DSC 680 Project 2 – Milestone 2

Final White Paper

Professor Catherine Williams

Presented by:

Sherry Kosmicki

May 7, 2023

Business Problem

Marketing efforts include enthusiastic individuals looking for that niche that can set them apart and give them just the energy they need in their respective market. If we can evaluate the current customer review data for a product and narrow it down to what is frequently said by the customer, it may be possible to design a campaign to build words that are commonly associated with positive customer experiences.

Background/History

Many of the review mechanisms used today to measure a products worth is based on a grading system. In addition, a free text can be added to enhance the graded response. While these scale values can determine a level of satisfaction, they also come with some inconsistency related to how one individual scores a product verses the next person.

Data Explanation (Data Prep/Data Dictionary/etc.)

The data is derived from Amazon reviews. This data has tens of thousands of review responses to products purchased. Additionally, a corresponding metadata record provides categorization of the reviews and simple information about the product. The review data is large and maintained in json zip files.

This data is large. It needs to be evaluated for what is useful, what categories are needed by the user, and where it can be stored and processed. Decisions need to be made by those closest to the project to determine what areas will be evaluated and to what extent it will be evaluated.

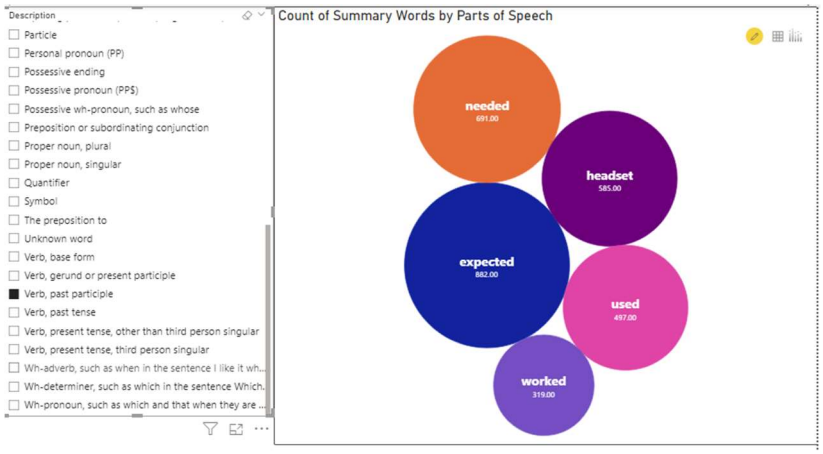
Methods

There are different ways to manage large amounts of data. When working in Python it's important to remember that you can load the data into a variable and will need to maintain the connection to the IDE to maintain the data loaded. Another option that was tested was to load the data into a SQLlite database. This allowed for easy storage and access to the data. With a background in SQL, it made it easier to query the data.

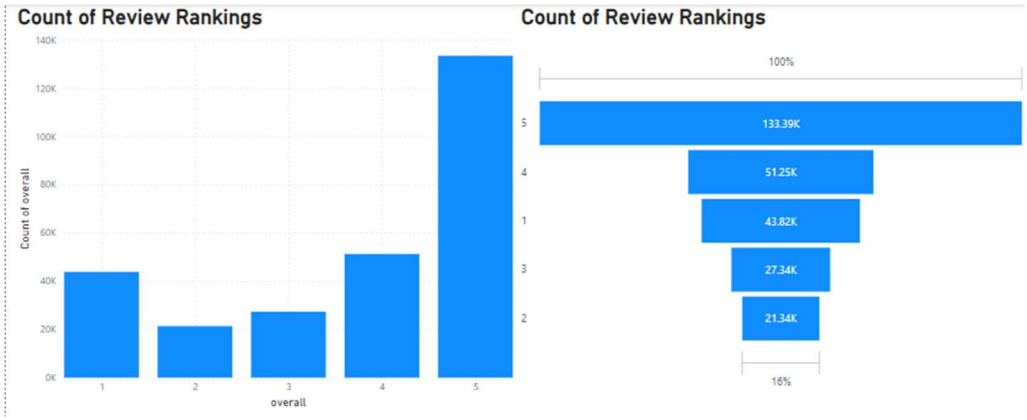
The word analysis can be done at many levels. The level of words management you want to do can define the results you want. In most every case it's important to remove the punctuation. The text processing capabilities in Python allow for common words to be removed. It also helps to look at the frequency of words and either remove those words that have the most frequency and those that have the least frequency.

By quantifying the data at the parts of speech level, it helps categorize the words to be evaluated by the end user.

I was able to use bubble charts to look at the data in a different way by parts of speech to look at the volumes.



I created some dashboard looks to include a visualization of the overall score of the reviews.



Analysis

After revamping the cleaning of the text data and applying the text processing functions, I was able to make better use of the data when creating visualizations. Interactive visualizations using the slicer option for each of the parts of speech allows a user to look at the words at each of those levels.

Conclusion

When evaluating this type of data, it's important to understand the data itself. The amount of data munging and cross reference is critical to make sure you are presenting valuable data to the end user. Once the interactive evaluation model and visualizations are built, it can lay on top of other Amazon review data.

Assumptions

I had no assumptions about the size of the data and the capacity I had to work with on my laptop. I assumed that PowerBI desktop would work with the larger data frame that did fit on my computer. PowerBI struggles with processing the text-to-words data manipulation using Python.

Limitations

The capacity of my individual laptop kept me from being able to work with the entire dataset of reviews. I could have removed several hours of data management if I had a server or larger capacity drive to work from.

Challenges

The biggest challenge with this project is the data itself. I spent a large amount of time trying to break down the data so I could work with it on a small scale. Once I had success on a small scale, the idea was to apply the methods to the larger scale of data.

Integrating the use of Python and PowerBI was more difficult than what was portrayed in the original method. The PowerBI desktop refers to the installed version of python, which sounds good on the surface. After much effort, I determined that I needed to run PowerBI executable through the IDE used for Python to get a much more seamless functionality from the product.

Python errors in PowerBI were hard to manage. It was easier to munge the data in Python instead of trying to manage in the desktop product. Small scale data functions could be done through Python but so far, the experience on a large scale is not proving to be functional within Power BI.

Future Uses/Additional Applications

Once tested, this application could be used with any review data that is formatted to fit within the dataset. Once it's refreshed, the dashboard can be dynamic in its evaluation.

Recommendations

To build this interactive dashboard for the end user, storage location will need to be determined. Since the Power BI desktop is unable to manage the Python-passed data, it will be important to evaluate the usefulness of Power BI. The report server version of Power BI could help with the gap usage of the large amount of data.

Python or R also have good visualization functionality that could be used instead of Power BI. The advantage of Power BI or an Excel product makes it easier to empower the users to work with the data.

Implementation Plan

Implementation of this dashboard would be a presentation to marketing personnel, make any needed changes, and build a roll out phase. The group together would determine an ongoing plan for how the data needs to be updated and processed.

Ethical Assessment

Ethical considerations are always changing in a world where we use words. Words can change meaning just by how they are used in society. Take "beautiful soup" for example. I would make the comment that chicken noodle is a "beautiful soup". Python named a component of it's programming language "beautiful soup" which is used to scrape data from a website that is in HTML format. Once words are identified as frequently used, the marketing area then needs to determine the validity of that word group to ensure the best possible use for their campaign.

References:

Oelshlager, M. Dec 2021. *Ethics in Natural Language Processing*. Website: https://dida.do/blog/ethics-in-natural-language-processing

Enterprise DNA., Sep 2022. *How To Do Text Analysis Using Python to Identify Parts of Speech in Texts in Power BI*. Website: https://www.youtube.com/watch?v=UkW5DvK9rqM

Shubham, J. Feb 2018. Ultimate guide to deal with Text Data (using Python) – for Data Scientists and Engineers. Website: https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/

10 Audience Questions

1. What methods were used to identify the words that were most frequently used by customers?
   Once the data was able to be downloaded, a join was created between the review data and general metadata. This allowed the review data to have a related category. Text extraction tools in Python were used to remove punctuation and basic words from the review data. The data was divided by the parts of speech as one way to analyze the data. The bubble chart visualization in Power BI helps to identify the prevalence of a word based on the size of the bubble.
2. How is evaluating the reviews by parts-of-speech beneficial to the campaign?
   Parts of speech evaluation gives the data of "words" a categorization that can be looked and reviewed by the marketing team.
3. What steps have been taken to keep the reviewer information anonymous?
   A review id and name are stripped from the data to keep it from being used in the evaluation.
4. Can this analysis be duplicated for other categories of Amazon reviews?
   Since the amazon data has a metadata field called "main_category". Several of those areas can be evaluated. A separate analysis on the title field could allow for a more granular evaluation of more specific areas. For example, I may want to look at earbuds within the cell phone and accessories category.
5. Are the words identified considered positive or negative?
   Sentiment values were assigned based on the overall 1-5 star review column. I then divided out the reviews at the 4-5 star value to be used in the text evaluation methods.
6. Is there any value in doing analysis of negative data?
   An effort to look at the negative text words can have value in determining areas of opportunity or evaluate if we may be using any of those "words" in the current campaign.
7. If you were to do sentiment analysis, could you then identify the positive verses negative reviews?

Sentiment analysis could be done to help predict the overall evaluation when considering the reviews. A model could be trained based on the existing reviews in a effort to consider good verse bad reviews.

8. How do we continue to evaluate this data on a monthly, quarterly, or yearly time frame?
   The review data presents a very large amount of data to be managed. Questions needs to be answered around how much data and for what data frame can be evaluated to consider it valuable. If certain products being evaluated changed, do we want to consider reviews prior to the product change? Once some data questions can be answered, a solution could be created to be refreshed on an as needed basis.
9. Does the Power BI application support the needs of this evaluation?
   In the initial evaluation of using Power BI along with the use of Python, Power BI was not able to handle the large volume of text data passed to it through Python. Power BI had to be executed through the Python IDE environment to get the appropriate versions of libraries to process the Python code. The Power BI Report server version may be a better candidate than the desktop version.
10. Is there another solution that has better functionality other than Power BI?
    Power BI had a user-friendly front end visualization capability which makes it easier to give to end users for them to be able to work with the data. Python or R both have the visualization capability and are more functional when it comes to handling the larger text processing. Even though Power BI has a script running capability via Python or R, it's more efficient to do the text data processing on the backend and push the data to the front-end Power BI or other product to aggregate and display the data.