Sentiment Analysis on Amazon Customer Reviews

ACRSent - Team Name

**Data Science Capstone Project   
Launch Report**

Date:

04/08/2021

Team Members:

Name: Kevin Fitzpatrick

Name: Tessa Karlovitz

Name: Dustin Luchmee

*[The purpose of this report is initiating a new project. It provides an overview description of the project. It includes three major sections: The System/Product, The Team, and The Project Plan.]*

**The System/Product**

**System/Product Name: Sentiment Analysis on Amazon Customer Reviews (E-commerce)**

**Introduction:**

*[Describe the background information, motivation, and goals of your Data Science capstone project. What are the deliverables of your project?]*

We are all interested in sentiment analysis and how that can be used in consumer retail. Many companies use sentiment analysis to analyze customer feedback on their products. We have decided to pull Amazon review data to perform sentiment analysis on the customer reviews of their products. We would like to pull data for two types of products: electronics and books.

The Amazon reviews include a review body which is the text the customer entered and a review rating of 1-5 stars also entered by the customer. Our first basic approach will be to build a model that predicts the rating from the review body. Secondly, we can pull the top 100 ranked books from Amazon by year and build a model that predicts whether the book will be in the top 100 per year based off the reviews. There is a lot of data preparation, cleaning, and exploration needed before either will be possible.

We want a system or model that can read a customer review and assign a sentiment rating and/or if the book will be in the top 100 ranking. The goal for this capstone course is to collect, review and process our data. Therefore, the goal of this project is to get our data into a state where these models can be run. We would also like to answer some of the questions listed below in our report. Hopefully by the end of the quarter we are able to run a basic model on the data and continue our efforts in DSCI592.

We can pull several tar files from Amazon. Using several files as the training data, we can then test on similar data. We can also pull the top 100 ranked books by year from Amazon or from a dataset on Kaggle. Before that some questions to answer are ‘How many reviews were given in 2010 compared to 2015?’ ‘How do the reviews change over time?’ ‘Do the numbers change?’ ‘Does the length change over time?’ ‘Do the ratings change over time?’ ‘How often is the rating missing? What to do in those cases?’ ‘How often is the body missing? What to do in those cases?’ ‘Are there groupings of the data that will help the analysis? For example, group short reviews together.’ ‘Is there any correlation between review length and star rating?’ ‘Which products had the biggest incline/decline in user satisfaction over time?’ ‘Can the earliest reviews be used to predict if a book will be in the top 100 ranked books per year?’ ‘Are there any similarities between the books that are in the top 100 each year?’

*Notes from chat with professor*

*EDA PART: Top 20 items on amazon based on high overall ratings, top 10 companies that their products have overall ratings greater than 4.5, top 10 items in terms of number of ratings.*

*Talk about how they find the 4.6 out of 5 rating per item what amazon says they do doesn’t seem to be any different than a weighted average*

*What we should do in the next 2 weeks*

* *Create github for this project and push all code there*
* *Prepare our pitch presentation* 
  + *Mention what we have done till week 5 and then details about our plan from week 5 to week 10 (detailed descriptions)*
  + *Prepare so that general audiences can follow and understand. All the info we’ve gathered through week 5*
* *Continue data cleaning and collecting*

*Week 6 is the next report that is due*

*Week 7 we will present statistical findings/things we can share*

*Check syllabus for more detailed schedule*

We will want to read a random selection of reviews and ratings ourselves by hand because there may be issues that we are not aware of before we review the data.

The deliverable for DSCI 591 will be two clean datasets with features engineered through the insights gained from our exploratory analysis. The datasets will be ready to be fed into machine learning models with perhaps a few preliminary models already implemented.

**Highlighted Features:**

*[Write it in bullet forms.]*

* Prediction of a star rating based on reviewer’s text
* Prediction of a positive, negative, or neutral sentiment
* Prediction of part of the top 100 book rankings based on reviewer’s text

**Sponsor or Proxy User:**

*[If you have an external sponsor or real user, please describe who they are and how you will work with them.]*

Not applicable...no sponsor.

**Issues:**

[Are there any potential issues in data acquisition for the project? Do you have the expertise or stakeholder who can help you to understand and interpret the insights extracted in the project?]

We don’t foresee any issues in gathering the data because Amazon has an organized dataset url from Amazon Web Services that contains all of the hundreds of millions of customer reviews available to the public stored in tsv format. Kaggle also provides a neat downloadable format for us to use.

The next steps of filtering and cleaning the data will present some potential issues since we are working with unstructured textual data. Some issues include inconsistencies in formatting, missing values, and the general ambiguity of analyzing each reviewer’s tone.

**The Team**

**Team Name:**

*[You will be assigned a team number in the first week of class, e.g. G1. You may also pick a creative name such as SmartHealth. In this example, your team name will be G1-SmartMedia. The team name will be used in all reports and presentations in this quarter.]*

Group - ACRSent

**Team Members and their specialties:**

*[Discuss with your team members about your individual experience, strength, and interest. Identify the role of each member. For example, one of you will be the team leader who coordinates the whole team. Some of you may focus on data acquisition, some may be doing data-preprocessing and cleaning, some may do exploratory data analytics, modeling etc.]*

Tessa Karlovitz – My strengths lie in coding in Python and working with Tableau. I’ve also had a lot of experience with data pre-processing. I’m very interested in machine learning and deep learning, particularly with tools such as PyTorch.

Kevin Fitzpatrick – My biggest strengths are data mining, manipulation, and pre-processing in Python. I have some experience with data visualizations using matplotlib and seaborn and would like to gain more experience building machine learning algorithms.

Dustin Luchmee – My strengths lie in data mining, manipulation, pre-processing, and the use of machine learning algorithms in Python. I have experience using Python, as well as experience using tools such as RawGraphs, Tableau, Orange, and IBM for visualization purposes. I would like to gain experience using NLP for practical purposes. I am also concurrently taking DSCI-691 this quarter as well.

**Team Communication:**

*[How do you communicate with each other? Provide some details about the schedules of meetings, communication tools, documents sharing, etc. It is particularly important if your team have both on-campus and online students.]*

We mainly communicate via text. We are going to talk at the beginning of each week to determine what day/time is best to meet. We all work full time so scheduling at the beginning of the week is best once we all know our work schedule.

**Team Issues:**

*[Identify any issues such as skill levels and scheduling problem that constitute risks for the project and discuss how you will address these issues.]*

We all balance full time jobs as well as school so meeting on nights and weekends is a limitation, not so much an issue. Kevin does not have extensive NLP experience but is looking forward to learning from the other team members and from the project.

Table of Contributions

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Project** | **All** | **All** |
| **2** | **Team** | **All** | **All** |
| **3** | **Plan** | **All** | **All** |

Note: We all contributed and edited each other's input into this document

**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.

**Data Collection & Data Selection**

Data would be collected from two sources: The Amazon Customer Reviews Dataset (1995-2015) available on s3.amazonaws.com and the Amazon Reviews for Sentiment Analysis available on Kaggle.

For our analysis, we chose to focus on data pertaining to mobile and stationary electronics. We chose this category of the data to work with because customers often review features of the device such as battery life or screen size. We are curious to see how mobile electronics and stationary electronics are reviewed and what differences exist between handheld devices versus non-mobile devices. We would focus only on the US marketplace, as there may be cultural differences that exist in expression and review style for other English-speaking countries.

Some features available:

1. Marketplace - 2 letter country code of the marketplace where the review was written.
2. Customer ID - Random identifier that can be used to aggregate reviews written by a single author.
3. Review ID - The unique ID of the review.
4. Product ID - The unique Product ID the review pertains to. In the multilingual dataset the reviews for the same product in different countries can be grouped by the same product ID.
5. Product parent - Random identifier that can be used to aggregate reviews for the same product.
6. Product title - Title of the product.
7. Product category - Broad product category that can be used to group reviews (also used to group the dataset into coherent parts).
8. Star rating - The 1-to-5-star rating of the review.
9. Helpful votes - Number of helpful votes.
10. Total votes - Number of total votes the review received.
11. Vine - Review was written as part of the Vine program.
12. Verified purchase - The review is on a verified purchase.
13. Review headline - The title of the review.
14. Review body - The review text.
15. Review date - The date the review was written.

**Analysis Plan**

For our analysis, we aim to perform an aspect-based sentiment analysis. Our first goal is to see if we can predict the overall rating of a review from the textual review body, but further analyses could be interesting. With all mobile and non-mobile electronic reviews, we observe that many customers leave reviews regarding specific features of a device such as battery life or screen size that they either like or dislike. We are interested to see whether customers focus on one feature more than another regarding device type when reviewing, what features are most likely to be complained about, and which features appear to be the most liked by clients. In addition, we also seek to see what words and review features may indicate a poor, average, and superb star rating (1, 3, 5). The second part of our analysis is to pull Amazon review data on books as well as their top 100 ranked books by year. We can use the review data to see if we can determine what books will be in the top 100 by year based on the reviews. We can also run the same star-based analysis as described for electronics on the books. Then, we can compare the two predictions and see which works better.

**Data Sources:**

* Amazon Customer Reviews Dataset (1995-2015): <https://s3.amazonaws.com/amazon-reviews-pds/readme.html>
  + <https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt> (where we can get all the tar files from)
* Amazon Reviews for Sentiment Analysis (Kaggle): <https://www.kaggle.com/bittlingmayer/amazonreviews>
* Amazon Top 100 Ranked Books by Year (Kaggle):

<https://www.kaggle.com/jiyoungkimpf/amazon-best-sellers-of-20102020-top-100-books>