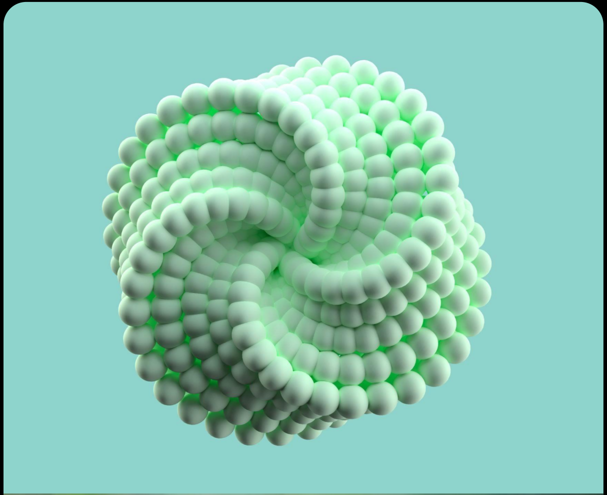


December 10
Introduction to Data Management and Processing

Amazon Electronic Product Reviews for Sentiment Analysis

Nishtha Sawhney Yi Zu

Introduction

01 Objective and Goals

02 Project Scope

03 Relevant Existing Work

04 Relevance of Project with Previous Work

Objectives and Goals

- Conduct sentiment analysis on Amazon Electronics product reviews using NLP techniques
- Analyze and compare sentiment distribution in product reviews
- Use VADER sentiment analysis model for processing informal text
- Create interactive visualizations of sentiment trends and patterns
- Generate comprehensive insights into customer feedback

Project Scope

- Dataset Parameters:
  - Analyze 5,000 product reviews from Amazon Electronics category
  - Focus on reviews up to July 2014
  - Using publicly available Amazon product reviews dataset
- Key Deliverables:
  - Processed and cleaned dataset
  - Sentiment analysis results and distribution
  - Keyword analysis and trends
  - Interactive visualizations
    1. stacked bar charts, word clouds, line graphs

Existing Work: Amazon Reviews Dataset Analysis (McAuley, J., & Leskovec, J. (2013)

- Combined hidden rating factors with review text analysis to reveal deeper patterns in how customers evaluate products
- Established the largest public Amazon review dataset (1996–2014), now an industry standard for e–commerce research
- Proved that analyzing both ratings and review text together leads to better understanding of customer opinions than ratings alone
    - Better understand WHY users gave certain ratings

Analysis Tools & Libraries

- Builds on their established Amazon electronics review dataset, focusing on 5000 recent reviews up to 2014 for targeted sentiment analysis
- Extends their combined rating–text analysis approach by applying VADER sentiment analysis to understand emotional content in technical reviews
- Advances their work through interactive visualizations and temporal analysis, making sentiment patterns more accessible and actionable

Methodology: Data Processing & Analysis

01 Data Exploration

02 Data Preprocessing

03 Sentiment Analysis

04 Data Visualization

Data Exploration

Data Collection & Parsing

1. Data Collection & Parsing
   - Custom GZIP JSON parser for efficient data loading
   - Pandas DataFrame for structured data organization
   - Initial sample: 5,000 reviews from Electronics dataset
2. Data Processing Pipeline
   - Timestamp conversion (Unix to datetime
   - Text length analysis
   - Helpfulness ratio calculation
   - Rating distribution analysis

Core Data Processing Logic

1. Parse GZIP JSON Data
2. Process Reviews：
   - Extract metadata (ID, rating, time)
   - Calculate text metrics
   - Compute helpfulness ratio
   - Convert timestamps
3. Analysis Pipeline：
   - Calculate review statistics
   - Generate distribution metrics
   - Perform time series analysis

Time Complexity Analysis

1. Major Operations:
   - Data Loading: O(n) where n = number of reviews
   - Text Processing: O(n * m) where m = avg review length
   - Statistical Analysis: O(n)
   - Time Series Grouping: O(n log n)
2. Space Complexity:
   - Primary DataFrame: O(n)
   - Auxiliary Data Structures: O(k) where k = unique categories

Data Structures Utilized

Primary Structures:

- Pandas DataFrame
  - Core data organization
  - Efficient column operations
  - Built-in indexing
- NumPy Arrays
  - Statistical computations
  - Numerical analysis

Supporting Structures:

- Python dictionaries for metadata
- Lists for sequential processing

Analysis Tools & Libraries

Core Libraries:

- pandas: Data manipulation and analysis
- numpy: Numerical computations
- matplotlib: Data visualization
- seaborn: Statistical visualization

Data Processing:

- json: JSON parsing
- gzip: Compressed file handling
- datetime: Time series management

Data Preprocessing

Preprocessing Methodology

Data Loading

- Custom JSON parser for compressed GZIP files
- Initial dataset: 1.6M Amazon Electronics reviews
- Final sample: 5,000 most recent reviews up to July 2014

Text Processing Pipeline

- ReviewPreprocessor class implementation
- Handles text cleaning, stopwords, dates
- Feature engineering and statistical analysis

Preprocessing Implementation Details

1. Text Cleaning
   - Lowercase
   - Remove HTML tags, URLs, special characters, extra whitespace
2. Stopword Removal
   - Custom stopwords list
   - Word tokenization
     - Breaking down text into individual words
     - Removal of unwanted words (“the”, “a”, “is”)
   - Filtered word rejoining after tokenization for sentiment analysis
3. Feature Creation
   - Review length calculation

Time & Space Complexity

Loading & Initial Processing:

- Data Loading: O(n) - linear with number of reviews
- Text Cleaning: O(n * m) where m = avg text length
- Feature Creation: O(n)

Memory Requirements:

- Primary DataFrame: O(n) where n = number of reviews
- Text Processing: O(m) temporary storage per review
- Feature Storage: O(k * n) where k = number of features

Data Quality Checks

Verification Steps

1. Data Integrity：
   - Missing value detection and handling
   - Duplicate review removal
   - Empty review filtering
   - Format validation
2. Statistics：
   - Review length distribution
   - Word count analysis
   - Rating distribution verification
   - Date range confirmation

Final Processing Outcomes

1. Dataset Statistics:
   - Initial reviews: 5,000
   - Final cleaned reviews: 4,966
   - Duplicates removed: 32
   - Average word count: 93.7
2. Quality Metrics:
   - Clean text coverage: 100%
   - Missing value resolution: Complete
   - Date range: July 15–23, 2014
   - Rating distribution preserved

Sentiment Analysis

Methodology

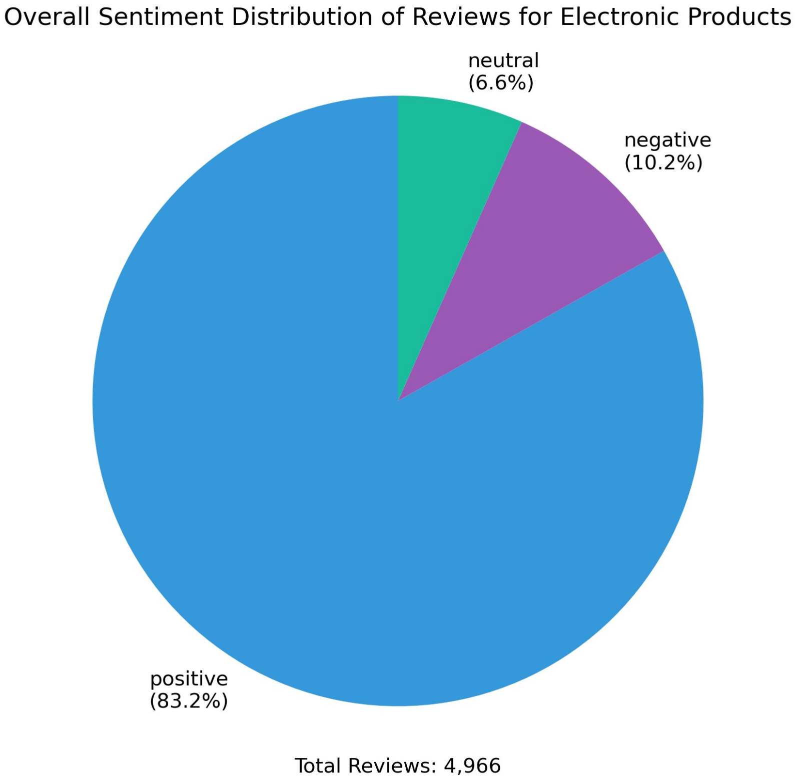
- Retrieve the metadata for products
  - which contains product name, category, and other relevant information
- Merged the asin number from the processed table with the metadata
- Used VADER to do sentiment analysis
  - Label each review one of the following:
    - Positive
    - Negative
    - Neutral

Sentiment Analysis Strategy

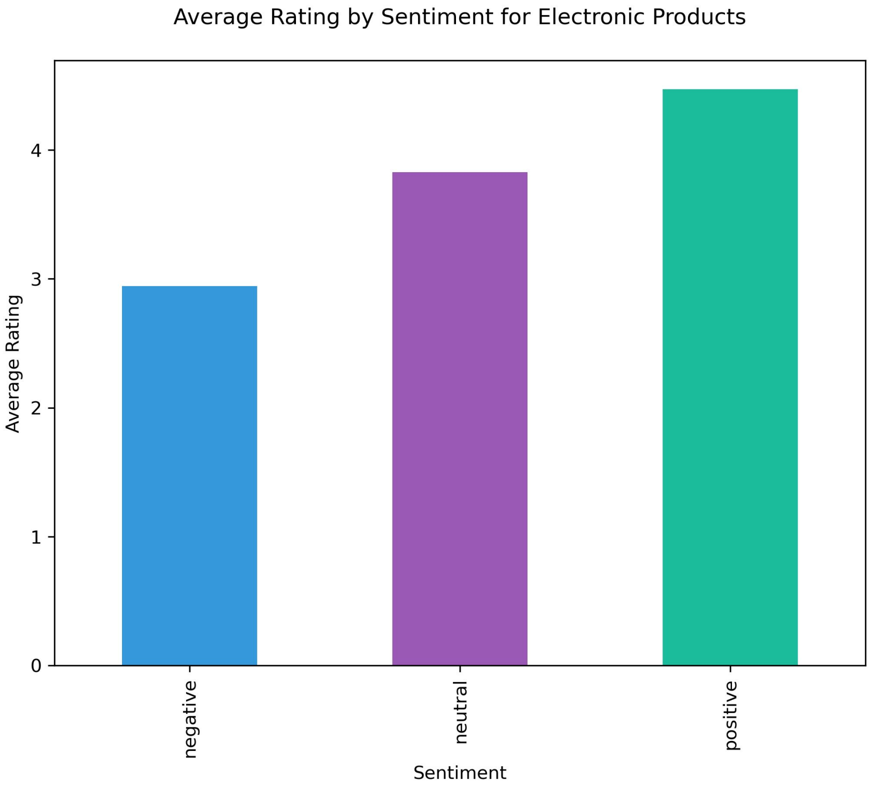
Why use VADER?

- Designed to handle informal language, ideal for analyzing reviews and other use–generated content
- Also, it is already a pre–trained model
  - No training needed
  - Quick to implement
- Has polarity with intensity
  - Goes from –1 (most negative) to (most positive).
- Is context–sensitive aspects

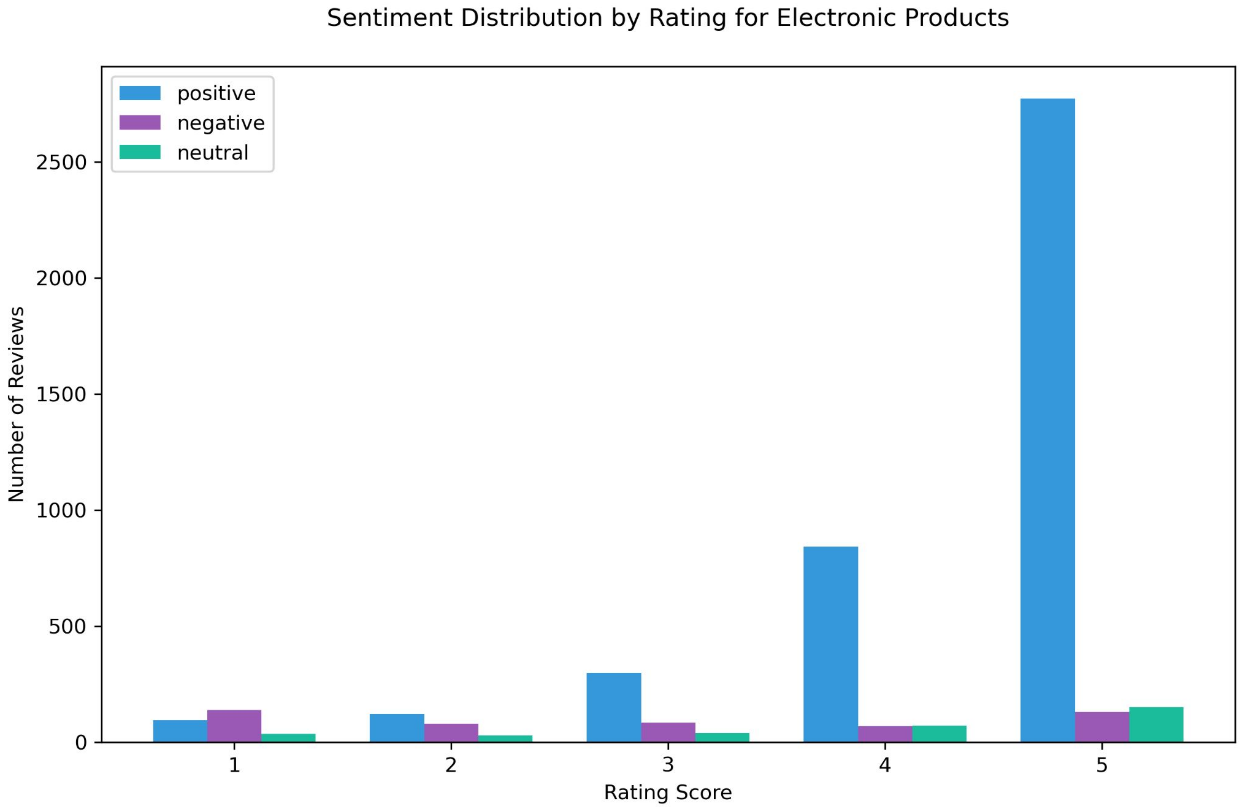
Data Visualization



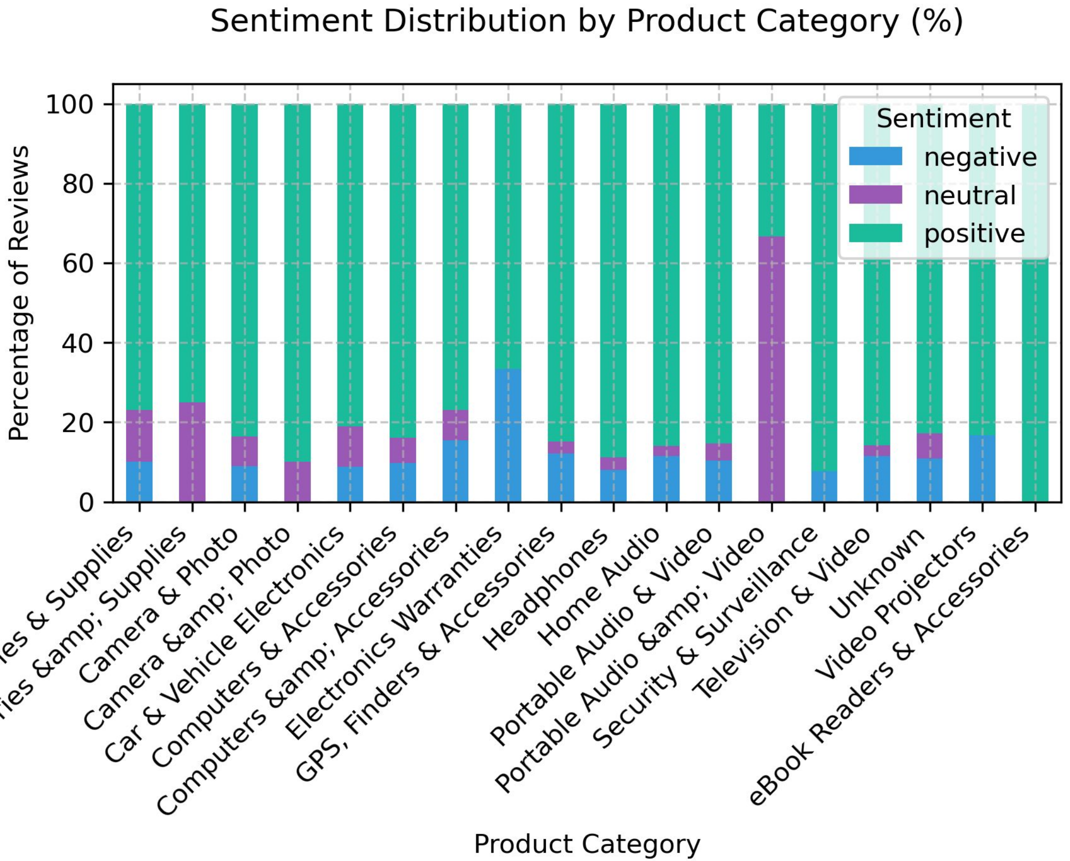
This pie chart shows the sentiment distribution of 4,966 reviews for electronic products. Most reviews are positive (83.2%), while 10.2% are negative, and 6.6% are neutral.



This bar chart shows the average ratings for electronic product reviews based on sentiment. Positive reviews have the highest average rating (around 4+), neutral reviews are slightly lower, and negative reviews have the lowest average rating (around 3).



This chart breaks down the number of reviews by sentiment (positive, negative, neutral) for each rating score (1 to 5). Positive reviews dominate at higher ratings (especially 5), while negative reviews are more common at lower ratings (1 and 2). Neutral reviews are spread out but less frequent overall.



The category with the most negative reviews seems to be Electronics Warranties, and the one with the most neutral reviews looks like Television & Video. Positive reviews dominate in all categories, making up the majority.