GMMA 865 – Amazon Review Project

GMMA 865 – Big Data Analytics

Team New York

Smith SCHOOL OF BUSINESS | Queen's University

Challenge: Conduct Analytics on an Amazon Big Dataset using Databricks

Hi, Team New York
Students since 2020

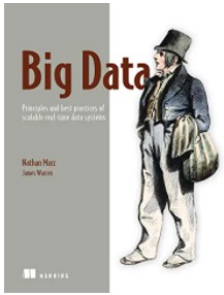
Two main objectives

1. Understand Review Habits
2. Predict helpfulness of reviews

★★★★☆ & Up
★★★☆☆ & Up
★★☆☆☆ & Up
★☆☆☆☆ & Up

Helpful

Recently viewed



Amazon Dataset
3,487,331 Reviews

1998 to 2018

Games (487,419)

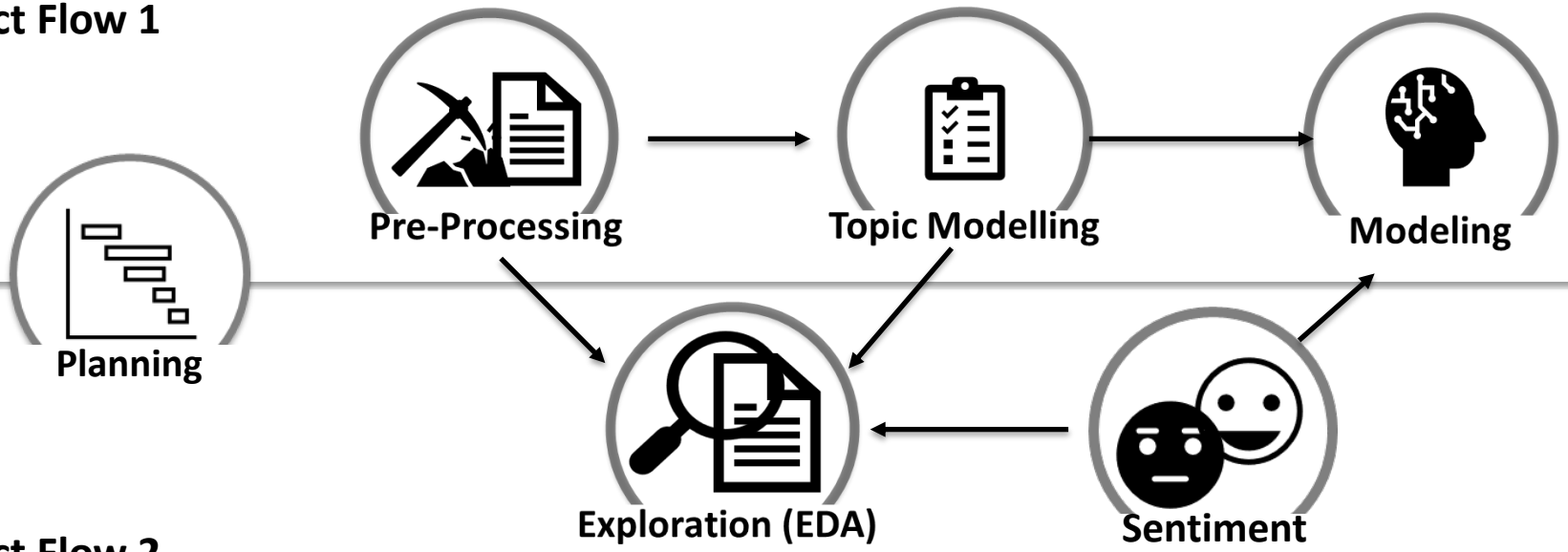
Books (993,913)

Homes (1,999,999)

Strategy

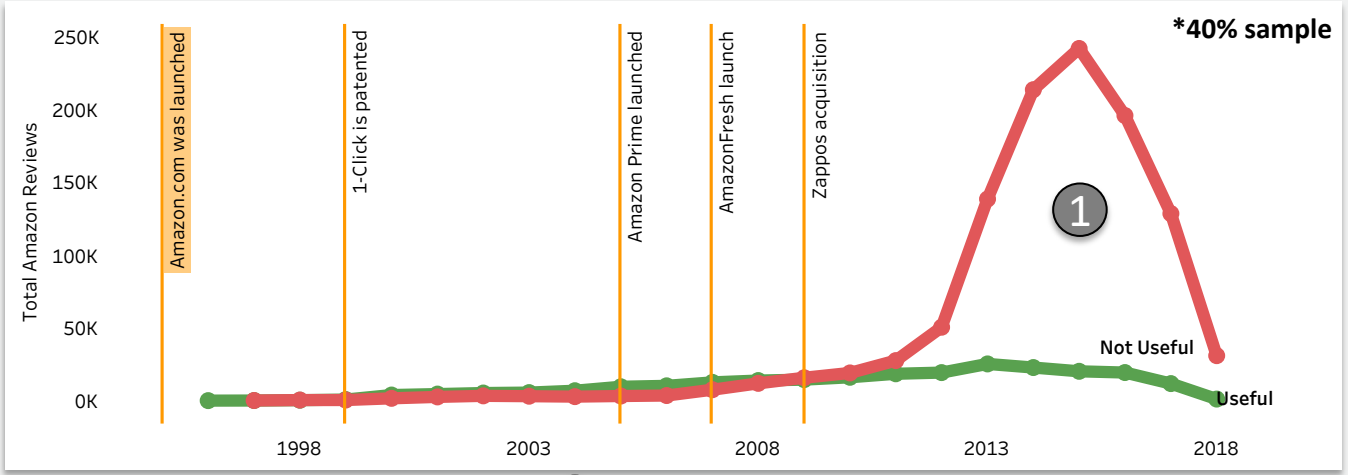
- Parallel team approach to coding and research
- Lower computing power by focusing on Spark compatible libraries
- Leverage ML Ops techniques and features in Databricks

Project Flow 1

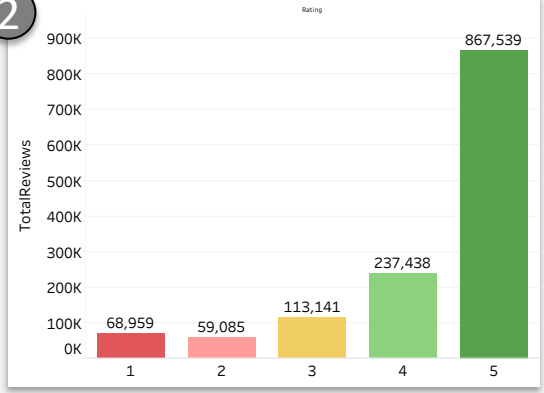


Project Flow 2

	Not Useful					Useful				
	1	2	3	4	5	1	2	3	4	5
Average Polarity	-0.04	0.05	0.16	0.27	0.40	-0.01	0.06	0.12	0.18	0.25
Average Word Count	60	75	78	76	47	142	198	230	238	179
Avg. Word Count Normalized	30	38	39	39	25	72	101	117	123	92
Total Reviews	38,704	39,560	86,053	193,340	746,682	30,255	19,525	27,088	44,098	120,857
% of Total Reviews	2.9%	2.9%	6.4%	14.4%	55.5%	2.2%	1.5%	2.0%	3.3%	9.0%



2



3

Was the review useful?

1996-2012	Not Useful	Useful	Total
False	23%	28%	51%
True	29%	20%	49%
Total	52%	48%	100%

Was the review useful?

2013-2018	Not Useful	Useful	Total
False	8%	1%	10%
True	82%	8%	90%
Total	90%	10%	100%

Key Takeaway

- **"Non-useful" reviews make up 82% of total reviews**
- **Useful reviews continue to grow after 2011:** however we noticed a spike in "non-useful" reviews
- **Dataset Imbalance:** After 2012 only 10% of all reviews were classified as "useful"

Reviewers

1.13 M

14% of reviewers are Anonymous

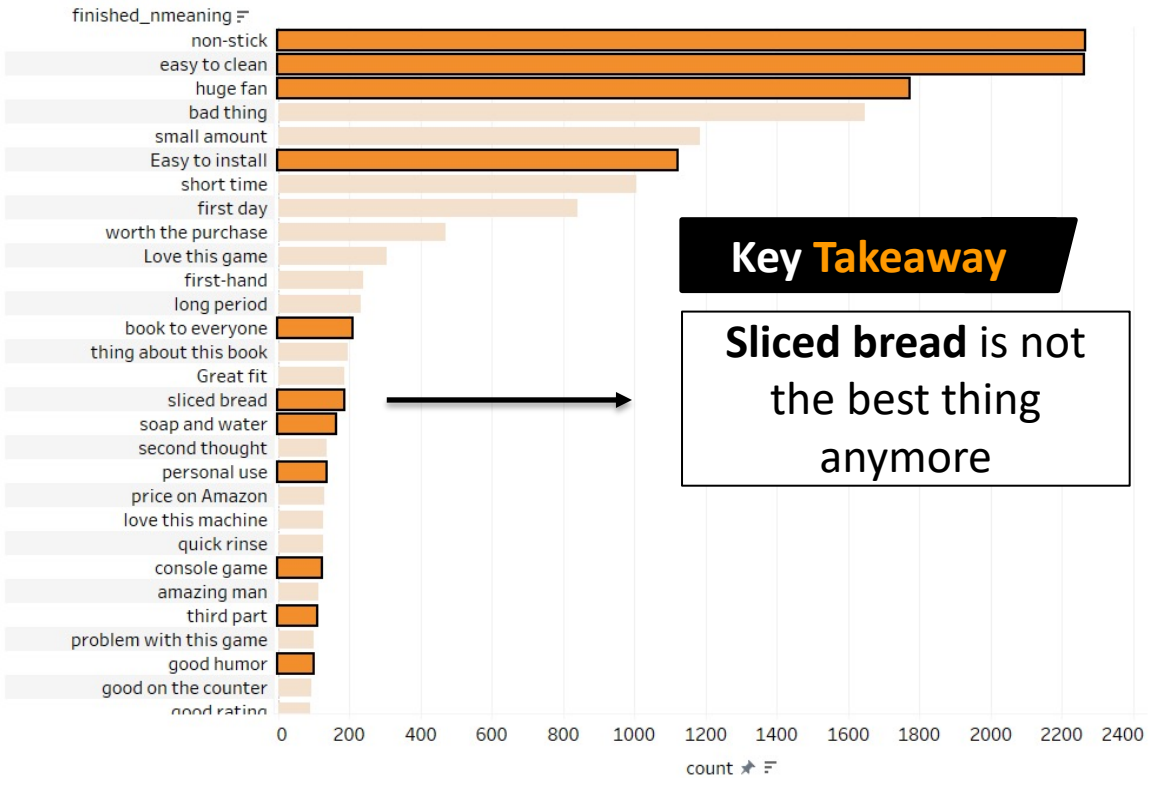
Most Reviewed

1.6% of total reviews

Top Reviewer Count

	reviewerID	count
1	A3V6Z4RCDGRC44	862
2	AJKWF4W7QD4NS	805
3	A1D2C0WDCSHUWZ	734
4	A2F6N60Z96CAJI	623
5	A3W4D8XOGLWUN5	536
6	A2QHS1ZCIQOL7E	471
7	A2TCG2HV1VJP6V	450

Most frequent N-Grams



Key Takeaway

Sliced bread is not the best thing anymore

Rating	Summary	Review Text
5	PS2 Memory Card 128MB for Playstation 2	Hello, i must say this card does just what it says, it holds all of my PS2 game saves, i have always wanted to use one "Memory Card" for all saves. The card holds 128MB and after saving all my saves i have 110MB left and i have well over 60 Playstation 2 games, it`s like having a "PC Hard Drive" in the PS2 and now that i know the card really do work, i will be ordering quite a few more before they are no longer in stock. It does not take much to make me happy and this makes me very happy, now i must get more PS2 games, i have 2 PS2, 2 PS3, 1 PSP, 1 Tablet, 2 Laptops and 2 Desktops all for gaming, i love gaming and proud to say i am a 56 year old gamer, started in 1992, love it. Thanks<a data-hook="product-link-linked" class="a-link-normal" href="/PS2-Memory-Card-128MB-for-Playstation-2/dp/9629971372/ref=cm_cr_arp_d_rvw_txt?ie=UTF8">PS2 Memory Card 128MB for Playstation 2

Key Takeaway

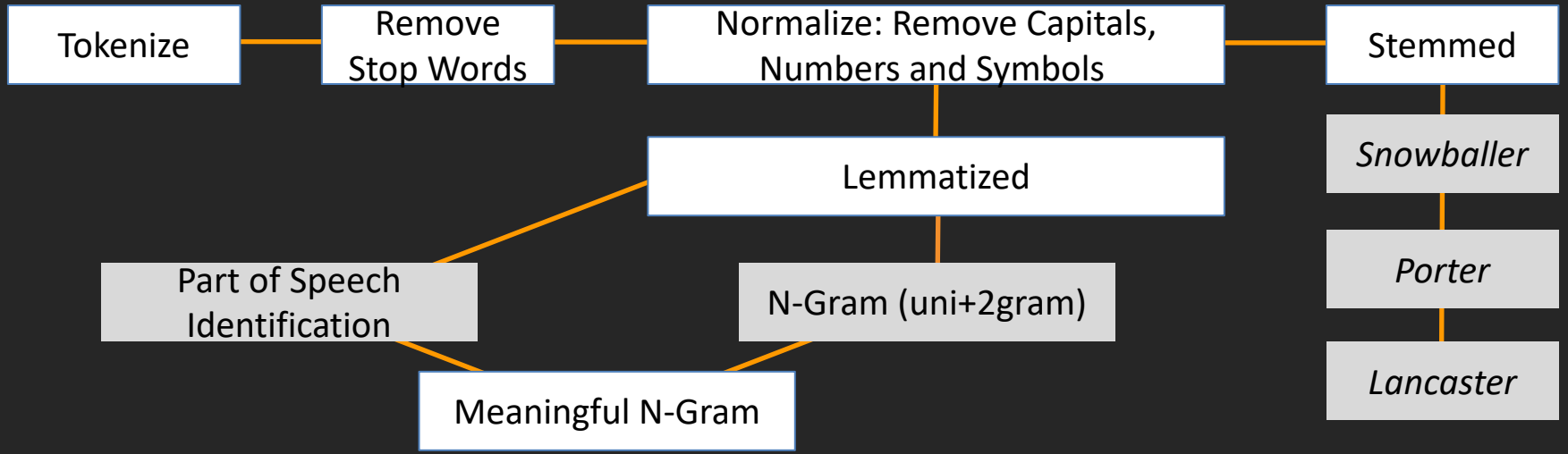
Reviews that mention

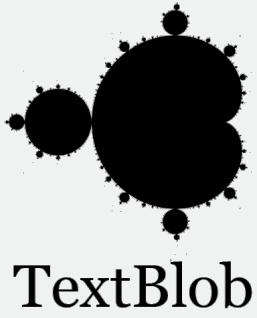
- lots of numbers
- multiple negation
- html data
- acronyms and shortcuts
- spell check

Cleaning the Reviews

- Combined Summary and Review
- Converted Date and Time Format
- Dropped Incomplete Rows
- Filled NAs with 0

Preprocessing Pipeline





TextBlob: Simplified Text Processing

Visit the Text Analytics Store

#1 Best Seller in Sentiment Analysis

- TextBlob packaged was used to generate the polarity and subjectivity
- Initially tried the JohnSnowLabs NLP; couldn't get the code to work; we attempted ViveknSentimentDetector and SentimentDetector

Tableau: Interactive Data Visualization Software

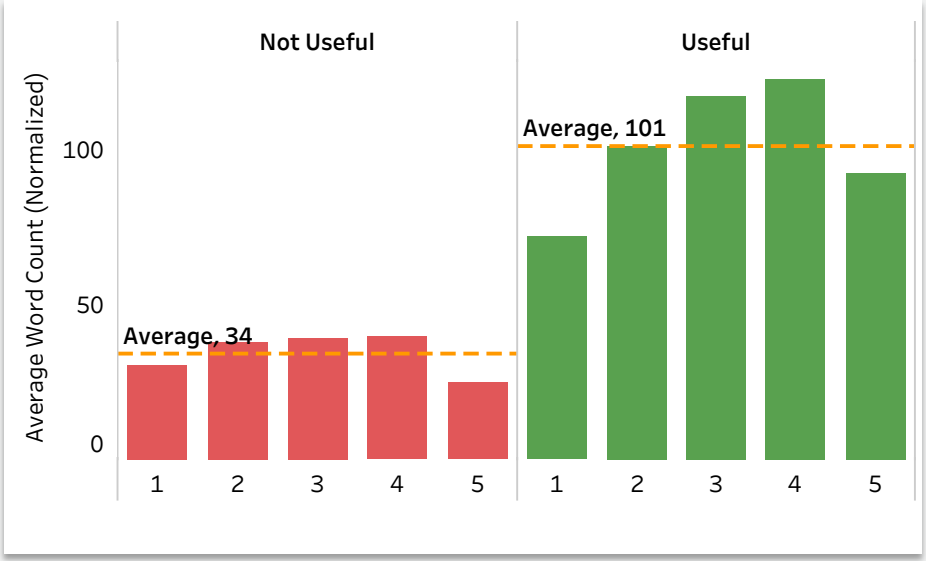
Visit the Text Analytics Store

#1 Best Seller in Data Visualization

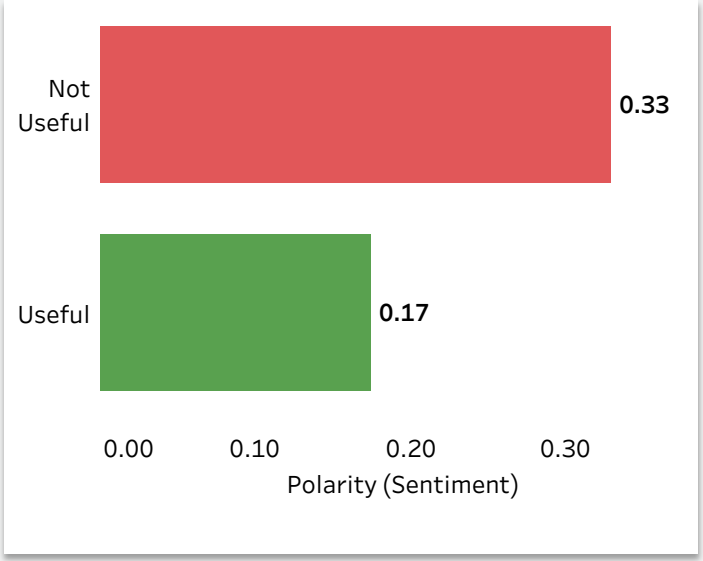
- Started at a macro level to understand distributions and trends like the number of reviews by category, over time, useful/non-useful
- Drilled-down into other measures like average word count, average polarity, average subjectivity to understand the impact of useful or non-useful reviews
- Analyzed the review content to understand what was said and the types of phrases that resonated in each category

Question 1: What makes a Review Helpful?

Words Matter



Feelings Matter

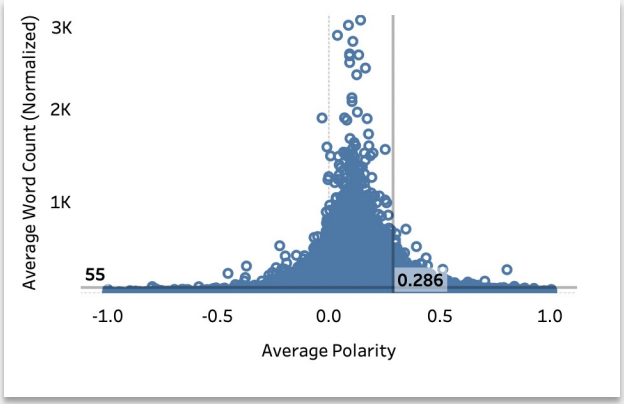


Key Takeaway

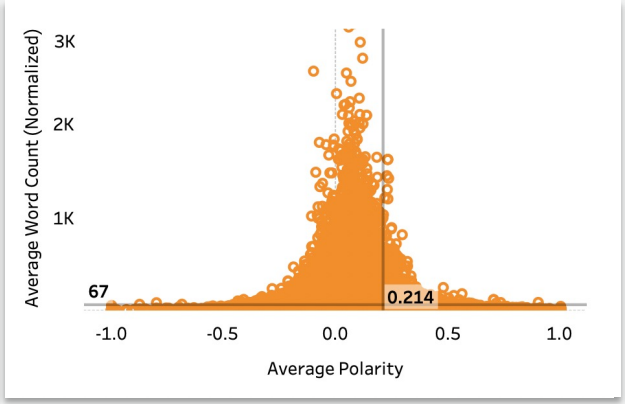
- **Useful reviews have a higher average word count but are more negative**
- **Useful 3 and 4-star reviews have 3x the average word count giving a Real and Genuine" feel**
- **Useful reviews tend to be more neutral in sentiment (closer to 0) than non-useful**

Question 2 and 3: Habits of reviewers, and do they change by category?

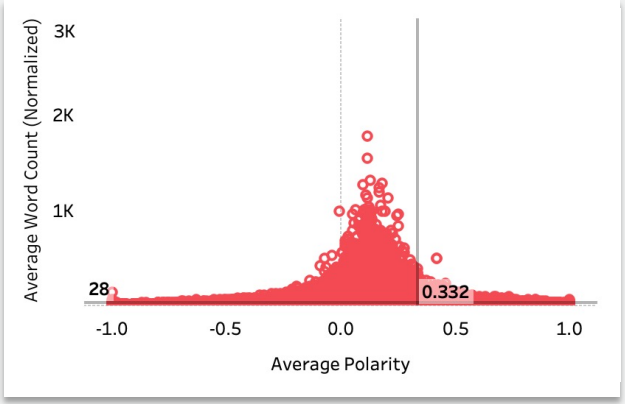
Books



Games



Home



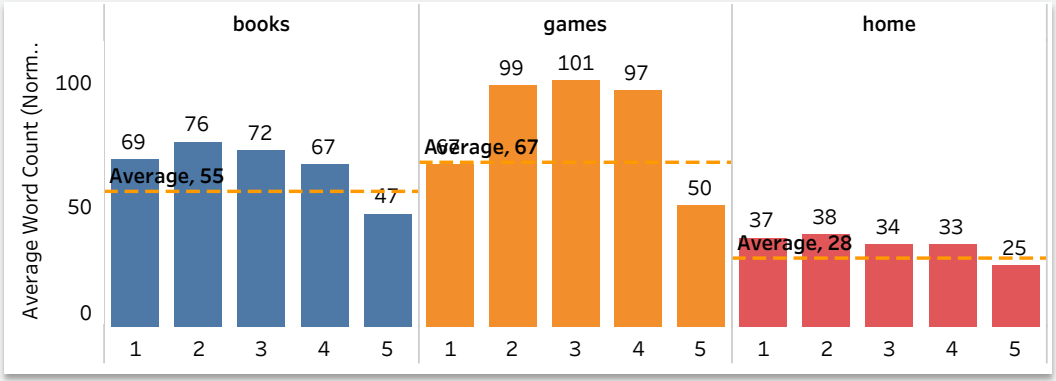
Key Takeaway

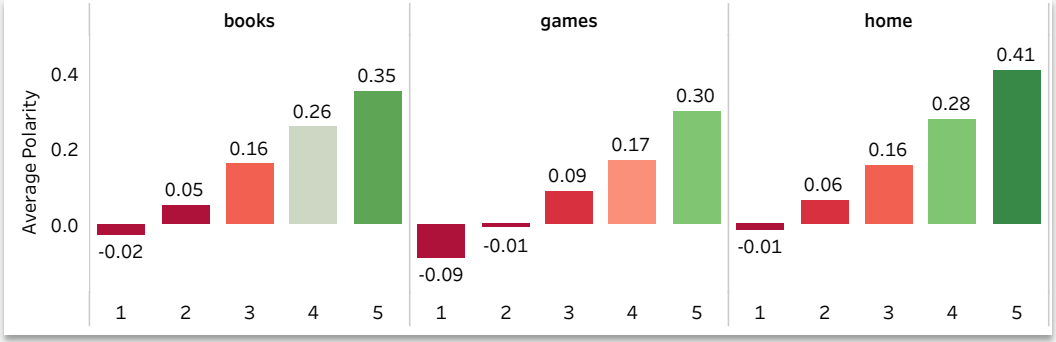
- Book reviewers write a fair amount and are **neutral** in their reviews relative to Gamers and Home reviewers

- Gamers appear to be write the most and are more **critical** in their reviews

- Home reviewers write shorter reviews and are the **most positive**

Question 2 and 3: Habits of reviewers, and do they change by category?





Key Takeaway

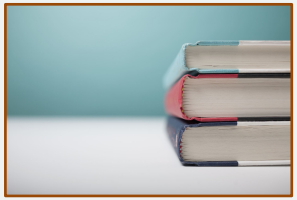
- Game reviews have the highest average word count, but less distinct words compared to other categories
- Game reviews tend to skew more negative in sentiment for all ratings when compared to the other categories
- Home reviews have the lowest average word count and contain more distinct words

Sponsored ⓘ

Home

Baking/Cooking: Bread, scale, bake, loaf, grind, grinder, weigh, poster, bean, flour
Mixer, beater, love, dough, refreshing, five, get, kitchen, book, star
Tea/Coffee: Pot, tea, water, lid make, use, cup, like, one size

---



Books

Content: Link, book, class, hook, product, war, data, use, normal, read, story
First Impression: Excellent book, Excellent read, Good Product, amazing story, first use, able to put, Great way, Alex Cross

---

Games

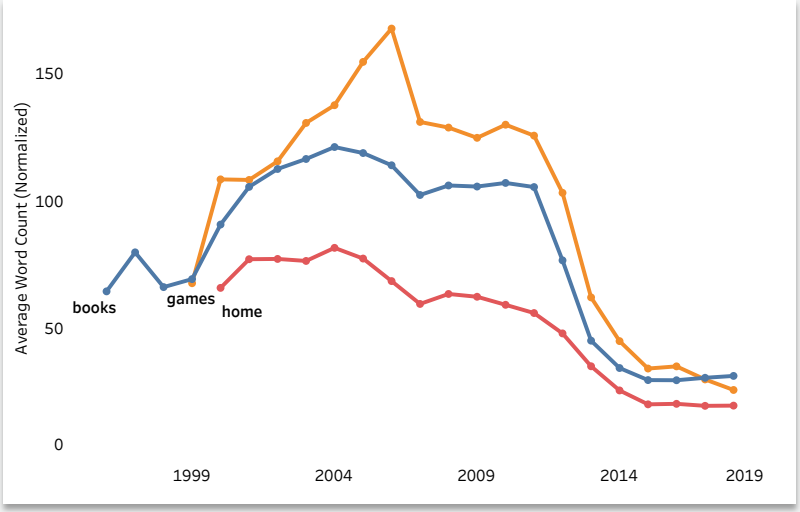
Console and Accessories: Controller, game, headset, cable, Xbox, resident, console, pc, play, headphone, Mass Effect, piece of junk

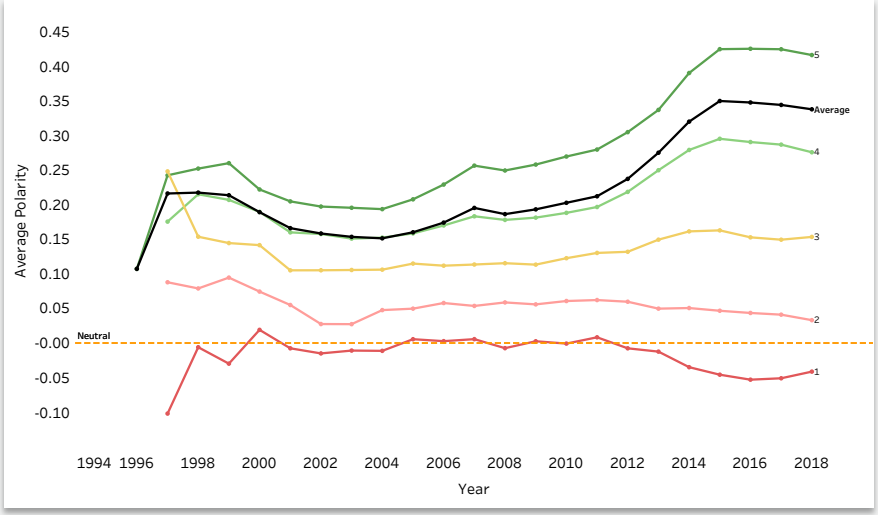
---

Key Takeaway

- **Tea/Coffee and Baking/Cooking are clear Topics in the Home area**
- **Gamers are critical reviewers: “piece of junk” likely referring to gaming accessories or the Xbox console itself**

Question 4: Do customer reviewing habits change over time? Has the overall content of reviews changed over time?





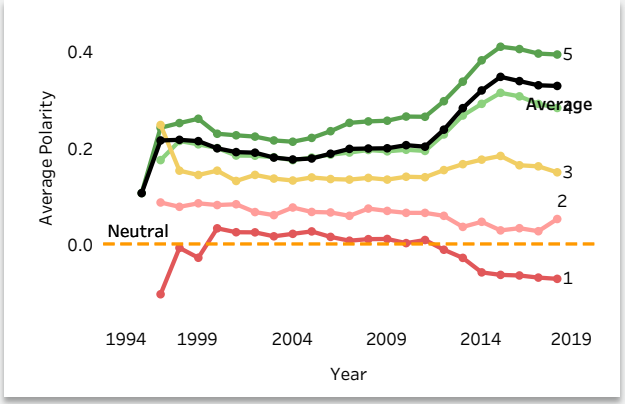
Key Takeaway

- Reviews have been getting shorter since 2012, largely driven by non-useful reviews
- Customer sentiment has also evolved over time, with 4 and 5-star reviews getting increasingly positive, again driven by non-useful reviews

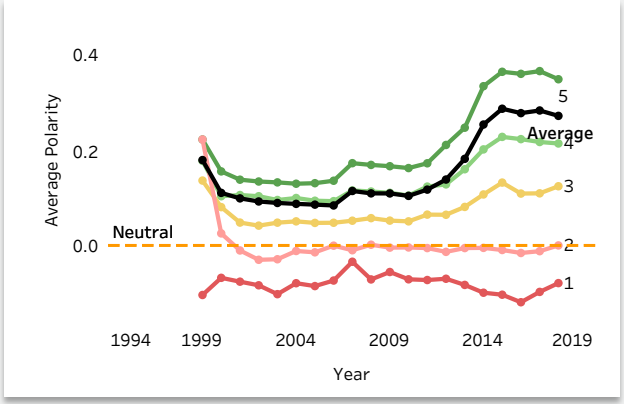
Question 4: Do customer reviewing habits change over time? Has the overall content of reviews changed over time?

What customers are **feeling** in their reviews...

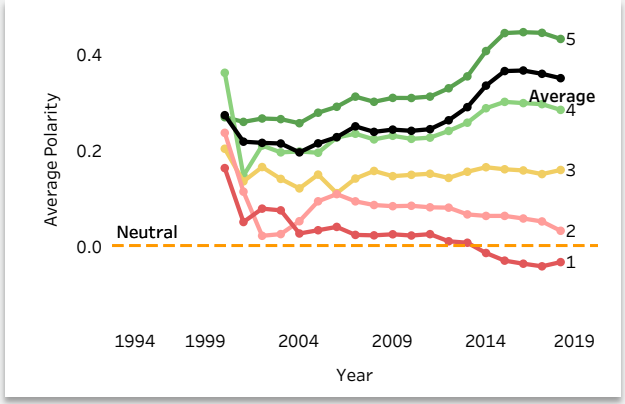
Books



Games



Home



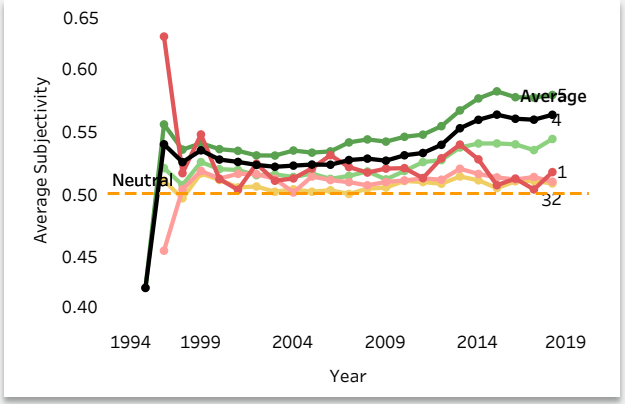
Key Takeaway

- **After 2014, Polarity stabilized across ratings**
- **The 3-star review sentiment in Games is trending positively; these reviews are positive, objective, and fair**

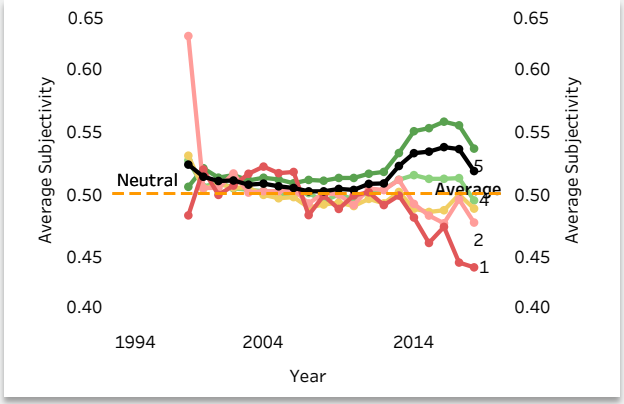
Question 4: Do customer reviewing habits change over time? Has the overall content of reviews changed over time?

Are the customer reviews more **opinion-based** or **factual-based**?

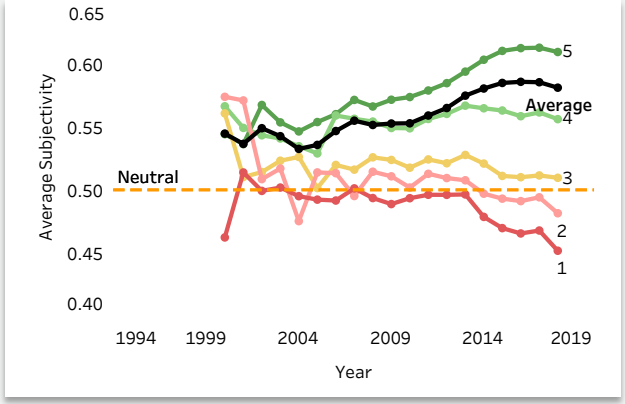
Books



Games



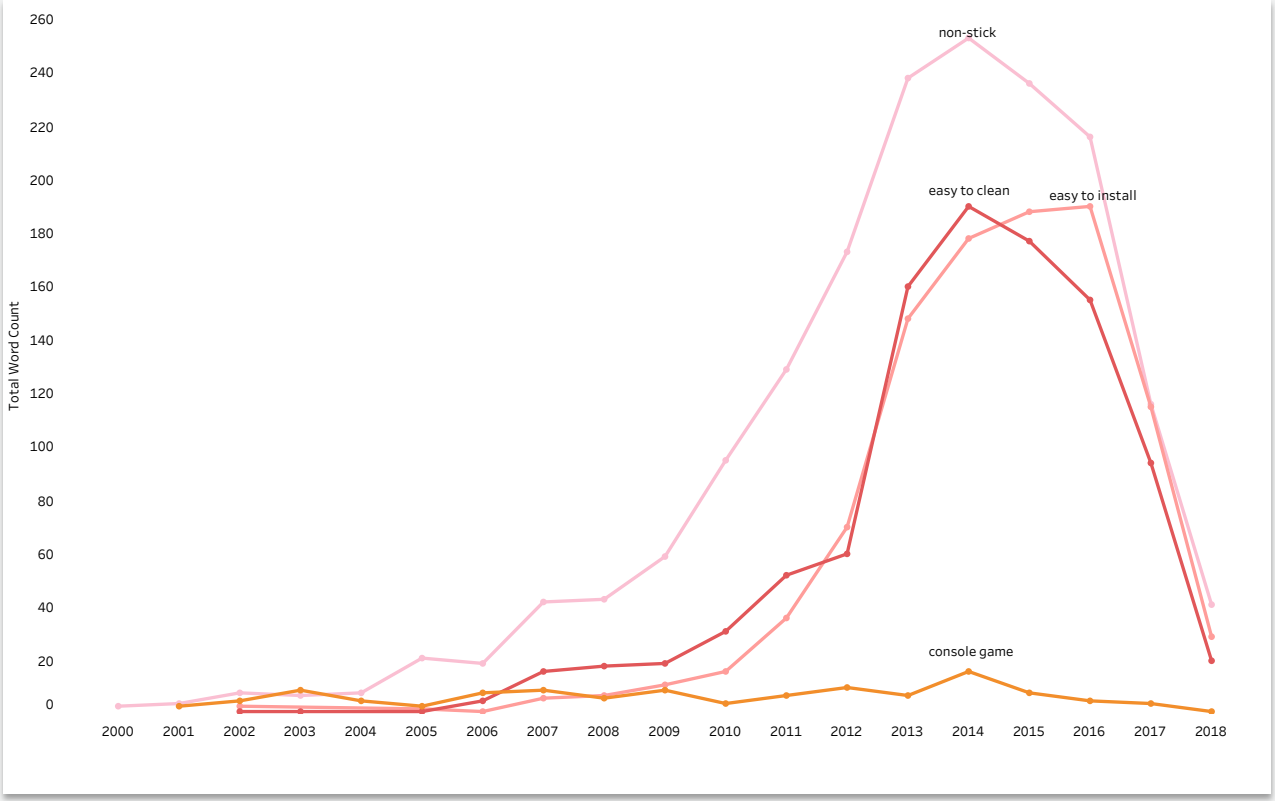
Home



Key Takeaway

- **Overall subjectivity for Books and Home are becoming more subjective over time**
- **Game reviews tend to be more neutral**
- **Home reviews are becoming more differentiated over time, 5-star reviews are more subjective, while 1-star reviews are more objective**

Question 4: Do customer reviewing habits change over time? Has the overall content of reviews changed over time?



Key Takeaway

- The term ‘Non-Stick’ and ‘Easy to Clean’ peaked in 2014, likely referring to a craze for frying pans
- The term ‘easy to install’ peaked in 2016, likely referring to a home product
- The term ‘console game’ peaked in 2014 but has trended downward since then. The rise of PC (Master Race) gaming could have impacted the term’s frequency

Latent Dirichlet Allocation (LDA) algorithm for topic modelling

- Ran two different ways: one for visuals and one for accuracy of the modelling (using 40% of sample)
- Set K = 50 (number of topics)

Model for Visuals:

- Used the the CountVectorizer for the option to pull words/terms (from lemmatized terms) to be able to create the visuals for the topic modelling
- CountVectorizer runs slower, so only used for visuals, set max terms to 2,000

Model for Kaggle:

- Used TFHasher for the modelling, since it was runs quicker because it doesn't convert to text (numeric), also performed better (perplexity score was lower)
- Top Best Topics under Home, Games, Books: there are some clear "topics" from outcome of the LDA (lemmatized version)

Some Limitations:

- Ran LDA with many different parameters: with more than 50 topics (80-100) , >40% of dataset, and with more iterations however runtime was too slow
- Could not run NMF topic modelling because of 'newness' of the software (doesn't exist) in PySpark
- Not as much availability to functionality in LDA in PySpark, limitations on publications of examples as well

Choose Text Analytics Modelling Techniques to Return?

- Peeler, cooker, book, shakespeares, game, aqua, series, pollution, make, love
- great story, first game, Stainless Steel, original game, end of the book, Call of Duty, audio version, top rack, first page, Great fun

Key Takeaway

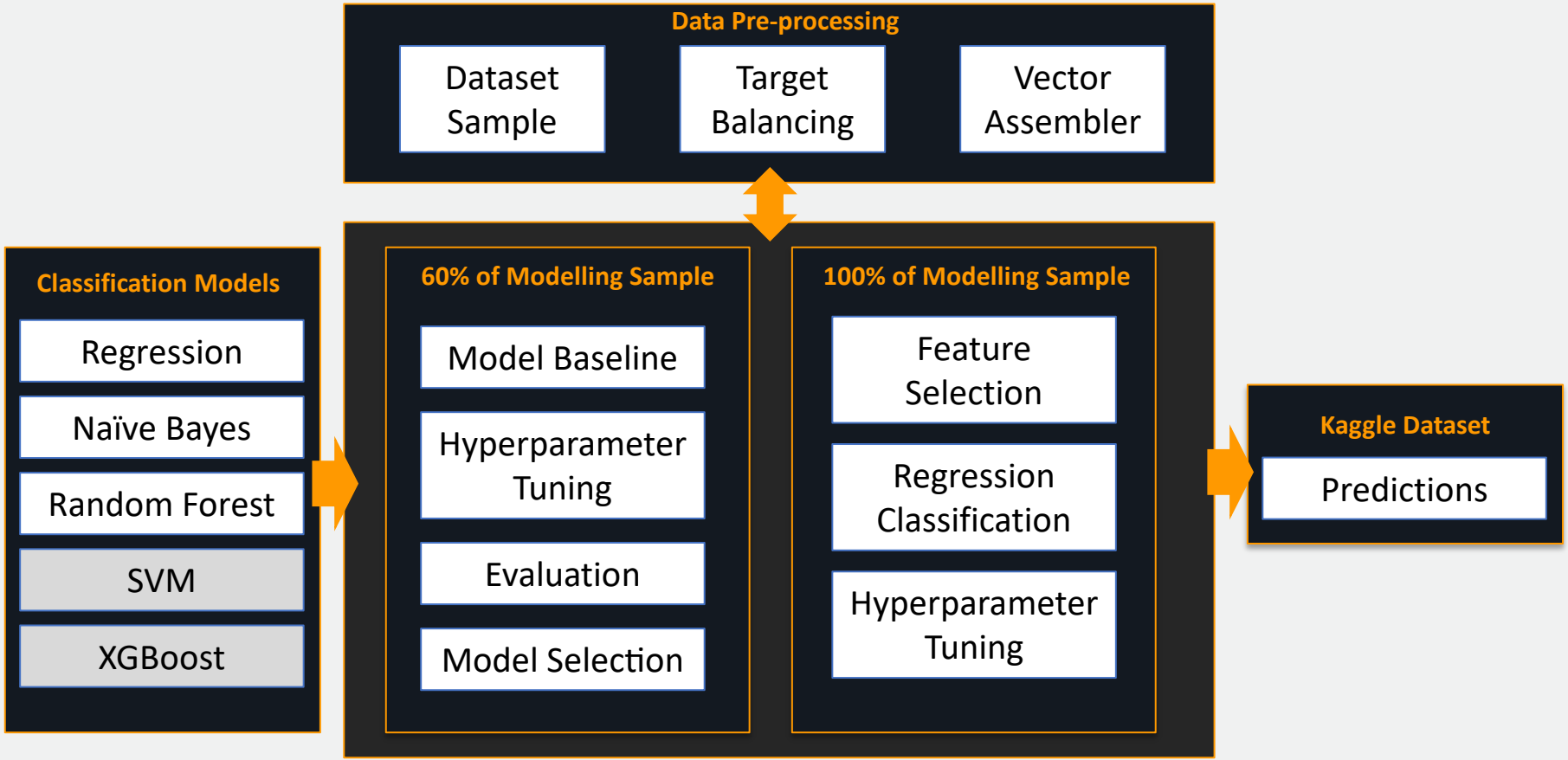
Sometimes LDA doesn't work as expected:

- **What is the topic?**
- **Result topic-terms are a mix of all areas: Home/Games/Books!**
- **Meaningful n-gram resulted in "weaker" topics than the lemmatized terms**

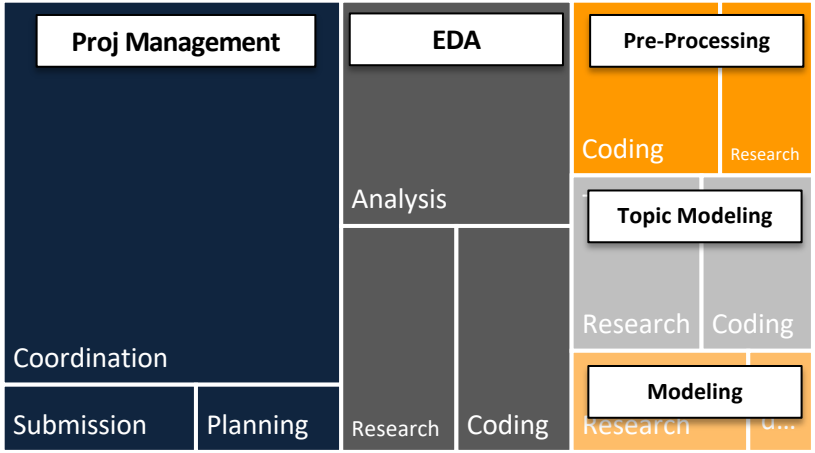
⚠ **This technique is no longer eligible for return.**

The return window closed on Oct 6, 2020.

Continue Presentation


Data Pre-processing
Dataset Sample
Target Balancing
Vector Assembler
Classification Models
Regression
Naïve Bayes
Random Forest
SVM
XGBoost
60% of Modelling Sample
Model Baseline
Hyperparameter Tuning
Evaluation
Model Selection
100% of Modelling Sample
Feature Selection
Regression Classification
Hyperparameter Tuning
Kaggle Dataset
Predictions

Time Distribution



Human Hours

Planning	Exploration (EDA)	Pre-Processing	Topic Modelling	Modeling
104	48	26	26	11

Research	Coding	Analysis
47	48	32

Total 215

databricks D12 V2: 4 vCPU(s), 28GB RAM, 1 DBU ($0.40 per DBU/hrs) + platform charge

- Development $185.57 (241 hrs)
- EDA runtime $1.00 per Jobs - (1 hrs per GB)
- Modeling runtime $12.00 per Jobs (7hrs per GB)

Total Estimated

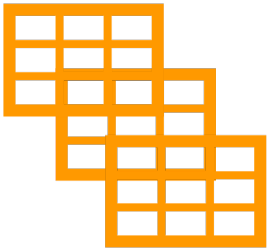
$ 185.57

*Estimated Costs

Key Takeaway

Limitation of PySpark vs. Python Libraries

Dataset Management



Python Visualization Expertise

Notebook, CPU, and Memory Management

amazon

Presentation Confirmation

Presentation #3: Team New York

Hello Steve,

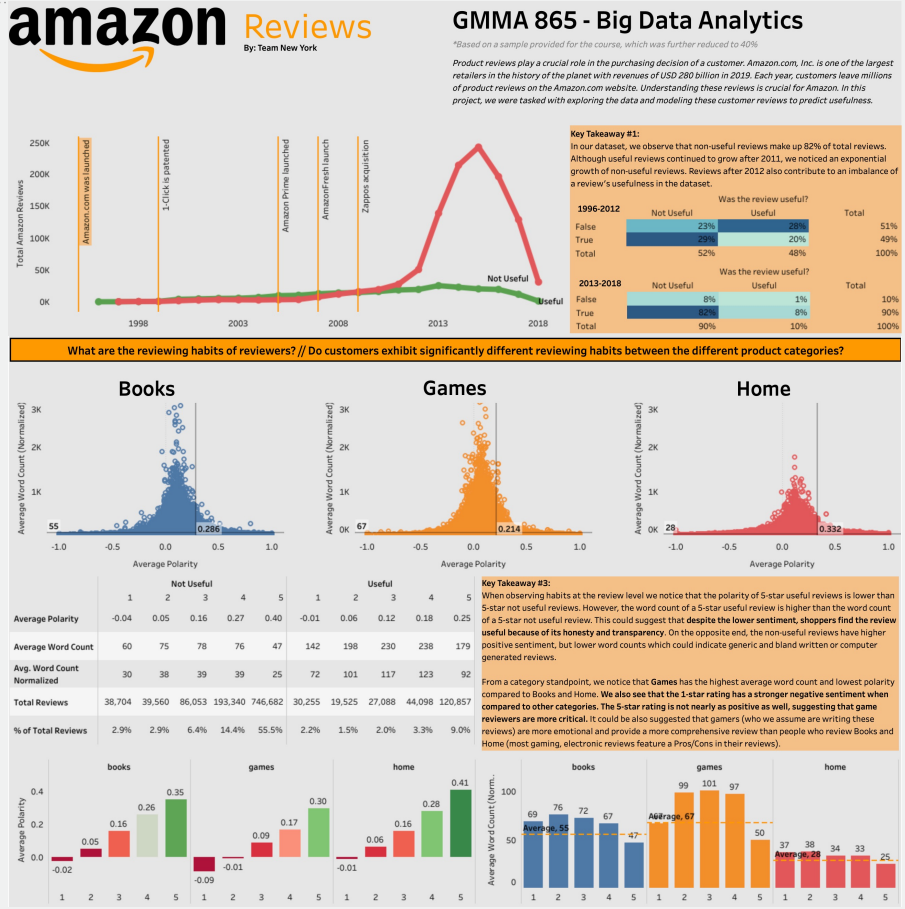
Thank you for your engagement throughout the presentation. We'd be happy to take any questions you may have.

Your estimated presentation date is:
Friday, October 9 12:50-1:15pm ET

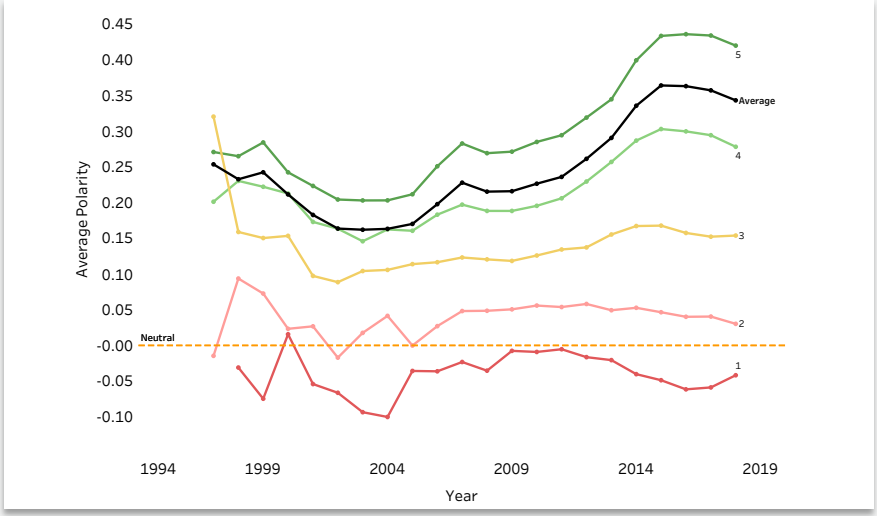
Your shipping speed:
Order Details

Your presentation will be sent to:
Stephen Thomas
Goodes Hall 328B
Kingston, Ontario
Canada

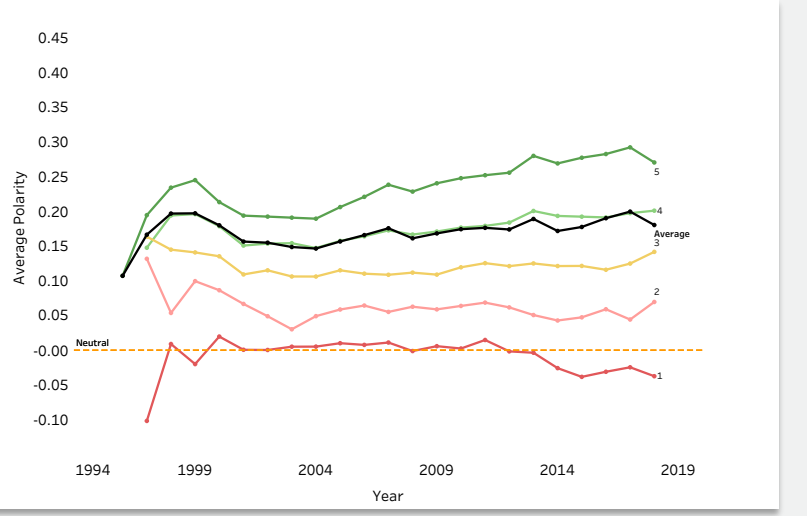
Appendix – Amazon Reviews Dashboard



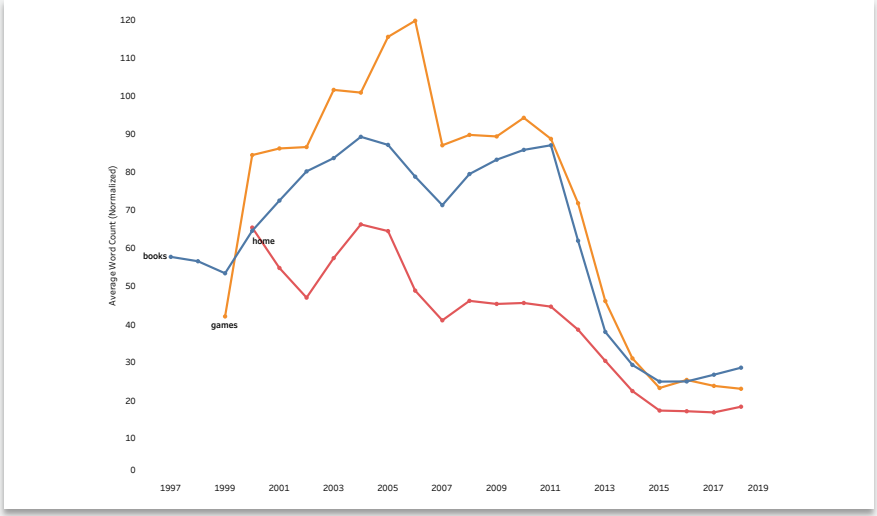
Non-Useful



Useful



Non-Useful



Useful