



Aaran

## Project 3 Instructor Feedback

### Problem Statement: 2.5

**Grows:** The overall goal and bar for success is clear. The scope of the project is also specific.

**Grows:** Explain why the goal was 90% accuracy. You might include more specific technical information in the problem statement: What techniques will be implemented and evaluated (what modeling algorithms, text processing, etc.).

### Data Collection: 2

**Grows:** A good amount of data was collected over several days from several different reddit streams. Good use of a function for processing data into a dataframe.

**Grows:** More optimization could be done to streamline rerunning the api calls and combining them with existing results using more custom functions.

### Data Cleaning and EDA: 2.5

**Grows:** Duplicates and missing values are handled. A good amount of attention given to cleaning text in individual posts. Unigrams, bigrams, and trigrams investigated. Nice summary!

**Grows:** Defend your choice of combining title and selftext in the context of your problem statement. Comparison of ngrams would be nice as side by side bar charts for comparing differences between subreddits easily.

### Preprocessing and Modeling: 2.5

**Grows:** Methodical process for modeling and tuning. Pipelines, gridsearch, and custom functions streamline process well. Good summaries of results.

**Grows:** Avoid switching metric used in gridsearch as it makes those models hard to compare to the others. Include the metrics in your discussion of the model performance in the summary to make it a little easier to follow.

## Evaluation and Conceptual Understanding: 3

**Glows:** Baseline is made clear. Good use of metrics. Coefficients are used for inference. Nice work!

**Grows:**

## Conclusion and Recommendations: 3

**Glows:** Specific, clear, following from problem statement and work done. Vocabulary differences between the subreddits based on modeling discussed in good detail.

**Grows:**

## Project Organization: 2.5

**Glows:** Good repo organization. File names make it easy to navigate. Great use of markdown in the notebooks. Great, detailed readme.

**Grows:** Use .gitignore to exclude pycache when .py files are included in the project. It is a bit subjective, but consider moving code to a code folder.

## Visualizations: 2.5

**Glows:** Overall visuals are nicely formatted. Good use of visuals in EDA to explore relationships. Good use of visuals in model evaluation to demonstrate results. Nice interpretations.

**Grows:** Make sure all visuals that make it into your final report have titles and labels.

## Python Syntax and Control Flow: 2.5

**Glows:** Code is clear and error free. Nice use of a custom module.

**Grows:** Add all your helper functions to your custom module and include docstrings to make them really polished.

## Presentation: 3

**Notes:** Clean presentation, clear intro with problem statement and subreddit reasoning stated. Why was your goal 90% accuracy? Good review of cleaning/preprocessing and detailing differences in subsets with key metrics and visualizations. Good exploration of additional stop word development and deep dives using domain knowledge.

---

**Total: 26**

*Avg: 2.6*