AMOD 5610 Project Prototype

# Team Member(s) and Project

|  |
| --- |
| **List your team member(s) and a one sentence description of your project (this is just to help us remember which is which, not any sort of thing we need to evaluate).**  **Team Members: Aiswarya Nellayi Venkitaraman (0667620)**  **Rakhi Konar (0670029)**  **Social Media Analysis and Netflix Show Recommendation System :-** The main idea of our project is to analyse the social media platform ‘Twitter’ and collect tweets related to Netflix shows, build a Netflix show recommendation system based on twitter user interests. |

Prototype

|  |
| --- |
| Below are the end-to-end functional elements for our project:   1. Reading data from Mongo DB 2. Pre-processing in pyspark: this includes  * cleaning the tweets * removing duplicates * extracting Netflix show names from tweets using regex * matching extracted show names with actual show names (from Netflix show names csv file (Kaggle dataset)) * Fuzzy string matching - Similarity check of strings using ‘levenshtein’ approach * Web scraping of IMDB ratings for each title. * Consolidating final data for sentiment analysis.  1. Sentiment Analysis of tweets for each user – Vader sentiment Analysis 2. Exploratory data analysis and pre-processing of data for model building: this includes  * Label Encoding * Test Train split * Normalization and Scaling * KBindiscretization  1. Model creation – Keras Sequence model -> Single Layer Fully connected model and Multilayer perceptron model 2. Model fitting and model evaluation 3. Prediction of sentiment scores 4. Recommendation based on predicted scores 5. Model testing 6. Results |
|  |

# Installation

Explain how to install/setup/run your project

|  |
| --- |
| Please refer the attached document. |

Testing of Functionality

|  |
| --- |
| We created two models a Single layer fully connected model and MLP model.  Initially the models were overfitting as testing losses were higher than training loss. We were able to control the overfitting by adding dropout of 40% to the models and by discretizing the target variable. Regularization methods were also applied.  After controlling the over fitting of the model,    Single layer fully connected model  MSE = 0.88  Accuracy = 0.62  MLP model  MSE = 0.13  Accuracy for training data = 0.79  The model predicts sentiment scores with the input data given and top 10 recommendation are made by selecting 10 highest prediction scores.  As part of testing, we have done below offline evaluation measures   1. MSE 2. MAE 3. RMSE     From the above result we can see that our model predicted values are more similar to the actual values. The error rate is very low for 3 metrics. Since our deep learning model is based on a regression problem, these offline tests based on the user’s historical data works well to test the prediction results. However, we have further tested the result using cross validation. Since cross validation process is computationally expensive, we have taken a sample of our model data to test the model response. For each iteration, we have seen that the root mean squared value has decreased. We don’t have access to get particular user real time behaviours and rank our Netflix recommendation system , hence we did offline testing using the above metrics only. |