Amazon Customer Review Sentiment Analysis

ACRSent

**Data Science Capstone Project   
Data Acquisition and Pre-Processing Report**

Date:

5/6/2021

Team Members:

Name: Kevin Fitzpatrick

Name: Tessa Karlovitz

Name: Dustin Luchmee

[The purpose of this report is to describe the data of your project. It includes three major sections: Data Sources, Data-Processing, and Appendix]

**Identifying Data**

**Data Sources:**

[Identify the data sources of your project. It may have more than one data source. Describe each of them and explain why you select the data sources.]

Our Amazon customer reviews come from tsv files that we retrieved from the cloud. These files are hosted by Amazon themselves to aid academic researchers and the progression of NLP / machine learning. The reviews span from 1995 – 2015 and include reviews from products that were for sale on Amazon’s marketplace. The dataset has one or more tsv files for each product category that a user can choose from. For our purposes, we are choosing to implement our project using reviews from electronics, mobile electronics, and books.

Read Me:

<https://s3.amazonaws.com/amazon-reviews-pds/readme.html>

Index:

<https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt>

**Acquisition Process:**

[Describe the data acquisition process. Is the dataset ready for download? How do you download the data? Do you need to write your own code to acquire the data from a public or private source? Describe how you do it. Are there multiple data sources? How do you integrate the data from multiple sources? Any other process involved in the acquisition process?]

**Mobile Electronics**

For mobile electronics, there is one tsv file with about 105,000 rows. Since the dataset is already available to download, I loaded the tsv using the Python pandas package. The file did error out so I had to use error\_bad\_lines = False so it could skip the bad rows and load everything else. In doing that, only 2 rows got skipped.

**Issues:**

[Are there any potential issues in data acquisition that have not be solved yet?]

**Data-Processing**

[Examine the data you have acquired and understand the data properties. Is there any pre-processing you need to do before you can start analyzing the data? For example, missing data, sparsity, noise, veracity, ambiguity, interoperability, etc. Describe each data issue in a sub-section and explain how you clean up the data.]

**Mobile Electronics Processing:**

Null Values

When investigating the mobile electronics dataset, there were 5 total rows that had at least one null value in them. The columns that had a null value varied but included the star\_rating and review\_body which are important variables to use in future analysis. Using pandas, we easily dropped all rows with any null values since there were so few occurrences.

Verify Dates are Datetime

To perform analysis where we look at reviews over time, we made sure to change the review\_date column to a datetime that pandas can easily read and filter by using pandas built-in to\_datetime function.

Convert Emojis into Text

Having emojis in the review body presented a challenge since we didn’t want to lose the emotion/sentiment behind the emoji, but obviously Python couldn’t read emojis. To solve the problem, we turned to a Python package called “demoji” which converts an emoji to simple text. For example, a smiling face emoji would be translated to a text string such as “smiling face” so we can read into the emotional intent for sentiment purposes.

Clean Up Bad Characters and Dealing with Special Characters

When doing some initial EDA, we discovered some characters that needed to be cleaned. These included extra spaces, blanks, html tags, and punctuation. We wanted the spaces, blanks, and tags removed, but decided to keep the punctuation in case it was needed later. To clean this up, we used BeautifulSoup to get the text from the review body. This had the effect of removing html tags (mostly line breaks) and cleared up any blanks. Finally, regex was used to remove any instances of 2 or more spaces. For how punctuation was handled, see the Tokenize section below.

Expand Contractions

The thought was that expanding contractions would be the route to take when doing further analysis so we installed a Python package called “contractions”. This handy package has a function called fix where you can feed it a text string and it will expand all the contractions it finds.

Lowercase All Text

We want all of the words to be consistent when grouping or counting so using Python’s built-in lower function did the trick here.

Tokenize

Tokenizing text means to split text up into smaller parts, either sentences from a paragraph or words from a sentence. In our case, we wanted a list of words so we went with the Python package NLTK and used the function word\_tokenize to split up the words in the review body. This allowed us to get our list of words, and it also separates out the punctuation in each sentence. We’re not sure if punctuation will be used in our analysis but we wanted to make sure it was there just in case.

Removal of Stop Words

Some text…

Lemmatization

Some text…

**Appendix**

[Provide the code or pseudo code, data definition, sample data, and any other information in the appendix here.]

Data Dictionary:

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| marketplace | string | 2 letter country code of the marketplace where the review was written. |
| customer\_id | string | Random identifier that can be used to aggregate reviews written by a single author. |
| review\_id | string | The unique ID of the review. |
| product\_id | string | The unique Product ID the review pertains to. In the multilingual dataset the reviews |
| product\_parent | string | Random identifier that can be used to aggregate reviews for the same product. |
| product\_title | string | Title of the product. |
| product\_category | string | Broad product category that can be used to group reviews |
| star\_rating | integer | The 1-5 star rating of the review. |
| helpful\_votes | integer | Number of helpful votes. |
| total\_votes | integer | Number of total votes the review received. |
| vine | string | Review was written as part of the Vine program. |
| verified\_purchase | string | The review is on a verified purchase. |
| review\_headline | string | The title of the review. |
| review\_body | string | The review text. |
| review\_date | datetime | The date the review was written. |

**Mobile Electronics Code:**

Mobile Electronics Sample Data (excluding columns not used):

Table

Description automatically generated

Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Data Sources** |  |  |
| **2** | **Data Pre-Processing** |  |  |
| **3** | **Appendix** |  |  |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.