Classifying Pet Product Reviews Rubric

DS4002 - Fall 2025 - Haley Mitchell
Due: December 8th
Submission Format: Upload link to GitHub repository on UVA Canvas

Why am I doing this?
This case study will give you hands-on experience using text data to solve a real-world problem. This is an opportunity to use the skills that you have accumulated from your classes within the Data Science minor.

What am I going to do?
The GitHub repository that outlines this project can be found using this link: https://github.com/halemitch/CS3-Project/tree/main.

The following steps are as follows:

- Read supplemental documents on background information and the model
- Follow the directions within the Data Folder in the GitHub to download the data
- Perform the EDA
- Clean the data
- Create a logistic regression model
- Use TF-IDF technique
- Collect performance statistics

Tips for success

- The data may take a long time to download (because it is large), so be patient with it
- As the data file is large, it is easily corruptible. We provide two methods to read the data into the Colab file
    - 1: You can run the first cell until it goes through
        - This typically takes about 2-5 times (about 2-3 minutes)
        - You should end up with 2,097,208 rows
    - 2: You can uncomment out the second cell of code and run that
        - This will just give you a subset of data up until it runs into the error

How will I know I succeeded?
You will meet the expectations for this case study when you have completed all of the criteria in the rubric below

Spec Criteria	Spec Detail

Formatting	• Create a new GitHub Repository (submitted via link on Canvas)
	• This repository should contain
	  ○ A README.md file (which auto displays)
	  ○ A LICENSE.md file (use MIT as default)
	  ○ A SCRIPTS folder
	  ○ A DATA folder
	  ○ AN OUTPUT folder

README.md	• The goal is for someone to grasp your project and understand where to go to find

	more information to be able to replicate it ○ Include a brief summary of the project ○ Include a map (tree) of your documentation
Scripts folder	● This folder will contain all of the coding for this project ○ You need to code to… ■ Read in your data ■ Clean the data ■ Perform an EDA ■ Create a logistic regression model ■ Use TF-IDF Techniques ■ Evaluate the model
Data Folder	● This folder will contain the information on how to access your data ○ As the data folder is very large, you will be unable to put it into GitHub ○ Use this folder to direct people to the website, so they can download it themselves ○ Include any further information that is needed/would be useful for someone trying to replicate this project
Output folder	● Create a folder that highlights your findings ○ Must contain a confusion matrix and accuracy ○ Must contain the top predictors of good and bad reviews found by using TF-IDF ○ Include any additional graphs you find that would be interesting to the viewer
References	● All references should be listed at the end of the document ● Use IEEE Documentation style (link)

Acknowledgements: Special thanks to Jess Taggart from UVA CTE for coaching on making this rubric and to Professor Alonzi for letting us use it. This structure is pulled from Streifer & Palmer (2020).