Sentiment Polarization in Celebrity vs. Non-Celebrity Makeup Reviews Case Study

DS 4002 – Spring 2024 – Grayson Fiveash

Submission format: Upload link to GitHub repository to Canvas

Why am I doing this? This case study is your opportunity to leverage your knowledge you've gained throughout your data science journey at UVA, specifically in conducting sentiment analysis and logistic regression. It is offering you a hands-on way of applying data science to real-world situations—such as product marketing. The deliverable you will create is a GitHub repository and a PDF document that covers your results and conclusions/learnings.

Preparatory Assignments – Read attached resources and preparatory materials.

What am I going to do? You will access the GitHub repository for the Sephora Review Sentiment Case Study (https://github.com/gffiveash/DS4002_CS3/tree/main). From there, you have the opportunity replicate the project your peers created and refined for you—specifically, in determining whether celebrity branded makeup products have more polarized reviews in order to inform your product marketing strategy. This will first involve cleaning and merging the product data, followed by conducting VADER sentiment analysis and building a logistic regression model. Deliverables include:

- GitHub repository – containing all data, figures, and scripts
- A PDF document detailing process, results, and conclusion/learnings

Tips for success:

- Familiarize yourself with Python and the VADER Sentiment Analysis article in the Supporting Materials folder.
- Familiarize yourself with the distribution and makeup of the final dataset, especially noting celebrity vs. non-celebrity makeup brands.
- Talk to your fellow students and instructors in order to seek guidance.

How will I know I have Succeeded? You will meet expectations on this Case Study when you follow the criteria in the rubric below.

Spec Category	Spec Details
Formatting	● Repository – A GitHub repo containing all materials ○ Create a new GitHub repository for this assignment titled 'CS2_SephoraReviewSentiment' that contains: ■ README.md ■ LICENSE/md

	▪ A DATA folder ▪ A SCRIPTS folder ▪ An OUTPUT folder ● Your PDF document should detail the process, results, and conclusions
Repository	● Goal: This is a GitHub repository containing all digital materials for the case study. From this repo, individuals should be able to reproduce and follow every step of your work, from data acquisition to results. ● The README should orient the user to the repo, link to every important folder in the repo, and offer a brief summary of what you've produced. ● Use folders to organize the major categories of materials, code, etc.
Data	● Goal: This folder contains all 18 cleaned datasets, as well as the "combined_clean_data.csv" for analysis. ● The data should be in .csv format.
Scripts	● Goal: This folder should include all the source code for the case study. ● All script files should include detailed header comments at the beginning of a script to introduce the materials, as well as comments describing the commands or sequence of commands themselves. ● Name each script accordingly, so anyone can reproduce the results. ● The scripts should include: ○ 1_Combination_VADER_Analysis ○ 2_EDA_Plots ○ 3_Logistic_Regresion ● These should be in .ipynb format.
Output	● Goal: This folder contains all of the output generated by your project, e.g. figures, tables, etc. ● Use informative names for your files. ● These should be in .png or .jpg format.
PDF Document	● Goal: Discuss your process reproducing this case study, specifically noting how you met the deliverables. ● This 1-page PDF should, in addition to detailing the process, should outline your results and conclusions. ● Reflect on any challenges in completing this case study and how you overcome them. Did you change the code? Troubleshoot?