**DATA 602 FINAL PROJECT PROPOSAL**

**Topic: Movie reviews based on IMDB Ratings:**

Internet Movie Database (IMDB) is the online portal that gives ratings for all the movies irrespective of the language based on the user reviews posted. These days number of movies upstreaming are increasing, so it has been quite difficult task for the viewer to decide which film to watch. The viewer may decide based on the rating that the IMDb provide for the movie. So, we decided to run analysis on the model to check whether the ratings were accurate or not. That is the main reason which had driven us to choose this topic as our final project.

**Problem that we will be dealing with:**

* Developing interesting visualization to interactively explore this dataset.
* Understanding and modeling of elements within a variable context.
* Handling Ambiguity of reviews mentioned.
* Semantic and context understanding is essential for reviews mentioned.

**Reason to choose this project:**

Since most of the viewers who watch the movies rely on the rating, so it is required to build a model to predict the user inputs and give the final rating for the movie. Our main aim of the project is to perform various classification techniques using the models mentioned below and figure out the best model among them.

**Dataset:**

The dataset that we are considering for the project is taken from IEEE-dataport and that contains various reviews collected from the viewers for more than 1000 movies across different genres. It contains two files where one file deals with the genre of the film and on the other hand the other file deal with the reviews from the users.

Following is the URL from where we are retrieving the data for our project.

* <https://ieee-dataport.org/open-access/imdb-movie-reviews-dataset#files>
* <https://ieee-dataport.org/open-access/imdb-users-ratings-dataset>

**Target Variable:** Ratings based on weightage of reviews from multiple users.

**Input Variable:** Reviews, title, date of review post.

**Approach that we follow for the completion of the project:**

* Loading the data set.
* Check for missing values, unidentified characters (emoji ,)
* **Data Cleaning**:

1. Analyze punctuation, emoji, stop words and remove them accordingly
2. Remove hyperlinks and numbers
3. Remove HTML tags, Email ids using Regex
4. Correct misspelled words
5. Look for words in other languages

* **Visualizing Input data**:

1. Word Count stats (distribution, count) of reviews
2. Word Frequency
3. Word Count stats (distribution, count) of reviews
4. N-gram Analysis (Bigram, Trigram)
5. Sentence Length distribution
6. Word Cloud Visualization
7. Reviews embedding clustering

* **Convert text data to its equivalent numeric vector embeddings options**

1. Glove Library
2. Word2Vec

For better quality of word embeddings, using pretrained models from HuggingFace transformer library will be beneficial

* Considering the sample data to run classifiers.
* Implementing Standard Scalar ()
* Following are the models that we are considering running over the sample data that we generate.

Models:

SVM (from Supervised Learning model), Random Forest Classifier (from Ensemble methods), KNeighbors Classifier (from Supervised Learning model), XgBoost Classifier (from Ensemble methods) and Neural Networks.

* Generate the classification report for the results.
* Identifying the best model and running that model over the entire data to check for the accuracy for rating the movie based on the user reviews.
* Running various Regression tests (OLS Linear Regression) among the user reviews and the IMDb rating data and run residual analysis, unit test to generate the results.

Note: Once after we start working on the project, based on the data and requirement of the project we may make required changes if required like adding some other classifiers and running over the data.

**Members of the Project:**

Sai Teja Avadhootha - YT41135

Sushanthik Reddy Poreddy - TY99095

Swapan Gupta Chollati - RJ29255