Emmys Predictions Case Study Rubric

Due: TBD
Submission format: Upload a pdf and a GitHub repository link to Canvas

General Description: Investigate how much of an effect reviews have on the nominees for the Emmy awards by utilizing sentiment analysis to see if the Emmy-nominated drama series that achieves the highest public sentiment score will end up winning the Emmys.

Why am I doing this? The goal of this assignment is for you to develop and demonstrate foundational text analysis and NLP skills relating to text data in order become familiar with how to aid in real-world tasks.

- Course Learning Objective: learn the basics of NLP by conducting text analysis in the context of a real-world situation.

What am I going to do? You will research how to use Python's VADER package on text data, and then create a model that will use VADER to predict the winner of the next Emmys best drama.

Tips for success:

- Make sure to spend enough time researching topics such as text analysis, NLP, and VADER to help you know where to begin
- If you feel stuck, don't worry this is normal; Ask your data science professor/TA for help anytime
- Have a clear outline that discusses what you will work on for a given session; A good tip is to have an original dataset that is untouched in case you need to go back and start from scratch

How will I know I have Succeeded? You will have succeeded on this case study when you follow and complete the criteria outlined in the rubric below:

Spec Category	Spec Details
Formatting	Submit each component listed in the rest of this rubric as advised below. • Submit a link to the repo ○ Everything is contained in the repo or linked to it if appropriate.
Written Portion	Discuss the process of creating your model, include any challenges or setbacks you encountered along the way. **Data frame Establishment Details:** • Summarize the data you use for the project • Present data dictionary • State all questions you explored and answered about the data **Analysis Plan:** • Use a paragraph to describe each step in the plan. You could include the following things:

	○ Preprocessing: How missing values will be handled, any feature engineering to perform, etc. ○ Methodology: ▪ If a statistical analysis, specify the significance level, whether this is a one or two sided analysis, the type of test to use, covariates to include, etc. ▪ If a predictive modeling analysis, specify data splitting, algorithm selection, cross-validation procedure, etc. ▪ Evaluation: metrics (e.g., Mean Absolute Error, R-squared) and the criteria to evaluate your analysis (e.g. $p < 0.05$; R-squared > 0). ○ Include a specific quantifiable goal that you can use as a finish line for your analysis. (Essentially this is your goalpost, it helps you to know when you have achieved your goal and can switch to preparing your presentation. **Executive Summary:** • Summarize the process of your project • Describe an overview of the packages used, as well as websites you scraped reviews from
DATA	Contains all data needed for replicating your work **Initial Data:** - Original data used **Final Data:** - Final data used after data cleaning and preprocessing
CODE	Contains all code needed to replicate your analysis