final implementation plan.	
	4	Revisit/review Sprint 1 results for optimizations	
3	1	Construct first pass at final product	Oct. 12 – Nov. 8
	2	Revisit/review Sprint 1 results for optimizations	
4	1	Retrospective of entire process for final presentation	Nov. 9 - 23