Master
Project Management
and Data Science

20/03/2024

Cultivating Consumer Insights from Customer Reviews: A Comprehensive Analysis Using Topic Modeling in Natural Language Processing

QUYNH PHAM

Master
Project Management
and Data Science

Agenda

1. Research Background & Motivation
2. Literature Review
3. Methodology
4. Business Insights
5. Evaluation & Limitations
6. Conclusion and Future Scope

Master
Project Management
and Data Science

1. Research Background and Motivation
2. Literature Review
3. Methodology
4. Business Insights
5. Evaluation & Limitations
6. Conclusion and Future Scope

RESEARCH BACKGROUND

Nowadays, customer reviews have become an important source of information for both consumers and businesses (Krishnan, 2023)

- For customers
  - Make informed decisions
  - Share experiences
- For businesses
  - Continuous product improvement
  - Attract new customers → increase revenue & reduce marketing cost

Master
Project Management
and Data Science

CHALLENGE & SOLUTION

CHALLENGE

Large amounts of customer reviews
→ impossible to manually sifting through all reviews to extract relevant information.

→

SOLUTION

Topic modeling - a technique that automatically identify the hidden themes in large textual datasets.

RESEARCH OBJECTIVES

Identify 2 prominent methods for topic modeling from recent studies (2016 -2023)

Replicate the experiment on the Amazon Reviews dataset.

Evaluate and compare the results of the selected methods

Derive business insights from the identified topics.

RESEARCH QUESTIONS

Q1 : Can the two chosen topic models successfully identify general topics mentioned in customer reviews? Can each review be accurately categorized into different areas of interest?

Q2 : Do the identified topics offer any valuable insights for businesses?

Q3 : Between the two topic modeling methods, which method performs better on the Amazon Reviews dataset?

1 Research Background and Motivation

2 Literature Review

3 Methodology

4 Business Insights

5 Evaluation & Limitations

6 Conclusion and Future Scope

Topic Modeling Techniques

Statistical Methods

- Latent Semantic Indexing (LSI)
- Non-negative Matrix Factorization (NMF)
- Probabilistic Latent Semantic Indexing (pLSI)
- Correlation Explanation (CorEx)
- Latent Dirichlet Allocation (LDA)

machine learning-based methods

- lda2vec
- Stochastic Block Model (SBM)
- deepLDA
- Top2Vec
- BERTopic

- Latent Dirichlet Allocation (LDA) is the most popular topic modeling method in recent studies.
- BERTopic, a newer method developed in 2022, has shown impressive results in topic modeling applications.
- Topic modeling is used in various fields such as health, e-commerce, transportation, education, finance, social network opinion analysis, etc. However, the application of topic modeling in analyzing customer reviews is underexplored (Krishnan, 2023).
- Existing literature often focuses on identifying topics and comparing metrics (coherence score, accuracy, precision) without delving into practical business implications.

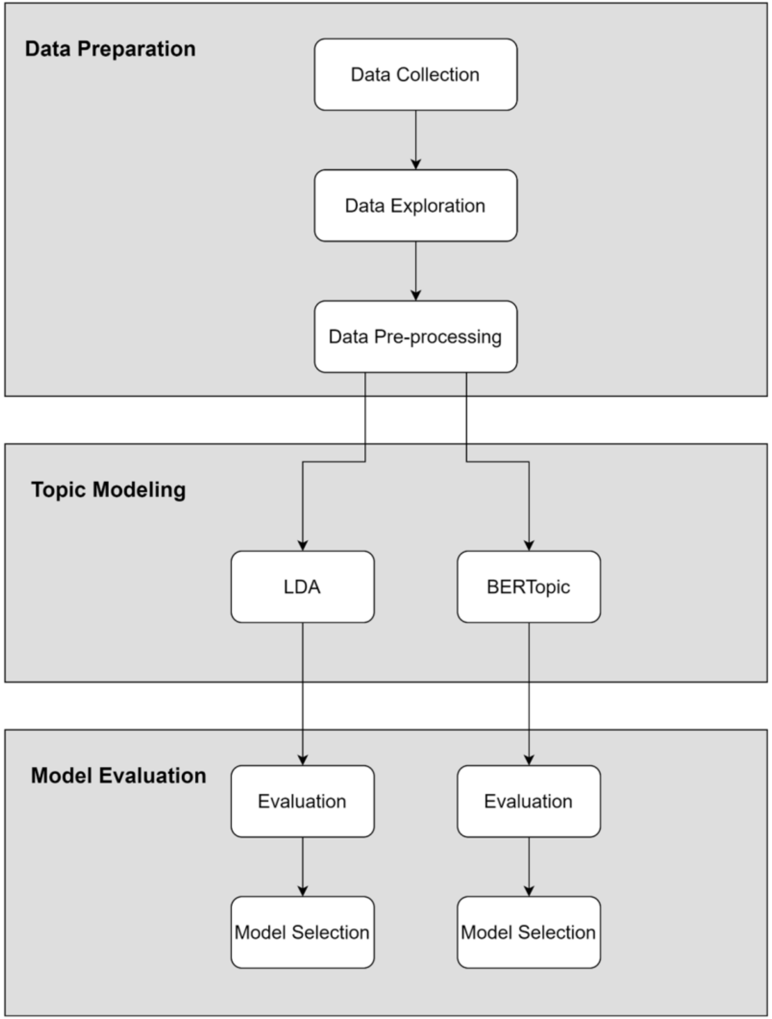
- **Addressing Literature Gaps**: Apply and evaluate LDA vs. BERTopic on the Amazon Reviews dataset for practical business insights by answering some key questions:
  1. Which topics do customers talk about most and least in their reviews?
  2. What topics are becoming more or less popular over time?
  3. Which topics are linked to the highest and lowest customer ratings, and which products are connected to these topics?
  4. For any given product, what topics are customers discussing the most?

Master
Project Management
and Data Science

1. Research Background and Motivation
2. Literature Review
3. Methodology
4. Business Insights
5. Evaluation & Limitations
6. Conclusion and Future Scope

Master
Project Management
and Data Science

Workflow Overview

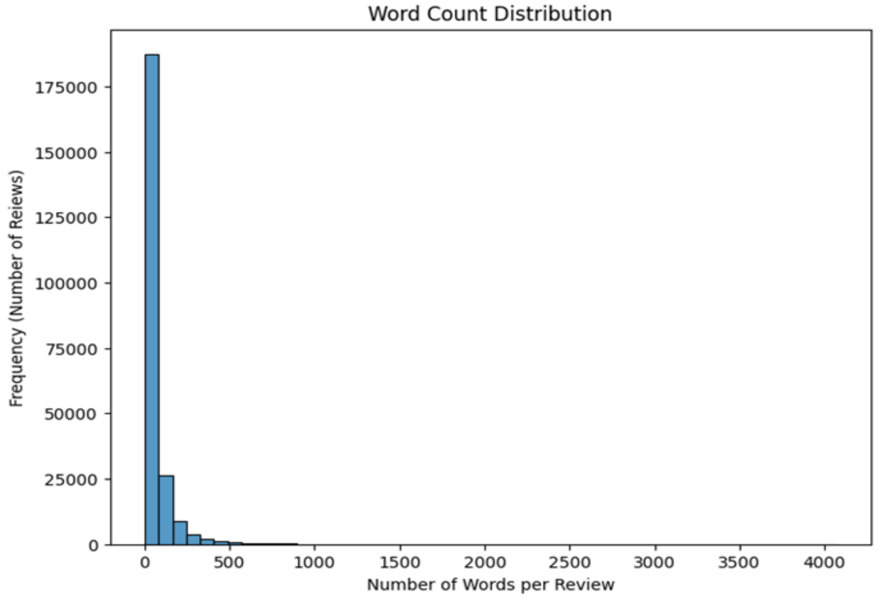


Master
Project Management
and Data Science

DATA COLLECTION & EXPLORATION

Amazon Reviews Dataset → consisting of 231,392 customer reviews from Amazon in the musical instruments sector between 10/2003 and 09/2018. This data was collected by Ni et al. (2019).

- The dataset has both long and short text.
- Longest review has 4,069 words.
- Shortest review has 0 or 1word.
- Average review length is 57 words.



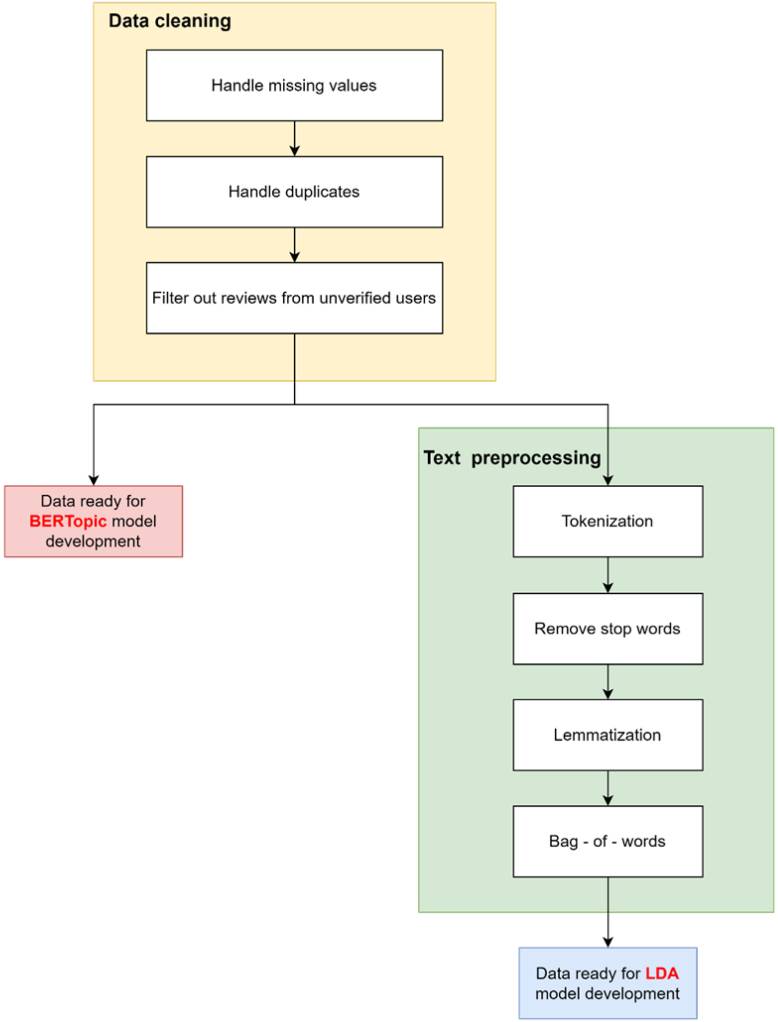
Some issues with the dataset:

- Most reviews are in English, with some are written in other languages
- 48 reviews with missing text
- 18,571 duplicate reviews.
- 23,015 reviews from 9,075 unverified users

Master
Project Management
and Data Science

DATA PREPROCESSING

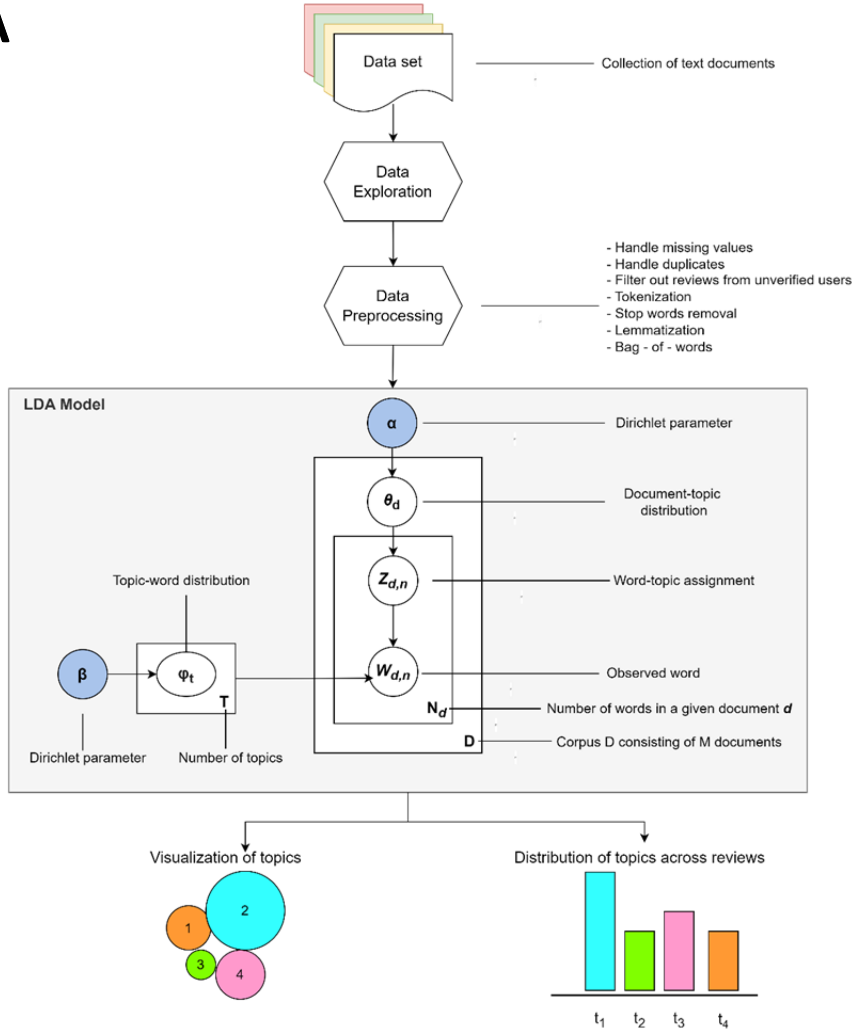
- **Data cleaning**
  - Remove 48 reviews with missing text
  - Remove 18,571 duplicate reviews
  - Remove 23,015 reviews from 9,075 unverified users
  - Dataset after cleaning: 198,940 reviews
- **Text preprocessing**



Master Project Management and Data Science

TOPIC MODELING - LDA

- LDA, developed by D. M. Blei et al. in 2003, is the most popular topic modeling algorithm.
- After the data preprocessing, perform a grid search with different numbers of topics (2 to 20) and alpha values (‘auto’, 0.01, 0.1, 1) to find the optimal LDA model.



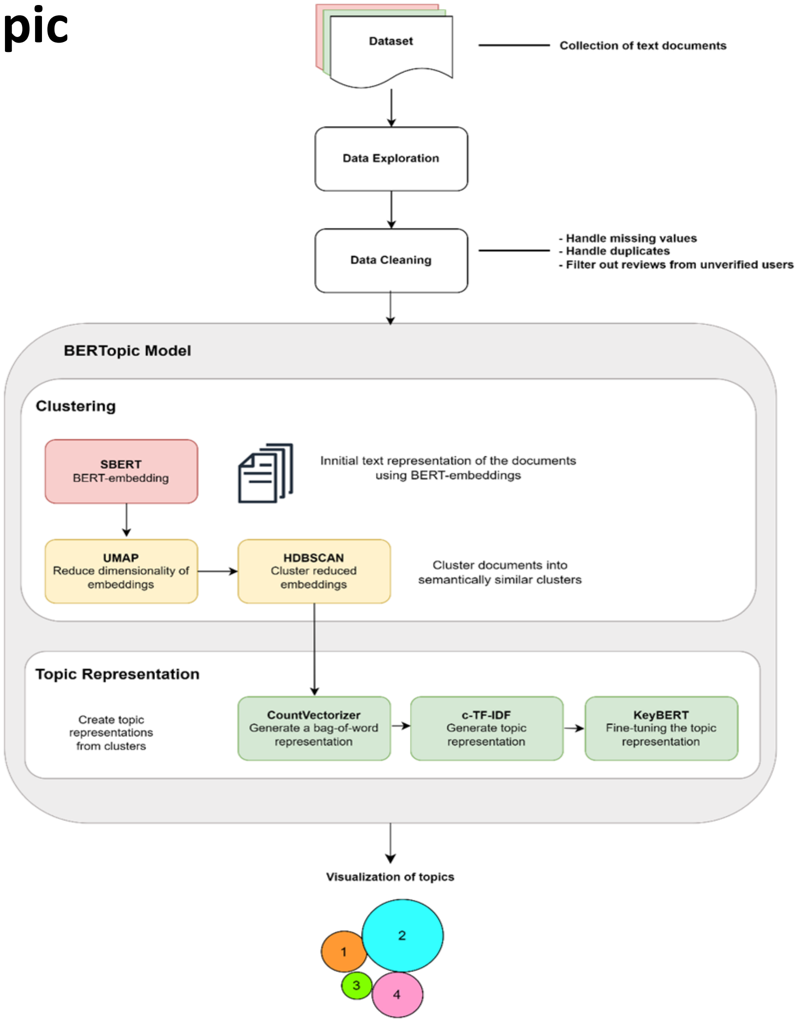
TOPIC MODELING - LDA

- Final LDA model has 11 topics with alpha value 1 and coherence score 0.531

TOPIC MODELING - BERTopic

Master
Project Management
and Data Science

- Introduced by Maarten Grootendorst in 2022, BERTopic leverages BERT for word embeddings and c-TF-IDF for keyword highlighting.
- The process has 5 steps:
    1. Embedding extraction
    2. Dimensionality reduction
    3. Clustering
    4. Tokenizer
    5. Weighting Scheme
    6. Fine-tuning Representation (optional)
- Its flexibility enables users to modify components, tailoring the model to specific use cases and datasets.



TOPIC MODELING - BERTopic

building a BERTopic model with default settings

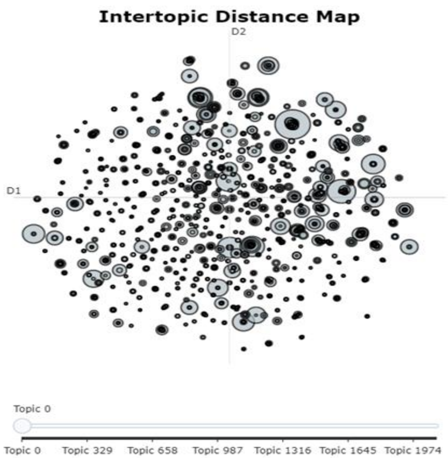
refining the BERTopic model by adjusting parameters to enable multilingual support and automatic merger of topics

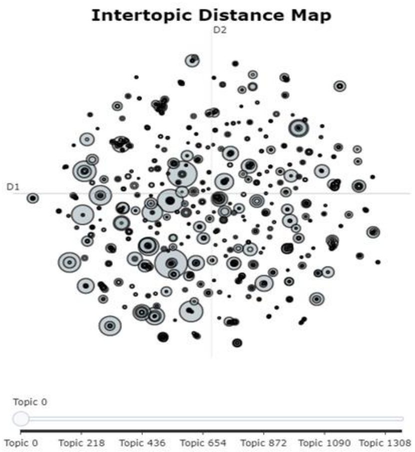
fine-tuning various hyperparameters

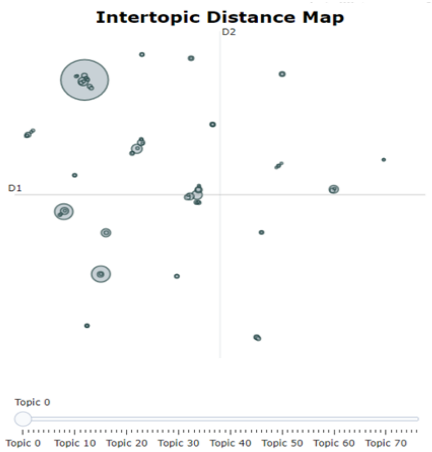
2,145 topics

1,371 topics

76 topics

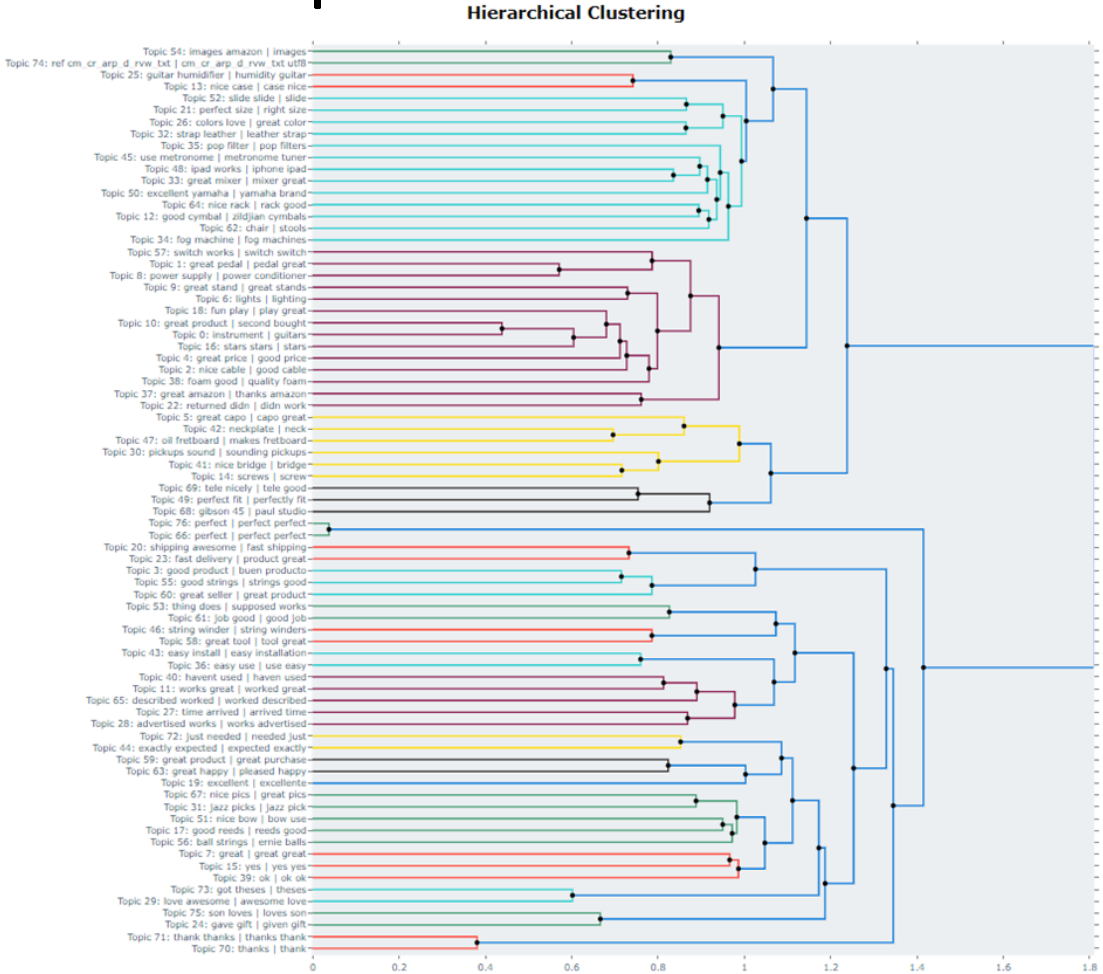






Master
Project Management
and Data Science

TOPIC MODELING - BERTopic



Master
Project Management
and Data Science

1 Research Background and Motivation

2 Literature Review

3 Methodology

4 **Business Insights**

5 Evaluation & Limitations

6 Conclusion and Future Scope

DERIVE BUSINESS INSIGHTS

Question 1

Which topics do customers talk about most and least in their reviews?

Question 2

What topics are becoming more or less popular over time?

Question 3

Which topics are linked to the highest and lowest customer ratings, and which products are connected to these topics?

Question 4

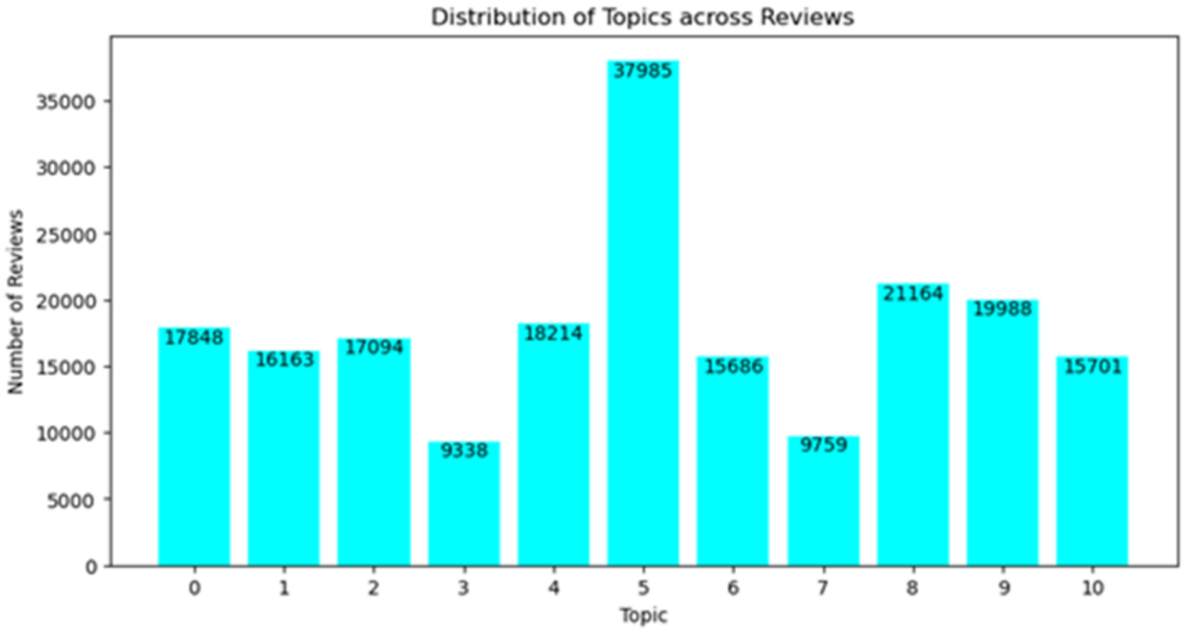
For any given product, what topics are customers discussing the most?

Master
Project Management
and Data Science

Question 1:

Which topics do customers talk about most and least in their reviews?

LDA



- Topic 5 (great, good, price, quality, product, look, love, awesome, deal, fast) is the most discussed topic, potentially focusing on "Value for Money"
- Topics 3 (string, set, high, bass, case, end, low, lot, replace, change) and Topic 7 (make, much, even, put, find, new, money, sure, hard, worth) are the least discussed topics.

Master
Project Management
and Data Science

Question 1:

Which topics do customers talk about most and least in their reviews?

BERTopic

Topic	Count	CustomName
-1	431	Topic: -1 good | excellent
0	75195	Topic 0: instrument | guitars
1	10004	Topic 1: great pedal | pedal great
2	9044	Topic 2: nice cable | good cable
3	6092	Topic 3: good product | buen producto
...	...	...
75	155	Topic 75: son loves | loves son
76	152	Topic 76: perfect | perfect perfect

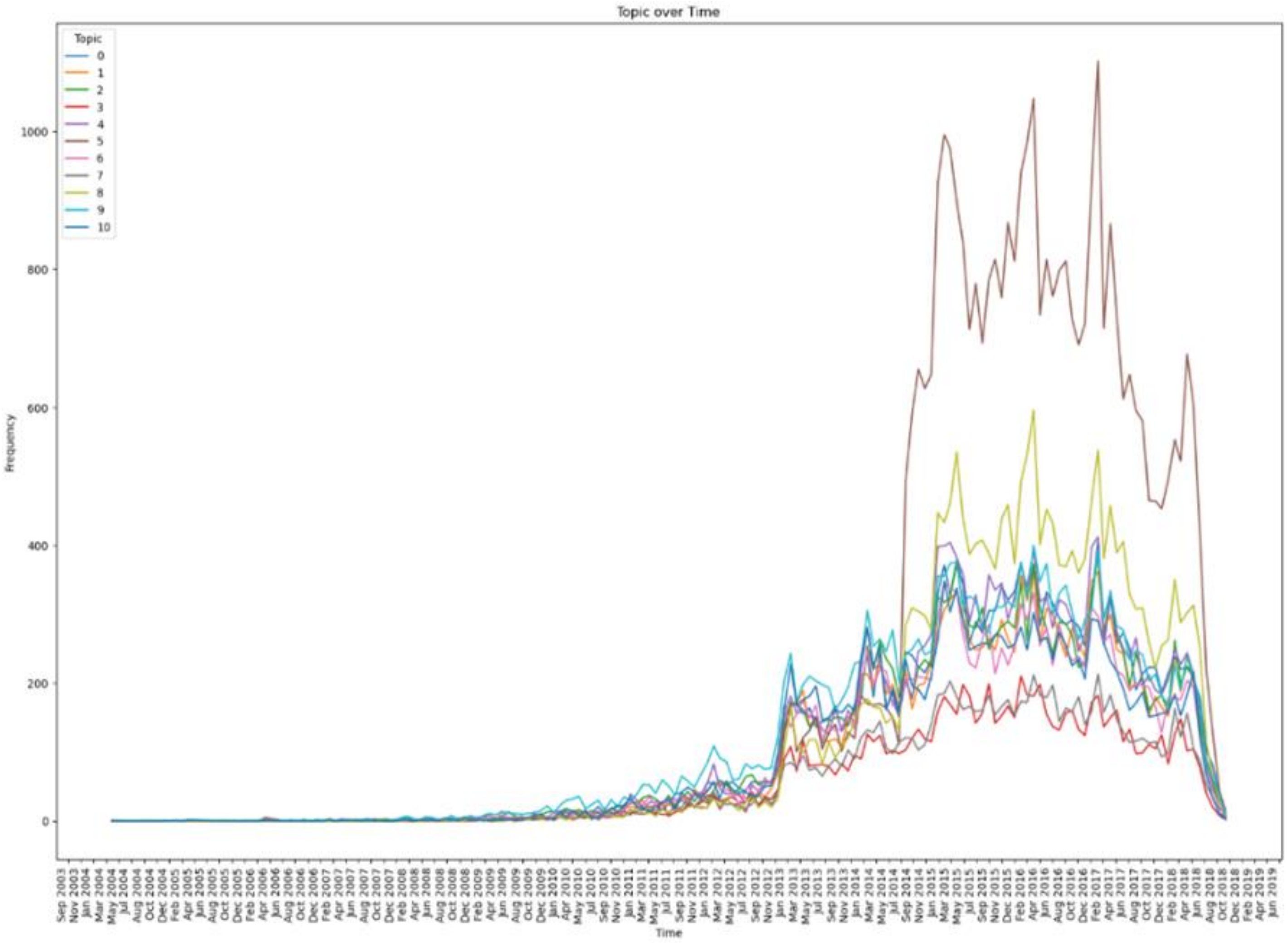
- BERTopic gives the results in the descending order where
  - Topic 0 (instruments, guitars) is the most discussed topic and
  - Topics 76 (perfect, perfect perfect) is the least discussed topics.

Master
Project Management
and Data Science

Question 2:

What topics are becoming more or less popular over time?

LDA

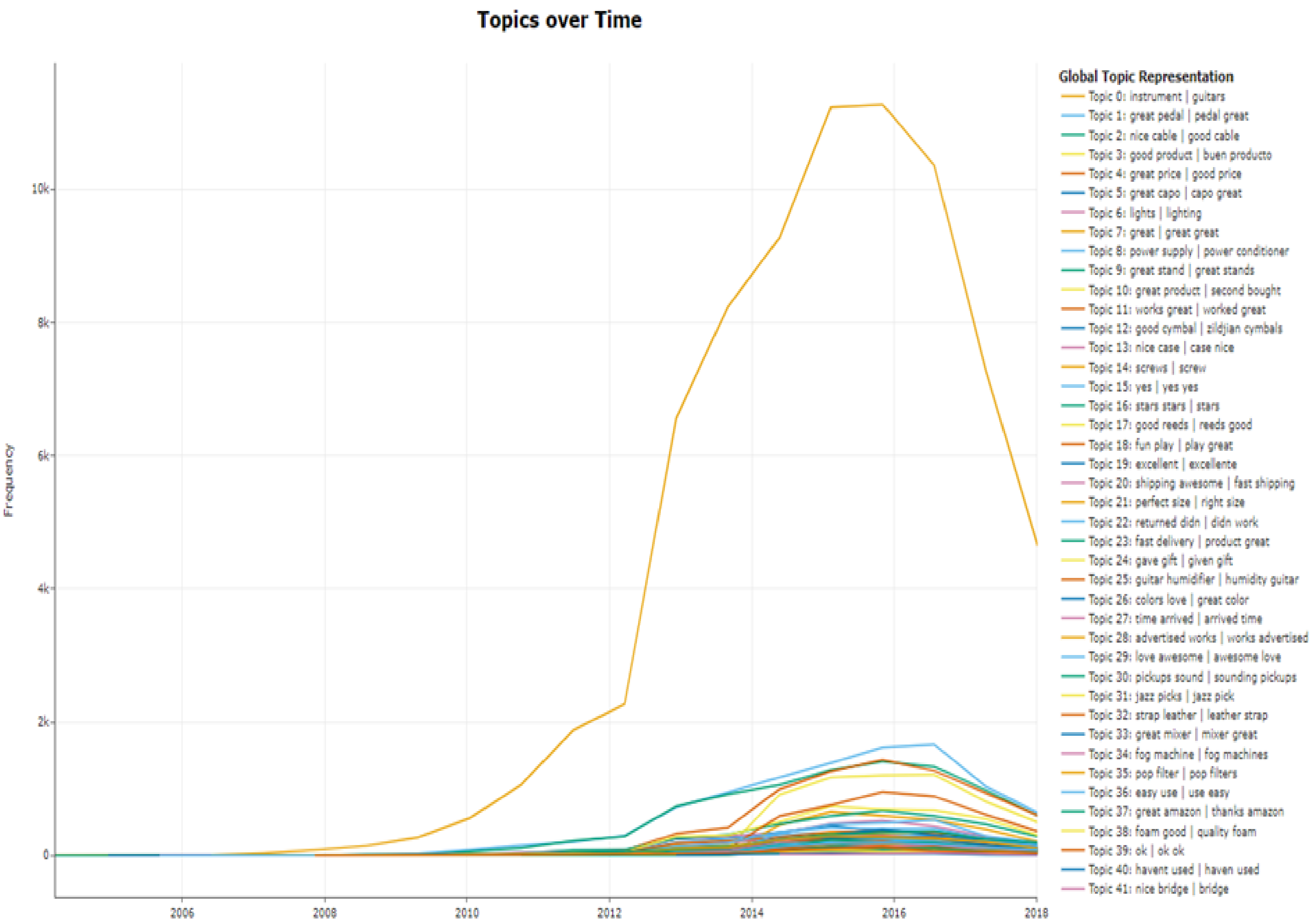


Master
Project Management
and Data Science

Question 2:

What topics are becoming more or less popular over time?

BERTopic

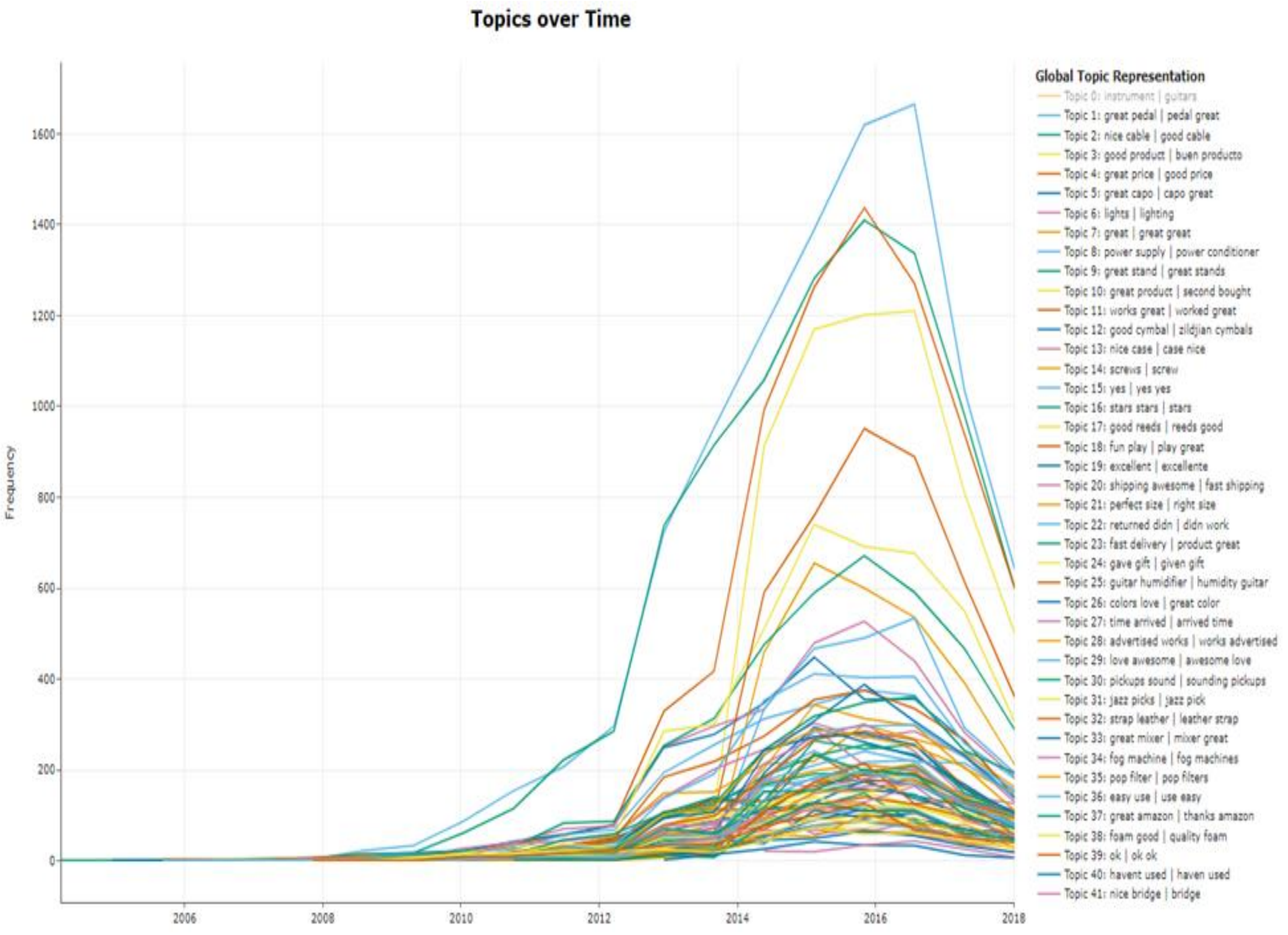


Master
Project Management
and Data Science

Question 2:

What topics are becoming more or less popular over time?

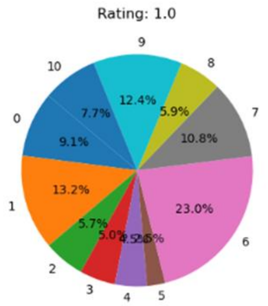
BERTopic

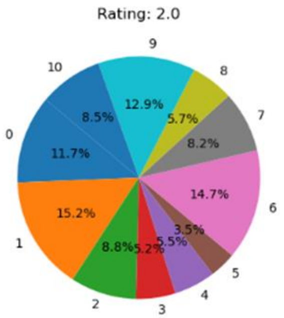


Question 3:

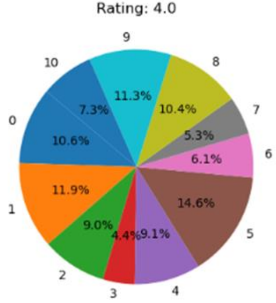
Which topics are linked to the highest and lowest customer ratings, and which products are connected to these topics?

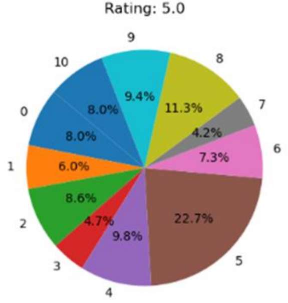
LDA





- Topic 6 (buy, go, come, say, cheap, see, purchase, first, think, know) more frequently in lower-rated reviews (1.0 and 2.0)





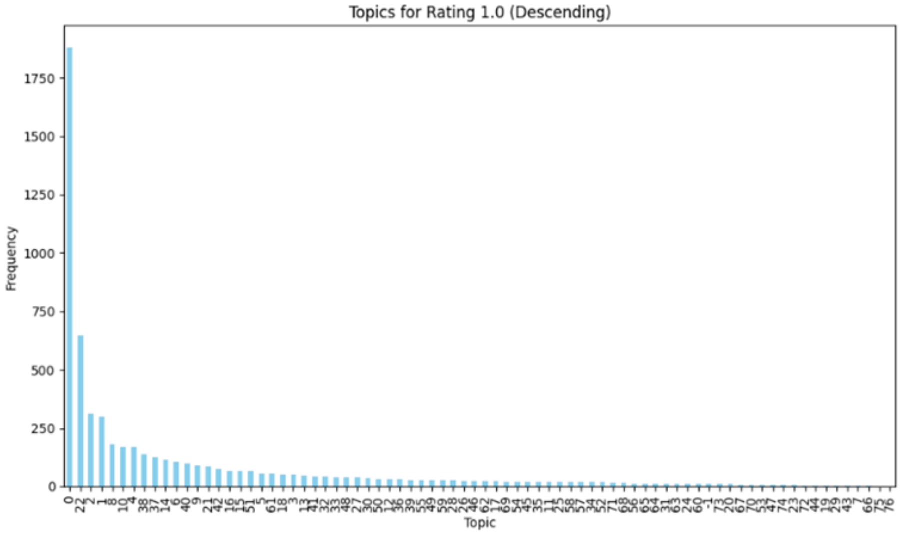
- Topic 5 (great, good, price, quality, product, look, love, awesome, deal, fast) is predominantly found in reviews with higher ratings (4.0 and 5.0),

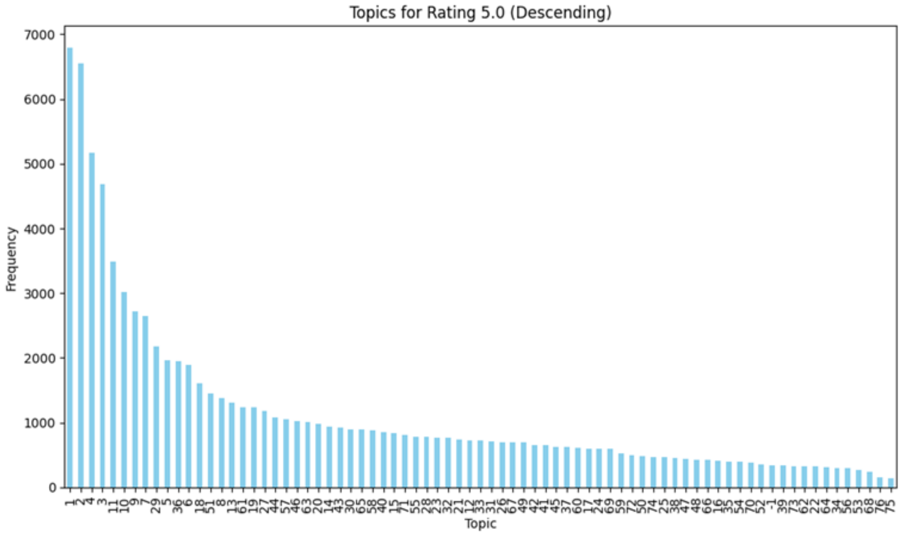
Master
Project Management
and Data Science

Question 3:

Which topics are linked to the highest and lowest customer ratings, and which products are connected to these topics?

BERTopic





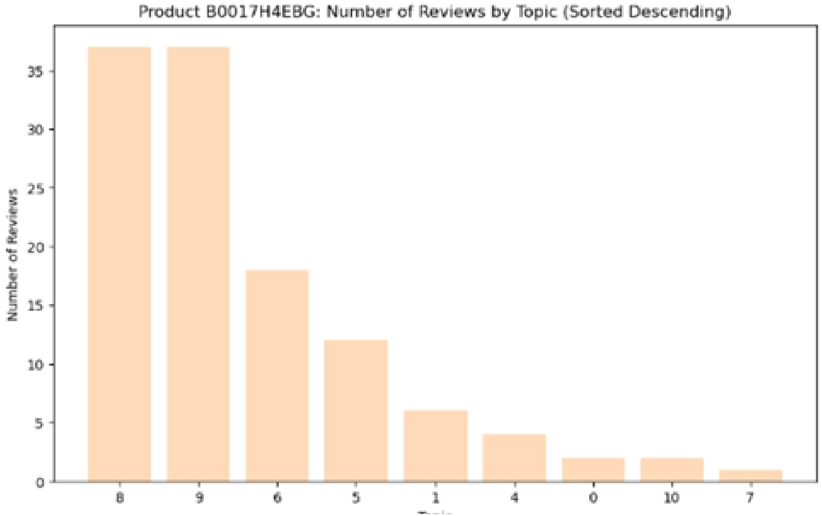
- Topic 22 (returned didn, didn work) is commonly associated with low ratings (1.0 and 2.0)
- Topic 3 (good product, buen producto) and Topic 4 (great price, good price) in higher ratings (4.0 and 5.0)

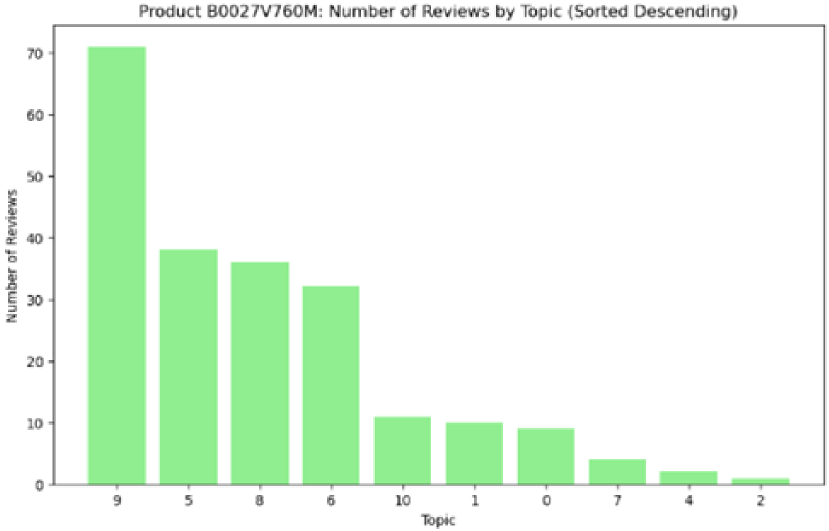
Master
Project Management
and Data Science

Question 4:

For any given product, what topics are customers discussing the most?

LDA





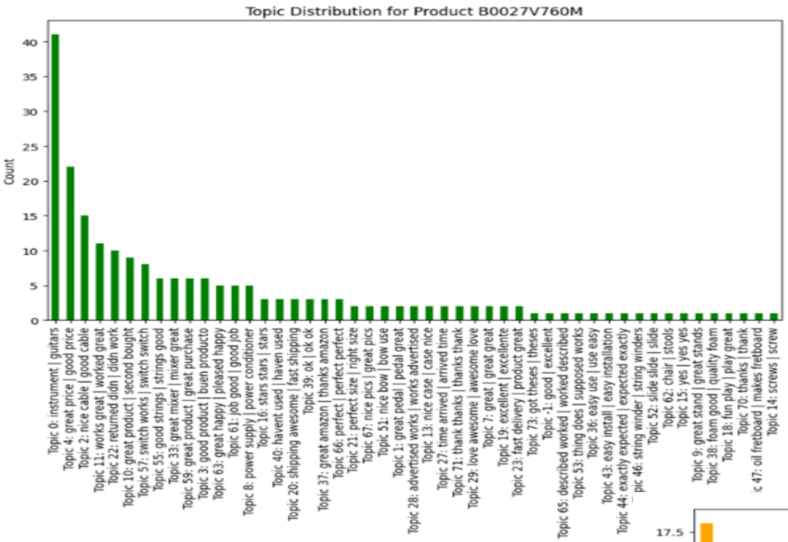
- The products most associated with Topic 6 and lower ratings are B0017H4EBG, B0027V760M, B0002GMH7G, B0002GMGYA, B00AZUAORE.

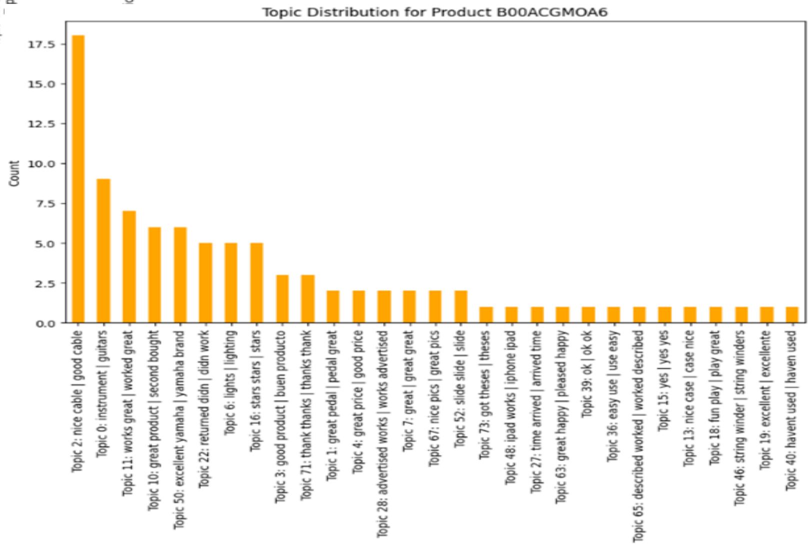
Master
Project Management
and Data Science

Question 4:

For any given product, what topics are customers discussing the most?

BERTopic





1. Research Background and Motivation
2. Literature Review
3. Methodology
4. Business Insights
5. Evaluation & Limitation
6. Conclusion and Future Scope

EVALUATION

Criteria	LDA	BERTopic
Coherence Score	LDA has higher coherence score (**0.531**)	BERTopic has lower coherence score (**0.436**)
Topic Quality	LDA give a clear snapshot of the general theme of the reviews with fewer topics	BERTopic model produces more topics, which are organized in a hierarchical order, featuring the main topics and their associated sub-topics
Computation Time	requires multiple iterations with a range of different numbers of topics, alpha and beta values → more time-consuming (**48 hours**) and resource-intensive	BERTopic automatically decide the optimal number of topics during training, which takes **6 hours** in total.
Data Preprocessing	Requires thorough and meticulous preprocessing steps	Minimal to no data preprocessing required
Visualization Tools	Limited to tools to pyLDAvis	Offers interactive intertopic distance maps similar to pyLDAvis, plus additional advanced options for analysis such as heatmap for topic similarity, visualizations for topic over time, topics per class, hierarchy topic, etc.

Limitations

- **Lack access to advanced computational resources (GPU)**
  - Limited local CPU capacity and lack of GPU resources extend training times and requiring a smaller dataset size, which impact efficiency and depth of analysis.
  - The computational constraints also hinder scalability, impede rapid iterations and optimization, affect the overall quality and speed of the topic modeling process.

- **Lack domain expertise in the field of musical instruments**
  - Topic modeling identifies patterns and clusters similar terms but lacks the ability to discern their true significance, especially in specialized fields like musical instruments.
  - Domain experts are crucial for interpreting the context of these terms, affecting topic labeling and insights into customer satisfaction and product design.

Master
Project Management
and Data Science

1 Research Background and Motivation

2 Literature Review

3 Methodology

4 Evaluation & Discussion

5 Limitations

6 **Conclusion and Future Scope**

Q1 : Can the two chosen topic models successfully identify general topics mentioned in customer reviews? Can each review be accurately categorized into different areas of interest?

Q2 : Do the identified topics offer any valuable insights for businesses?

Q3 : Between the two topic modeling methods, which method performs better on the Amazon Reviews dataset?

- **Future Scope:**
  - **Sentiment Analysis**
    - to gain a well-informed picture and support decision-making.
  - **Incorporating OpenAI's GPT-3.5 Turbo**
    - to automatically topic labeling in BERTopic.

Master
Project Management
and Data Science

THANK YOU !

www.mpmd.htw-berlin.de