James Yang

CSE 573

Project Proposal

Project Topic: Unsupervised Sentiment Analysis of Netflix Description

Motivation: Netflix has show descriptions for all of their movies and TV-shows on their platform. I'm trying to find a correlation between the sentiment found in these descriptions and the TV rating it has been labeled as. I am curious to see if descriptions are written in correlation with the audience in which they are being directed towards. For example, are TV-G shows written in a more sentimentally positive manner than rated R shows?

Definition: Sentiment analysis is an NLP (natural language processing) technique used to determine whether data is considered positive, negative, or neutral. Many libraries such as PyTorch have the ability of analyzing sentiment and generating a value on a scale of positive-negative sentiment. For example, "I hate Netflix" would generate a very negative value for sentiment because the word "hate" has a very negative sentiment value while "I" and "Netflix" remain more neutral.

Dataset: I am pulling from a Kaggle dataset generated from Shivam Bansal (https://www.kaggle.com/datasets/shivamb/netflix-shows?datasetId=434238&sortBy=voteCount) The dataset consists of a show_id, type, title, director, cast, country, date added, release year, rating, duration, listed in, and description. I intend on finding whether there is correlation between the rating and its description to find whether they are sentimentally correlated with their ratings.

Resources: There are many examples on using sentiment analysis. An article by Rafal Wojcik gives examples of extracting sentiment analysis from data without labels (https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483). I also intend on using Bhadresh Savani's sentiment analysis tutorial with PyTorch (https://bhadreshpsavani.medium.com/tutorial-on-sentimental-analysis-using-pytorch-b1431306a2d7).

Milestones/Timeline: Data cleaned (5/11), analyze sentimental analysis (5/22), begin working on report and find correlation significance (6/1), generate presentation (6/8), finish report (6/10).

Detailed Experiment Plan:

Data cleaned (5/11): Eliminate insignificant symbols in description text (for example: â€™). Remove misspelled words. Have proper dataframes established for descriptions, sentiments, and TV ratings.

Analyze sentimental analysis (5/22): Run Pytorch sentiment analysis, and clean any error values and misrepresented sentiments.

Correlation Significance/Report (6/1): Begin analyzing correlation values between tv-ratings and sentiment. Determine whether there is relation between the two and create a report about it.

Generate Presentation/Finish Report (6/8-6/10): Generate slides and finish report for project.