NOVA IMS Information Management School

BDA Project

Amazon Reviews (2022-23)

Presentation | 27.05.2025
UC | Big Data Analytics
Professor | Márcia L. Baptista (NOVA IMS)

Group W

André Silvestre	20240502
Filipa Pereira	20240509
João Henriques	20240499
Umeima Mahomed	20240543

TP1 & **P2**

DATA COLLECTION & PREPROCESSING

Collection

*Filtering the data collected from **Amazon Reviews's 23** dataset collected in **2023** by **McAuley Lab***

amazon UCSD

JSON

ETL

Data Pre-Processing

- Deduplicate
- Convert to the correct data types
- Clear timestamp

PySpark databricks

TXT

EDA

Exploratory Analysis & Verify Collection Quality

Visualize and Analyse the data obtained

Spark SQL seaborn Matplotlib

ANALYSIS | TEXT MINING

Text Mining

- *Tokenization*
- *Stop Words Removal*
- *N-Grams*

NLTK

word cloud

Spark MLlib

Sentiment Analysis

Transformers

Text Sentiment Classification

- ***citizenlab**/twitter-xlm-roberta-base-sentiment-finetunned*
- ***MoritzLaurer**/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7*

Topic Analysis

Transformers

Classification of Text Topics

- ***MoritzLaurer**/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7*

+ EXTRA

Clustering

K-Means

Spark MLlib

Graph Analysis

Graph Analysis **User-Product**

GraphX Vertices: Edges: Triplets: APACHE Spark

Streaming Simulation

Batch and **Advanced** Streaming Simulation in databricks

PySpark LIVE STREAMING

Figure 1 | Amazon box.
Source: Google Images

Figure 2 | Amazon headquarters
Source : SustainableJapan

Objective

Modelling Amazon Tech Reviews for Consumer Insight

Analysis of Amazon Electronics and Computers reviews (2022–2023) using Sentiment and Topic Modelling, Clustering, and Graph Analysis to uncover Consumer Purchasing Behaviour.

Project Management
Control Instruments

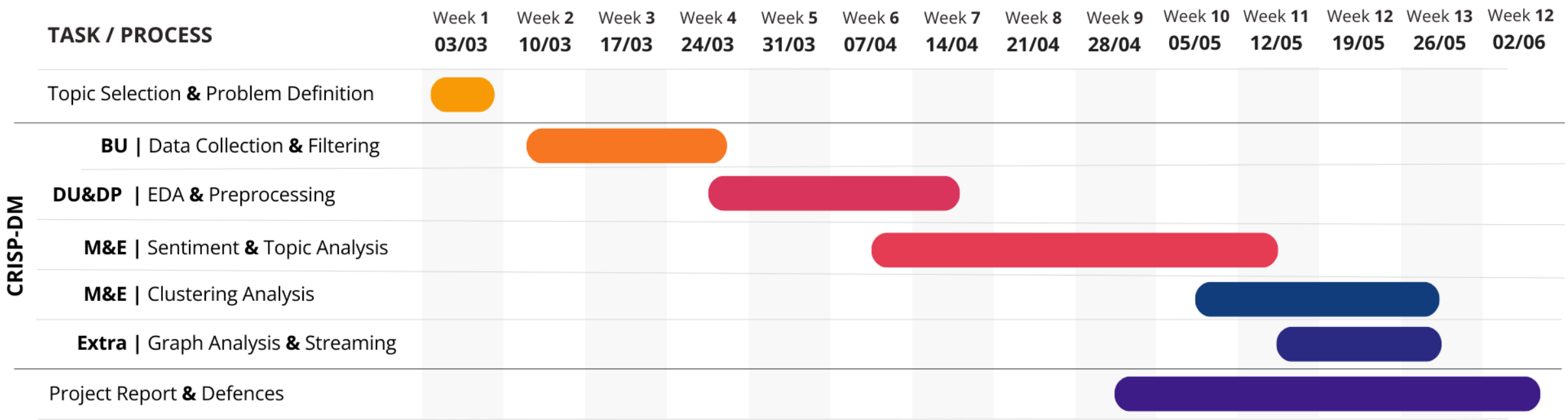
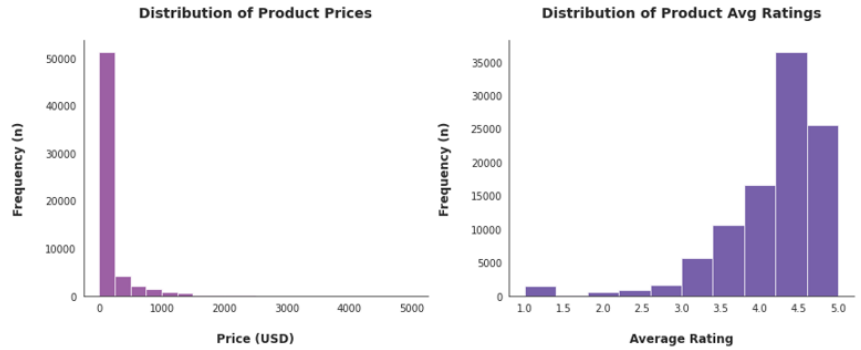


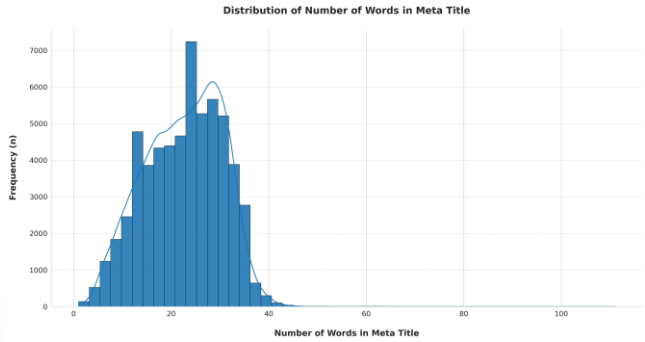
Figure 1 | Gantt Chart with Task Distribution.

NOVA IMS Information Management School

EDA

Products



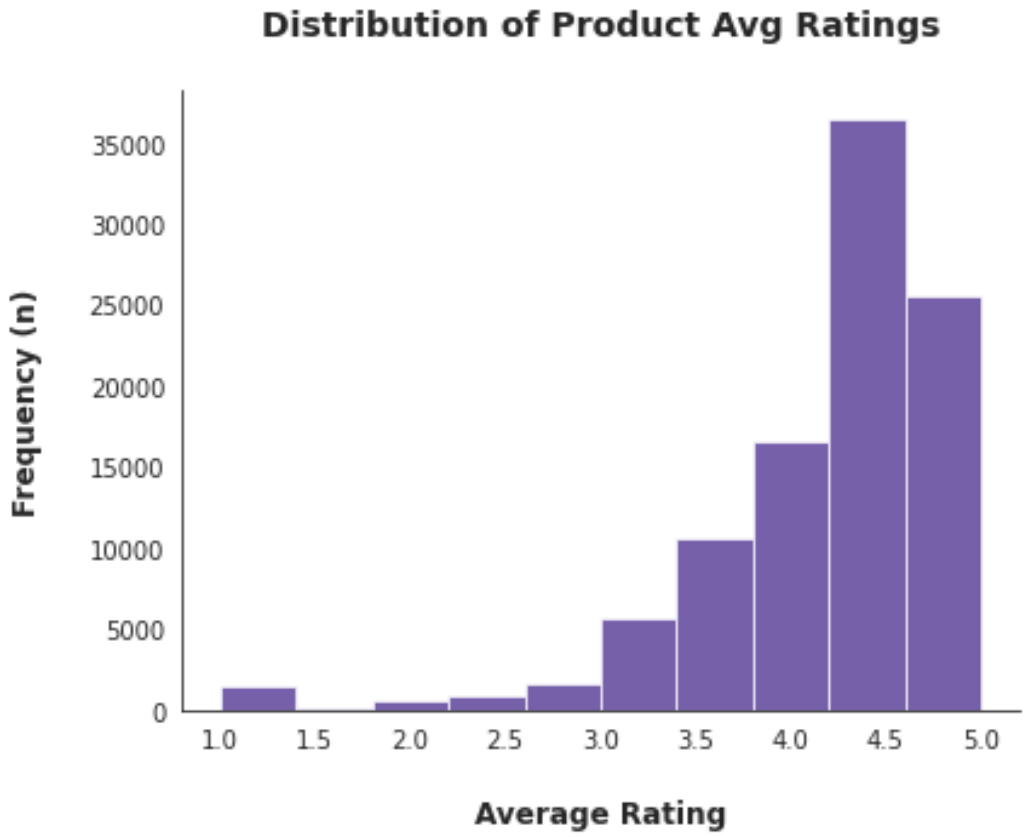


Store	n	%
Amazon Renewed	6402	6.4
HP	3699	3.7
Lenovo	2040	2.0
ASUS	1886	1.9
Dell	1444	1.4
Generic	948	1.0
SAMSUNG	830	0.8
MOSISO	744	0.7
SanDisk	725	0.7
Logitech	473	0.5

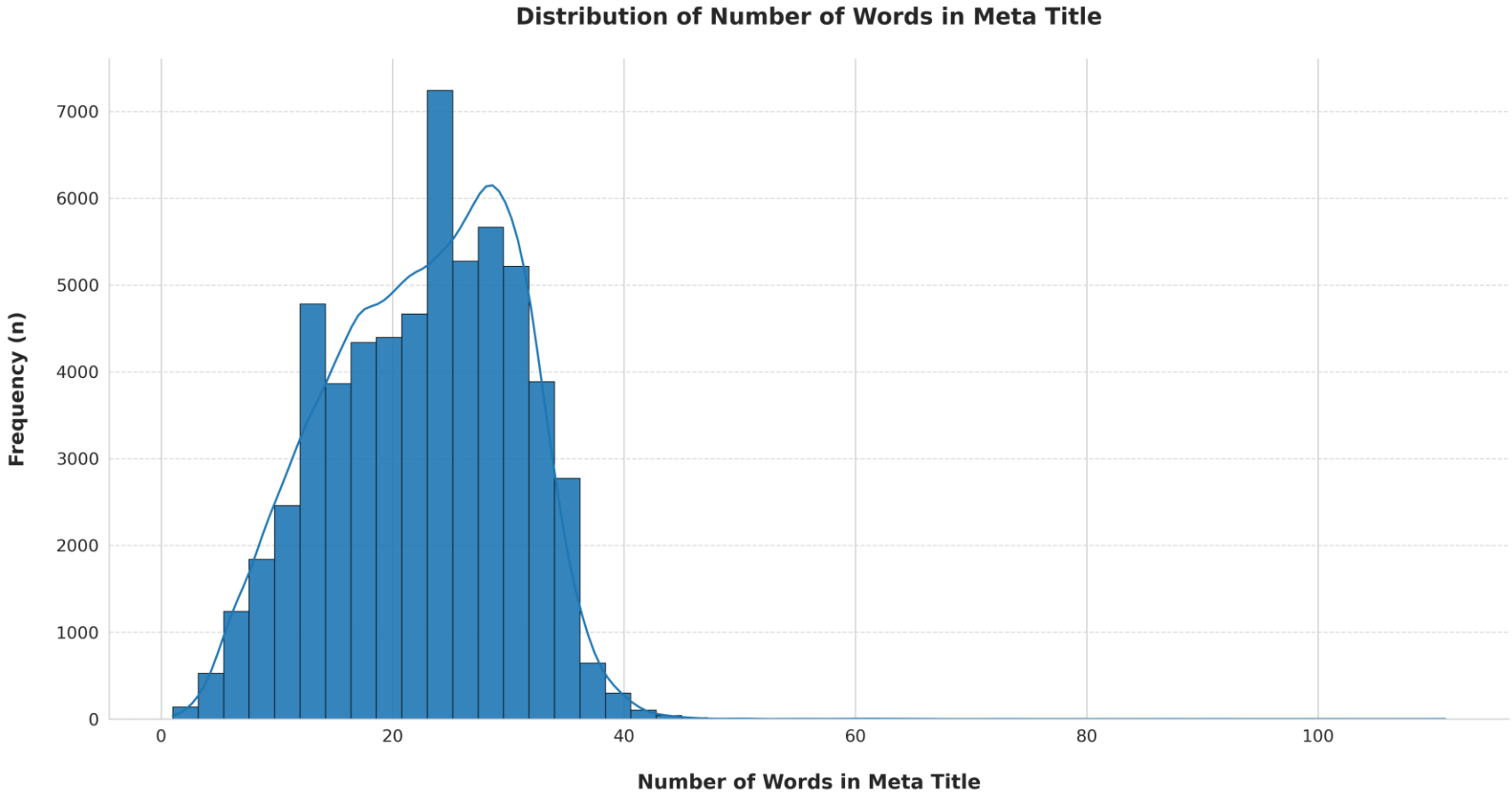
Price Category	n	%	Cumulative %
0 - 10	8774	8.8	8.8
10 - 50	28197	28.3	37.1
50 - 100	6958	7.0	44.1
100 - 250	7332	7.4	51.5
250 - 500	4205	4.2	55.7
500+	5612	5.6	61.3
Unknown	38557	38.7	100.0

Products

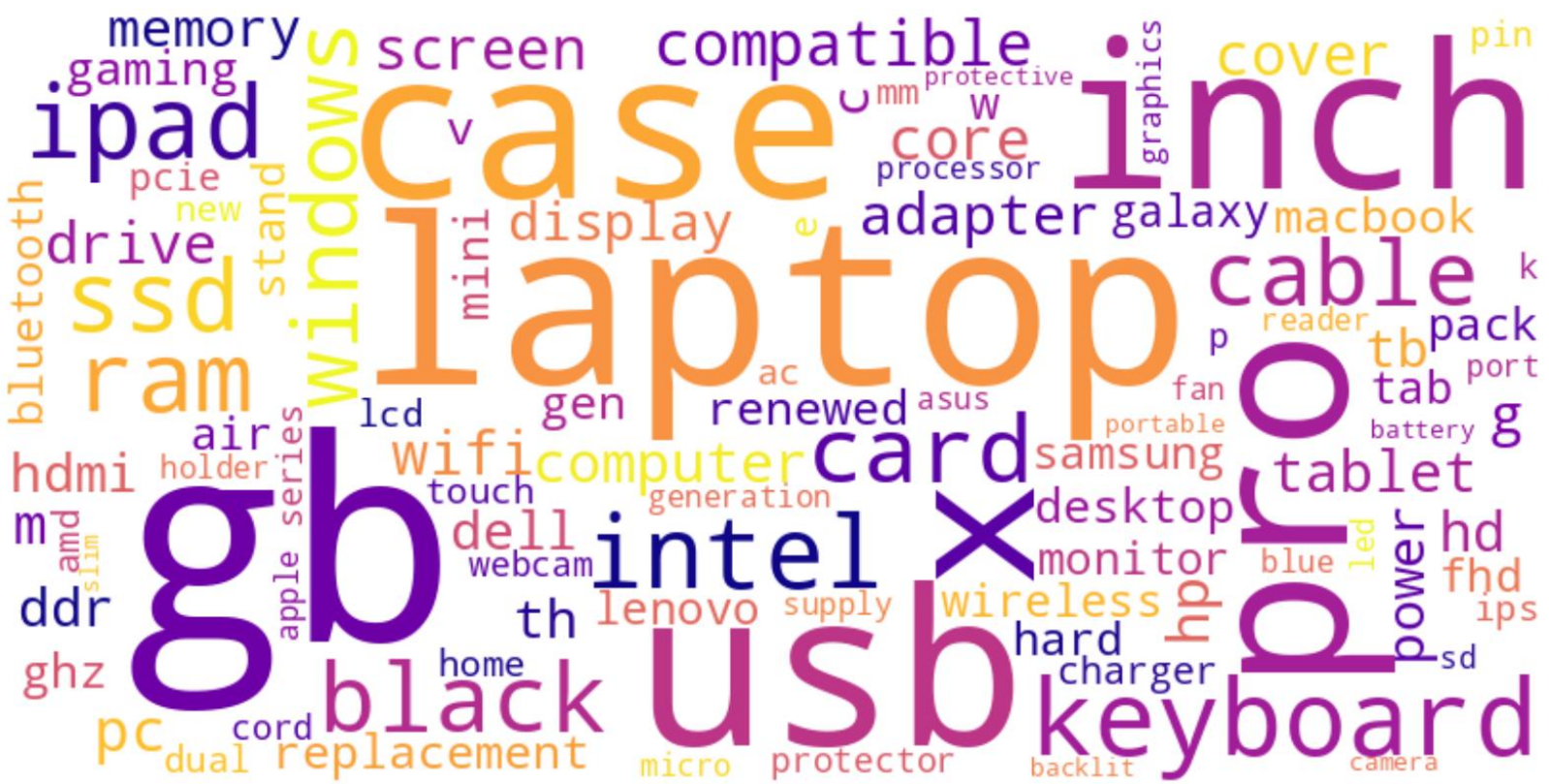

Distribution of Product Prices
Frequency (n)
50000
40000
30000
20000
10000
0
0
1000
2000
3000
4000
5000
Price (USD)


Distribution of Product Avg Ratings
Frequency (n)
35000
30000
25000
20000
15000
10000
5000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Average Rating

Products


Distribution of Number of Words in Meta Title
Frequency (n)
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
Number of Words in Meta Title

Products



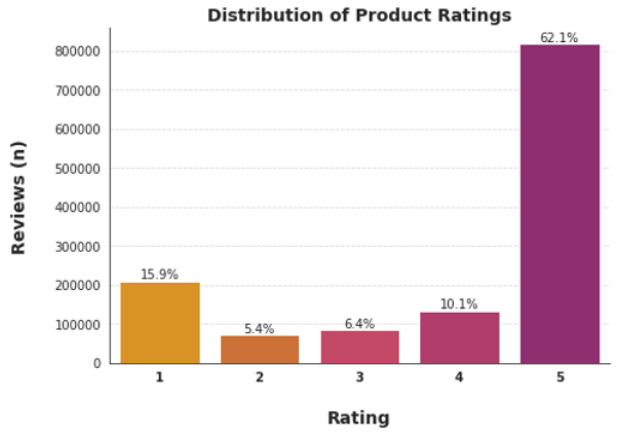
Unigrams

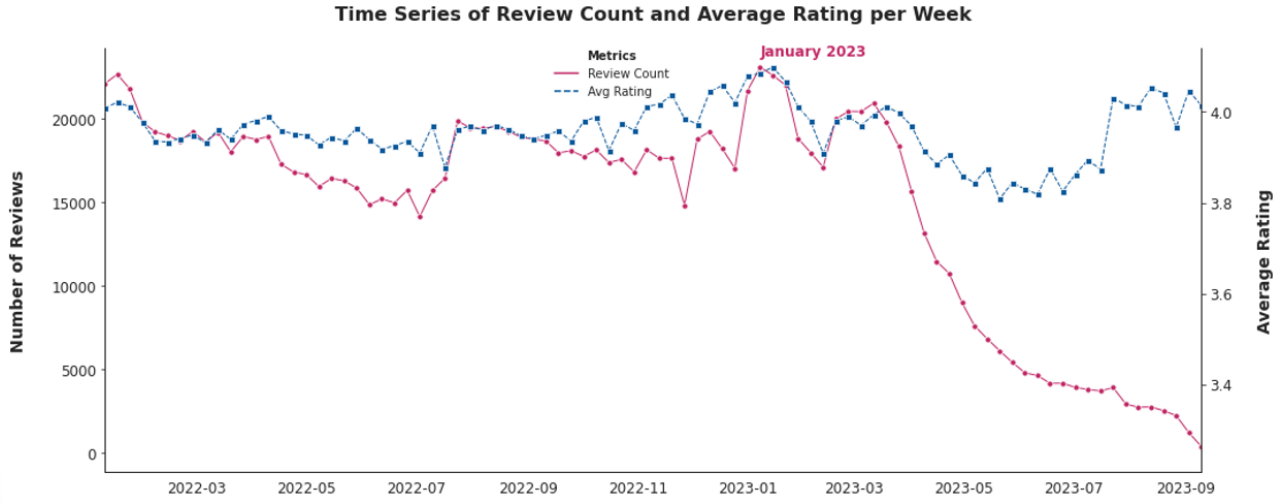
Products

Store	n	%
Amazon Renewed	6402	6.4
HP	3699	3.7
Lenovo	2040	2.0
ASUS	1886	1.9
Dell	1444	1.4
Generic	948	1.0
SAMSUNG	830	0.8
MOSISO	744	0.7
SanDisk	725	0.7
Logitech	473	0.5

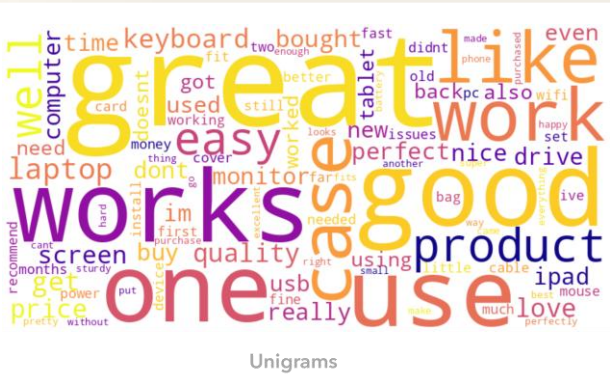
Price Category	n	%	Cumulative %
0 - 10	8774	8.8	8.8
10 - 50	28197	28.3	37.1
50 - 100	6958	7.0	44.1
100 - 250	7332	7.4	51.5
250 - 500	4205	4.2	55.7
500+	5612	5.6	61.3
Unknown	38557	38.7	100.0

Reviews

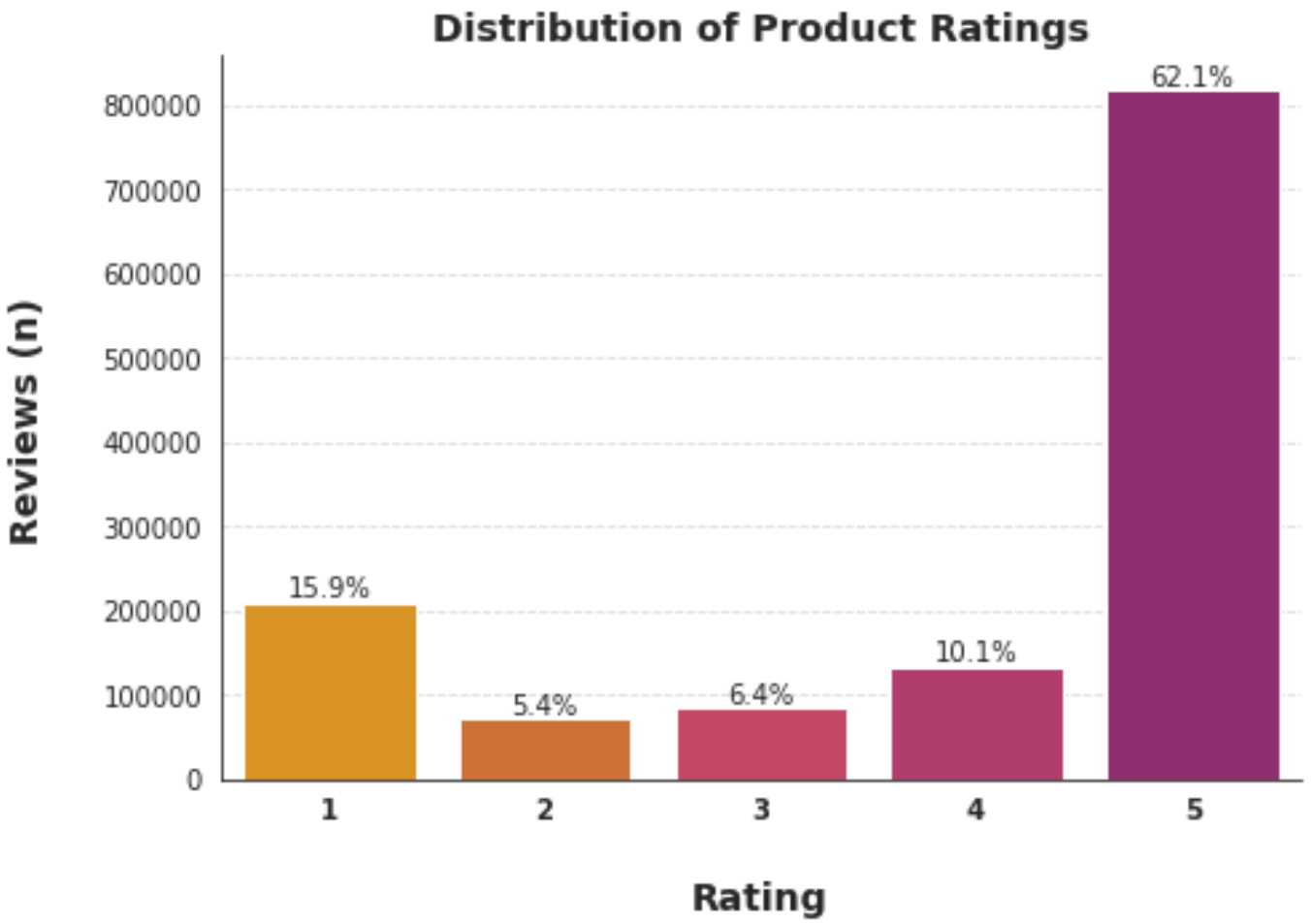




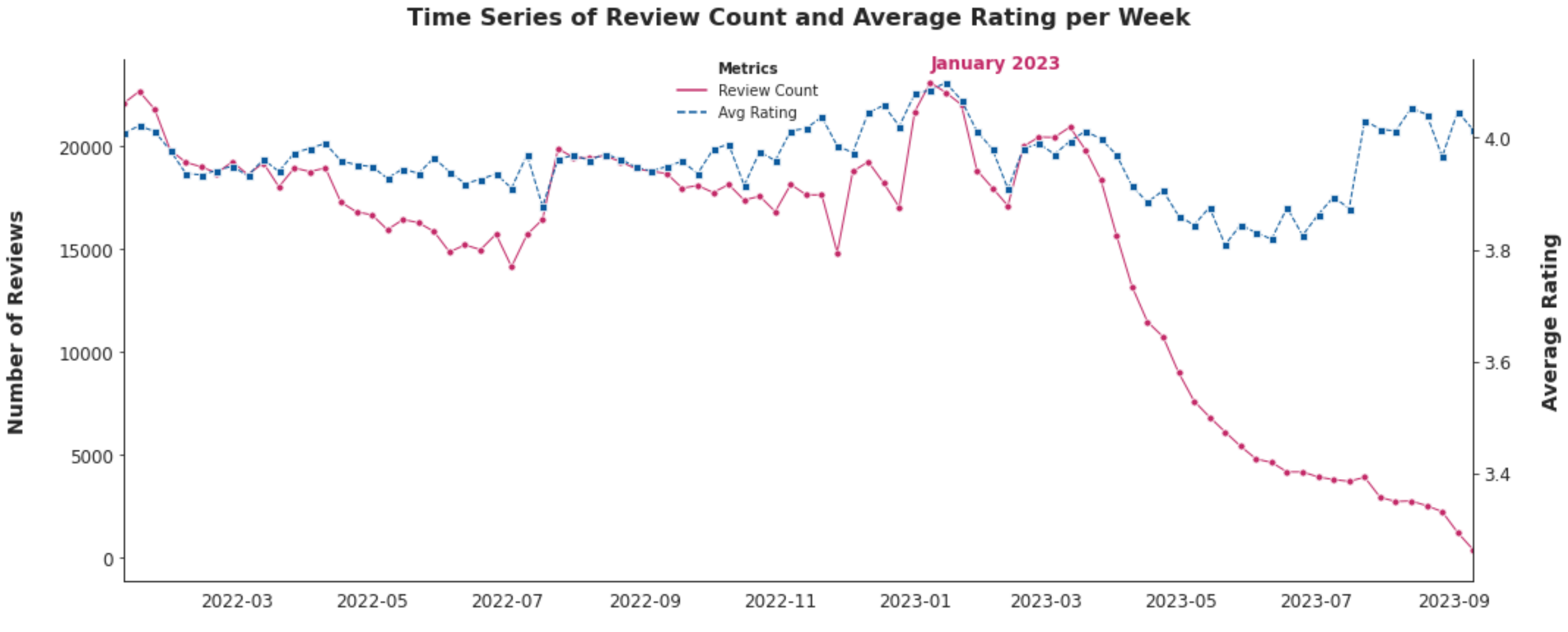
Verified Purchase	n	%	Cumulative %
True	1226888	93.21	93.21
False	89405	6.79	100.00



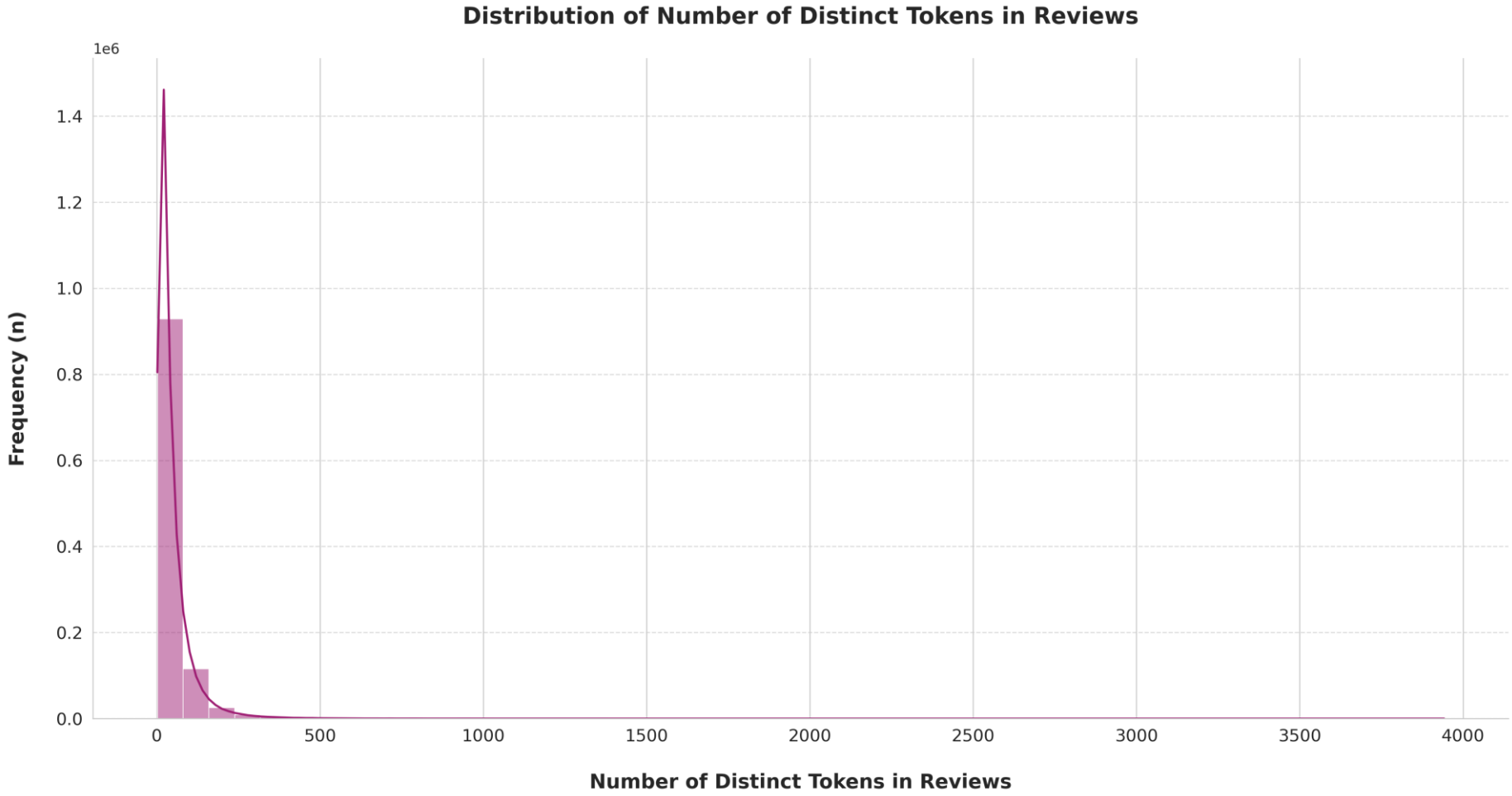
Reviews


Distribution of Product Ratings
Reviews (n)
800000
700000
600000
500000
400000
300000
200000
100000
0
15.9%
5.4%
6.4%
10.1%
62.1%
1
2
3
4
5
Rating

Reviews

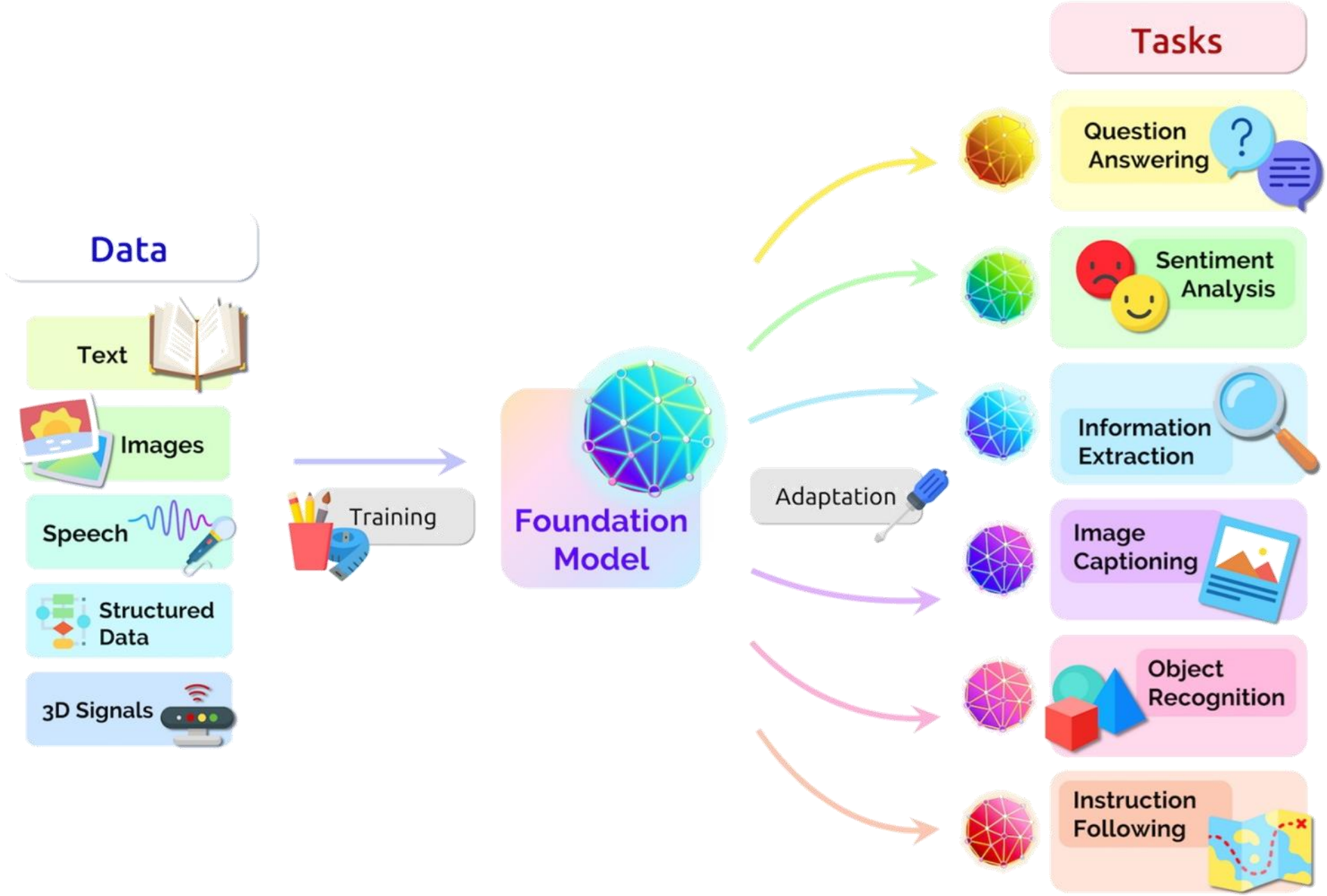

Time Series of Review Count and Average Rating per Week
Metrics
Review Count
Avg Rating
January 2023
Number of Reviews
Average Rating
0
5000
10000
15000
20000
3.4
3.6
3.8
4.0
2022-03
2022-05
2022-07
2022-09
2022-11
2023-01
2023-03
2023-05
2023-07
2023-09

Reviews


Distribution of Number of Distinct Tokens in Reviews
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency (n)
0
500
1000
1500
2000
2500
3000
3500
4000
Number of Distinct Tokens in Reviews

Reviews

Unigrams

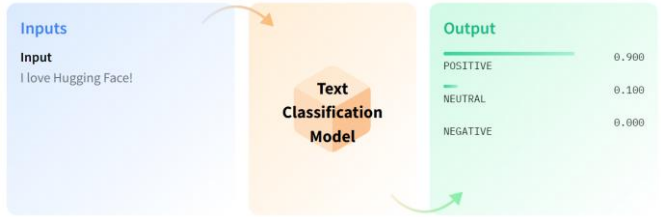


Source: NVIDIA (2022)

Hugging Face

Text Classification

Task of assigning a label or class to a given text, based on **predefined classes trained by the model.**



citizenlab/**twitter-xlm-roberta-base-sentiment-finetunned**

facebook + Google

- It is an **MLLM** (Multi-Lingual Language Model)
- Trained with **2.5TB** of *CommonCrawl* data filtered
- Adjusted in **~58M *tweets*** for *Sentiment Analysis*

Source: Cardiff NLP Group (2020)

Zero-Shot Classification

A task in which a model is trained on a set of labelled examples but is **then able to classify new examples of previously unseen classes**.

MoritzLaurer/**mDeBERTa-v3-base-xnli-multilingual-nli-2mil7**

Microsoft + facebook + Google

- It is an **MLLM** (Multi-Lingual Language Model) trained by *Microsoft* in +100 languages.
- Trained with **2.7M** *hypothesis-premise* pairs

Source: Microsoft (2023)

Sentiment Analysis

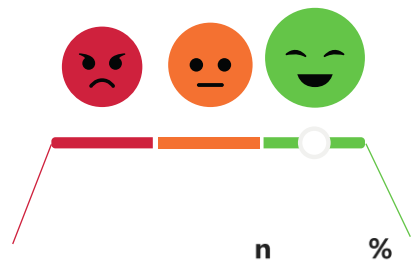
Amazon Reviews

NOVA IMS Information Management School

Hugging Face

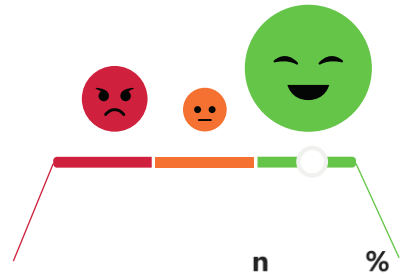
> **Beautiful, I really recommend it I liked everything no complains and I bought it for my son**
>
> *Amazon review*

citizenlab/twitter-xlm-roberta-base-sentiment-finetunned



TXRBSF Sentiment	n	%
Positive	530567	48.88%
Neutral	379199	34.94%
Negative	175665	16.18%

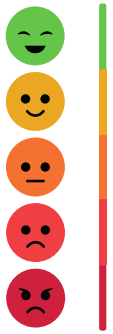
MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7



mDeBERTa Sentiment	n	%
Positive	799639	73.67%
Negative	279410	25.74%
Neutral	6382	0.59%

Classification Matrix of the Results of Both Models

mDeBERTa Model / TXRBSF Model	Positive	Neutral	Negative
Positive	525061	341	5165
Neutral	257855	5453	115891
Negative	16723	588	158354



Combined Sentiment	n	%
Positive	525061	48.37
Positive Tendency	258196	23.79
Neutral	27341	2.52
Negative Tendency	116479	10.73
Negative	158354	14.59

Topic Analysis
Amazon Products

Hugging Face

MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

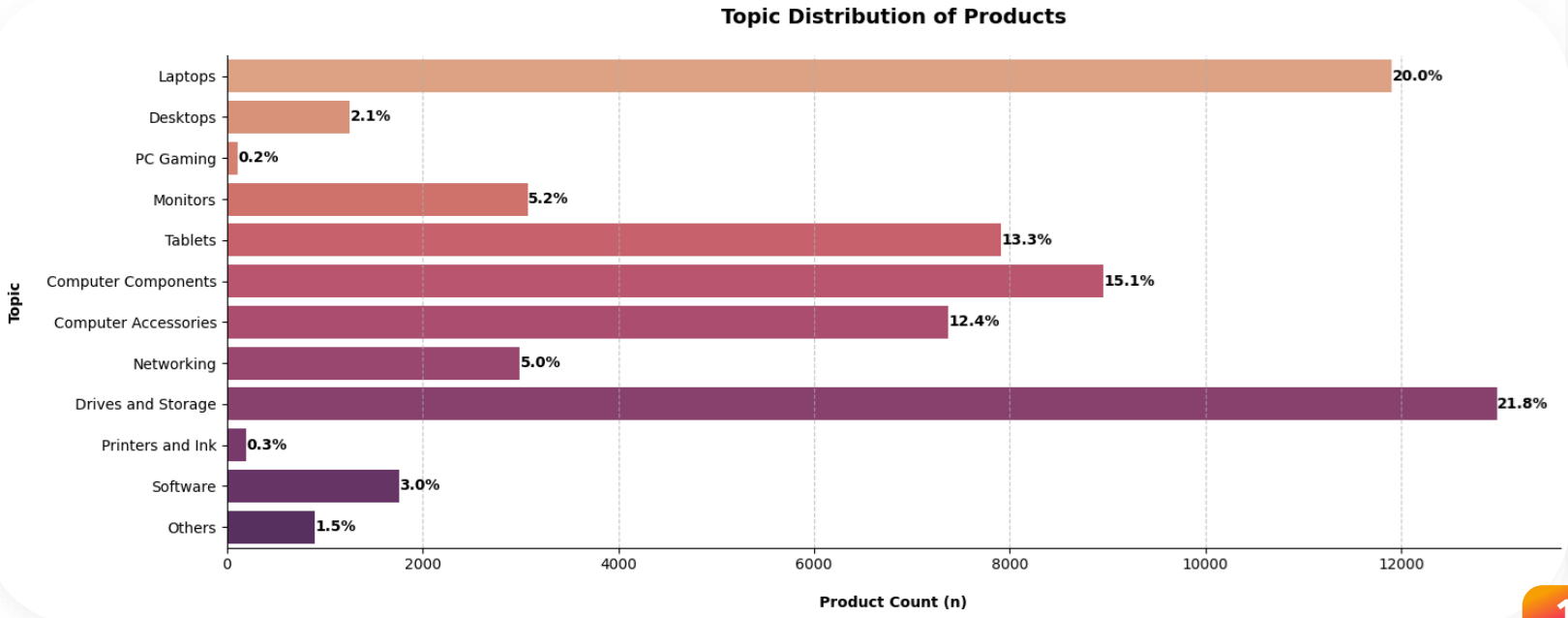
Predefined Topics [12]

- Laptops
- Desktops
- PC Gaming
- Monitors
- Tablets
- Computer Components
- Computer Accessories
- Networking
- Drives & Storage
- Printers & Ink
- Software
- Others

> **AOPEN by Acer 16PM1Q Bbmiuux 15.6" Full HD 1920 x 1080 IPS ...**
>
> *Example of an excerpt of* ***Amazon electronic product***

1st

Monitors
Score: 0.945



Clustering Analysis

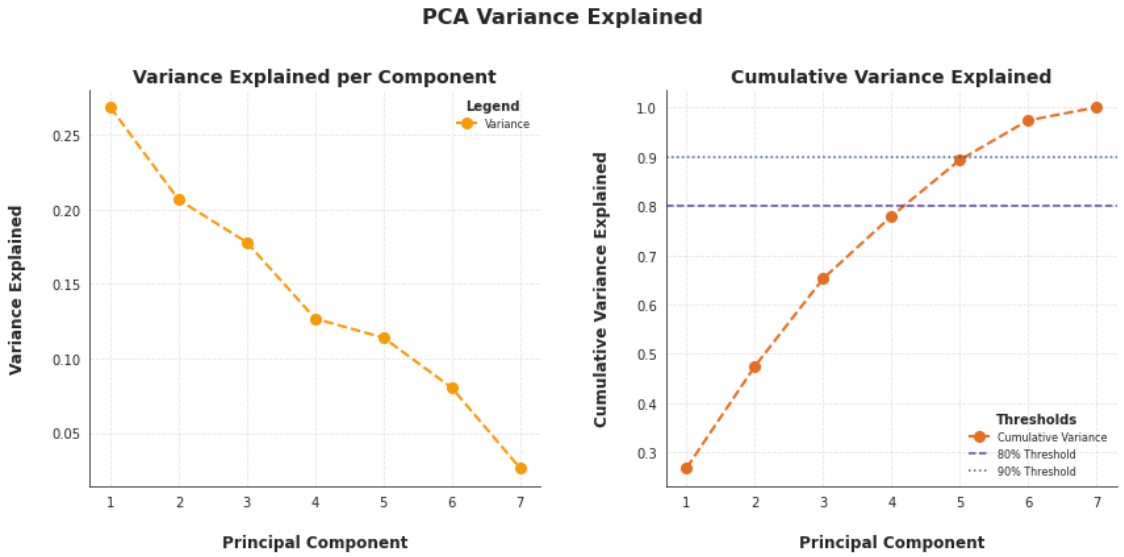
PCA & K-Means

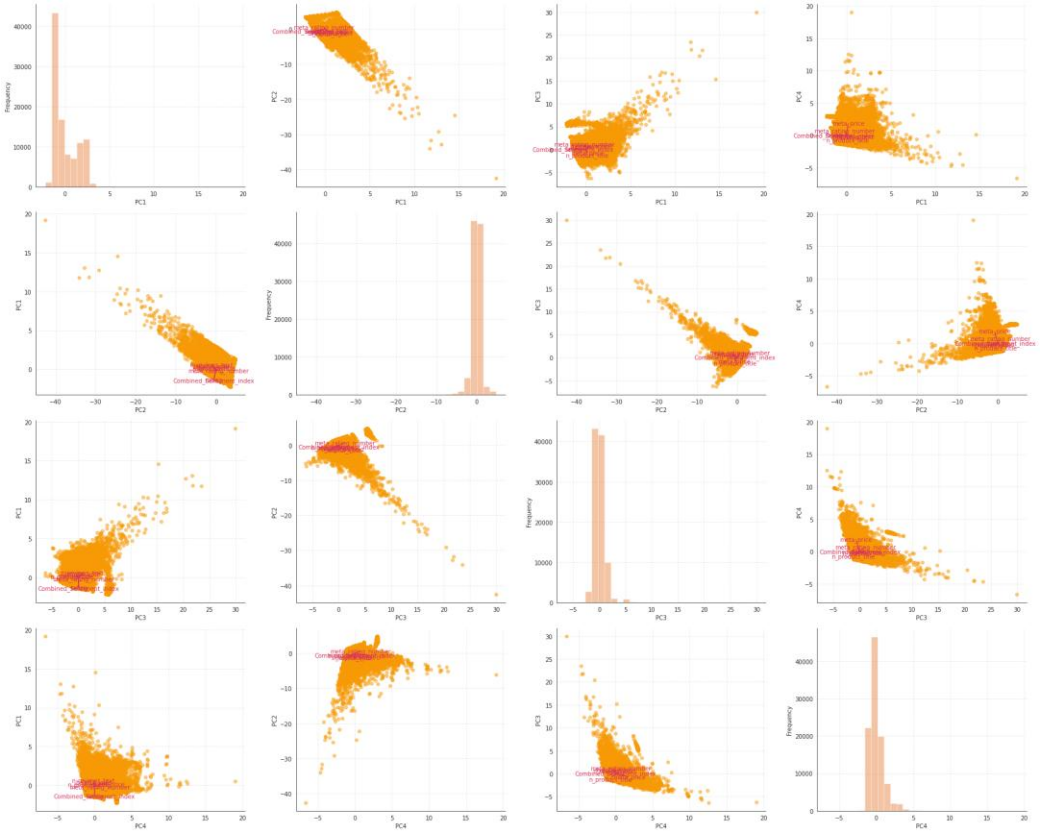
Combined_Sentiment Price Bucket	Negative	Negative Tendency	Neutral	Positive Tendency	Positive
0 - 10	15967 (14.7%)	11152 (10.3%)	2523 (2.3%)	28919 (26.7%)	49877 (46.0%)
10 - 50	83618 (14.1%)	57187 (9.7%)	14362 (2.4%)	137872 (23.3%)	299363 (50.5%)
50 - 100	20333 (15.5%)	15430 (11.7%)	3467 (2.6%)	32399 (24.7%)	59795 (45.5%)
100 - 250	22864 (15.0%)	18923 (12.4%)	4145 (2.7%)	36604 (24.0%)	70155 (45.9%)
250 - 500	9232 (15.2%)	7755 (12.8%)	1571 (2.6%)	13864 (22.9%)	28172 (46.5%)
500+	6340 (15.9%)	6032 (15.1%)	1273 (3.2%)	8538 (21.4%)	17699 (44.4%)

Products: ~10k
Reviews: ~100k

PCA

(Principal Component Analysis) – 7 Features





Clustering Analysis

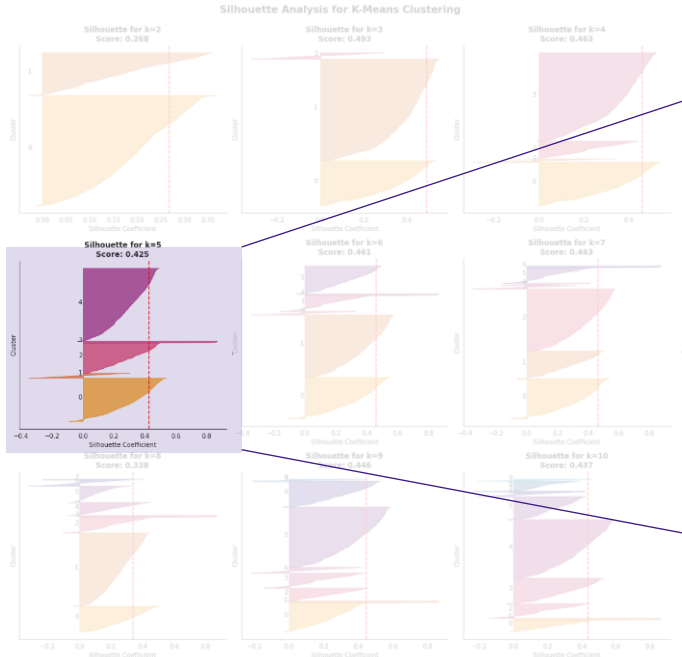
PCA & K-Means

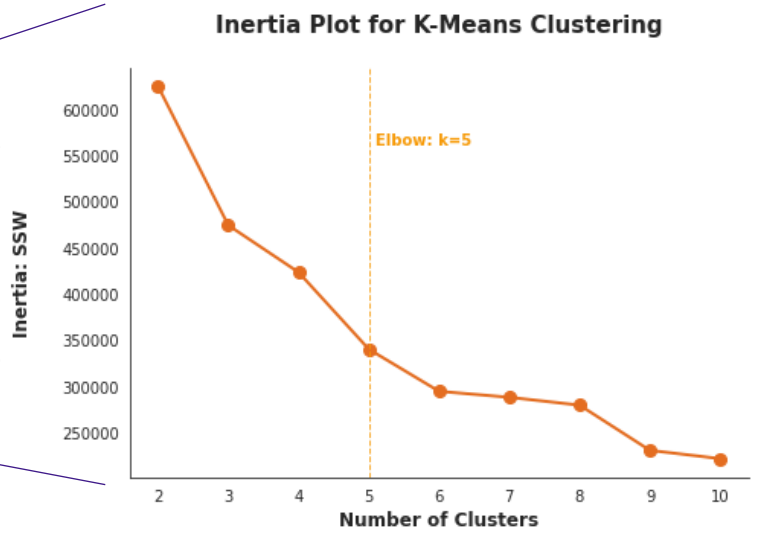
Combined_Sentiment Price Bucket	Negative	Negative Tendency	Neutral	Positive Tendency	Positive
0 - 10	15967 (14.7%)	11152 (10.3%)	2523 (2.3%)	28919 (26.7%)	49877 (46.0%)
10 - 50	83618 (14.1%)	57187 (9.7%)	14362 (2.4%)	137872 (23.3%)	299363 (50.5%)
50 - 100	20333 (15.5%)	15430 (11.7%)	3467 (2.6%)	32399 (24.7%)	59795 (45.5%)
100 - 250	22864 (15.0%)	18923 (12.4%)	4145 (2.7%)	36604 (24.0%)	70155 (45.9%)
250 - 500	9232 (15.2%)	7755 (12.8%)	1571 (2.6%)	13864 (22.9%)	28172 (46.5%)
500+	6340 (15.9%)	6032 (15.1%)	1273 (3.2%)	8538 (21.4%)	17699 (44.4%)

Products: ~10k
Reviews: ~100k

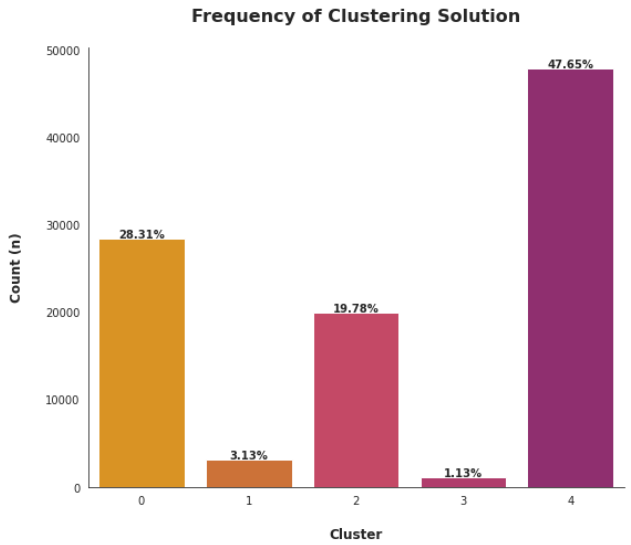
Clustering

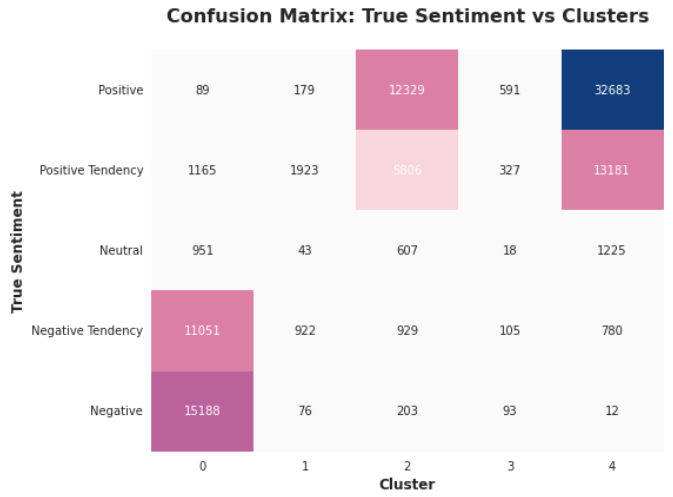
K- Means **&** Original 7 Features





Cluster	n	%
0	28444	28.31
1	3143	3.13
2	19874	19.78
3	1134	1.13
4	47881	47.65



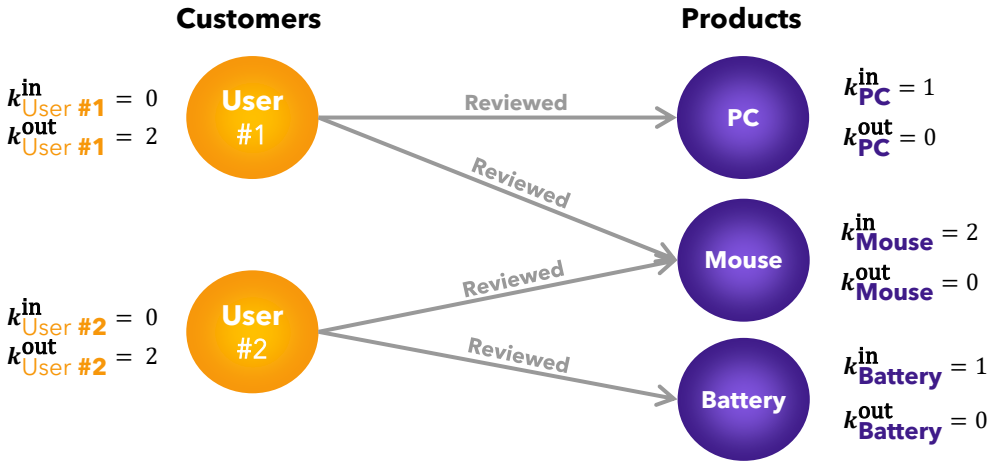


NOVA IMS Information Management School

Combined_Sentiment Price Bucket	Negative	Negative Tendency	Neutral	Positive Tendency	Positive
0 - 10	15967 (14.7%)	11152 (10.3%)	2523 (2.3%)	28919 (26.7%)	49877 (46.0%)
10 - 50	83618 (14.1%)	57187 (9.7%)	14362 (2.4%)	137872 (23.3%)	299363 (50.5%)
50 - 100	20333 (15.5%)	15430 (11.7%)	3467 (2.6%)	32399 (24.7%)	59795 (45.5%)
100 - 250	22864 (15.0%)	18923 (12.4%)	4145 (2.7%)	36604 (24.0%)	70155 (45.9%)
250 - 500	9232 (15.2%)	7755 (12.8%)	1571 (2.6%)	13864 (22.9%)	28172 (46.5%)
500+	6340 (15.9%)	6032 (15.1%)	1273 (3.2%)	8538 (21.4%)	17699 (44.4%)

Products: ~10k
Reviews: ~100k

Graph (GraphFrames)
Users-Products



Graph Created

Graph Analysis

Algorithms Applied

Page Rank

Higher PageRank Scores: For products, indicate popularity or frequent reviews.

Community Detection Label Propagation

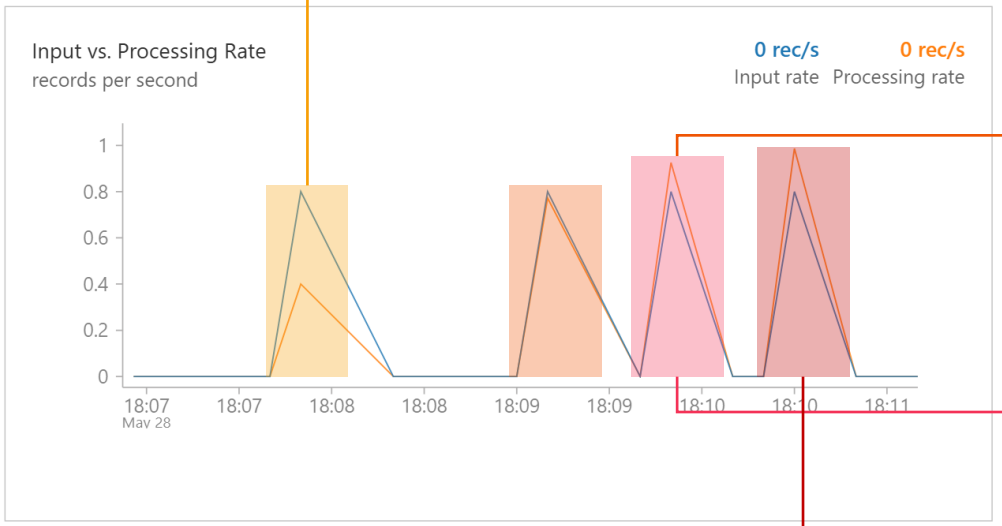
Larger Communities: indicate groups of users frequently reviewing similar high-value products, suggesting shared interests or preferences.

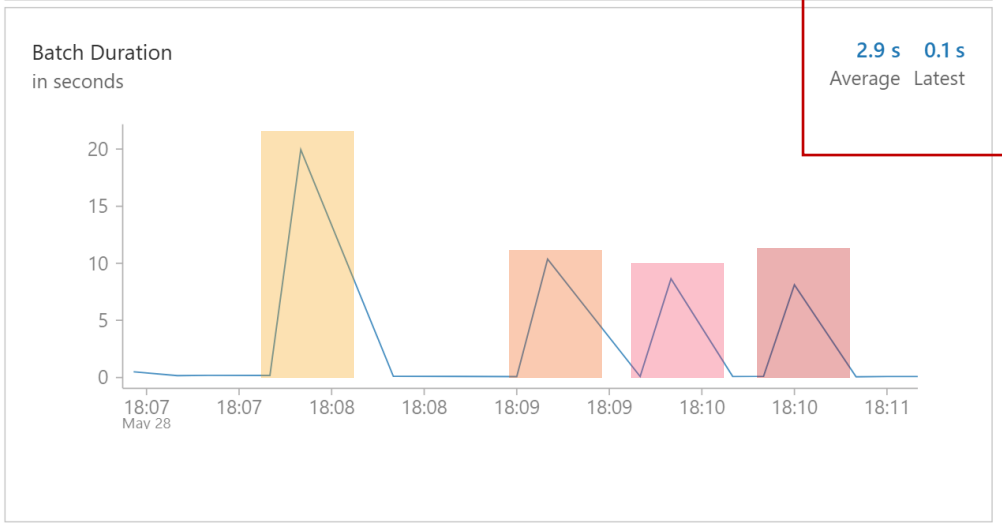
Smaller Communities: may represent niche markets or specialized product categories, which could be targeted for personalized marketing.

Degree Distribution Analysis

In-Degree: Products with high in-degrees are reviewed by many users, indicating popularity or visibility.

Out-Degree: Users with high out-degrees are prolific reviewers, contributing significantly to the review ecosystem.





	review_id	SentimentLabel	SentimentScore
1	new_review_001	Positive	0.9916028380393982
2	new_review_002	Negative	0.9736137986183167

	review_id	SentimentLabel	SentimentScore
1	new_review_001	Positive	0.9916028380393982
2	new_review_002	Negative	0.9736137986183167
3	new_review_005	Positive	0.9478766918182373
4	new_review_006	Negative	0.8762170672416687

	review_id	SentimentLabel	SentimentScore
1	new_review_001	Positive	0.9916028380393982
2	new_review_002	Negative	0.9736137986183167
3	new_review_005	Positive	0.9478766918182373
4	new_review_007	Negative	0.9797434210777283
5	new_review_008	Positive	0.987636148929596
6	new_review_006	Negative	0.8762170672416687

	review_id	SentimentLabel	SentimentScore
1	new_review_001	Positive	0.9916028380393982
2	new_review_002	Negative	0.9736137986183167
3	new_review_005	Positive	0.9478766918182373
4	new_review_007	Negative	0.9797434210777283
5	new_review_008	Positive	0.987636148929596
6	new_review_009	Negative	0.7507771253585815
7	new_review_010	Positive	0.9913573265075684
8	new_review_006	Negative	0.8762170672416687

Streaming Simulation

NOVA IMS Information Management School

Conclusion