Theodore Koby-Hercsky

01/11/2023

DSC680-T301 2233-1

Professor Catie Williams

Milestone 1 – Proposal

Github: Portfolio_TheodoreKoby_Hercsky/Sentiment Analysis Car Brand Reviews at main · TheodoreKoby-Hercsky/Portfolio_TheodoreKoby_Hercsky (github.com)

Topic - Describe and name your project in 1-2 sentences max

Choosing a car can be a big decision using sentiment analysis users can analyze mass amounts of reviews by determining positive and negative feedback. Creating a hassle-free way to narrow down a car company they would like to pursue.

Business Problem - Describe the business problem your project is trying to solve and/or the research questions you will explore

The business problem I plan on exploring is regarding car brands and the reviews they receive from customers. I will create a report that takes reviews from customers and conducts sentiment analysis to determine positive and negative reviews. Allowing users to determine which car brands and years of makes have the best and worst reviews. By further evaluating these car brands and the reviews received customers can make better choices when determining what brand to go with for their next car. Also, car companies can use this report as well to determine what is working for their company and what they need to improve on.

Datasets - where are you getting your data? Describe the data that you will use to solve the problem

Data set: Reviews of 5 Car Brands | Kaggle

The data set that I will be utilizing contains review data for five different car brands Audi, Lexus, Infiniti, BMW, and Mercedes-Benz. The data set includes five variables that can be seen below with a brief description of each.

- Rating - A numeric variable that gives the car brand a rating of one to five
- Car_Year – The year the car was built
- Brand_Name – Is the brand such as Audi, Lexus, Infiniti, BMW, or Mercedes-Benz
- Date – Is the date the owner of the car created the review
- Review – The review given by the costumer regard the car they purchased.

Methods - What analysis methods will you use to complete this project? Note: this is just a proposal, your project can adapt as you work on it

The project will include Exploratory Data Analysis to get a better understanding of what the data set is comprised of. Exploratory Data Analysis "is a technique to analyze data with visual techniques and all statistical results." (Gupta) I will also create visualizations on different variables within the data set to determine what the range of years the cars where made, percentage of ratings each car brand received, and the overall number of reviews received for

each brand. This analysis will help me get started and allow me to work towards natural language processing and sentiment analysis that I will conduct on the reviews for each brand.

Natural language processing is the first step an analyst will need to take before conducting any sentiment analysis. The steps needed to conduct a natural language processing pipeline include tokenizing sentences, removing stop words, normalizing words, and vectorizing text. The process of tokenizing a sentence is the breakdown of text into sentences, words, or even other units. While stop words that will be removed consist of "if", "but", "or", and so on. The process of normalizing words consists of condensing all forms of a word into a single form. Vectorizing text which is the process of "turning the text into a numerical representation for consumption by your classifier." (Stratis) Last, I will use VADER sentiment analysis which is a "lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media." (Ankthon) Which will be used to create polarity and rating that will produce a in depth score rating and a finer scale indicating positive, negative, or neutral for the Vader polarity.

The project will also require me to split my data in to test and training sets to create models such as Random Forest classifier and other models I decide to use. The random forest classifier is a "supervised machine learning algorithm used for classification, regression and other tasks using decision trees." (Amanda) This classifier is used to create a prediction by collecting votes from different decision trees from a set of randomly selected subset of training sets. The overall goal for this project is to determine the number of positive and negative reviews and predict the overall rating of each company. Which give customers and workers a better understanding of which car brand to go with and what the companies need to do to improve.

Ethical Considerations - What are some potential ethical concerns of this topic or analyzing the data?

An ethical consideration this data has is anonymity which means that a user does not who the participants and prevents analysist or viewers from linking an individual to the data provided. The data set that is being used for this project focuses on reviews for car brands but does not include any person information such as name, phone number, email address, or the IP address of the reviewer. Being that no person information was provided I do not see any ethical concerns with confidentiality or potential for harm.

Challenges/Issues - What are some issues and challenges do you think you might face?

A challenge I believe I might face is missing data which will have to be delt with in the start of the project using EDA to correctly locate and fix any missing values within each variable. While another challenge I might face would be during natural language processing as this step requires analysts to break down reviews to be able to conduct sentiment analysis on reviews the car brands received.

References - What sources will you use to validate your results and support your project topic?

- Exploratory data analysis article can be used for different tips and tricks for the EDA I will conduct in the beginning of the project.
    - https://www.geeksforgeeks.org/exploratory-data-analysis-in-python-set-1/

- Sentiment analysis will be used to validate the approach taken to tokenize the reviews, removing stop words, normalizing words, and vectorizing the text. While also splitting the data into test and training sets to create models.
    - https://realpython.com/sentiment-analysis-python/
- Next The article on Sentiment Analysis using VADER will help when creating a rating and polarity such as positive or negative for each review.
    - Python | Sentiment Analysis using VADER - GeeksforGeeks
- The Random Forest Classifier article from Geeks for Geeks will be used to understand how to create a model for the car model reviews.
    - Random Forest Classifier using Scikit-learn - GeeksforGeeks

Citation

Gupta, Mohit. "Exploratory Data Analysis in Python: Set 1." *GeeksforGeeks*, 21 Jan. 2019, https://www.geeksforgeeks.org/exploratory-data-analysis-in-python-set-1/.

Stratis, Kyle. "Use Sentiment Analysis with Python to Classify Movie Reviews." *Real Python*, Real Python, 10 Nov. 2022, https://realpython.com/sentiment-analysis-python/.

Ankthon. "Python: Sentiment Analysis Using Vader." *GeeksforGeeks*, 7 Oct. 2021, https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/.

Amanda. "Random Forest Classifier Using Scikit-Learn." *GeeksforGeeks*, 13 Dec. 2022, https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/.