

# Project Scope and Plan: Sentiment Analysis of Amazon Electronics Product Reviews

Team Members: Yi Zu, Nishtha Sawhney

October 14, 2024

## 1 Project Scope

The aim of this project is to conduct sentiment analysis on Amazon product reviews in the Electronics category. The project will analyze and compare sentiment distribution using Natural Language Processing (NLP) techniques. This analysis will help to capture and categorize user sentiment, providing insights into customer feedback.

The project's main objectives are:

- To analyze sentiment in Amazon product reviews, specifically within the Electronics category.
- To utilize the VADER sentiment analysis model due to its effectiveness in handling short, informal text such as user-generated content.
- To create interactive visualizations that display sentiment trends, frequent keywords, and overall sentiment distribution.

The expected outcomes include:

- A processed and cleaned dataset of Amazon product reviews.
- Comprehensive sentiment analysis results, including sentiment distribution and keyword analysis.
- Visualizations such as stacked bar charts, word clouds, and line graphs.
- A final project report and presentation.

## 2 Project Kickoff

### 2.1 What are the specific goals of this project?

**Overall Goals and Objectives:**

- Analyze and compare sentiment distribution in Amazon product reviews for the Electronics category.
- Apply NLP techniques for sentiment analysis.
- Create interactive visualizations to present findings.

### 2.2 How do we define the project scope clearly to avoid scope creep?

- We are limiting the project to 5000 product reviews to capture positive, neutral, and negative user sentiment.
- The time frame will include the latest 5000 product reviews from the Electronics dataset, focusing on reviews up until July 2014.

**For NLP analysis, we are going to use VADER because:**

- Amazon reviews are often short and informal, similar to social media text.

- VADER's ability to handle slang and emoticons is beneficial for user-generated content.
- Its compound score provides a single sentiment metric, which can be useful for analysis and visualizations.

**For analysis and visualizations, we will create the following:**

- Top 10 most frequent positive and negative keywords.
- Stacked bar chart for sentiment distribution.
- Word clouds for positive and negative reviews.
- Line graph for sentiment trends over time.

## 2.3 What deliverables must be completed at different phases?

**Major Deliverables:**

- Collected dataset of Amazon product reviews for Electronics.
- Processed and cleaned dataset.
- Sentiment analysis results.
- Data analysis report.
- Final presentation or report.

## 2.4 What are the major milestones, and what deadlines should we set?

The overall timeline for the project is divided into phases:

- **Week 1 (October 7 - October 13):** Define project scope, establish team roles, and outline skills/tools.
- **Week 2 (October 14 - October 20):** Collect and process 1000 product reviews for Electronics.
- **Week 3 (October 21 - October 27):** Begin development of the code. Start researching how to do sentiment analysis, and develop the skills required to begin coding.
- **Week 4 (October 28 - November 3):** Start with sentiment classification of product reviews.
- **Week 5 (November 4 - November 10):** Finalize sentiment classification of product reviews. Start analysis report with sentiment distributions, top keywords, and trends.
- **Week 6 (November 11 - November 17):** Complete analysis report with sentiment distribution. Develop interactive visualizations (stacked bar charts, word clouds, line graphs, heatmap)
- **Week 7 (November 18 - November 28):** Final presentation, report submission, and project closure.

Key milestones include:

- Project Scope Defined, Team Roles Established, and Initial Data Collection Completed (October 13 - October 20).
- Sentiment Analysis Research and Development Started (October 21 - November 3).
- Sentiment Classification Completed, Analysis Report and Visualizations Developed (November 4 - November 17).
- Final Presentation and Report Submission (November 28).

## 2.5 Do the team's capabilities align with these goals? Are there any gaps that need to be addressed early on?

Our team's capabilities align with the goals of the project as we are both proficient in Python. However, there could be gaps in sentiment analysis and NLP-related packages since this is one of the first times we are using such technologies.

## 2.6 Do you have a dataset ready to use for the current project?

There is a public dataset available containing Amazon Product Reviews from 1990 to 2014. We will be using this dataset for our project. The dataset can be accessed at:

[https://jmcauley.ucsd.edu/data/amazon/index\\_2014.html](https://jmcauley.ucsd.edu/data/amazon/index_2014.html)

# 3 Team Discussion

## 3.1 Team Roles

- **Yi Zu:** Responsible for data collection, preprocessing, and implementing sentiment analysis using NLP techniques such as VADER.
- **Nishtha Sawhney:** Responsible for sentiment classification and analysis, along with visualizing sentiment trends (e.g., sentiment distribution charts and keyword analysis).
- **Both Team Members:** Collaboratively responsible for writing documentation, preparing the final report using Overleaf, and managing the project repository on GitHub. Responsible for testing, ensuring the quality of the sentiment analysis results, and preparing the final project presentation.

These roles are flexible and may change as the project progresses.

## 3.2 What are the core skills each team member brings to the table?

**Yi's core skills:**

- Python programming
- Data cleaning
- Data analysis
- Data visualization

**Nishtha's core skills:**

- Data preprocessing
- Data analysis
- Data visualization
- Python programming

## 3.3 How will each person's expertise contribute to specific tasks?

We are both going to clean, analyze, and create visualizations for Electronics product reviews.

## 3.4 What skills are missing that may cause delays or challenges?

We are new to NLP sentiment analysis, which will require us to self-learn NLP concepts while doing the project.

### 3.5 What tools do we have experience with, and what do we need to learn?

We have experience with Python for data cleaning and visualization. We are going to learn tools like NLTK or spaCy for NLP tasks, specifically text processing and analysis. For sentiment analysis, we will learn either VADER (from NLTK) or TextBlob.

### 3.6 What programming languages and platforms should we select based on our project needs and team experience?

We will primarily use Python for our programming needs, given our combined expertise with its extensive libraries. This will simplify our data preprocessing, analysis, and visualization tasks. Specifically, we will utilize Pandas for data preprocessing and Matplotlib for data visualization.

## 4 Skills & Tools Assessment

### 4.1 Skills Assessment

Team members need the following skills to complete the project:

- Proficiency in Python for data preprocessing and sentiment analysis.
- Understanding of Natural Language Processing (NLP) techniques, including experience with libraries such as NLTK or spaCy.
- Familiarity with sentiment analysis models, specifically VADER or TextBlob, for sentiment scoring.
- Experience with data visualization techniques for presenting sentiment trends and key insights.
- Ability to collaboratively manage project documentation using Overleaf and version control using GitHub.

### 4.2 Skills Gaps & Resource Plan

We identified a few gaps in NLP techniques and sentiment analysis, as this is one of the first times we are working with these technologies. To address these gaps, we will rely on external resources such as Coursera and YouTube tutorials to improve our understanding of NLP concepts. Additionally, we will consult the official documentation for the libraries we are using, such as NLTK, VADER, to ensure we apply these tools effectively during the project.

### 4.3 Are there external resources or team members with expertise in the areas where we lack skills?

There are Coursera and YouTube channels that teach NLP concepts, so we will be referring to these resources to learn more about sentiment analysis. We will also consult the official documentation for the various libraries that we plan to use.

### 4.4 Which tools, frameworks, and libraries are most suitable for the project's scope?

**Data Processing:**

- **Tool:** Pandas
- **Purpose:** For data manipulation and cleaning.

**Natural Language Processing (NLP):**

- **Tools:** NLTK or spaCy
- **Purpose:** For text processing and analysis.

**Sentiment Analysis:**

- **Tools:** VADER (from NLTK) or TextBlob
- **Purpose:** For sentiment scoring.

**Data Visualization:**

- **Static Visualizations:** Matplotlib

#### 4.5 How can we ensure that each team member is comfortable with the tools selected?

Since we both have similar skills and are comfortable with Python, it makes sense to use what we already know while also learning new NLP tools together. We can make sure this works well by talking openly, making decisions together about what tools to use, and learning new skills side by side. This way, we can tackle any problems that come up and make our work together more efficient.

#### 4.6 Have specific tasks been assigned based on individual strengths, and are team members clear on their roles?

Since we both have similar skill sets, we have decided to collaborate closely on each part of our project. We haven't assigned specific tasks to each other because we prefer to do pair programming. This approach allows us to learn more effectively by working together rather than dividing the tasks.

## 5 Initial Setup

### 5.1 What development environment setup is necessary for this project?

For this project, we'll need to set up Anaconda and Jupyter Notebook. Anaconda helps manage our software packages easily, and Jupyter Notebook lets us write and test our code interactively. This setup will simplify our coding tasks.

**Programming Language:**

- Python 3.x (preferably 3.8 or later)

**Integrated Development Environment (IDE):**

- PyCharm, Visual Studio Code, or Jupyter Notebooks

**Version Control:**

- Git for version control
- GitHub or GitLab for repository hosting

**Virtual Environment:**

- **conda** for managing project-specific dependencies
- **pandas:** for data manipulation and analysis
- **numpy:** for numerical operations
- **nlTK:** for natural language processing tasks
- **matplotlib** and **seaborn:** for static data visualization
- **jupyter:** for using Jupyter Notebooks

### 5.2 Have we successfully configured version control (such as Git)?

Yes, we have.

### 5.3 Does everyone have access to the repository?

Yes.

**5.4 Have we installed and configured all required software, libraries, and tools?**

Yes.

**5.5 What testing can we run to ensure that the development environment is functioning correctly?**

We have run basic functions and print statements to ensure it is working correctly.

**5.6 What troubleshooting steps should we take if the setup does not work as expected?**

Read the error logs, use Stack Overflow, and paste error logs into ChatGPT to further debug.

## **6 Progress Review**

**6.1 What has been achieved so far? Have we completed the initial setup and repository configuration?**

We want to ensure that our project receives approval from the instructor before moving forward with the data collection and cleaning tasks.

**6.2 Have there been any issues or blockers, and how can we address them quickly?**

Not yet.

**6.3 Is each team member contributing as expected, and does everyone understand their role?**

Yes.

**6.4 Are we on track with the timeline and milestones, or do we need to adjust them?**

We need to receive approval on the project first.

**6.5 How does the progress align with the project's overall objectives?**

We need to receive approval on the project first to be able to move forward with our intended project plan.

## **7 Plan Revision**

**7.1 Based on progress so far, do we need to adjust the project timeline or milestones?**

No, we are ready to start retrieving the data required for data processing. We did thorough research, and we are prepared to start the project.

**7.2 Are any tasks delayed or requiring reassignment due to workload or skill gaps?**

We are seeking final approval of the project before moving forward with the rest of the project.

### **7.3 How can we ensure that all members are clear on the revised plan and their next steps?**

We will check in at the beginning of the week on Tuesday after class to monitor progress and hold ourselves accountable to the project's estimated timeline.

### **7.4 What communication strategies can we implement to avoid future delays or misunderstandings?**

We hold weekly meetings to set our goals for the week, ensuring that we're both on the same page. We also make sure to talk regularly about any new ideas or issues that arise, keeping our discussions open and proactive. To maintain clarity, we take notes during meetings and frequently check in with each other to update on progress. We give and receive feedback regularly to foster continuous improvement. These steps keep our communication clear and our project on track.

### **7.5 How will we track progress going forward and maintain alignment with the revised plan?**

We will use Excel to track our progress on the work completed so far.

## **8 Submission for This Iteration**

### **8.1 What specific tasks need to be documented for this iteration's submission?**

- PDF of the report.
- Excel tracking sheet of project progress.
- GitHub repository link: <https://github.com/yizucodes/Five-Star/tree/main>

### **8.2 Have we detailed the challenges faced, the solutions implemented, and any adjustments to the plan?**

Yes.

### **8.3 If your data is available online, please provide a link to access it.**

The data is available online at: [https://jmcauley.ucsd.edu/data/amazon/index\\_2014.html](https://jmcauley.ucsd.edu/data/amazon/index_2014.html). We will use data under the heading "Small subsets for experimentation."

### **8.4 Is the PDF using the Overleaf template, and does it reflect the team's actual progress?**

Yes, the PDF has been created using the Overleaf template provided and reflects the team's actual progress.

### **8.5 Has everyone tracked their progress using the Excel file, and are we submitting it along with the PDF?**

Yes.

### **8.6 Does the submission meet all the project requirements, and is it ready for review by stakeholders?**

Yes.

## 8.7 Please provide a link to your GitHub repository with the updated files, including the PDF uploaded there as well.

<https://github.com/yizucodes/Five-Star/tree/main>

## 9 Initial Setup Evidence (October 14)

### 9.1 Project Repository

The project repository has been created on GitHub, accessible by all team members. The repository can be found at <https://github.com/team/library-system>.

### 9.2 Setup Proof

The development environment has been successfully set up. Screenshots of the setup process are provided in the Appendix.

## 10 Progress Review (October 14)

### 10.1 Progress Update

The team successfully completed the initial setup of the development environment and repository.

### 10.2 Issues Encountered

We were previously planning to retrieve Reddit technology-based question and answer data. However, we were unable to retrieve this information due to technical limitations. As a result, we decided to use publicly available data for sentiment analysis, specifically Amazon product reviews.

## 11 Revised Project Plan (October 14)

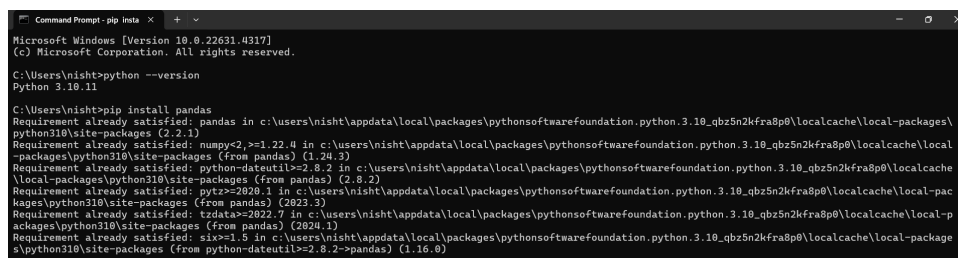
### 11.1 Updated Plan

After reviewing our progress, we updated the timeline to account for the change in data sources. We are now focusing on sentiment analysis using Amazon product reviews, and this adjustment has allowed us to streamline our data collection process.

### 11.2 Justification for Changes

The change in data source was necessary due to our inability to retrieve the Reddit data as initially planned. By shifting to Amazon product reviews, we ensure that we can proceed with sentiment analysis without further delays. This allows us to remain on track with the project milestones.

## 12 Appendix



```
Microsoft Windows [Version 10.0.22631.4317]
(c) Microsoft Corporation. All rights reserved.

C:\Users\nisht>python --version
Python 3.10.11

C:\Users\nisht>pip install pandas
Requirement already satisfied: pandas in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (2.2.1)
Requirement already satisfied: numpy<2, >=1.22.4 in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from pandas) (1.24.3)
Requirement already satisfied: python-dateutil<=2.8.2 in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from pandas) (2023.3)
Requirement already satisfied: tzdata>=2022.7 in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from pandas) (2024.2)
Requirement already satisfied: six>=1.5 in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from python-dateutil<=2.8.2->pandas) (1.16.0)
```

Figure 1: Python and Pandas Installation



```

C:\Users\nisht>
C:\Users\nisht>pip install nltk matplotlib vaderSentiment
Collecting nltk
  Downloading nltk-3.9.1-py3-none-any.whl (1.5 MB)
    1.5/1.5 MB 18.9 MB/s eta 0:00:00
Collecting matplotlib
  Downloading matplotlib-3.9.2-cp310-cp310-win_and64.whl (7.8 MB)
    7.8/7.8 MB 38.5 MB/s eta 0:00:00
Collecting vaderSentiment
  Downloading vaderSentiment-3.3.2-py2.py3-none-any.whl (125 kB)
    125/125 kB 9.3 MB/s eta 0:00:00
Requirement already satisfied: click in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from nltk) (8.1.3)
Requirement already satisfied: joblib in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from nltk) (1.2.0)
Collecting regex>=2021.8.3 (from nltk)
  Downloading regex-2024.9.11-cp310-cp310-win_and64.whl (274 kB)
    274/274 kB 16.5 MB/s eta 0:00:00
Collecting tqdm (from nltk)
  Downloading tqdm-4.66.5-py3-none-any.whl (78 kB)
    78/78 kB 2.1 MB/s eta 0:00:00
Collecting contourpy>=1.0.1 (from matplotlib)
  Downloading contourpy-1.3.0-cp310-cp310-win_and64.whl (216 kB)
    216/216 kB 13.7 MB/s eta 0:00:00
Collecting cycler>=0.10 (from matplotlib)
  Downloading cycler-0.12.1-py3-none-any.whl (8.3 kB)
Collecting fonttools>=4.22.0 (from matplotlib)
  Downloading fonttools-4.54.1-cp310-cp310-win_and64.whl (2.2 MB)
    2.2/2.2 MB 34.9 MB/s eta 0:00:00
Collecting kiwisolver>=1.3.1 (from matplotlib)
  Downloading kiwisolver-1.4.7-cp310-cp310-win_and64.whl (55 kB)
    55/55 kB 4.4 MB/s eta 0:00:00
Requirement already satisfied: numpy>=1.23 in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from matplotlib) (1.24.3)
Requirement already satisfied: packaging>=20.0 in c:\users\nisht\appdata\local\packages\pythonsoftwarefoundation.python.3.10_qbz5n2kfra8p0\localcache\local-packages\python310\site-packages (from matplotlib) (23.1)
Collecting pillow>=8 (from matplotlib)
  Downloading pillow-10.4.0-cp310-cp310-win_and64.whl (2.6 MB)
    2.6/2.6 MB 25.4 MB/s eta 0:00:00
Collecting pyparsing>=2.3.1 (from matplotlib)

```

Figure 2: Relevant Python Libraries for Sentiment Analysis and Data Visualization

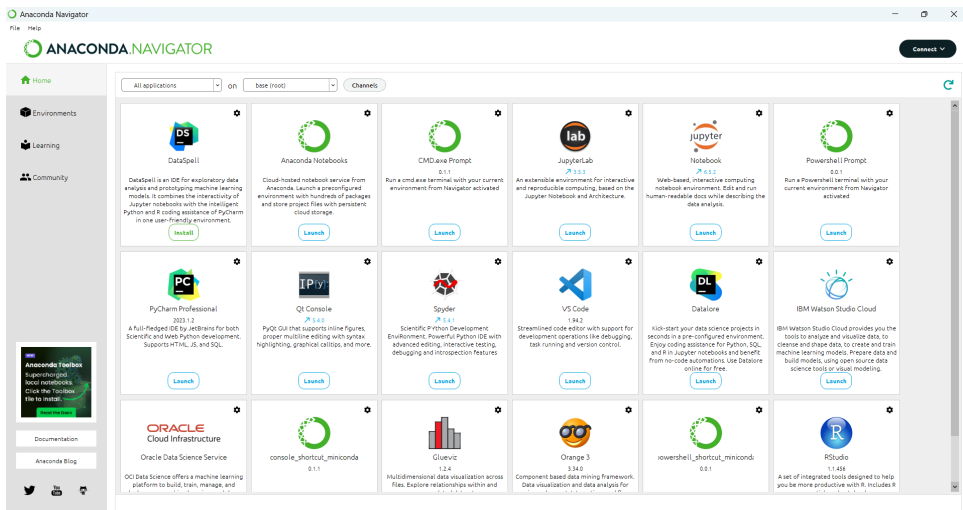


Figure 3: Anaconda Installation

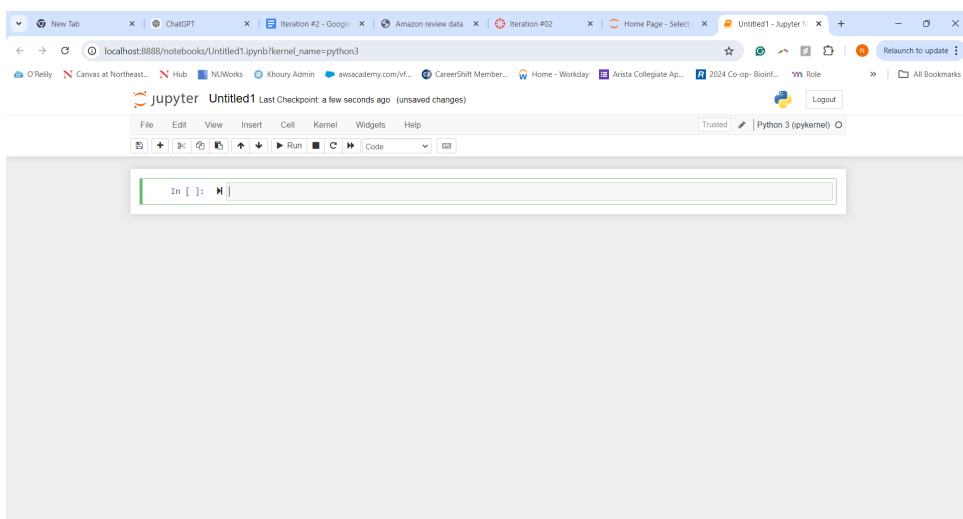


Figure 4: Jupyter Notebook Installation