Hochschule für Technik
und Wirtschaft Berlin
University of Applied Sciences

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

Master Thesis

Master Project Management and Data Science

Faculty 3

from

Quynh Pham

Date:

Berlin, 19.02.2024

1st Supervisor: ……..

2nd Supervisor: …….

Index

1 Introduction 7

2 Literature Review 10

3 Methodology 18

3.1 Data Collection 18

3.2 Data Preprocessing 19

3.3 Topic Modeling Techniques 24

3.3.1 Latent Dirichlet Allocation 25

3.3.2 BERTopic 29

3.4 Evaluation Metric 33

3.5 Chapter Summary 35

4 Experimental Results 37

4.1 Data Exploration 38

4.2 Data Preprocessing 41

4.3 LDA Result 44

4.4 BERTopic Result 55

4.5 Discussion 70

4.6 Chapter Summary 74

5 Conclusion 76

List of literature 78

Statutory Declaration ..... Error! Bookmark not defined.

List of figures

Figure 1: Timeline of the evolution of topic modeling (Churchill & Singh, 2022, p. 3-4).....11
Figure 2: Overview of topic modeling process 18
Figure 3: Data preprocessing workflow for topic modeling with BERTopic and LDA 20
Figure 4: An example of the tokenization process 22
Figure 5: The representation of BoW.....24
Figure 6: The intuitions behind LDA (Blei, 2012. p.78).....26
Figure 7: Graphical representation of LDA model 27
Figure 8: Overview of LDA topic modeling workflow 28
Figure 9: Visualization by pyLDAvis 29
Figure 10: BERTopic pipeline 30
Figure 11: Components of the default BERTopic Model 32
Figure 12: Customizable components for the personalized topic model in BERTopic 32
Figure 13: Overview of BERTopic workflow 33
Figure 14: Sample entries from the Amazon Reviews dataset.....39
Figure 15: Distribution of reviews in multiple languages in the Amazon Reviews dataset.....39
Figure 16: Distribution of reviews over time (2003 – 2018) 40
Figure 17: Distribution of product ratings in the Amazon Reviews dataset 40
Figure 18: Distribution of word counts for all reviews in Amazon Reviews dataset.....41
Figure 19: Data cleaning process for Amazon Reviews dataset 42
Figure 20: Tokenization example for the third review in the dataset.....43
Figure 21: Example of stop word removal for the third review in the dataset.....43
Figure 22: Lemmatization example for the third review in the dataset.....44
Figure 23: Example of corpus and frequency dictionary for the third review in the dataset 44
Figure 24: Coherence score trends across different LDA models.....46
Figure 25: Optimal number of topics for LDA model selection based on coherence scores.....46
Figure 26: Visualization of topic clusters for LDA models with 5, 7, and 11 topics.....47
Figure 27: Visualization of the final LDA model 50
Figure 29: Topic trends over time (2003 - 2018).....53
Figure 30: Distribution of topics across product ratings 54
Figure 31: Distribution of topics in product B0017H4EBG and B0027V760M 55
Figure 32: Result of the default BERTopic model.....56

Figure 33: Visualization of the default BERTopic model....56
Figure 34: Result of multilingual BERTopic model with automated topic merging....57
Figure 35: Visualization of multilingual BERTopic model with automated topic merging....57
Figure 36: Topic representations with KeyBERTInspired, POS, and MMR....58
Figure 37: Bar chart of word importance scores for top 10 topics....59
Figure 38: Term importance score across all topics....59
Figure 39: Result of refined BERTopic model after hyperparameter tuning....60
Figure 40: Visualization of refined BERTopic model after hyperparameter tuning....60
Figure 41: Dendrogram of topic relationships of refined BERTopic model....61
Figure 42: Pairs of similar topics with high similarity scores....62
Figure 43: Result of the final BERTopic model....62
Figure 44: Visualization of the final BERTopic model....63
Figure 45: Topic trends over time (2003 – 2018)....64
Figure 46: Topic trends over time, excluding Topic 0 (2003 – 2018)....65
Figure 47: Topic distribution across all product ratings in the final BERTopic model....65
Figure 48: Topic distribution across each product ratings....68
Figure 49: Topic distribution for product B0027V760M....69
Figure 50: Topic distribution for product B00ACGMOA6....69
Figure 51: Topic distribution for third review in the dataset....70

List of tables

Table 1: Summary of related works on topic modeling methods across various fields....15
Table 2: Details of Amazon Reviews dataset....19
Table 3: Field descriptions of the Amazon Reviews dataset....38
Table 4: Coherence scores across different LDA model configurations....45
Table 5: Topic labels and frequent words in LDA models with 5 and 7 topics....49
Table 6: Statistics of the final LDA model....50
Table 7: Statistics of the final BERTopic model....63
Table 8: Comparison between LDA and BERTopic models....72

Index of abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
BoW	Bag-of-Words
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CTM	Correlated Topic Model
CorEx	Correlation Explanation
DCN	Deep Clustering Network
GSDMM	Gibbs Sampling Dirichlet Mixture Model
GPU	Graphics Processing Unit
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
KATE	K-Competitive Autoencoder for Text
KGNMF	Kernelized Generalized Non-Negative Matrix Factorization
Kmeans	K-Means Clustering
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
LSTM	Long Short-Term Memory
MMR	Maximal Marginal Relevance
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NMF	Non-Negative Matrix Factorization
PAM	Partition Around Medoids
PCA	Principal Component Analysis
POS	Part of Speech
RP	Random Projection
SBM	Stochastic Block Model

UMAP	Uniform Manifold Approximation and Projection
VADER	Valence Aware Dictionary and sEntiment Reasoner
c-TF-IDF	class-based Term Frequency-Inverse Document Frequency
pLSI	Probabilistic Latent Semantic Indexing

1 Introduction

Nowadays, customer reviews have become an important source of information for both consumers and businesses (Krishnan, 2023, p. 1). They are feedback, critiques, and suggestions provided by consumers after using a company's products or services. With the advent of technology, the growing popularity of online platforms and the impact of social media, consumers are more motivated to share their experiences and opinions (Krishnan, 2023, p. 1). Some people choose to write about their experiences, while others want to include photographs depicting the items or services they have received. By analyzing this feedback, businesses can gain valuable insights into consumers' preferences and feelings towards their goods and services.

The emergence of e-commerce has had a significant influence on purchasing habits. As the internet has become an inseparable part of everyone's daily lives and business practices, this has fueled the demand for new services and applications to accommodate and improve the shopping experience for consumers worldwide (Albalawi et al., 2020, p. 1). This trend may be linked to the ease and efficiency of online purchases: items are delivered right to the customer's door, reducing the need to set aside time for shopping visits. This method saves people a lot of time and effort (Nogoev et al., 2011, p. 4). Furthermore, businesses are progressively supporting online shopping by offering attractive incentives such as affordable or free delivery options (Shehu et al., 2020, p. 3) and simple return policies (Liu & Du, 2023, p. 1), making the entire process hassle-free. The COVID-19 epidemic accelerated this trend since social distancing measures severely disturbed the usual shopping experience. During this time, there was a huge increase in internet purchasing, indicating a clear shift from physical to digital retail environments. This slow but dramatic shift from physical to online buying is transforming the world of retail and consumer behaviors (Gu et al., 2021, p. 2277).

Customer reviews are important to any business because they provide insights into the strengths, weaknesses, and overall quality of a product or service (Krishnan, 2023, p. 1). To begin, client input plays an essential role in the ongoing improvement of products or services. Reviews provide businesses the opportunity to hear directly from their consumers. By analyzing this feedback, companies can identify which aspects are resonating with consumers and which areas need improvement. This process allows for continuous product or service improvement, ensuring that they are more closely aligned with client preferences and expectations. Furthermore, customer feedback is instrumental in attracting new customers and optimizing expenses. The inability to visually examine products is a significant disadvantage of online shopping. As a result, potential customers often rely on reviews from others who have previously purchased the product to assist with their purchasing decisions (Duan et al., 2008, p. 1009). Positive customer feedback leads to positive reviews, which may contribute to a product's reputation. This, in turn, helps in attracting new customers. This means the potential for increased revenue and reduced marketing

expenditures for businesses since positive word-of-mouth functions as a powerful, cost-effective promotional tool (Pooja & Upadhyaya, 2022, p. 2).

However, the immense number of customer reviews available online might be overwhelming and intimidating for any business attempting to manually sort through this huge unstructured data and extract relevant information. Therefore, topic modeling emerges as a vital solution to this challenge (Albalawi et al., 2020, p. 2). As part of natural language processing (NLP), topic modeling is an effective technique for identifying the underlying themes in a large collection of textual data (Jelodar et al., 2017, pp. 1–2). Since 1990, there have been several approaches to topic modeling that yield good results from both long and short texts (Xie & Xing, 2013, p. 1; Cheng et al., 2014, p. 1). These methods mainly fall into two categories: statistical and machine learning-based methods. Initially, the traditional methods make use of statistics to discover the underlying topic in documents. Such methods are Latent Semantic Indexing (LSI), Non-negative Matrix Factorization (NMF), Probabilistic Latent Semantic Indexing (pLSI), Correlation Explanation (CorEx), Latent Dirichlet Allocation (LDA). Nevertheless, advancements in machine learning and artificial intelligence have led to the development of new methods utilizing these technologies have emerged and yielded promising results such as lda2vec, Stochastic Block Model (SBM), deepLDA, Top2Vec, BERTopic, etc.

This study aims to identify the two most effective methods for topic modeling from the recent literature and replicate the experiment using the Amazon Reviews dataset. The goal is to extract hidden topics, classify reviews into different areas of interest, and provide useful insights that can help with decision-making and business strategy. The following points shall be made in this study:

- Review journal articles about topic modeling in various fields from 2016 to 2023.
- Assess the popular topic modeling methods discussed in these studies and select two prominent methods for analysis on the Amazon Reviews dataset.
- Evaluate and compare the results of the selected methods using coherence score, visualization tools and human interpretation of topic quality.
- Derive potential insights from the identified topics.

The following questions are used to direct the study and the experimental process:

1. 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?
2. Do the identified topics offer any valuable insights for businesses?
3. Between the two topic modeling methods, which method performs better on the Amazon Reviews dataset?

This study is structured as follows: Section 2 reviews recent scientific journal publications on NLP and topic modeling with a focus on the application of topic modeling across various fields. Section 3 presents a detailed framework for implementing and evaluating the two chosen topic models. The experimental results, along with a discussion of the outcomes and limitations of this research, are detailed in section 4. Section 5 summarizes the key findings and offers recommendations for future work.

2 Literature Review

In the current digital era, an immense volume of data is generated from various sources, including blogs, social media platforms, and web pages. A significant portion of this data is in the form of unstructured text, such as tweets, blog posts, and reviews. This data, often reaching gigabytes in volume, is rich with information that companies can leverage. Analyzing and understanding this unstructured text is vital for companies to gain important insights and make well-informed decisions. However, manually processing such large datasets is both impractical and inefficient for businesses. As a result, there is a growing need for tools and methods that can efficiently transform this data into actionable information. NLP is introduced as a powerful solution in this context.

Natural language refers to the languages used by humans for everyday communication and information exchange (Stanisz et al., 2024, p. 4). English, Hindi, German, and many other languages are examples of natural languages. These languages have evolved over time and are difficult to define with specific rules. Human languages are inherently complex and characterized by nuances such as ambiguity, slang, and cultural references. In contrast, artificial languages include programming languages and mathematical notations, which are characterized by their logical structure and precision. These languages are designed for clarity and unambiguity, making them well-suited for machine interpretation and computation. However, the process of translating natural language into a form that machines can understand is not straightforward. Natural language cannot be directly converted into a precise set of mathematical operations. Instead, it must be transformed into data, comprised of numbers 0 and 1, which computers can use to learn and make sense of the world (Nagarhalli et al., 2021, p. 1531).

NLP serves as a bridge between human and machine communication. It combines linguistics, artificial intelligence, and computer science to help machines understand, interpret, and generate human language in a way that is natural for humans (Khurana et al., 2023, p. 3714). This capability allows for the development of systems that assist humans in managing and analyzing large volumes of unstructured data for a variety of tasks (Hannigan et al., 2019, p. 10), including language understanding, language generation, language translation, sentiment analysis, named entity recognition, text classification, speech recognition, chatbots, conversational agents, etc.

One of the tasks that can be solved by NLP is topic modeling. It is an unsupervised machine-learning approach that can help with the discovery of new, previously unknown information by automatically extracting topics from large text documents (Bisgin et al., 2011, p. 7; Guo et al., 2017, p. 10). Examples of documents include websites, books, emails, reviews, news articles, or scientific journals. Information can be obtained from the use of techniques and algorithms such as statistics and machine learning. Thus, in the case of customer reviews, topic modeling is widely used to explore the most frequent and influential topics discussed by consumers. Over the years,

extensive research has been conducted in this field, leading to the development of various methods.

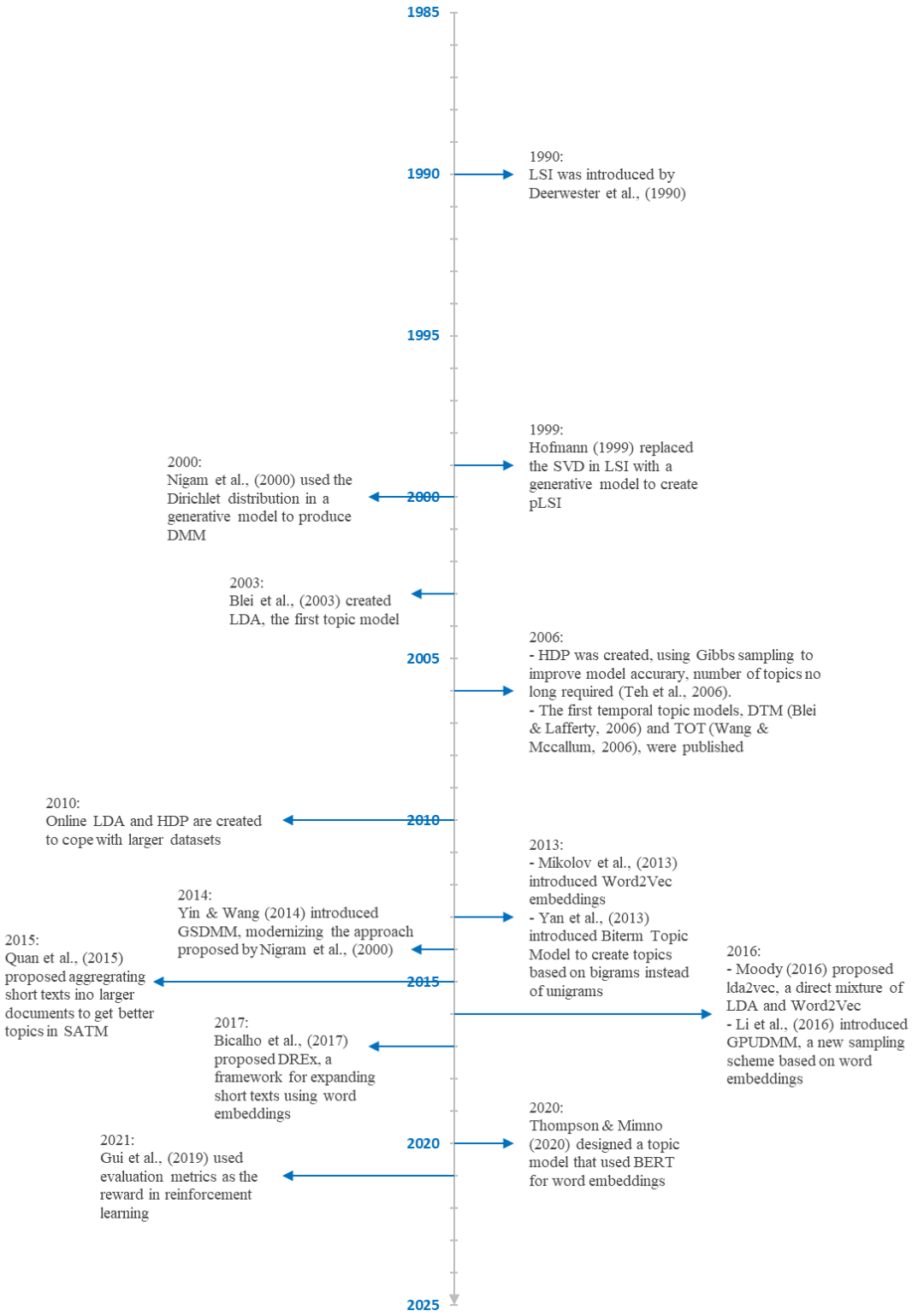


Figure 1: Timeline of the evolution of topic modeling (Churchill & Singh, 2022, p. 3-4)

The history of topic modeling dates back to the early 1990s when Deerwester et al. (1990) first introduced LSI, a non-probabilistic technique for automatically indexing and retrieving semantic

structures inside texts. In 1999, building on the foundation of LSI, Hofmann (1999) presented the first genuine probabilistic topic model pLSI. Blei et al. (2003) further expanded upon Hofmann's work by incorporating Bayesian ideas and developing the LDA model in 2003. This has subsequently been regarded as the most prominent method in topic modeling to this day. Since then, there has been an increase in studies on the subject of topic modeling. Since 2016, machine learning has gained major popularity due to the exponential growth and development in artificial intelligence and deep learning, which has resulted in the release of various sophisticated neural models such as lda2vec, Top2Vec, BERTopic, etc. The timeline for the evolution of topic modeling over time is shown in Figure 1.

In recent years, topic modeling has seen application across various fields such as health, e-commerce, transportation, education, finance, social network opinion analysis, etc. (Albalawi et al., 2020, p. 5). It is a useful tool for researchers and businesses to analyze and categorize large amounts of textual data. For example, a study by Maier et al. (2018) applied LDA to investigate food safety concerns across 186,557 web pages. In academia, Asmussen and Møller (2019) employed LDA on 650 research publications to explore the literature reviews and identified 20 distinct topics for different study areas, leading to a more optimized and effective review procedure. Similarly, García-Méndez et al. (2023) used LDA to recognize important financial events from 2,158 financial news stories, along with other techniques for better forecasting and analysis.

Social media content analysis also benefits from topic modeling, as demonstrated by various studies. For instance, Albalawi et al. (2020) compared five topic modeling approaches – LDA, NMF, Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), and Random Projection (RP) – on datasets from newsgroups and Facebook discussions and found that LDA and NMF were the most effective in extracting meaningful topics. Similarly, Egger & Yu (2022) analyzed 50,000 tweets about travel during the COVID-19 pandemic using LDA, NMF, Top2Vec, and BERTopic and found that NMF and BERTopic generated the most insightful results. Another study by Yin et al. (2022) analyzed 78,827 tweets about COVID-19 vaccines utilizing LDA for topic modeling and Valence Aware Dictionary and Sentiment Reasoner (VADER) for sentiment analysis to gauge public sentiment and discussion topics. Egger and Yu (2021) examined tourist experiences of dark tourism through 33,881 Instagram posts using LDA, CorEx, and NMF and discovered that CorEx offered the most in-depth uncovering of tourist experiences.

Since its introduction in 2018, Bidirectional Encoder Representations from Transformers (BERT) has revolutionized the field of NLP. Numerous researchers have studied and developed models that leverage the power of BERT. One such model is BERTopic, which has been proven in many studies to deliver much better results than the traditional methods for topic modeling. For instance,

Krishnan (2023) analyzed several topic modeling methods, including LSA, LDA, NMF, Partition Around Medoids (PAM), Top2Vec, and BERTopic on two datasets containing 1,600 reviews from TripAdvisor and Amazon Mechanical Turk and 29,200 reviews from the United Arab Emirates Ministry of Economy government website. The study showed that BERTopic outperformed other methods by efficiently identifying relevant topics with minimal preprocessing, high coherence scores, and reasonable computation time. Similarly, An et al. (2023) analyzed 500,000 customer reviews from 17 product categories on South Korea's top e-commerce platform, Naver Shopping. The study compared BERTopic with traditional methods such as Deep Clustering Network (DCN), LDA, and Kmeans. The results demonstrated that both BERTopic and DCN outperformed traditional methods in extracting market insights. Another experiment by Grootendorst (2022) evaluated BERTopic against LDA, NMF, Correlated Topic Model (CTM), and Top2Vec across three distinct datasets, including 20 NewsGroups, BBC News, and a collection of Donald Trump's tweets. The goal was to demonstrate that BERTopic outperformed other methods across various use cases. These findings highlight the potential of BERTopic as a leading tool for topic modeling, capable of providing significant advantages in various situations. However, it is important to recognize its limitations and to carefully consider when choosing BERTopic for specific applications. Table 1 provides a comprehensive summary of recent research on topic modeling methods applied in various fields.

Related Work	Topic Modeling methods	Field of study	Results
Mazarura & de Waal (2016)	- LDA - GSDMM	3 datasets: a collection of 495 Finweek news articles, a collection of 77,946 tweets about the weather made available by the CrowdFlower Open Data Library and 15,984 tweets about the August GOP debate that took place in Ohio, USA, in 2015.	GSDMM gives better results
Chen et al. (2017)	- NMF - PCA - LDA - KATE	dataset of 8-K and 10-K filings, from the years 2005–2016, of 578 bank holding companies	LDA gives the best result
Maier et al. (2018)	- LDA	websites on the internet regarding food safety issues	

Related Work	Topic Modeling methods	Field of study	Results
Anantharaman et al. (2019)	- LDA - LSA - NMF	3 datasets: PubMed 20k RCT, BBC News Dataset and Twenty Newsgroups Dataset	- LSA gives the best result for short text - LDA performs better on large text
Asmussen & Møller (2019)	- LDA	literature reviews of research publications	
Chen et al. (2019)	- LDA - NMF - KGNMF	5 datasets: snippets drawn from Google; news articles from Sina website; English news and its titles from popular English newspaper websites; short text dataset from StackOverflow.com	NMF gives better result than LDA
Ray et al. (2019)	- LSI - LDA - NMF	dataset of news articles in Hindi scraped from 2 newspaper websites Amar Ujala and Navbharat Times.	NMF gives the best result
Xu et al. (2020)	- LDA	dataset of top 500 short reviews for Douban movies	
Albalawi et al. (2020)	- LDA - LSA - NMF - PCA - RP	20newsgroup data and 20 brief discussions from Facebook	LDA and NMF give best result
Chakkarwar & Tamane (2020)	- LDA	550 abstracts of blockchain technology	
Egger & Yu (2021)	- LDA - CorEx - NMF	tourist experiences on dark tourism from Instagram posts	CorEx give best result

Related Work	Topic Modeling methods	Field of study	Results
Egger & Yu (2022)	- LDA - NMF - Top2Vec - BERTopic	English-language tweets about travel and the COVID-19 pandemic	NMF and BERTopic give best result
Grootendorst (2022)	- BERTopic - LDA - NMF - CTM - Top2Vec	3 separate datasets: 20 NewsGroups, BBC News, and Trump's tweets	BERTopic gives best result
Yin et al. (2022)	- LDA	tweets about the COVID-19 vaccine	
An et al. (2023)	- BERTopic - DCN - LDA - Kmeans	Customer reviews on Korea's leading ecommerce site, Naver Shopping	BERTopic and DCN give the best results
García-Méndez et al. (2023)	- LDA	important financial events from financial news	
Krishnan (2023)	- LDA - NMF - LSA - Top2Vec - BERTopic	2 datasets: Customer Satisfaction dataset from UAE Ministry of Economy and OpSam dataset containing reviews from popular platforms such as TripAdvisor and Amazon Mechanical Turk	BERTopic gives the best result

Table 1: Summary of related works on topic modeling methods across various fields

From the works listed in Table 1, it becomes apparent that LDA is extensively used in these studies, making it the most popular and frequently used topic modeling method in recent literature. Meanwhile, BERTopic, a more recent approach, has consistently produced impressive results. Despite significant exploration of topic modeling across various fields, its application to customer reviews remains underexamined (Krishnan, 2023, p. 4). This shortfall is particularly noticeable in how existing literature often stops at identifying topics and comparing quantitative performance metrics such as coherence score, accuracy, and precision. However, these studies

rarely explore the practical implications of topic modeling for businesses, particularly in utilizing customer reviews to extract actionable insights.

This research seeks to address these gaps by comparing the effectiveness of LDA and BERTopic using the Amazon Reviews dataset and illustrating how topic modeling can be used by businesses to ask and answer key questions for deeper insights. The Amazon dataset was purposefully chosen due to its status as a leading online marketplace with a wide range of product categories and users from diverse demographics and languages. Therefore, it provides a comprehensive foundation for analysis. This study aims to contribute a new dimension to the practical application of topic modeling in customer review analysis.

Summary

This chapter gives an overview of NLP and its extensive use in different fields, serving as a bridge between human communication and machine understanding. One of its applications is topic modeling, a technique that has become a beneficial tool for cultivating valuable business insights from unstructured textual data generated from various online sources. This chapter covers the following key points:

- Natural language is used by humans for daily communication and has evolved over time with its complexities, such as slang, ambiguity, and cultural nuances. Artificial languages, designed with precision for machine interpretation, rely heavily on numbers to process information and perform tasks. NLP is a field that situated at the crossroads of artificial intelligence, computer science, and linguistics. It equips machines with the capability to interpret and analyze human language and opens up a wide range of applications such as language understanding, language generation, language translation, sentiment analysis, named entity recognition, text classification, speech recognition, chatbots, conversational agents, etc.
- Topic modeling is introduced as an NLP technique for discovering hidden topics in a large collection of unstructured text. The development of topic modeling traces back to the early 1990s with the introduction of foundational probabilistic methods such as LSI, pLSI, and LDA to the modern advanced approaches incorporating machine learning and deep learning such as lda2vec, Top2Vec, and BERTopic. Figure 1 presents the timeline of topic modeling evolution, highlighting the milestones achieved and their significant impact on the field.
- A comprehensive summary of recent studies demonstrates the application of topic modeling across diverse fields such as health, e-commerce, finance, academic research, market analysis, and social media. These studies assess the performance of different topic modeling methods in various contexts, highlighting the strengths and limitations of each approach.

Among these methods, LDA and BERTopic stand out as the two most prominent methods. LDA is the most popular method, while BERTopic is the recent advanced technique leveraging the BERT model to achieve remarkable results. Table 1 displays the summary of recent literature on topic modeling methods across fields from 2006 to 2023.

- This study aims to fill the research gap in recent literature by evaluating the effectiveness of LDA and BERTopic in analyzing customer reviews from the Amazon Reviews dataset. The study shall demonstrate how businesses can use topic modeling to ask and answer important questions, providing deeper insights into customer review analysis. By doing so, the study aims to offer a new perspective on the practical application of topic modeling in business intelligence.

3 Methodology

This section presents an in-depth explanation of the fundamental concepts behind two topic modeling techniques, LDA and BERTopic, as well as the framework for implementing and evaluating these methods on the Amazon Reviews dataset. Figure 2 illustrates the entire process, including data collection, data preprocessing, topic modeling, model selection, and model evaluation. Each approach, however, shall require a slightly different process for data preprocessing. The detailed workflow for each method shall be explained further in Sections 3.3.1 and 3.3.2.

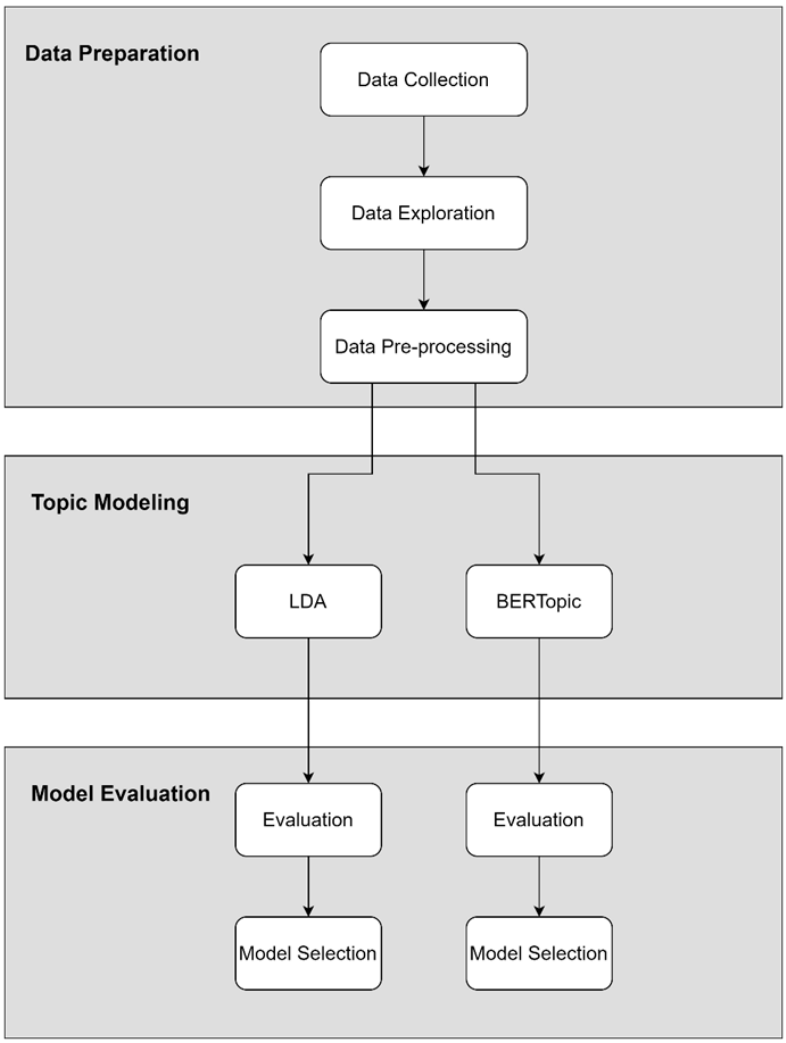


Figure 2: Overview of topic modeling process

3.1 Data Collection

One of the most important aspects of developing any model is to choose an appropriate dataset for the research. Exploring the dataset gives a clear picture of the characteristics of each data point, thereby enabling the selection of suitable processing techniques and models that align with

the study objectives. This study aims to conduct a research experiment on a collection of unstructured data consisting of customer reviews from Amazon in the musical instruments sector. This data was collected by Ni et al. (2019). Instead of using the complete set of 1,512,530 reviews in the musical instruments category, this experiment shall use the subset of that metadata that contains a total of 231,392 reviews, all of which relate to musical instrument items. Table 2 shows the statistics of this dataset. Due to many observations with issues such as missing review text, unverified users, etc., it is essential to thoroughly process and clean the dataset before modeling. While LDA necessitates specific preprocessing steps, BERTopic has different preprocessing requirements and may inherently handle some of these issues.

Dataset	**Description**
Customer Review in Amazon Musical Instruments[1]	231,392 total reviews from 27,530 users Average review length: 57 words Verified users: 27,222 There are some missing review texts, duplicated reviews, unverified users that need to be processed before modeling

[1] *https://nijianmo.github.io/amazon/index.html#subsets*

Table 2: Details of Amazon Reviews dataset

3.2 Data Preprocessing

Data preprocessing plays a crucial role in topic modeling since the quality of input shall determine the quality of output. The extent of data preprocessing varies greatly depending on the chosen method. BERTopic requires little to no data cleaning because the algorithm has built-in text preprocessing during model development. On the other hand, LDA requires extensive and thorough data cleaning and text preprocessing to ensure that the input data is as refined as possible. Data preprocessing consists of two main stages: data cleaning and text preprocessing. Figure 3 outlines the tasks involved in each phase in detail. The data cleaning process includes handling missing values in the review text, dealing with duplicate reviews from the same user for an identical product, and filtering out reviews from unverified users. Tokenization, lemmatization, stop word removal, and the construction of bag-of-words representation are the core elements of text preprocessing.

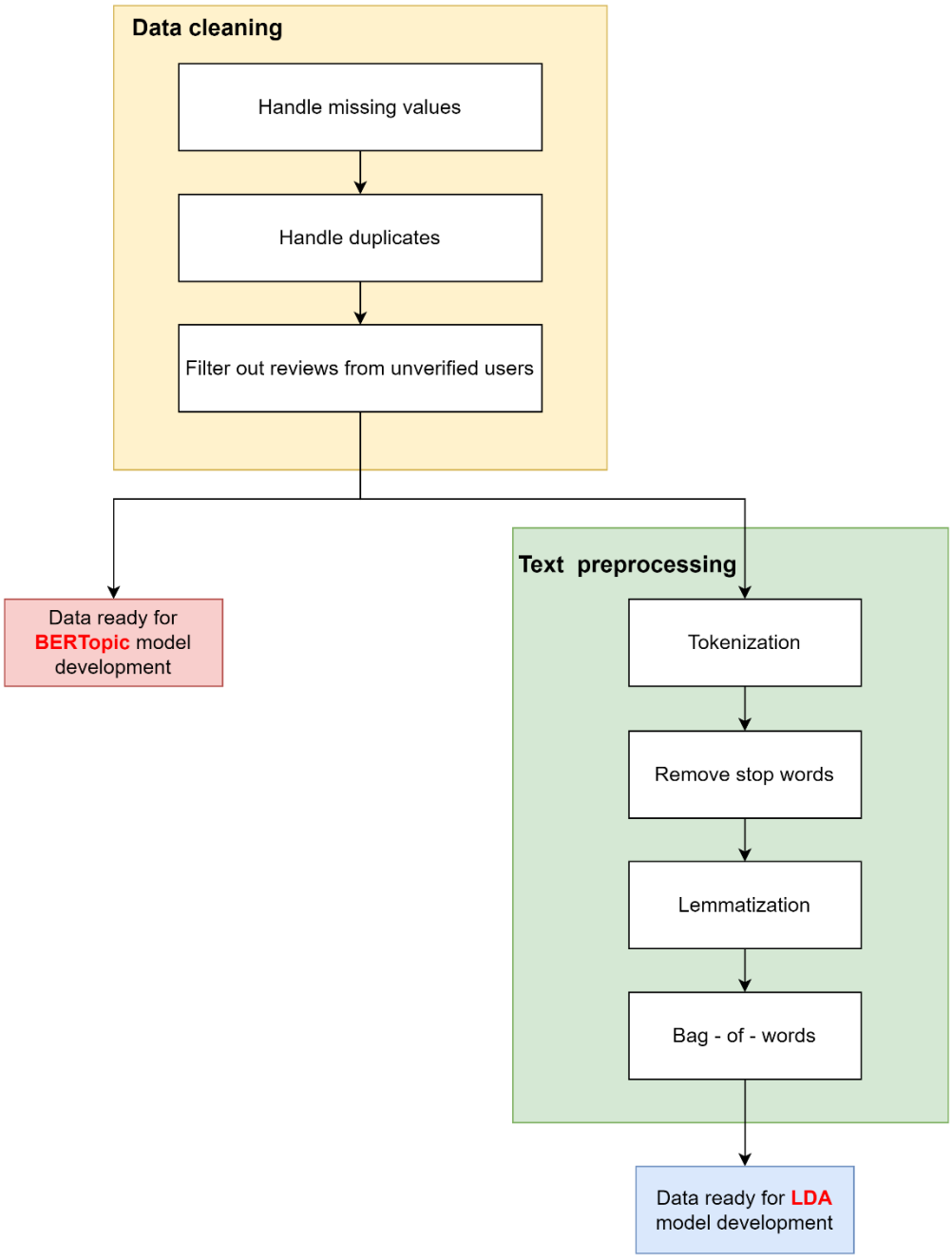


Figure 3: Data preprocessing workflow for topic modeling with BERTopic and LDA

Handle Missing Values

It is crucial to identify and handle missing review text when performing topic modeling, as it can significantly impact the dataset's quality. There are various methods that can be used to address this issue, such as removing entries without text or replacing missing data with appropriate alternatives. In particular, for methods like LDA, where textual input is the backbone of the topic modeling, the presence of any missing text can severely undermine the model construction. Therefore, it is recommended to identify and remove any instances of missing review text.

Handle Duplicates

Duplicate reviews pose another concern as they are repeated reviews that appear more than once within the dataset. Such redundancies do not bring additional meaningful insights but rather create noise that can distort the analytical outcomes. This can be problematic for LDA, which is an approach that relies heavily on the term frequency in documents to determine possible topics. Duplicates can skew the modeling outcomes, which is why they should be treated accordingly.

Filter out Reviews from Unverified Users

Another issue that should be addressed is dealing with reviews from unverified users. In today's world, there are artificial interventions like chatbots, which can generate fake feedback, leading to questionable authenticity of reviews (Luo et al., 2023, p. 2). This could drastically affect the reliability of the data. Hence, it is recommended to remove reviews from unverified sources to ensure that the dataset accurately reflects genuine user experiences and sentiments.

Customers often express their opinions using special characters or repeating phrases to emphasize their thoughts. Many users often use abbreviations and emoticons for shorter sentences, ignoring grammatical rules and spelling checks. As a result, it can be challenging to understand the content when analyzing the feedback. Therefore, the subsequent stage of text preprocessing is crucial to managing noisy data, improving accuracy, and ensuring the model's effectiveness.

Tokenization

Tokenization is a process of dividing a series of strings into separate elements such as words, phrases, keywords, symbols, and other components referred to as tokens. This is the first step in translating human-readable language into tokens that computers can understand. Tokenization eliminates punctuation marks and transforms all tokens into lowercase. The tokens serve as input for various processes, such as parsing and text mining (Haque et al., 2018, p. 3; Mullen et al., 2018, p. 1). There are various methods for tokenization proposed in the literature. It is a crucial stage in BoW feature extraction, where words are divided into tokens (Nayak & Kanive, 2016, p.16878). It is essential to NLP pipelines because it makes modeling and analysis of textual data easier. The simplest way to tokenize a sentence or paragraph is to divide it into individual words, as shown in Figure 4.

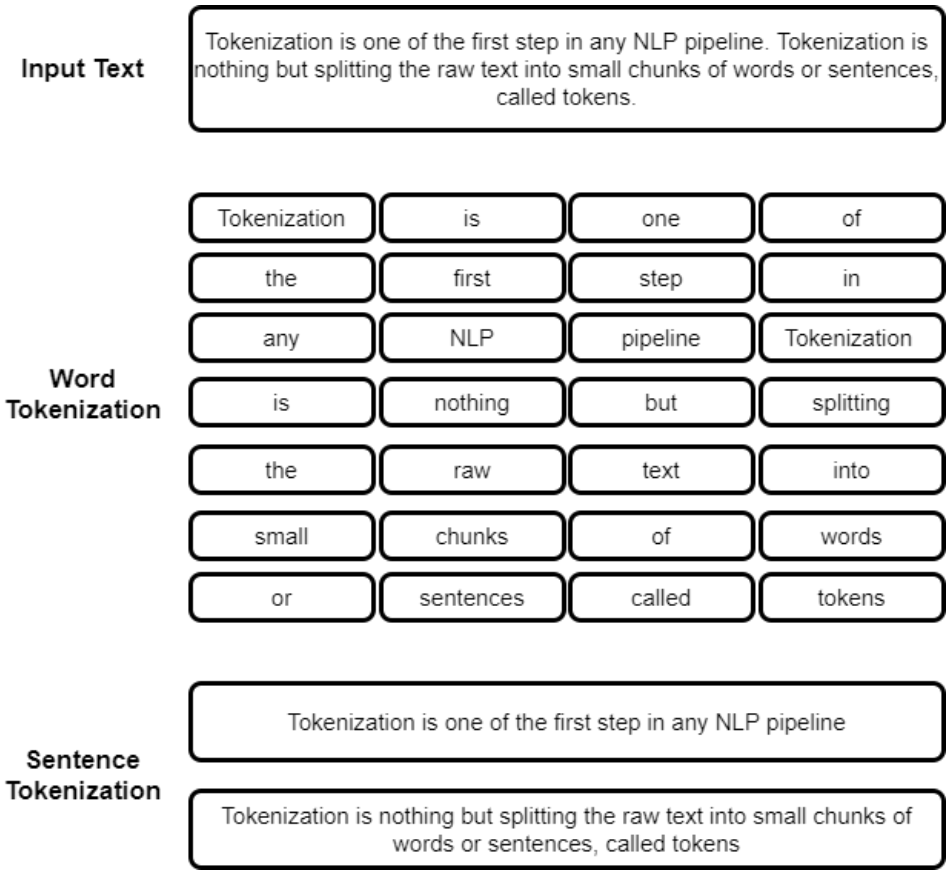


Figure 4: An example of the tokenization process

Removing Stop Words

In NLP, it is common practice to discard certain low-impact words from the text to optimize the analysis. These words, such as "the," "an," "is," and "and," are often redundant, and their removal can streamline computational processes. This helps to focus on words that carry more meaning. Studies conducted by Anees et al. (2020) and Patra and Singh (2013) have both demonstrated that this technique can significantly enhance the precision of operations such as sentiment analysis and data retrieval by filtering out words that contribute little to distinguishing different texts.

LDA is a useful technique for identifying topics in a text document. However, due to the complexities of language, LDA often mistakenly identifies irrelevant word pairings. For instance, common words such as "the" are frequently used to support important nouns that researchers want to highlight. Since LDA relies on the idea that words that appear together are significant, it assigns too much importance to these stop words. In previous studies aimed at improving the quality of topics, researchers have typically ignored stop words in favor of other removal techniques. However, even advanced methods like hyperparameter optimization cannot entirely eliminate the impact of domain-specific stop words and other frequently used words on the quality of topics (Fan et al., 2017, p. 13).

A recent study by Miyajiwala et al. (2022) investigates the effects of making minor modifications to input data on deep learning models used in text classification. Specifically, the study looks at the impact of including or removing insignificant tokens like stop words and punctuations. Three models are analyzed in the study: BERT, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). The findings suggest that these models can handle a few minor changes without much impact on their performance. However, if the number of these perturbations increases, their performance significantly deteriorates, even for the advanced BERT model. Additionally, the study reveals that removing tokens has a more significant impact on these models than adding them. Models with trainable word embeddings are more resilient to these changes. Among the models, CNN is found to be more robust compared to LSTM. Overall, the study demonstrates that even simple alterations in input data can have a significant impact on the performance of these deep learning models.

Lemmatization

Lemmatization is a process that involves grouping various inflected forms of a word into a single category. This process is essential in NLP as it helps in reducing data complexity and is vital for many NLP tasks that rely on lexical analysis (Müller et al., 2015, p. 2268). Unlike stemming, which merely removes a word's suffix, lemmatization uses a vocabulary to derive the root form of a word. For example, while stemming might reduce "driving" to "driv," lemmatization identifies its base form as "drive" or "drived," depending on the context (Khurana et al., 2023, p. 3717). Parts-of-speech (POS) tagging assigns a grammatical category, such as noun, verb, adjective, etc., to each word in a sentence. This step is crucial for lemmatization, as the process of reducing a word to its base form – its lemma – can depend on its grammatical role. After words are tokenized and tagged with their POS, the lemmatization algorithm uses a lexicon or linguistic rules to determine the lemma of each word. Unlike the word's root, the lemma represents its dictionary form, which may vary for irregular verbs or words with multiple lemma options. By leveraging linguistic rules and patterns, lemmatization algorithms can accurately identify the correct base form for such words. The result is a collection of words in their simplest form, making it easier to analyze and understand the underlying meaning of a text (Müller et al., 2015, p. 2268). Lemmatization notably enhances the interpretability of topic models and is particularly useful for languages with rich morphological structures like Russian (May et al., 2016, p. 6).

Bag-of-Words

The bag-of-words (BoW) model is a popular method for representing text in various fields. It supports linear classifiers, which are efficient, robust, and interpretable, allowing for accurate predictions. However, challenges arise with extensive vocabularies, significant variations in word frequencies, numerous classes, and particularly with brief texts such as titles or single sentences. In these situations, the BoW model often leads to excessive sparsity that lowers classifier

accuracy. The problem becomes worse when uncommon or rarely used terminology is included in brief texts that do not provide enough context for classification (Heap et al., 2017, p. 2).

The BoW model is used to extract features by analyzing word frequencies within a document. It counts how often each word appears without considering the order or context of the words. For instance, as shown in Figure 5, the word "it" is mentioned six times, while "I" appears five times, among other word counts. In this model, every unique word in the training corpus becomes a feature, represented by its frequency of occurrence.

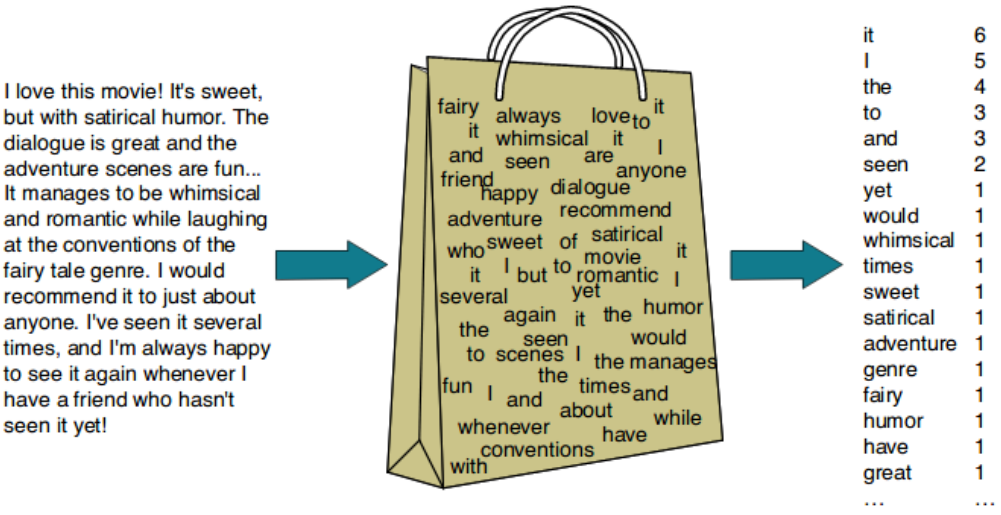


Figure 5: The representation of BoW

There are some problems with using the BoW approach for text analysis. As the text corpus gets larger, the number of unique words also increases, resulting in a mostly zero-filled and sparse representation. Moreover, the BoW model treats documents simply as collections of words without taking into account their syntactic structure. This can lead to misunderstandings of meaning. For example, the sentences "I love apple but I hate pineapple" and "I love pineapple but I hate apple" have the same vector representations in BoW, even though they convey opposite feelings about apples and pineapples. However, despite this limitation, BoW can be improved by using techniques like stop word removal and lemmatization, which can significantly increase its accuracy (HaCohen-Kerner et al., 2020, p. 3).

3.3 Topic Modeling Techniques

Due to the vast amount of textual data available, it is impractical to manually go through each document to make sense of them. As a result, many businesses turn to artificial intelligence tools, such as topic modeling, to extract topics from these documents. Topic modeling does not provide a predetermined topic label but instead produces a group of words that are considered to be

representative of that topic. It is up to the user to interpret and determine what topic the group of words pertains to. For instance, topic modeling may generate words like "space," "launch," "orbit," and "lunar," and the user can conclude that the topic is probably related to space travel based on those words.

3.3.1 Latent Dirichlet Allocation

There are several algorithms available to perform topic modeling, but one of the most widely used algorithms is LDA. It was developed by D. M. Blei and his colleagues in 2003 and has since found various applications, including document classification (Blei et al., 2003, p. 1008), sentiment analysis (Liang et al., 2014, p. 511), and even bioinformatics (Juan et al., 2020, p. 4758). LDA is a three-level hierarchical Bayesian model consisting of three levels - corpus level, document level, and word level. Blei et al. (2003) defines the following important terminologies for LDA:

- A $word$ is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \dots, V\}$. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for $w \neq v$.
- A $document$ is a sequence of N words denoted by $\boldsymbol{w} = (w_1, w_2, \dots, w_N)$, where w_N is the nth word in the sequence.
- A $corpus$ is a collection of M documents denoted by $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_M\}$ (p. 995)

LDA is a statistical model for discovering the underlying topics that are present in a collection of documents. It assumes that each document is made up of a distribution of a fixed number of topics, and each topic is made up of a distribution of words. For example, an article about selling microscopes shall typically include two topics: science and business. Similarly, if the topic is law, the words "code," "decree," and "circular" might occur frequently, whereas the words "house," "goods," and "trains" might appear less often. Figure 6 shows the basic idea behind LDA.

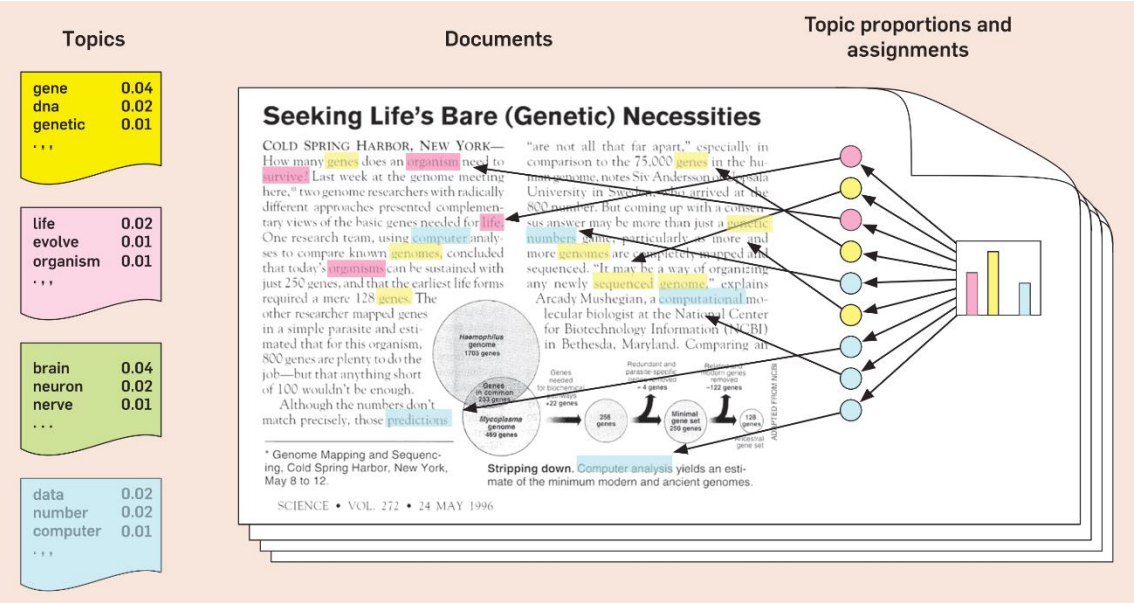


Figure 6: The intuitions behind LDA (Blei, 2012. p.78)

Mathematically, LDA assumes that each document d in a corpus D consisting of M documents have N_d words ($d \in \{1, \dots, M\}$). The generative process for LDA is described by Zhao et al. (2015) as follows:

(1) Choose a multinomial distribution φ_t for topic $t \in \{1, \dots, T\}$ from a Dirichlet distribution with parameter β.

(2) Choose a multinomial distribution θ_d for document $d \in \{1, \dots, M\}$ from a Dirichlet distribution with parameter α.

(3) For a word w_n ($n \in \{1, \dots, N_d\}$) in document d:

   a. Choose a topic z_n from θ_d

   b. Chosose a word w_n from φ_{zn}, a multinomial probability conditioned on the topic z_n

In above generative process, words in documents are the only observed variables while others are latent variables (φ and θ) and hyper parameters (α and β). In order to infer the latent variables and hyper parameters, the probability of observed data D is computed and maximized as follows:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)\right) d\theta_d \text{ (p. 2)}$$

In simple words, the LDA model works in the following way: Initially, all the text in the text set is considered empty without any words. Suppose the texts are a collection of topics. For each text, a topic is selected from the set of topics of that text, and then a word is selected from the set of words of the selected topic. This action is repeated until the probability distribution of the topic

has been determined. This sequence of actions is performed with all documents in the corpus. The graphical representation of the LDA model is shown in Figure 7.

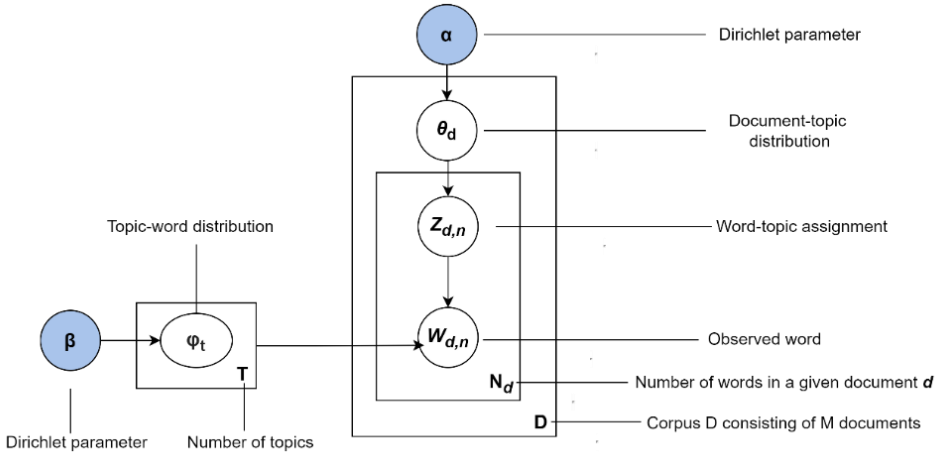


Figure 7: Graphical representation of LDA model

However, reality presents a contrasting picture: For a given collection of documents, all the words presented in these documents are known. The model takes in textual data with different writing styles, vocabulary, and grammar. It presumes that the words in the same document are related and that any document can be described by a formula specifying a combination of various topics and the proportion of each topic that it contains. The model then generates such a document by selecting the appropriate amount of words from specific topics and mixing them together. The result is a file that contains all the topics made up of all the words with their respective probabilities of belonging to each topic. To accomplish this, LDA requires two inputs: the number of topics in the corpus and some additional rules for constructing them – referred to as hyperparameters alpha and beta.

Both alpha and beta are parameters of the Dirichlet distribution. Alpha determines the document-topic distribution, whereas beta is responsible for the topic-word distribution. A high alpha value means that every document is likely to contain a mixture of several topics rather than just one single topic exclusively. A low alpha value indicates that a document is more inclined to be represented by only a few topics. Similarly, a high beta value means that each topic is likely to include a wide range of words rather than being focused on specific phrases. Conversely, a low beta value suggests that a topic might include only a few words. In short, a high alpha value shall make documents appear more similar to one another, and a high beta value shall increase the similarity between topics.

LDA is a BoW model that does not consider syntax rules in the text. If one were to remove the syntax from any given text, it would result in a significant loss of information, making the text unreadable to most people. Nevertheless, people can still process the missing data and infer the

probable topic from the limited set of keywords. The adage "garbage in, garbage out" highlights that the output's quality is contingent on the quality of the input. The model processes the text as a BoW and only understands the words that are included in the model training. Therefore, data preprocessing and dictionary construction are crucial steps in the LDA model. The workflow for the LDA model is shown in Figure 8.

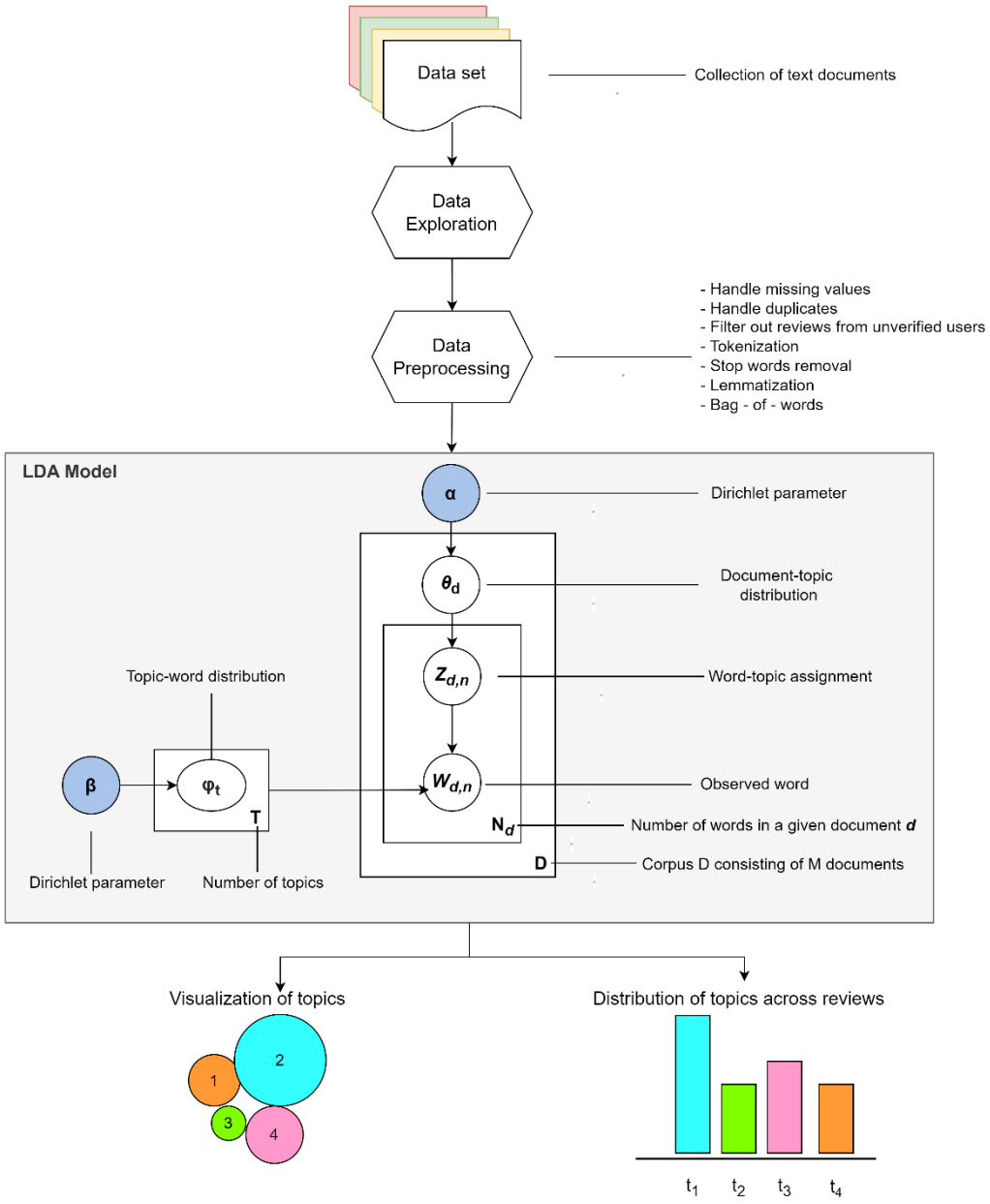


Figure 8: Overview of LDA topic modeling workflow

When using LDA, determining the optimal number of topics and alpha value can be challenging. Hyperparameter tuning is essential to achieve the best LDA model. A grid search is conducted to identify the optimal combination of hyperparameters (Belete & Huchaiah, 2021, p. 1), which systematically explores every possible pairing of the number of topics and alpha values. Although

this process may consume a considerable amount of computational resources and lead to long execution time, it ensures the selection of the final model with the best hyperparameter setup. The number of topics and alpha values shall be selected from a range of values. During the process, one variable is modified while the other remains constant, and the coherence scores are recorded and compared at the end. The coherence score is a metric that measures the similarity of words within each topic (Stevens et al., 2012, p. 954). The higher the coherence score, the more accurate the results are.

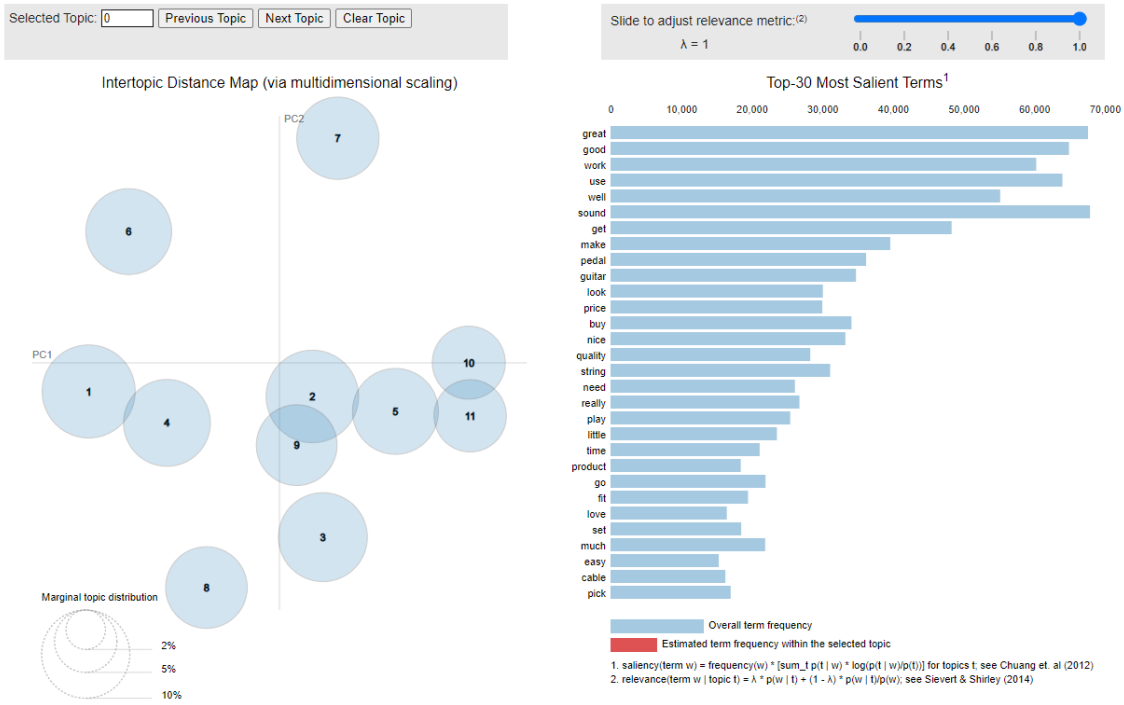


Figure 9: Visualization by pyLDAvis

To get a better understanding of the models, pyLDAvis shall be used to show the intertopic distance map (Islam, 2019, p. 5). Figure 9 provides an example of a pyLDAvis map. Each bubble represents a topic, with the size of each bubble indicating the frequency of the topic. The distance between the bubbles corresponds to the degree of similarity between the respective topics. If the bubbles are far apart, it means that there is a high degree of independence between the topics. On the contrary, if the bubbles are in close proximity or even intersecting, it indicates that the topics are highly similar in nature. An optimal LDA model should have an adequate number of topics, as shown by a reasonable distribution of topics in the pyLDAvis visualization.

3.3.2 BERTopic

Introduced by Maarten Grootendorst in 2022, BERTopic is a sophisticated method designed to identify themes within large sets of text data. It leverages the power of BERT for word embeddings, enabling it to grasp the meanings of words and their relationships in a dataset (Levy

& Goldberg, 2014, p. 304). Additionally, it also uses a class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) technique to highlight keywords while minimizing attention to less important ones. Despite being relatively new, BERTopic has proven to be more effective than traditional topic modeling approaches, such as LDA, achieving higher coherence scores and better interpretability (Krishnan, 2023, p. 11). It is a versatile tool designed for analyzing and interpreting vast amounts of text.

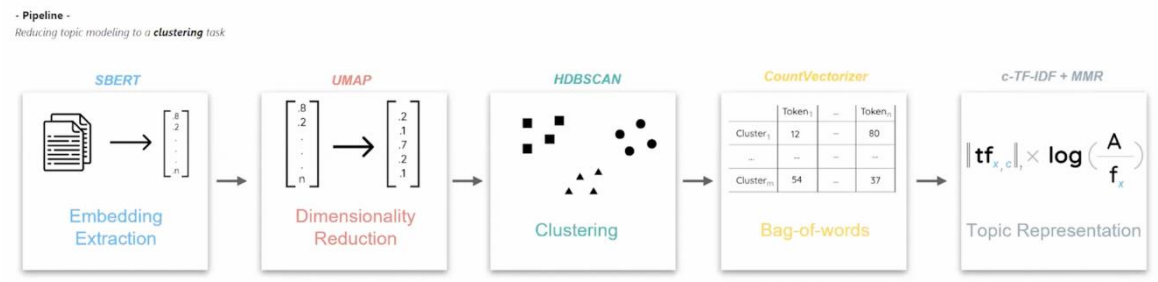


Figure 10: BERTopic pipeline

BERTopic approaches a topic modeling task by reducing it to a clustering task, which is detailed through a five-step process, as illustrated in Figure 10. These steps include:

1. **Embedding Extraction**: This step involves using Sentence Transformers to convert text segments into dense vector representations. The Sentence Transformer library provides a user-friendly platform for generating sentence embeddings by enhancing the SBERT approach. The HuggingFace Model Hub offers access to a variety of pre-trained models for different applications. BERTopic, by default, employs the 'all-MiniLM-L6-v2' model for processing English texts and the 'paraphrase-multilingual-MiniLM-L12-v2' model for handling texts in multiple languages. Despite their smaller size, these models deliver relatively good performance compared to larger models and require less computing power and time for processing. As a result, they are well-suited for environments with limited computational resources, such as Central Processing Units (CPUs).
2. **Dimensionality Reduction**: Large text data is often complex because of the high dimensionality, where each word is mapped as a vector within a vast dimensional space. To simplify this complexity, the dimensionality reduction method Uniform Manifold Approximation and Projection (UMAP) is used. Introduced by McInnes and his colleagues in 2018, it is a non-linear approach for reducing dimensions that effectively maintains the original structure of high-dimensional data at both local and global levels. This technique simplifies embeddings without compromising significant structural information.
3. **Clustering**: After dimensionality reduction, the embeddings are organized into clusters that group documents with similar semantic contents. There are several clustering

algorithms available, such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), k-Means, and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). However, HDBSCAN is the preferred method in the default BERTopic setting. It is a clustering technique that excels in analyzing large text collections by utilizing a density-based approach to form clusters. During this process, it is common to identify outliers or documents that do not comfortably fit into any cluster. These outliers are designated with a topic label of -1, indicating documents not allocated to any specific topic.

4. **Tokenizer**: As clusters can vary greatly in density and shape, a method that relies on a predefined structure of clusters becomes less applicable. In this case, the BoW method, which assumes minimal structure within clusters, is preferred for its flexibility. This approach aggregates all documents within a cluster, treating them as a single document, and then tallies the occurrence of each word across the cluster. CountVectorizer facilitates this process by transforming text collections into a matrix of token frequencies. This allows for the extraction of features from text by representing each document as a vector of word frequencies. This process includes several text preprocessing steps, such as tokenization and stop word removal, to refine the text data further.
5. **Weighting Scheme**: The c-TF-IDF approach is used to give topical labels to each cluster, assigning a unique topic to every one of them. It evaluates the significance of a word in a document by comparing it to an entire corpus. This helps in identifying specific terms that can serve as topical markers for each cluster.
6. **Fine-tuning Representation (optional)**: The field of NLP is evolving rapidly, with the frequent introduction of new methodologies that can improve model performance. KeyBERT, developed by Maarten Grootendorst in his article "Keyword Extraction with BERT," published in Towards Data Science, is a notable method for extracting the most pertinent keywords and phrases from a document's text. Within the BERTopic framework, there is a representation package called KeyBERTInspired that is inspired by KeyBERT and follows the same fundamental concepts. It incorporates optimizations to accelerate the processing time while ensuring faster extraction of keywords and keyphrases without compromising accuracy.

BERTopic is a linear process where each step is designed to be independent of the others and, therefore, does not affect the subsequent steps. Each phase of the process offers a selection of sub-models (Jiang et al., 2023, p. 6), which enables users to customize their topic modeling approach according to their specific needs. As shown in Figure 11, the default configuration of BERTopic includes Sentence Transformers, UMAP, HDBSCAN, CountVectorizer, and c-TF-IDF.

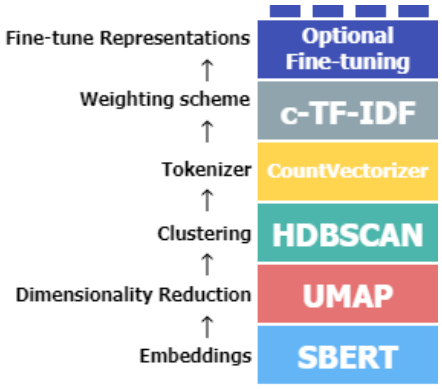


Figure 11: Components of the default BERTopic Model

However, BERTopic's true strength lies in its flexibility, which allows users to modify any part of the process with alternative methods that best align with their specific use case and objectives. This versatility is clearly showcased in Figure 12. For example, a user can opt to replace the Sentence Transformer in the first step with a spaCy model. In case users find UMAP difficult to understand, they can choose to use a PCA model instead. Additionally, some non-western languages may not be tokenized on Spacy, and therefore, user may need to use a different tokenizer.

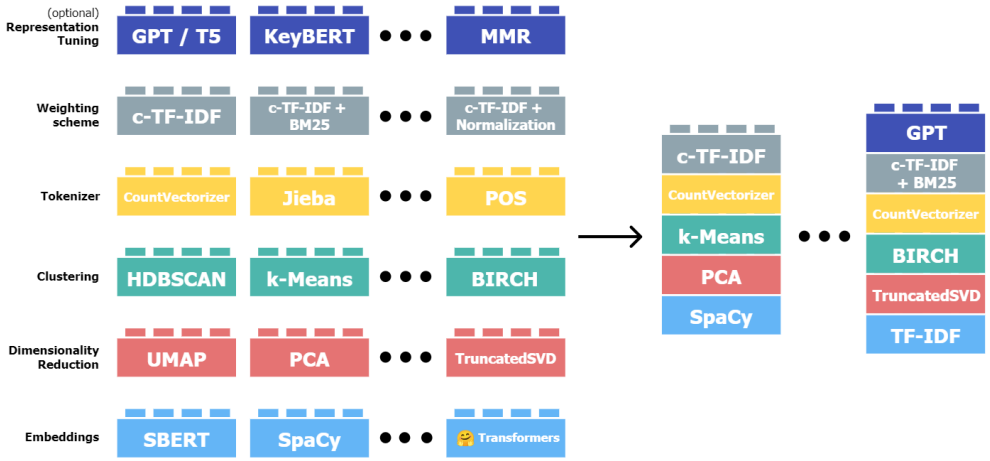


Figure 12: Customizable components for the personalized topic model in BERTopic

BERTopic is a powerful tool that has the capability to address diverse problems across different areas. However, it is important to keep in mind that it may not be suitable for every use case since it is not a one-size-fits-all solution. Therefore, by giving the users the ability to customize, it allows them to pick and choose the most suitable components for their use case, enhancing its utility across various contexts and datasets. The BERTopic modeling process, marked by this customizable nature, is visually summarized in Figure 13.

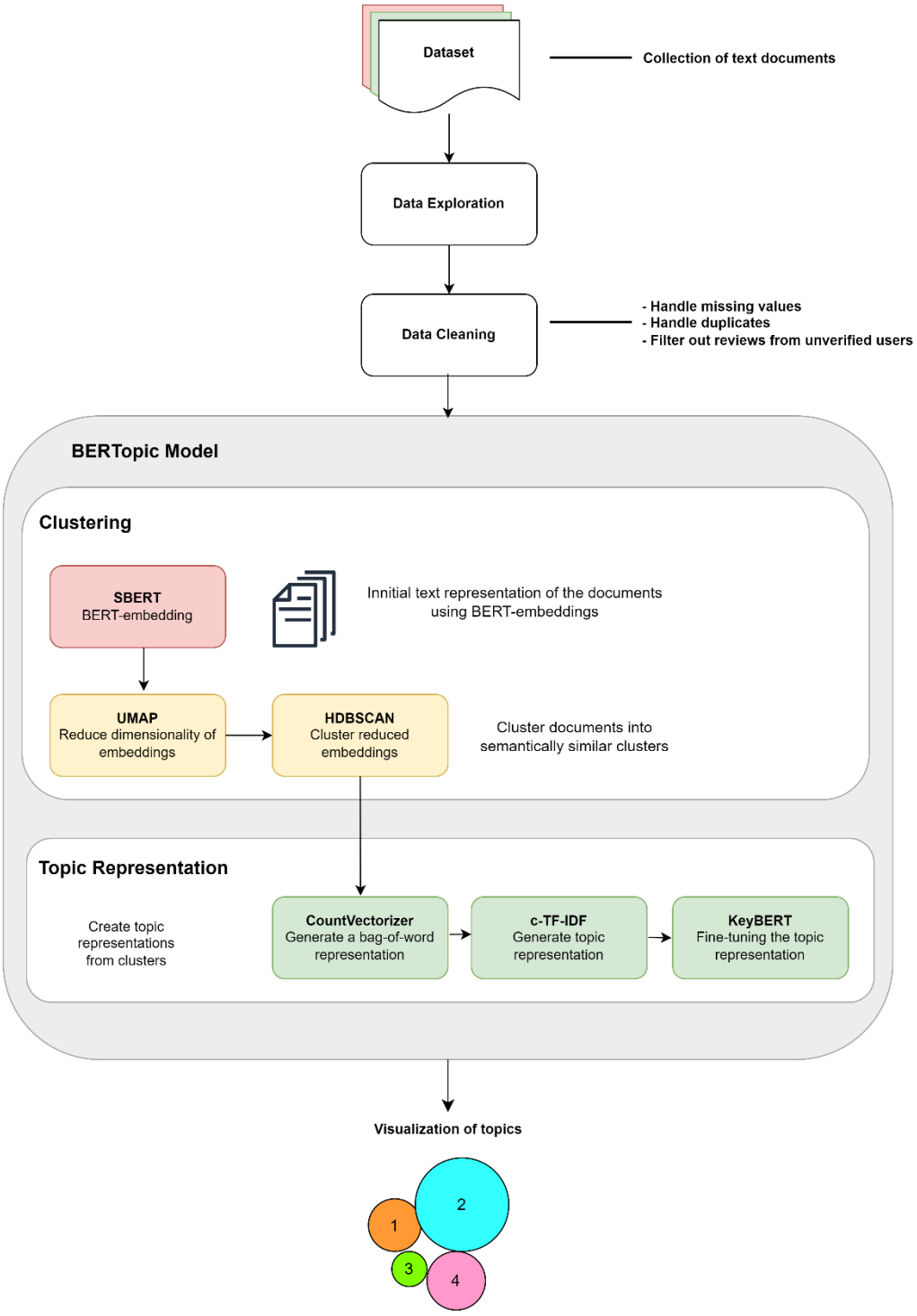


Figure 13: Overview of BERTopic workflow

3.4 Evaluation Metric

To assess and compare different topic modeling techniques, both quantitative and qualitative approaches are necessary. The coherence score is the most commonly used quantitative metric (Rosner et al., 2014, p. 1). Visualization tools and human interpretation of topic quality are

examples of qualitative metrics. The combination of these metrics shall be used to evaluate and compare both LDA and BERTopic.

Coherence Score

Coherence score is a quantitative metric used to evaluate the semantic similarity between high-scoring words within each topic (Stevens et al., 2012, p. 954). A higher coherence score indicates that the topic is more interpretable and consists of words that are more meaningful when considered together (Syed & Spruit, 2017, p. 167). For LDA, coherence scores are essential in optimizing the number of topics and the hyperparameters alpha and beta, which influence topic distribution and word distribution within topics, respectively. In the context of BERTopic, the coherence score helps assess the quality of topics generated through its sophisticated embedding and clustering approach.

Visualization Tools

Visualization tools like pyLDAvis for LDA and the different visualization capabilities of BERTopic provide an intuitive and interactive means to inspect, understand, and interpret the topics generated by each model. These tools allow for the exploration of the distance and relationships between topics, providing insights into topic distribution across the dataset. For LDA, pyLDAvis helps in understanding topic dominance and the distribution of terms within topics. On the other hand, BERTopic, with its advanced algorithm, offers a range of visualization methods that can enhance the user's ability to understand and evaluate the model's output effectively.

Human Interpretation of Topic Quality

Topic modeling inherently grapples with the nuances of subjectivity. As an unsupervised learning technique, it lacks a definitive "correct" answer or ground truth because there is no clear metric to evaluate a topic model's quality (Oelke et al., 2014, p. 205). The content that constitutes a topic is open to interpretation. What one person considers a topic may not resonate as such with another. The absence of a clear-cut metric for evaluating a topic model's quality means that human judgment is essential to validate the relevance of the results for the intended application. While it is possible to train and fine-tune any model to get the highest coherence score, it would be pointless if the results fail to represent the topics. To fully understand the context, it is essential to examine the raw reviews that contribute to each topic (Egger & Yu, 2022, p. 10). Additionally, opinions from domain experts or researchers are also needed to determine the relevance, distinctiveness, and interpretability of the identified topics (Zhou et al., 2021, p. 601). This is to ensure the topics are not only statistically significant but also reflect meaningful and coherent themes relevant to the domain of study.

3.5 Chapter Summary

This section provides a detailed explanation of the fundamental principles that underlie two topic modeling techniques, namely LDA and BERTopic. It also outlines the framework for implementing and evaluating these methods on the Amazon Reviews dataset. The entire process consists of the following steps: data collection, data preprocessing, development of topic modeling, model selection, and model evaluation. The following are the key insights that have been discussed:

- This study uses a subset of Amazon customer reviews in the musical instruments sector provided by Ni et al. (2019). The dataset contains 231,392 reviews in raw text format, which have several issues, such as missing review texts, unverified users, etc. These issues need to be thoroughly addressed before moving to the model development phase.
- Data preprocessing involves two main stages: data cleaning and text preprocessing. Data cleaning deals with missing values, duplicates, and reviews from unverified users. Text preprocessing, on the other hand, includes tokenization, stop word removal, lemmatization, and the creation of BoW.
- LDA and BERTopic require different preprocessing steps. LDA demands extensive data cleaning and text preprocessing to ensure the best data quality before model development. On the other hand, BERTopic requires minimal to no data cleaning, depending on the dataset.
- LDA is a three-level hierarchical Bayesian model developed by D. M. Blei and his colleagues in 2003. Even after two decades since its first introduction, LDA remains the most popular approach for topic modeling. The algorithm assumes documents are mixtures of topics, where topics are distributions over words. The generative process of LDA is influenced by the number of topics and the hyperparameters alpha and beta. Since it cannot predetermine the number of topics, optimizing the LDA model involves a grid search over various numbers of topics and hyperparameter values. The evaluation of the LDA model is based on the coherence score, pyLDAvis visualization inspection, and human interpretation of topic quality.
- BERTopic is a recent advanced approach to topic modeling that utilizes transformers. The process involves five steps: embedding extraction, dimensionality reduction, clustering, tokenizer, and weighting scheme. One significant advantage of BERTopic over other traditional methods is its high degree of flexibility and customizability. Users can switch out the methods used in each step to create a tailored model that suits their specific domain knowledge and use cases. Moreover, BERTopic offers a wide range of user-friendly visualization functions to effectively evaluate the model's output.

- When evaluating and comparing topic modeling techniques, such as LDA and BERTopic, a combination of quantitative and qualitative metrics is used. The quantitative metrics include coherence score, while the qualitative metrics include visualization tools and human interpretation. The goal is to identify meaningful topics that are both statistically significant and relevant to the domain.

4 Experimental Results

This chapter presents the experiments based on the models proposed in chapter three. Two experiments shall be conducted to identify and classify common topics from unstructured customer reviews on the Amazon Reviews dataset. The experiments are as follows:

- Experiment 1: Building a topic model using the LDA method
- Experiment 2: Building a topic model using the BERTopic method

Based on the experimental results, the evaluation, comparison, discussion, and future directions shall be provided.

The first experiment carries out the following tasks:

(1) Data exploration

(2) Data preprocessing, including data cleaning and text preprocessing

(3) Building different LDA models with the number of topics varying between 2 to 20 and alpha set to values from [0.01, 0.1, 1, 'auto']

(4) Evaluating the topic quality by coherence score, pyLDAvis map and human interpretation

(5) Selecting the final LDA model

(6) Delivering business insights derived from the outcomes of the final LDA model

The second experiment carries out the following tasks:

(1) Data exploration

(2) Data cleaning

(3) Building the BERTopic model with the default setting

(4) Hyperparameter tuning for the final BERTopic model

(5) Delivering business insights derived from the outcomes of the final BERTopic model

The experiments are conducted using Python on a local CPU environment. For the construction of LDA models, the Gensim library is used, as it is known for its efficiency and ease of use in processing texts for topic modeling. The text preprocessing for LDA incorporates spaCy for the lemmatization process to reduce words to their dictionary form and the comprehensive stop words list provided by the Natural Language Toolkit (NLTK) to filter out irrelevant words and enhance model accuracy.

For the development of the BERTopic model, the BERTopic library is used, along with many sophisticated libraries designed to optimize topic modeling performance. These included the SentenceTransformer library for generating dense vector representations of sentences, UMAP for

dimensionality reduction, HDBSCAN for clustering data points, and CountVectorizer from scikit-learn for converting text data into a matrix of token counts. Furthermore, the BERTopic model's representation and relevance measures were enriched with KeyBERTInspired, MaximalMarginalRelevance (MMR), and POS tagging to fine-tune the model's ability to discern and articulate topics with higher relevance and coherence.

4.1 Data Exploration

The experiments utilized data from the musical instruments category in the Amazon Reviews dataset. The dataset consists of 231,392 observations, each of which contains relevant information such as product details, review content, user ID, images, etc. Each entry in the dataset represents a review of a particular product. The data was extracted and stored in a DataFrame called ‘df’, along with its corresponding features. Table 3 and Figure 14 provide a comprehensive list of attributes and their corresponding descriptions.

Column Names	**Description**
overall	ID of the reviewer
verified	verified user or not
reviewTime	time of the review (raw)
reviewerID	ID of the reviewer
asin	ID of the product
reviewerName	name of the reviewer
reviewText	text of the review
summary	summary of the review
unixReviewTime	time of the review (unix time)
vote	helpful votes of the review
style	a disctionary of the product metadata
image	images that users post after they have received the product

Table 3: Field descriptions of the Amazon Reviews dataset

	overall	verified	reviewTime	reviewerID	asin	reviewerName	reviewText	summary	unixReviewTime	vote	style	image
0	5.0	True	10 30, 2016	A3FO5AKVTFRCRJ	0739079891	francisco	It's good for beginners	Five Stars	1477785600	NaN	NaN	NaN
1	5.0	True	06 30, 2016	A3UCGC1DHFMBCE	0739079891	Eb Jack Murray	I recommend this starter Ukulele kit. I has e...	Five Stars	1467244800	NaN	NaN	NaN
2	5.0	True	05 9, 2016	A2S9SLRYLPGYZB	0739079891	Clara LaMarr	G'daughter received this for Christmas present...	Learning new songs to play regularly	1462752000	NaN	NaN	NaN
3	4.0	True	04 10, 2016	A15RTJWPG8OKOE	0739079891	Eagle80	According to my order history, I bought this t...	A bargain-bin good-enough ukulele that's held ...	1460246400	NaN	NaN	NaN
4	1.0	True	02 6, 2016	A12ET1WO3OAVU7	0739079891	Amazon Customer	Please pay attention better than I did to the ...	Poor Quality product.	1454716800	NaN	NaN	NaN

Figure 14: Sample entries from the Amazon Reviews dataset

This dataset contains customer reviews written in different languages, as shown in Figure 15, indicating a diverse international user base. The dataset predominantly consists of 213,976 reviews in English, while 30 other languages also contribute significantly, such as Afrikaans (3,001 reviews), Romanian (2,486 reviews), Catalan (1,942 reviews), Somali (1,687 reviews), German (923 reviews), and French (667 reviews). Additionally, there are 132 entries in the 'Error' category, which suggests potential issues in language detection or data collection that may require attention. The linguistic diversity of the dataset highlights the necessity for incorporating multilingual analysis during the model development phase.

```
language
en       213976
af         3001
ro         2486
ca         1942
so         1687
de          923
fr          667
es          543
pl          499
no          497
sl          466
sk          452
cy          445
da          430
pt          381
it          345
et          320
hu          320
tl          297
nl          226
sv          224
vi          190
cs          169
sq          142
fi          141
id          137
hr          136
Error       132
tr          106
sw           66
lv           32
lt           14
Name: count, dtype: int64
```

Figure 15: Distribution of reviews in multiple languages in the Amazon Reviews dataset

The reviews were collected from October 2003 to September 2018. Most of the reviews were written between 2013 and 2017. However, there is missing data from November 2003, December 2003, January 2004, February 2004, April 2004, May 2004, May 2005, and July 2005. Normally, missing values in other analytical models would be replaced or removed before the model development phase. However, for topic modeling, the missing dates do not affect the development of a topic model since the textual data is the primary factor that affects the outcome of a topic model. Therefore, there is no need to eliminate or substitute them. Nevertheless, it is essential to

keep this in mind and revisit the issue once the topics are defined. Figure 16 displays the distribution of all reviews over time from 2003 to 2018.

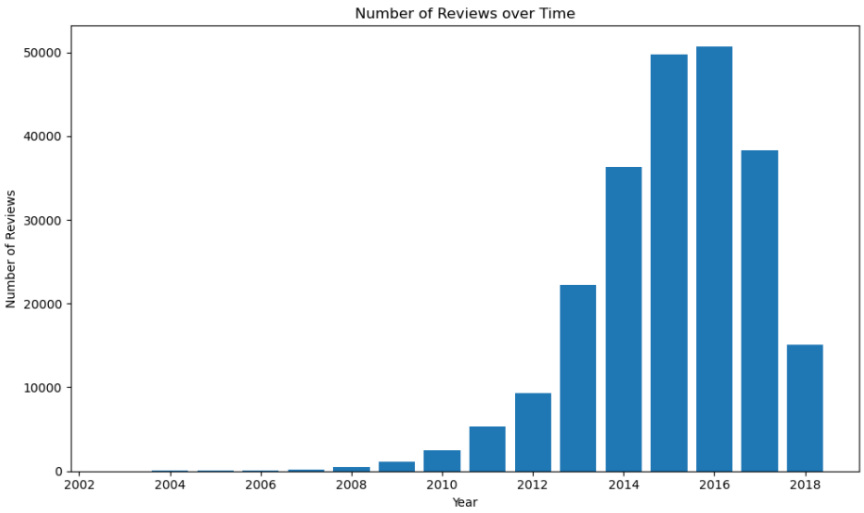


Figure 16: Distribution of reviews over time (2003 – 2018)

According to the product rating scale, 1.0 is the lowest rating, and 5.0 is the highest rating. Based on the data shown in Figure 17, most customers gave ratings of 5.0 and 4.0, accounting for 86.7% of all reviews. On the other hand, ratings of 1.0 and 2.0 only make up for a little over 6% of all reviews. This indicates that customers are generally satisfied with their purchases. This information shall be useful later when conducting a more in-depth analysis of which features or areas require improvement or continuing development after the topic modeling is completed.

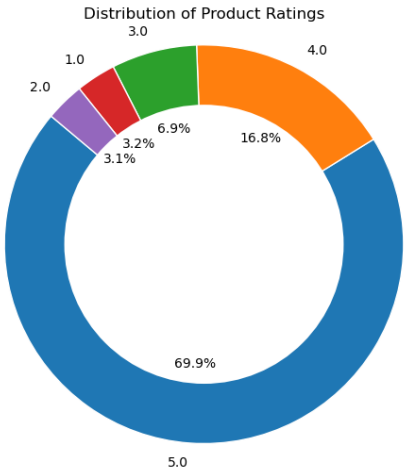


Figure 17: Distribution of product ratings in the Amazon Reviews dataset

This dataset contains a mixture of long and short text, with an average review length of 57 words. The longest review is 4,069 words, while the shortest review has either none or just one word. As shown in Figure 18, the word count distribution reveals that most of the reviews are short text, with less than 200 words.

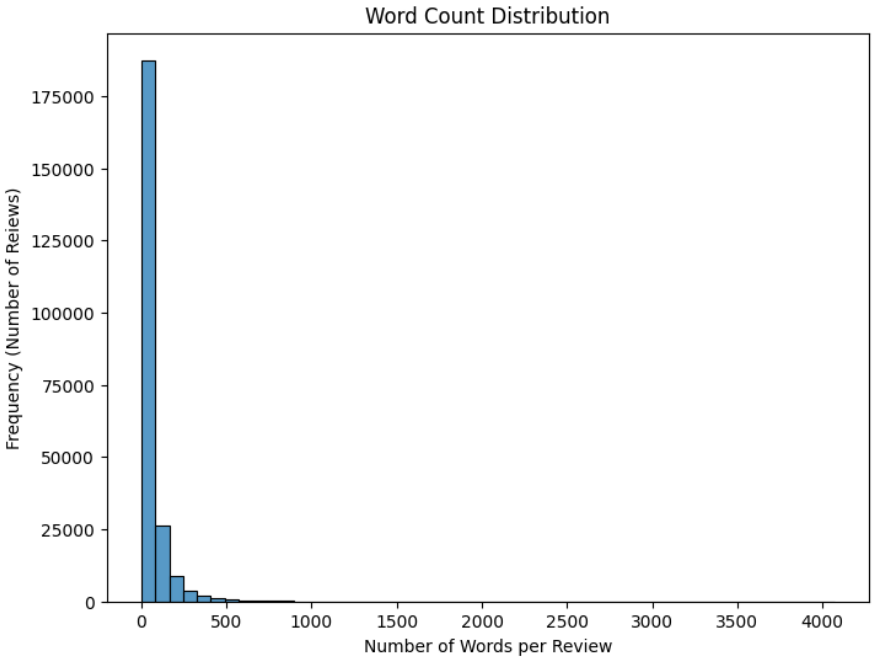


Figure 18: Distribution of word counts for all reviews in Amazon Reviews dataset

After performing an exploratory analysis, some issues are found that need to be addressed during the next step of data preprocessing:

- **Missing values**: There are missing values (NaN) in several columns, namely: reviewerName, reviewText, summary, vote, style, and images. To be precise, there are 25 missing values in reviewerName, 48 in reviewText, 51 in summary, 196,615 in vote, 110,082 in style, and 227,503 in image. Out of these numbers, only the missing values (empty string) in the reviewText column shall severely impact the process of building a topic model. Besides, because 48 out of 231,392 is a relatively small percentage, these reviews shall be excluded from the dataset before building both LDA and BERTopic models.
- **Duplicates**: 18,571 rows were identified as duplicates and, hence, need to be eliminated from the dataset.
- **Unverified users**: Out of 27,530 total users, there are 9,075 unverified users. As text bots and fake reviews are a significant issue in customer reviews, it is advisable to remove these reviews to reduce the noise in the dataset.

4.2 Data Preprocessing

Based on the findings from data exploration, the subsequent preprocessing steps are undertaken to refine the dataset for topic model development. The original dataset of 231,392 entries is reduced to 231,344 entries by eliminating 48 rows with missing review content. Subsequently, an

additional 18,571 duplicates are discarded since they do not contain significant information for the topic modeling process. This step substantially reduces the number of reviews to 221,955 entries. The final cleaning measure involves excluding the reviews from unverified users due to their potential to compromise the model's reliability by introducing inauthentic reviews from unverified sources, which may be generated by automated bots. From the dataset, a total of 23,015 reviews written by 9,075 unverified users were removed. Following the completion of the data cleaning procedure, the remaining dataset has 198,940 reviews in total. This marks the removal of 32,452 entries, accounting for approximately 14.02% of the initial dataset. Figure 19 summarizes the data cleaning process. This data can now be fed into BERTopic for topic modeling development. For LDA, additional text preprocessing steps are required after this.

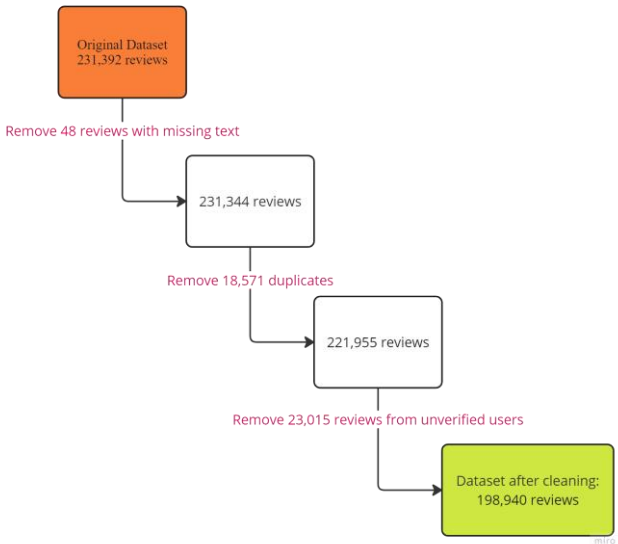


Figure 19: Data cleaning process for Amazon Reviews dataset

Next, the tokenization process uses the ‘simple_preprocess’ method from the Gensim library to tokenize the review content. This method processes each review, converting them into lists of words and removing punctuation. By iterating over each review in the dataset and applying this method, the raw review texts are transformed into a format that is better suited for an NLP task. The result is a list of lists, where each sub-list contains the tokenized words of a corresponding review. An example of the output from the third entry in this list is displayed in Figure 20. It illustrates a review text before and after the tokenization process, in which the review text is stripped of punctuation and broken down into words.

```
Before tokenization:

G'daughter received this for Christmas present last year and plays if often.


After tokenization:

['daughter', 'received', 'this', 'for', 'christmas', 'present', 'last', 'year', 'and', 'plays', 'if', 'often']
```

Figure 20: Tokenization example for the third review in the dataset

Following that is the removal of stop words from the tokenized reviews. Stop words are frequently encountered words in a language that are generally considered irrelevant for data analysis, such as "the," "is," and "in." The elimination of stop words involves using a pre-defined list of standard English stop words from the NLTK library for Python. This list is further extended by including the set of punctuation marks. Afterward, a function is constructed to filter out these stop words from the list of tokenized words that are previously obtained from the reviews. This function iterates over each review in the dataset and removes words that appear in the stop words list. The result is a sublist of the tokenization list above, where the filtered tokens now contain no stop words and punctuations. The result of this process is illustrated in Figure 21.

```
Before removing stop words:

['daughter', 'received', 'this', 'for', 'christmas', 'present', 'last', 'year', 'and', 'plays', 'if', 'often']


After removing stop words:

['daughter', 'received', 'christmas', 'present', 'last', 'year', 'plays', 'often']
```

Figure 21: Example of stop word removal for the third review in the dataset

The text preprocessing continues with lemmatization – a process that transforms words to their base or dictionary form, known as lemmas. Utilizing the spaCy library in Python, a lemmatization function was defined to process the previously filtered tokens. This function considers only words that are nouns, adjectives, verbs, or adverbs since they greatly contribute to the semantic structure required for topic modeling. After the completion of the lemmatization process, a list of lists is obtained, where each sublist contains the lemmatized tokens. Figure 22 shows the example of lemmatization for the third review in the dataset.

```
Before lemmatization:

['daughter', 'received', 'christmas', 'present', 'last', 'year', 'plays', 'often']


After lemmatization:

['daughter', 'receive', 'present', 'last', 'year', 'play', 'often']
```

Figure 22: Lemmatization example for the third review in the dataset

Once the lemmatization process is done, the next step involves constructing a dictionary and a corpus that are necessary for LDA. The Gensim library offers the ‘corpora.Dictionary’ function, which assigns each unique lemmatized word to a specific integer ID, to help with the creation of a dictionary. This procedure is applied to the lemmatized data, resulting in an indexed dictionary. Afterward, a corpus is generated, which consists of a list of BoW representations of each document. It converts each document into a combination of word IDs and their corresponding frequencies. An example is shown in Figure 23. This concludes the data preprocessing steps for LDA, setting a foundation for the next model development phase.

```
Lemmatized text:
['daughter', 'receive', 'present', 'last', 'year', 'play', 'often']


Corpus:
[(5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1)]


Human-readable format of corpus (term-frequency):
[[('daughter', 1),
  ('last', 1),
  ('often', 1),
  ('play', 1),
  ('present', 1),
  ('receive', 1),
  ('year', 1)]]
```

Figure 23: Example of corpus and frequency dictionary for the third review in the dataset

4.3 LDA Result

The evaluation metric for choosing the best LDA model is the coherence score (c_v). The coherence score increases proportionally with the number of topics in the model. Since building these models requires significant computing power, the optimal LDA model should have a high coherence score and minimal number of topics. Additionally, the final LDA evaluation involves human interpretation of topic quality by analyzing the top 10 frequent words in each topic and inspecting the topic visualization by pyLDAvis. This is done to ensure that the topics provide

meaningful insights and add value to the business. In order to achieve this, a grid search is run using different combinations of numbers of topics {2, …, 20} and alpha values {'auto', 0.01, 0.1, 1}. The resulting coherence scores are recorded and shown in Table 4 and Figure 24 below. Some of the potential candidates for final LDA models are highlighted in green in Table 4 and presented as the local peaks in each line in Figure 24.

Number of topics	Coherence Scores with alpha = auto	Coherence Scores with alpha = 0.01	Coherence Scores with alpha = 0.1	Coherence Scores with alpha = 1
2	0.528927632	0.453106404	0.459132372	0.473389578
3	0.554293096	0.482753595	0.481716651	0.519850014
4	0.542902811	0.445091207	0.469870877	0.535936902
5	0.505212963	0.485354891	0.509068126	0.552197782
6	0.531604294	0.476353278	0.494843127	0.536505927
7	0.50242819	0.474368964	0.509302314	0.540786036
8	0.49198595	0.47992104	0.495307265	0.538659689
9	0.485872779	0.477443554	0.487585017	0.530576087
10	0.44860492	0.473387439	0.481743323	0.519918912
11	0.474476123	0.446825703	0.473889612	0.530945086
12	0.452897743	0.440051196	0.450047471	0.525573151
13	0.436704762	0.44273628	0.450429861	0.496960026
14	0.417463406	0.456615433	0.455213231	0.489994458
15	0.433435626	0.463255895	0.453897953	0.495868455
16	0.416314396	0.441962317	0.445329694	0.486415613
17	0.416323846	0.448450245	0.447687051	0.487833239
18	0.413733296	0.428189076	0.457395815	0.482372603
19	0.433795517	0.440633216	0.458031007	0.477500705
20	0.415855787	0.430001225	0.444565117	0.473783715

Table 4: Coherence scores across different LDA model configurations

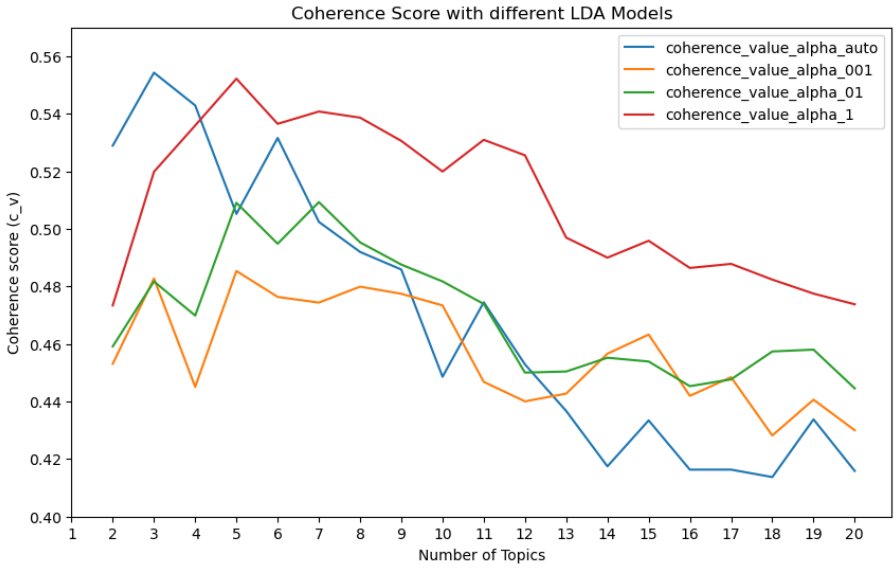


Figure 24: Coherence score trends across different LDA models

The line chart in Figure 24 illustrates the coherence scores of different LDA models with varying alpha values across a range of topics. From this graph, it is clear that LDA models with an alpha value of 1 obtain the best coherence scores. However, to determine which of these models provides the optimal results, further analysis is needed to assess the coherence score of these models. The ideal LDA model should have the lowest possible number of topics while simultaneously achieving the highest coherence score. Figure 25 shows the details of LDA models with the alpha value of 1.

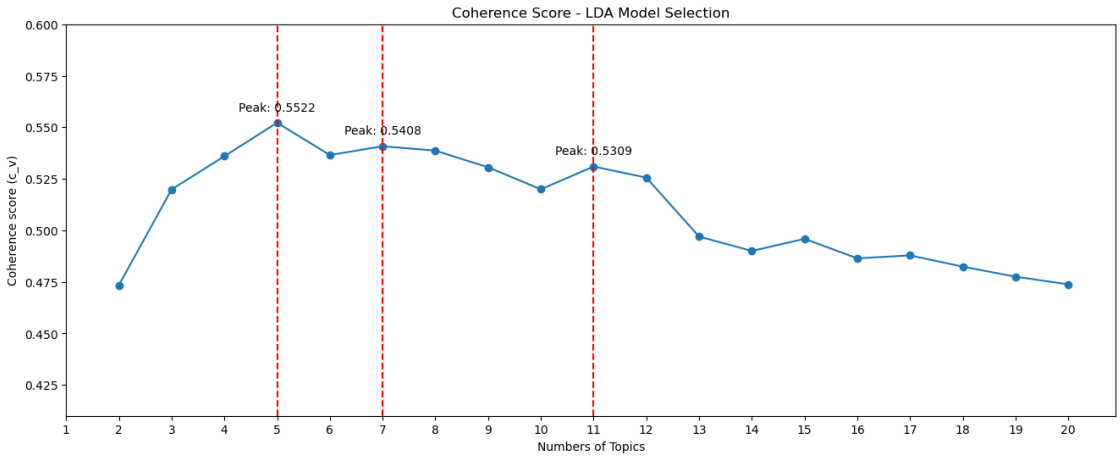
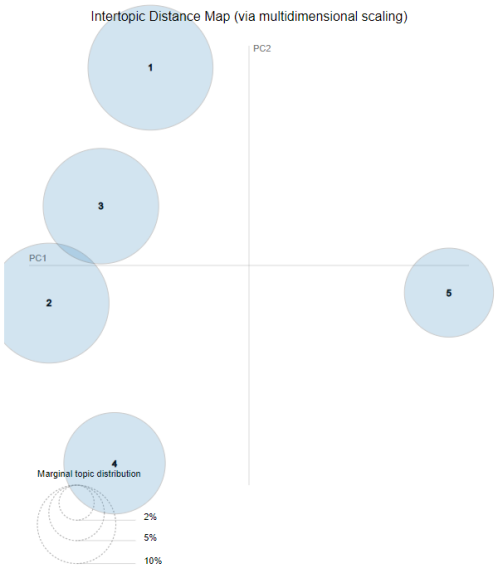


Figure 25: Optimal number of topics for LDA model selection based on coherence scores

In Figure 25, the vertical red lines mark the local maximum points of coherence scores, suggesting where the potential optimal models could be found. These peaks occur at 5, 7, and 11 topics, with the highest peak at 5 topics. As the higher coherence score indicates the extent to which the words in a topic are related to each other, the highest peak at 5 topics suggests that the model with 5

topics has the most coherent and interpretable topics, making it an ideal candidate for final model selection.

However, it is important not to rely solely on coherence scores to select the final model. A qualitative review of the topics is equally important. This process involves analyzing the most frequent words within each topic and examining topic visualizations. The ideal model should have topics that are distinct from each other with minimal overlap and are evenly distributed across the visualization space. Hence, while the 5-topic model initially stands out based on coherence, the models with 7 and 11 topics may also offer other valuable insights and should be examined through visual and qualitative analysis before making a final decision.



(1) LDA model with 5 topics

(2) LDA model with 7 topics

(3) LDA models with 11 topics

Figure 26: Visualization of topic clusters for LDA models with 5, 7, and 11 topics

Upon carefully examining the intertopic distance maps for models with 5, 7, and 11 topics in Figure 26, the following observations on the topic clusters are made:

- The model with 5 topics forms clear, distinct topic clusters with very little overlap. However, the clusters do not occupy a large portion of the surface. This raises the question that this model may not fully capture the diversity of the data and oversimplify the underlying topics.
- The model with 7 topics has excessive overlaps between topics 1, 2, 3, and 4. Such overlaps are not ideal because topics could be too closely related, which could cause ambiguity when classifying reviews into these topics.
- The model with 11 topics has some overlaps, especially between topics 9 and 2, and 1 and 6. While having overlaps is generally not preferred, it is less extensive than in the seven-topic model. Besides, the topic clusters are relatively equal in size and more evenly distributed across the surface. This may indicate a better representation of the data's complexity and diversity.

Based on these observations, the decision for the final LDA model narrows down to the model with 5 or 11 topics. To make an informed decision, it is necessary to examine the words associated with each topic in these models. In order to streamline the analytical process, each topic is labeled based on its top frequent words. This list of words and their labels is shown in Table 5. It is important to note that these labels are provisional and intended primarily to make it easier to analyze the data as their accuracy is not guaranteed. The real assignment of topic labels demands the input of domain experts, who can draw on their knowledge to recognize specific topics and issues within their field.

	Model with 5 topics		**Model with 11 topics**	
Topic	**10 frequent words**	**Topic labels**	**10 frequent words**	**Topic labels**
0	use, sound, nice, string, well, really, make, love, little, easy	Musical Instrument Quality and Performance	small, light, stand, enough, hold, keep, size, build, bag, solid	Portable Equipment
1	fit, strap, guitar, make, case, stand, neck, hold, size, need	Guitar Accessories	get, little, thing, way, bit, take, pretty, want, right, star	Preferences and Satisfaction
2	get, play, guitar, go, buy, pick, try, say, want, bass	Guitar Playing	sound, pedal, tone, really, pickup, effect, volume, knob, clean, color	Audio Effects and Adjustments

	Model with 5 topics		Model with 11 topics	
Topic	**10 frequent words**	**Topic labels**	**10 frequent words**	**Topic labels**
3	pedal, sound, use, cable, mic, switch, effect, power, need, volume	Audio Equipment and Sound Effects	string, set, high, bass, case, end, low, lot, replace, change	Guitar Accessories and Maintenance
4	great, good, work, price, quality, well, product, look, buy, perfect	Value for Money	guitar, nice, play, strap, easy, neck, tuner, instrument, fret, tune	Guitar Features and Playability
5			great, good, price, quality, product, look, love, awesome, deal, fast	Value for Money
6			buy, go, come, say, cheap, see, purchase, first, think, know	Purchasing Decisions
7			make, much, even, put, find, new, money, sure, hard, worth	Financial Considerations
8			work, well, need, fit, perfect, recommend, fine, expect, perfectly, exactly	Product Performance and Fit
9			cable, use, mic, switch, power, amp, unit, board, speaker, plug	Audio Equipment and Connections
10			use, time, pick, seem, try, feel, long, still, far, year	Time and Experience

Table 5: Topic labels and frequent words in LDA models with 5 and 7 topics

The model with 5 topics offers a broader and simpler representation of the data, making it easier to interpret at a high level. On the other hand, the 11-topic model provides a greater variety in the top words, indicating more nuanced and specific topics. This granularity provides a more in-depth understanding of the dataset and can be particularly useful in categorizing documents into more precise topics. In addition, it also has the third highest coherence score. Therefore, after careful deliberation, the most suitable choice for the final LDA model is the 11-topic model as it can provide more insights from the data. Figure 27 and Table 6 contain the details of the final LDA model.

Figure 27: Visualization of the final LDA model

Final LDA Model	
Alpha value	1
Number of topics	11
Coherence score	0.530945086

Table 6: Statistics of the final LDA model

Selecting the final LDA model and identifying the topics are only the beginning of the process. The real value lies in turning these topics into business insights, which require further analysis. With LDA, each review is made up of distributions of many topics. To simplify the analysis, each review is assigned to the single most probable topic. This approach makes it easier to handle the

data and allows for different types of analysis based on business needs and requirements. Examples of some key questions that can be answered with additional analytics include:

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?

To conduct a complete analysis, it is crucial to take both quantitative and qualitative measurements into account. This requires delving into the actual reviews that form the core of each topic and consulting with industry experts. These experts can provide valuable insights on interpreting the data and ensuring that decisions made are logical, both in terms of numbers and real-world context. Once this comprehensive analysis is complete, suitable solutions and strategies can be formulated.

Regarding the most discussed topics, an overview of the topic distribution is shown in Figure 28. It appears that topic 5 is the most discussed among all users with 37,985 reviews. A closer look at this topic reveals keywords such as "great," "good," "price," "quality," "product," "look," "love," "awesome," "deal," and "fast." These terms suggest that topic may revolve around "Value for Money." While words like "price," "quality," "product," and "deal" might relate to monetary aspects, other words like "great," "good," "love," and "awesome" are indicative of positive customer experience and satisfaction. In contrast, topics 3 and 7 exhibit the least engagement with 9,338 and 9,759 reviews respectively. These topics may either represent a niche area of interest within the dataset or not well-captured hidden themes during the model development process.

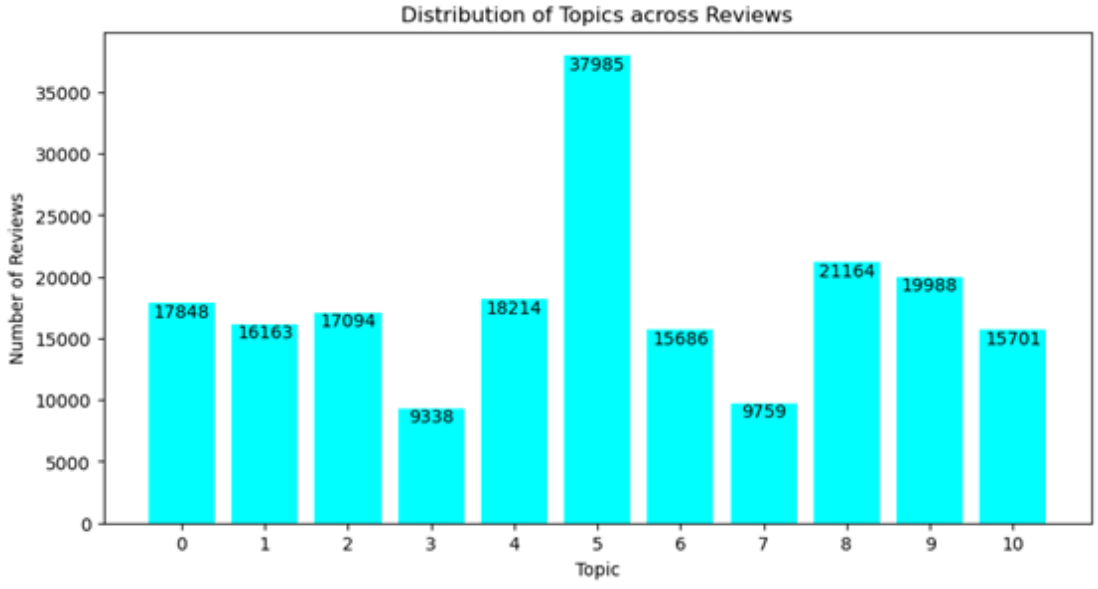


Figure 28: Topic distribution across all reviews in final LDA model

Based on these observations, some suggestions can be drawn:

- Businesses should further go through the content written in reviews associated with topic 5 to pinpoint what drives customer satisfaction. Additional sentiment analysis on this topic might help to better understand the nuances of customer feedback, including positive, neutral, and negative sentiments.
- The low discussion around topics 3 and 7 presents an opportunity for businesses to increase focus in these areas. This could involve offering additional information or improved product offerings that could increase customer engagement and satisfaction in these areas.

To understand how customer interests and market trends shift over time, it is helpful to look at topic trends. Normally, some topics might become less relevant as others gain traction. According to the data exploration phase, most of the dataset covers the years 2013 to 2017, which coincides with the observed fluctuations in topic popularity, as presented in Figure 29. In general, every topic experienced some level of increased attention over the years. However, the period from July 2014 to June 2018 witnessed particularly notable spikes in the frequency of discussion of topic 5 and topic 8. These sudden increases could be attributed to major events or shifts within the music industry, which require expert knowledge to fully assess. For instance, there could be advancements in musical technology or influential endorsements of specific instruments or accessories during this period that spurred interest in these topics. Additionally, broader economic trends, like increased disposable income or the introduction of cost-effective product lines, could have contributed to the rise in these discussions. For businesses, understanding these fluctuations is beneficial to investigate the underlying causes of the heightened interest in topics 5 and 8. This can drive decisions to make adjustments in product development, marketing strategies, or customer engagement.

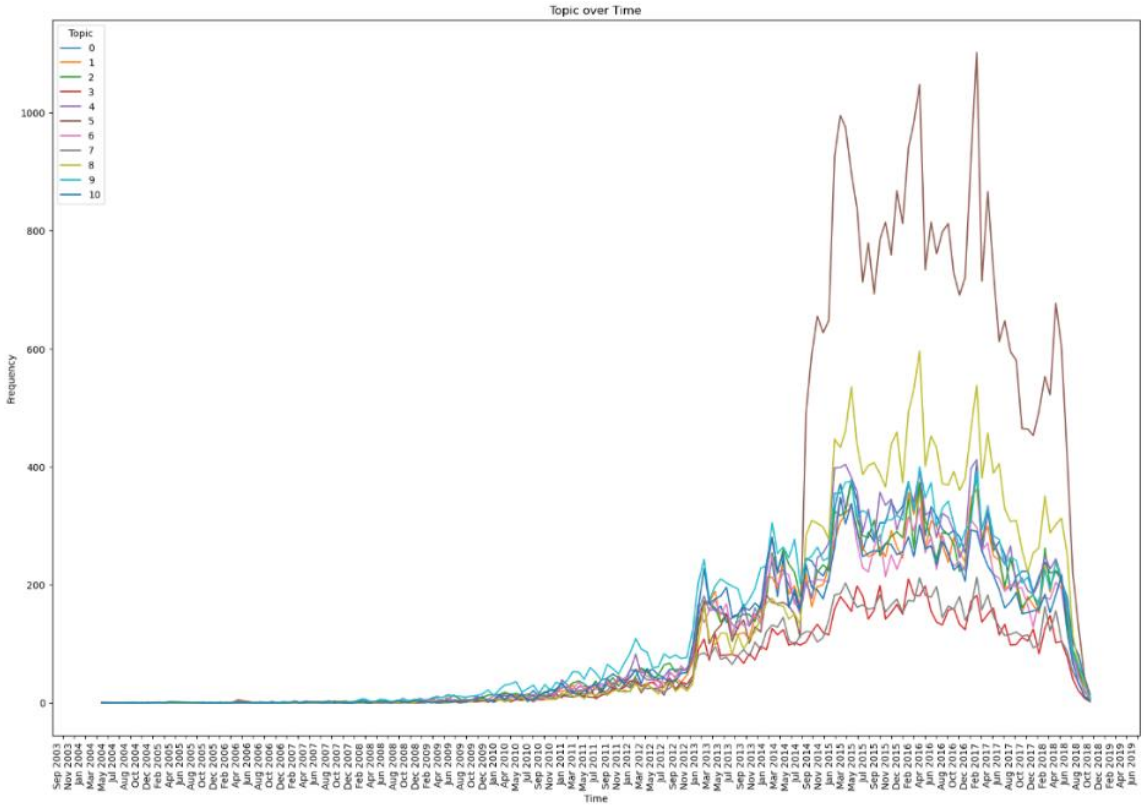


Figure 28: Topic trends over time (2003 - 2018)

To further analyze the topics, Figure 30 shows how different topics are represented across various rating levels. Upon initial inspection, it is evident that topic 5 makes up a significant portion of the reviews in the higher rating categories (4.0 and 5.0). This may indicate that customers who believe they are getting good value for their money tend to give higher ratings. It could also be a reflection of the company's successful pricing strategies and product quality that meet customer expectations. To identify the products aligning with this positive perception, the analysis can focus on the five leading products that consistently receive high ratings in conjunction with topic 5. These products are: B0006LOBA8, B0002E3CK4, B0002H05BA, B0002H03YY, B0007Y09VO.

On the other side of the spectrum, topic 6 is more frequently mentioned in reviews with lower ratings (1.0 and 2.0). The prominence of this topic highlights areas that businesses may need to pay attention to, such as customer service, purchasing processes, or product expectations. Addressing the issues associated with topic 6 could potentially improve customer satisfaction and lead to an improvement in ratings. The products most commonly associated with topic 6 and receiving lower ratings include B0017H4EBG, B0027V760M, B0002GMH7G, B0002GMGYA, B00AZUAORE.

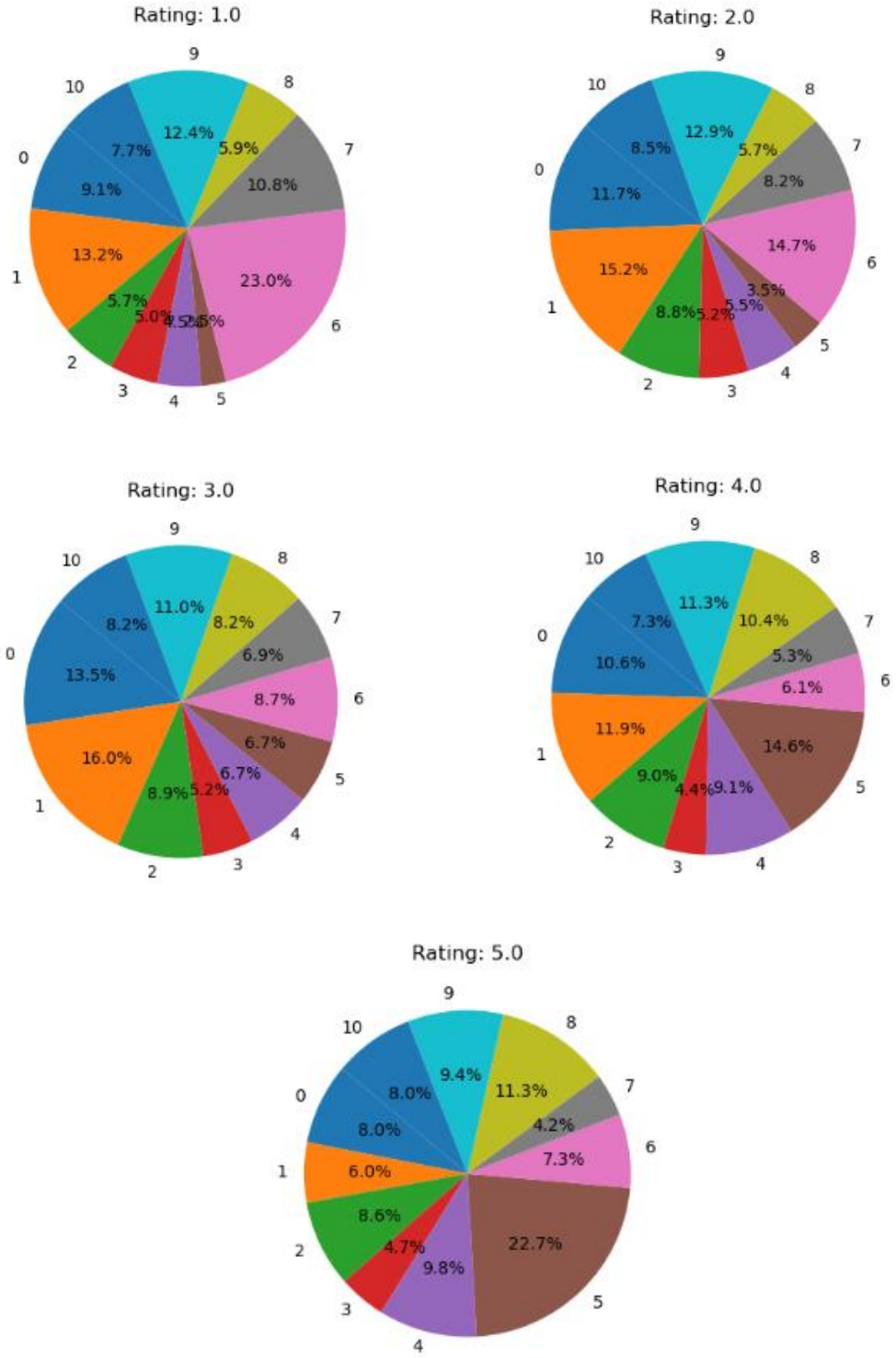


Figure 29: Distribution of topics across product ratings

For detailed insights on any specific product, the distribution of topics for that product can be identified, as displayed in Figure 31. By examining the actual content of the reviews, deeper questions can be addressed. For example, are there specific features or benefits that are consistently highlighted? Alternatively, are there persistent issues or shortcomings that are

frequently brought up? This type of in-depth analysis could be incredibly useful for product managers, marketing and customer service teams.

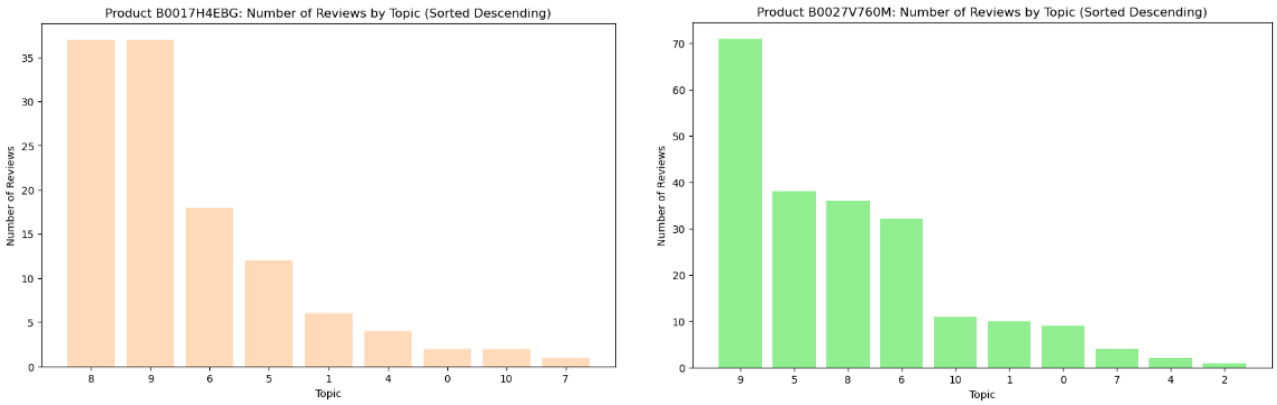


Figure 30: Distribution of topics in product B0017H4EBG and B0027V760M

4.4 BERTopic Result

After cleaning the data, a dataset consisting of 198,940 reviews is used to build and refine the BERTopic model. To evaluate the relationships and distances between topics, BERTopic offers the ‘visualize_topic’ function that generates an intertopic distance map, similar to the pyLDAvis tool used in LDA with Gensim. The distance between the bubbles reflects their similarity, with closer topics being more similar. The experiment goes through three stages of model configuration, each aimed at optimizing the thematic representation and interpretability of the resulting topics.

The initial phase of model configuration is building a BERTopic model with default settings. This results in the extraction of 2,145 topics, including outliers (topic -1) that do not fit into any category, as seen in Figure 32. The configuration at this stage has several problems. Firstly, although having an excessive number of topics can allow for fine-grained analysis of the data, it can be more difficult to comprehend the results. Secondly, the substantial presence of outliers (76,169 reviews in topic -1) implies that a considerable amount of the data is not well represented in the current topic structure. Besides, a scattered distribution of topics with significant overlap, as shown in Figure 33, indicates the lack of clear, distinct topic clusters. Furthermore, the terms used to describe each topic are quite broad and generic, making it difficult to understand the unique essence of each topic. For instance, in topic 2, the terms "action," "guitar," “neck” and "finish" are not specific enough to clearly define the context in which they are being discussed. Therefore, these issues suggest a need for further model refinement.

	Topic	Count	Name	Representation	Representative_Docs
0	-1	76169	-1_pedal_mic_use_to	[pedal, mic, use, to, it, with, and, my, on, for]	[I purchased this because even though I use mo...
1	0	2527	0_strap_locks_straps_leather	[strap, locks, straps, leather, schaller, lock...	[The best strap locks. I put them on all of my...
2	1	2517	1_amp_tubes_tube_amps	[amp, tubes, tube, amps, vox, practice, jj, wa...	[I have been playing for many years and have l...
3	2	1340	2_action_guitar_neck_finish	[action, guitar, neck, finish, beginner, epiph...	[I love this guitar and was so surprised at ho...
4	3	1294	3_pickup_pickups_duncan_coil	[pickup, pickups, duncan, coil, humbucker, sey...	[Play well on Neck Pickups, Play well on Neck ...
...	...	...	...	...	...
2140	2139	10	2139_patches_diffrent_whit_scrolling	[patches, diffrent, whit, scrolling, patch, um...	[Can't beat these patches for the money. Well...
2141	2140	10	2140_cage_rack_yep_nuts	[cage, rack, yep, nuts, screws, plasticnylon, ...	[Best rack screws money can buy., Rack screws,...
2142	2141	10	2141_pocket_mildewy_lysol_arevmuch	[pocket, mildewy, lysol, arevmuch, demensions,...	[Great for the pocket,($$$), Great product, gr...
2143	2142	10	2142_lightening_board_modular_stole	[lightening, board, modular, stole, pedalboard...	[Very good pedal board..., Perfect. Stole it o...
2144	2143	10	2143_eflat_assure_insane_killer	[eflat, assure, insane, killer, lasting, super...	[Killer sound, better price point. I play the...

2145 rows × 5 columns

Figure 31: Result of the default BERTopic model

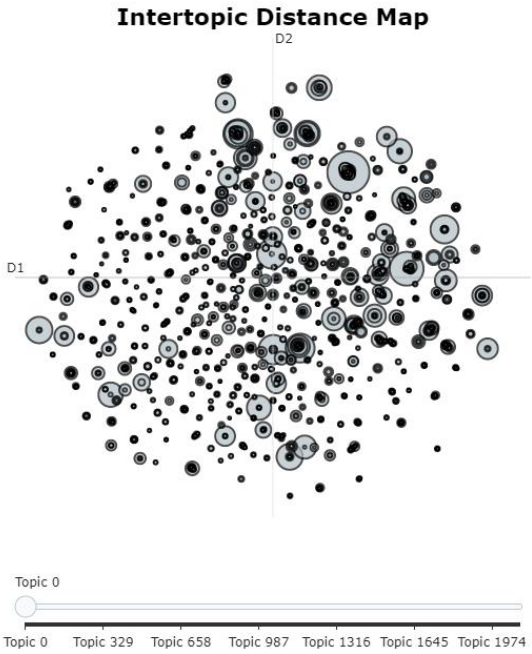


Figure 32: Visualization of the default BERTopic model

The second phase of model configuration is refining the BERTopic model by adjusting parameters to enable multilingual support and automatic merger of topics. Recognizing the presence of multiple languages in the dataset, the ‘language’ parameter is set to 'multilingual.' This allows the model to recognize and process topics in multiple languages more effectively, as opposed to just focusing on the English language. To address the challenge of excessive topic identification in the first phase, the model is adjusted to automatically combine similar topics. This was done by setting the ‘nr_topics’ parameter to ‘auto.’ These modifications result in a substantial reduction of topics from 2,145 to 1,371, as seen in Figure 34. Furthermore, the topic representative terms become more specific. However, despite this progress, challenges remain. The intertopic distance map, seen in Figure 35, still shows issues with clarity. The map displays that many topics are still

overlapping, which means the topics are not as distinct and well-defined as they could be. Therefore, more model refinement is still needed.

	Topic	Count	Name	Representation	Representative_Docs
0	-1	89760	-1_pedal_sound_it_to	[pedal, sound, it, to, the, and, but, with, of...	[Been using these strings for 20 years. I've u...
1	0	3094	0_mic_microphone_mics_boom	[mic, microphone, mics, boom, shure, sm58, spe...	[Great mic, Great mic for the $$$, Great mic]
2	1	2497	1_capo_capos_kyser_shubb	[capo, capos, kyser, shubb, spring, pressure, ...	[nice capo., Good capo., This is a good capo....]
3	2	1940	2_cable_cables_connectors_xlr	[cable, cables, connectors, xlr, quality, mono...	[Great cable!, Great Cable!, great cable]
4	3	1591	3_them_these_bought_they	[them, these, bought, they, ordered, second, t...	[Bought two of them!, These are a great produc...
...	...	...	...	...	...
1366	1365	10	1365_pics_pic_tongue_death	[pics, pic, tongue, death, cat, lovely, intend...	[great value on great guitar pics, Great pics ...
1367	1366	10	1366_converters_outboard_midnight_saffire	[converters, outboard, midnight, saffire, outs...	[The RV-7 is a great unit; generating a nice v...
1368	1367	10	1367_gs8_valuable_measure_gold	[gs8, valuable, measure, gold, changed, classi...	[These are my favorite acoustic strings. I re...
1369	1368	10	1368_beautiful_crispy_sounds_looks	[beautiful, crispy, sounds, looks, unique, rou...	[Beautiful and sounds so crispy good., Beautif...
1370	1369	10	1369_insalates_snark_aunt_useing	[insalates, snark, aunt, useing, isolates, spi...	[After seeing a friends guitar top crack, from...

1371 rows × 5 columns

Figure 33: Result of multilingual BERTopic model with automated topic merging

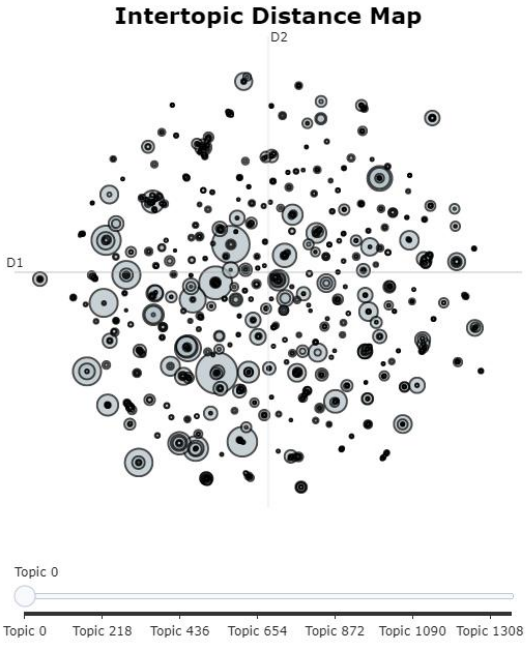


Figure 34: Visualization of multilingual BERTopic model with automated topic merging

To further improve the performance of the BERTopic model, the final stage of model configuration involves fine-tuning various hyperparameters. The modifications made at each step are detailed below:

- **Embedding Model**: The ‘paraphrase-multilingual-MiniLM-L12-v2,’ which is a pre-trained embedding model supporting over 50 languages, is selected for Sentence Transformers. It is a great model for multiple languages as it provides a balance between speed and performance.

- **Dimensionality Reduction**: For UMAP, the ‘n_neighbors’ parameter is set to 15 to help create larger and more coherent topic clusters. Also, to ensure that the model gives the same results with each run, the ‘random_state’ parameter is set to 42.
- **Clustering**: With HDBSCAN, the ‘min_cluster_size’ is set to 150 instead of the default value of 10. This helps reduce the number of clusters by increasing their sizes and relevance.
- **Tokenizer**: With CountVectorizer, the ‘stop_words’parameter is set to ‘english’ to ignore English stopwords and the ‘min_df’ parameter is set to 2 to eliminate words that occur less than twice in the reviews. Besides, the ‘ngram_range’ is set to (1, 2) to include words that are made up of one or two words. For example, ‘New York’ shall be considered as one word instead of two separate words ‘New’ and ‘York.’
- **Topic Representation**: The default labels for each topic concatenates the top ten words of each topic with underscores. This is not the best way to describe each topic. To improve the clarity and relevance of the labels, various topic representations such as KeyBERTInspired, POS, and MMR are used and compared together, as seen in Figure 36. Among these, KeyBERTInspired gives the most relevant keywords that best represent the topic’s essence and, therefore, shall be chosen as the method for customizing topic names.

```
{'Main': [('pedal', 0.04693467050789381),
  ('pedals', 0.02550389278167272),
  ('board', 0.012395296754020053),
  ('delay', 0.010394007125829215),
  ('pedal board', 0.008786508627477974),
  ('power', 0.0084668213741224),
  ('distortion', 0.008099511945701396),
  ('tone', 0.007766315851218255),
  ('sounds', 0.00765613574801289),
  ('like', 0.00759338562905861)],
 'KeyBERT': [('great pedal', 0.8537107),
  ('pedal great', 0.80301553),
  ('pedals', 0.79635787),
  ('pedal', 0.7734407),
  ('pedalboard', 0.7599253),
  ('pedal board', 0.7378403),
  ('pedal does', 0.6889464),
  ('knob', 0.37498826),
  ('knobs', 0.37187117),
  ('tones', 0.36463752)],
 'MMR': [('pedal', 0.04693467050789381),
  ('pedals', 0.02550389278167272),
  ('board', 0.012395296754020053),
  ('delay', 0.010394007125829215),
  ('pedal board', 0.008786508627477974),
  ('power', 0.0084668213741224),
  ('distortion', 0.008099511945701396),
  ('tone', 0.007766315851218255),
  ('sounds', 0.00765613574801289),
  ('like', 0.00759338562905861)],
 'POS': [('pedal', 0.04693467050789381),
  ('pedals', 0.02550389278167272),
  ('board', 0.012395296754020053),
  ('delay', 0.010394007125829215),
  ('power', 0.0084668213741224),
  ('distortion', 0.008099511945701396),
  ('tone', 0.007766315851218255),
  ('sounds', 0.00765613574801289),
  ('sound', 0.007445613431229465),
  ('effects', 0.007205943724909495)]}
```

Figure 35: Topic representations with KeyBERTInspired, POS, and MMR

Another important question arises: How many key terms are needed to accurately represent a topic? An examination of the most commonly used words within the top ten topics (as illustrated in Figure 37) reveals that while a list of ten words provides a comprehensive view, most topics can be described using just the top two or three words.

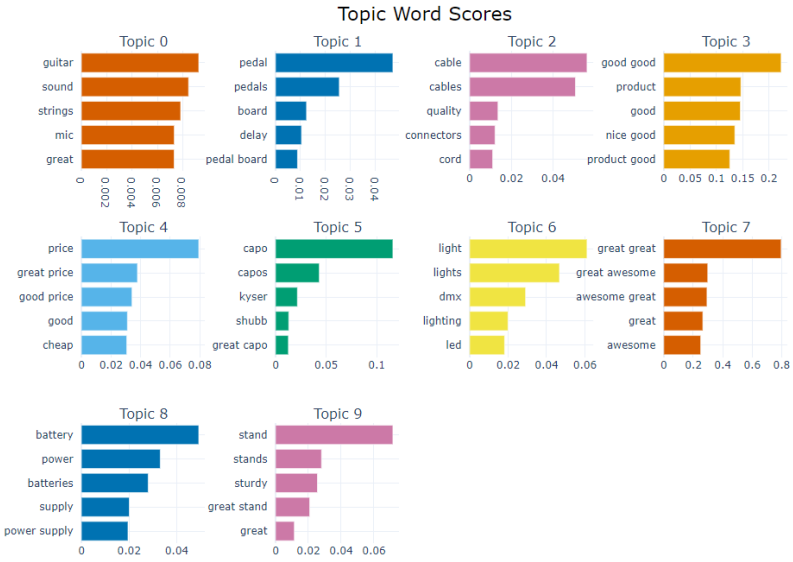


Figure 36: Bar chart of word importance scores for top 10 topics

To assess the importance of each word in a topic label, the term score is graphed. Similar to the 'elbow method' used in k-means clustering to find the optimal number of clusters, the significance of each additional word decreases as more terms are included in the topic representation. From Figure 38, it is clear that the top two words from each topic generally suffice to convey the main idea. While some topics might require up to five or six words to fully capture what the topic is all about, adding more words beyond that does not make any big difference in most cases. This understanding leads to the creation of new custom labeling using KeyBERTInspired with the top two words for each topic. The resulting custom label is displayed in the CustomName column of the dataset.

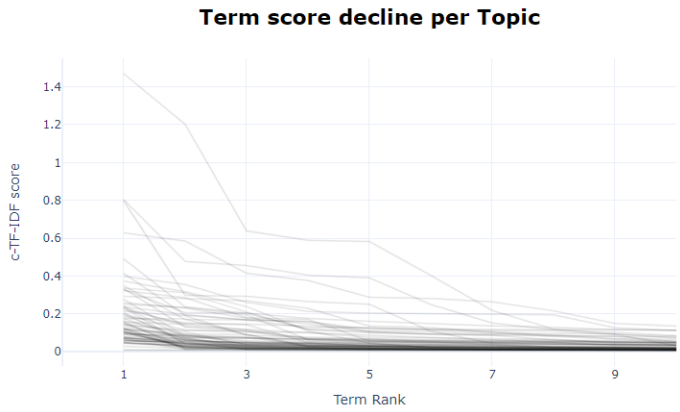


Figure 37: Term importance score across all topics

The changes made have significantly improved the model's results. The number of topics is reduced to 78, including outliers, with more defined topic labels, as shown in Figure 39. Furthermore, the intertopic distance map in Figure 40 shows that the model has succeeded in forming several distinct topic clusters. However, there are still some overlapping sections among the bubbles, which suggests that merging some topics further may be necessary.

	Topic	Count	Name	Representation	KeyBERT	MMR	POS	Representative_Docs
0	-1	81503	-1_strings_great_guitar_good	[strings, great, guitar, good, sound, use, jus...	[good, excellent, great, nice, works great, pe...	[strings, great, guitar, good, sound, use, jus...	[strings, great, guitar, good, sound, use, pro...	[I have a Yamaha acoustic guitar. I love it, b...
1	0	57708	0_guitar_sound_strings_mic	[guitar, sound, strings, mic, great, like, goo...	[instrument, guitars, guitar, bass, fender, mu...	[guitar, sound, strings, mic, great, like, goo...	[guitar, sound, strings, mic, great, good, str...	[I can't even begin to describe how beautiful ...
2	1	8883	1_pedal_pedals_board_delay	[pedal, pedals, board, delay, pedal board, pow...	[great pedal, pedal great, pedals, pedal, peda...	[pedal, pedals, board, delay, pedal board, pow...	[pedal, pedals, board, delay, power, distortio...	[Always been a fan of the boss pedals. Glad to...
3	2	8545	2_cable_cables_quality_connectors	[cable, cables, quality, connectors, cord, xlr...	[nice cable, good cable, great cables, great c...	[cable, cables, quality, connectors, cord, xlr...	[cable, cables, quality, connectors, cord, xlr...	[I love these cables there awesome!, Excellent...
4	3	2882	3_good good_product_good_nice good	[good good, product, good, nice good, product ...	[good product, buen producto, product good, pr...	[good good, product, good, nice good, product ...	[product, good, good product, great product, n...	[Good, good quality product., Good product at ...
...	...	...	...	...	...	...	...	...
73	72	166	72_just needed_needed_needed just_exactly needed	[just needed, needed, needed just, exactly nee...	[just needed, needed just, needs just, just ne...	[just needed, needed, needed just, exactly nee...	[thanks, , , , , , , ,]	[just what i needed, just what i needed, just ...
74	73	161	73_love_things_love things_things great	[love, things, love things, things great, got ...	[got theses, theses, things awesome, love thin...	[love, things, love things, things great, got ...	[things, great, theses, awesome, need, feel, h...	[Love these!, Love these!, Love these things. ...
75	74	158	74_link_nbsp_ref cm_cr_arp_d_rvw_txt_link normal	[link, nbsp, ref cm_cr_arp_d_rvw_txt, link nor...	[ref cm_cr_arp_d_rvw_txt, cm_cr_arp_d_rvw_txt,...	[link, nbsp, ref cm_cr_arp_d_rvw_txt, link nor...	[link, data, normal, guitar, stage, electric, ...	[I have mine attached to one of the support ar...
76	75	155	75_loves_son_loves son_son loves	[loves, son, loves son, son loves, wife, loved...	[son loves, loves son, son likes, son loved, s...	[loves, son, loves son, son loves, wife, loved...	[son, wife, daughter, husband, granddaughter, ...	[My son loves it., my son loves!!!!, Son loves...
77	76	152	76_perfect perfect_perfect__	[perfect perfect, perfect, , , , , , ,]	[perfect, perfect perfect, , , , , , ,]	[perfect perfect, perfect, , , , , , ,]	[, , , , , , , ,]	[Perfect, Perfect, Perfect]

78 rows × 8 columns

Figure 38: Result of refined BERTopic model after hyperparameter tuning

Figure 39: Visualization of refined BERTopic model after hyperparameter tuning

Determining which topics to merge further involves a thorough examination of topic hierarchies and similarities. A hierarchy tree, as presented in Figure 41, visually shows the relationships between topics. This tree uses color coding to group similar topics, showing potential sub-topics nestled within larger topics. Hovering over the black nodes within the tree provides a representation of the topic at that level of the hierarchy. In addition to the visual assessment, a quantitative analysis to examine the similarity scores between pairs of closely related topics is

presented in Figure 42. The decision regarding which topics to merge is made by taking into account both the visual analysis of the hierarchy tree and an understanding of the semantic content represented by each topic. The goal is to reduce redundancy within the topic model and create a set of topics that are both distinct and comprehensive. Based on the insights from the hierarchy tree and the similarity scores, merging topics 66 and 76, as well as 70 and 71, shall enhance the model's clarity and coherence. As a result, the total number of topics is reduced to 76, including the outlier category denoted as '-1'.

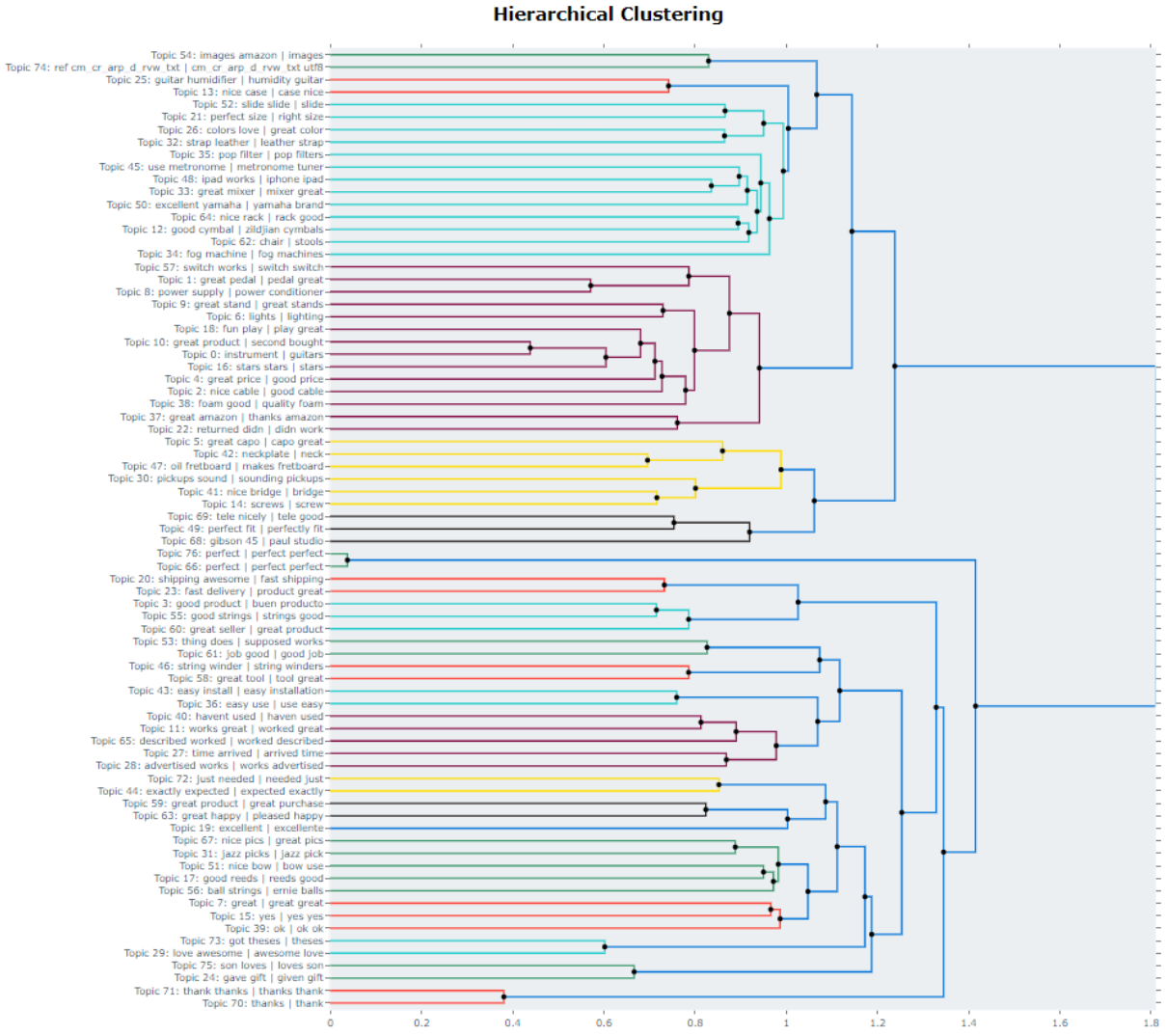


Figure 40: Dendrogram of topic relationships of refined BERTopic model

	topic1	topic2	distance
5303	66_perfect perfect_perfect_needed perfect_perf...	76_perfect perfect_perfect__	0.992983
1568	19_excellent excellent_excellent_excellente_sy...	7_great great_great awesome_awesome great_great	0.908317
5610	70_thanks thanks_thanks_thank thanks_thanks thank	71_thank_thanks_thank thank_thanks thank	0.902259
1288	15_yes_meh_aok_yes yes	39_ok ok_ok_thks_ok tuner	0.849745
368	3_good good_product_good_nice good	55_good strings_great product_strings good_pro...	0.841516
1564	19_excellent excellent_excellent_excellente_sy...	3_good good_product_good_nice good	0.836263
1662	20_shipping_fast shipping_fast_shipping great	23_delivery_service_fast delivery_fast	0.833974
4741	59_buy_purchase_product buy_happy purchase	60_seller_great seller_seller great_product	0.821984
1637	19_excellent excellent_excellent_excellente_sy...	76_perfect perfect_perfect__	0.802772
1627	19_excellent excellent_excellent_excellente_sy...	66_perfect perfect_perfect_needed perfect_perf...	0.800054
2416	29_love_love love_like love_cool	75_loves_son_loves son_son loves	0.791610
450	4_price_great price_good price_good	59_buy_purchase_product buy_happy purchase	0.790370
6014	76_perfect perfect_perfect__	7_great great_great awesome_awesome great_great	0.784148
4429	55_good strings_great product_strings good_pro...	60_seller_great seller_seller great_product	0.783749
320	3_good good_product_good_nice good	7_great great_great awesome_awesome great_great	0.783399
451	4_price_great price_good price_good	60_seller_great seller_seller great_product	0.779231
5234	66_perfect perfect_perfect_needed perfect_perf...	7_great great_great awesome_awesome great_great	0.779014
940	11_great works_works great_works_great	3_good good_product_good_nice good	0.775520
1933	23_delivery_service_fast delivery_fast	60_seller_great seller_seller great_product	0.767832
3090	38_plastic_foam_cheap_product	47_fretboard_fret_frets_fret board	0.767297

Figure 41: Pairs of similar topics with high similarity scores

It is important to manage outliers when refining the BERTopic model. Outliers refer to the reviews that do not fit into the model's defined topics. In this case, a significant portion of the reviews – 81,503 out of 198,940 reviews, equating to 40.97% – is classified as outliers. This presents a challenge because simply excluding these outliers can result in losing valuable insights from a large segment of the data. To mitigate this issue, the ‘reduce_outliers’ function is used to reassess and reassign these outliers to existing topics and the ‘update_topics’ function is utilized to update the topic representation. The result is a remarkable decrease in outliers to only 431, as seen in Figure 43. The overall coherence score is 0.43627. This is the final BERTopic model with details presented in Figure 44 and Table 7.

	Topic	Count	Name	CustomName	Representation	KeyBERT	MMR	POS	Representative_Docs
0	-1	431	-1_aaaaaaa_none_made_must	Topic -1: good | excellent	[aaaaaaa, none, made, must, well, wicky, resid...	[good, excellent, great, nice, works great, pe...	[strings, great, guitar, good, sound, use, jus...	[strings, great, guitar, good, sound, use, pro...	[I have a Yamaha acoustic guitar. I love it, b...
1	0	75195	0_the_and_to_it	Topic 0: instrument | guitars	[the, and, to, it, for, of, is, my, this, on]	[instrument, guitars, guitar, bass, fender, mu...	[guitar, sound, strings, great, mic, like, goo...	[guitar, sound, strings, great, mic, good, str...	[I can't even begin to describe how beautiful ...
2	1	10004	1_pedal_pedals_it_this	Topic 1: great pedal | pedal great	[pedal, pedals, it, this, the, you, is, of, bo...	[great pedal, pedal great, pedals, pedal, peda...	[pedal, pedals, board, delay, pedal board, pow...	[pedal, pedals, board, delay, power, distortio...	[Always been a fan of the boss pedals. Glad to...
3	2	9044	2_cable_cables_quality_and	Topic 2: nice cable | good cable	[cable, cables, quality, and, the, to, connect...	[nice cable, good cable, great cables, great c...	[cable, cables, quality, connectors, cord, xlr...	[cable, cables, quality, connectors, cord, xlr...	[I love these cables there awesome!, Excellent...
4	3	6092	3_product_good_nice_very	Topic 3: good product | buen producto	[product, good, nice, very, excellent, price, ...	[good product, buen producto, product good, pr...	[good good, product, good, nice good, product ...	[product, good, good product, great product, n...	[Good, good quality product., Good product at ...
...	...	...	...	...	...	...	...	...	...
73	72	562	72_deal_things_needed_what	Topic 72: just needed | needed just	[deal, things, needed, what, great, love, just...	[got theses, theses, love things, loved things...	[love, things, love things, things great, got ...	[things, great, theses, awesome, need, feel, h...	[Love these !!!, Love these!, Love these thing...
74	73	429	73_link_nbsp_linked_cm_cr_arp_d_rvw_txt	Topic 73: got theses | theses	[link, nbsp, linked, cm_cr_arp_d_rvw_txt, href...	[ref cm_cr_arp_d_rvw_txt, cm_cr_arp_d_rvw_txt ...	[link, nbsp, link linked, ref cm_cr_arp_d_rvw_...	[link, data, normal, guitar, stage, electric, ...	[I have mine attached to one of the support ar...
75	74	592	74_link_nbsp_cm_cr_arp_d_rvw_txt_href	Topic 74: ref cm_cr_arp_d_rvw_txt | cm_cr_arp_...	[link, nbsp, cm_cr_arp_d_rvw_txt, href, linked...	[son loves, loves son, son likes, son loved, s...	[loves, son, loves son, son loves, wife, loved...	[son, wife, daughter, husband, granddaughter, ...	[My son loves this!!!, My son loves it., Son l...
76	75	155	75_loves_son_loved_wife	Topic 75: son loves | loves son	[loves, son, loved, wife, daughter, husband, l...	NaN	NaN	NaN	NaN
77	76	152	76_perfect___	Topic 76: perfect | perfect perfect	[perfect, , , , , , , , ,]	NaN	NaN	NaN	NaN

78 rows × 9 columns

Figure 42: Result of the final BERTopic model

Figure 43: Visualization of the final BERTopic model

Final BERTopic Model	
Number of topics	75 (exclude outliers)
Coherence score	0.43627

Table 7: Statistics of the final BERTopic model

Similar to LDA, the BERTopic model can also provide insightful answers to key business questions, such as:

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?

From Figure 43, identifying the most and least discussed topics by customers is straightforward with BERTopic outcome. The topic distribution is arranged in descending order of popularity from top to bottom, with topic 0 (instruments, guitars) being the most frequently discussed topic with 75,195 reviews, and topic 76 (perfect, perfect, perfect) being the least popular topic with only 152 reviews.

Regarding the topic popularity over time, BERTopic has a ‘topics_over_time’ function that offers a clear visualization of how topics trend over different periods. This feature is interactive, allowing users to focus on specific topics or filter out others. This can be particularly useful for users to track the rise or decline in popularity of specific topics. As illustrated in Figure 45 and Figure 46, a pronounced increase in activity for all topics starts from 2013, reaches a peak in 2016, and then experiences a decline after mid-2016.

The prominent topic, topic 0 (instrument, guitars), shows a remarkable surge in interest from early 2012, peaking in 2015, before witnessing a sharp decline from 2017 onwards. By excluding the dominant topic 0, a more detailed picture reveals that other topics follow a similar pattern. They experience a rise at the beginning of 2012, with several topics experiencing a secondary surge towards the end of 2013. However, starting from mid-2016, all topics generally decline. Certain topics stand out due to their more pronounced popularity, such as topic 1 (great pedal, pedal great), topic 2 (nice cable, good cable), topic 3 (good product, buen producto), topic 4 (great price, good price), topic 6 (lights, lightning), topic 7 (great, great great), topic 9 (great stand, great stands), topic 10 (great product, second bought), topic 11 (works great, worked great).

The analysis indicates that customers have shown significant interest in certain product features, such as the quality of pedals and cables, or the functionality of stands. With this knowledge, it would be advantageous for businesses to examine the content of reviews from these peak periods to identify the underlying factors contributing to these shifts in topic popularity.

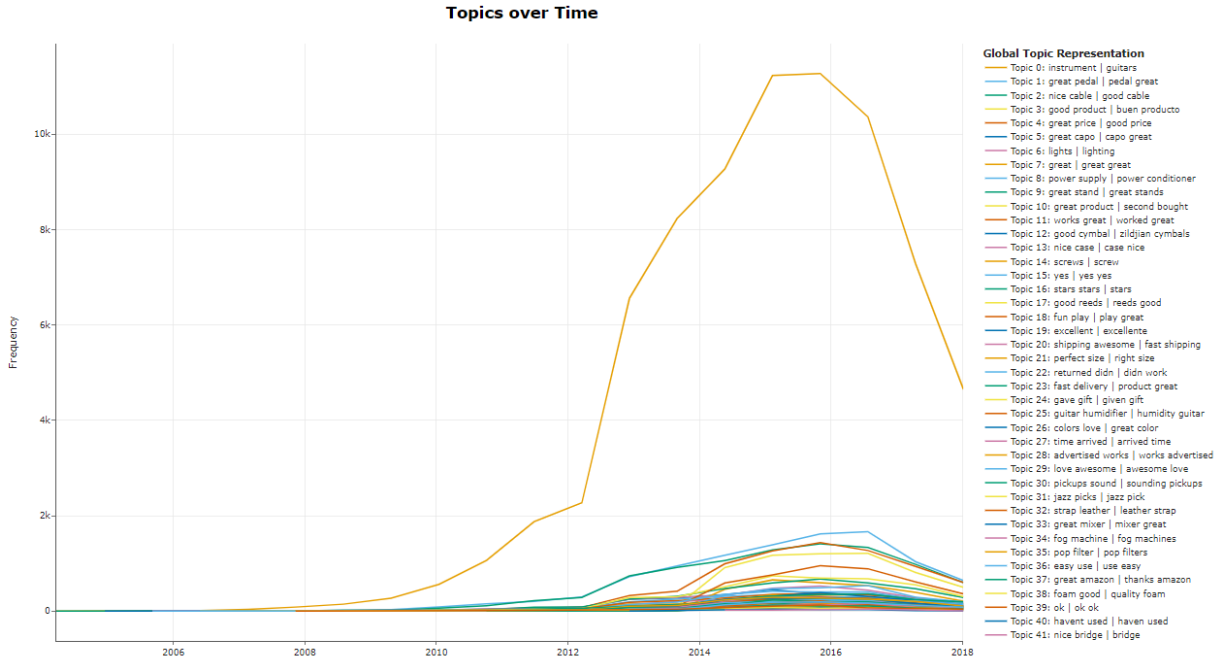


Figure 44: Topic trends over time (2003 – 2018)

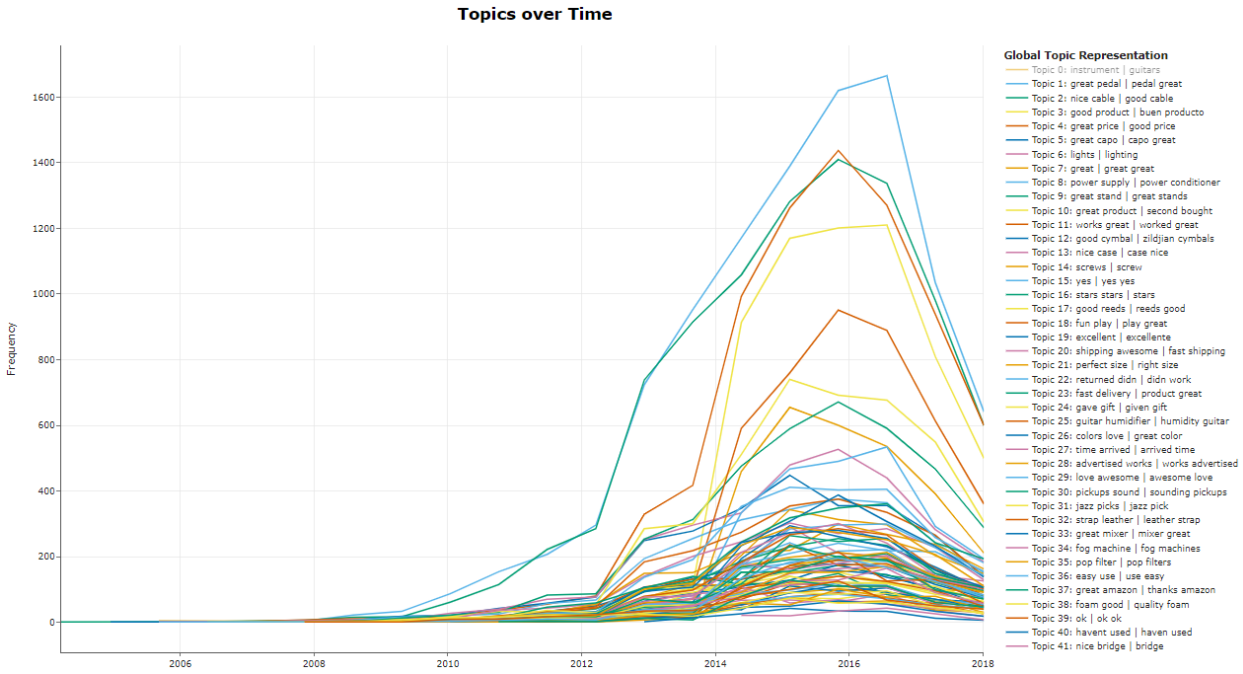


Figure 45: Topic trends over time, excluding Topic 0 (2003 – 2018)

By linking topics to customer ratings, the BERTopic model helps identify which subjects are correlated with higher or lower satisfaction levels, as well as the specific products related to these topics. This analysis can pinpoint the product strengths and areas for improvement. From Figure 47, it is observed that topic 0 (instrument, guitars) is a significant topic across five different ratings. This suggests that most of the discussions in the dataset are about instruments, with a particular emphasis on guitars. Excluding topic 0 allows for the acquisition of more detailed insights.

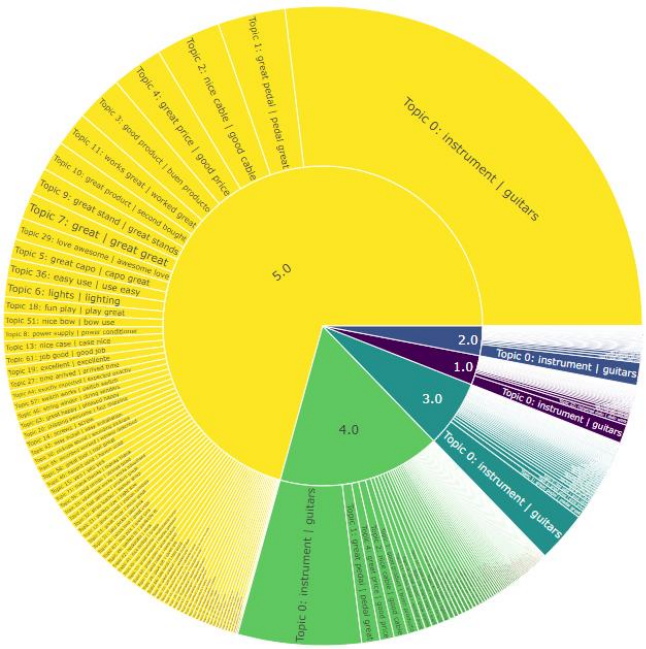
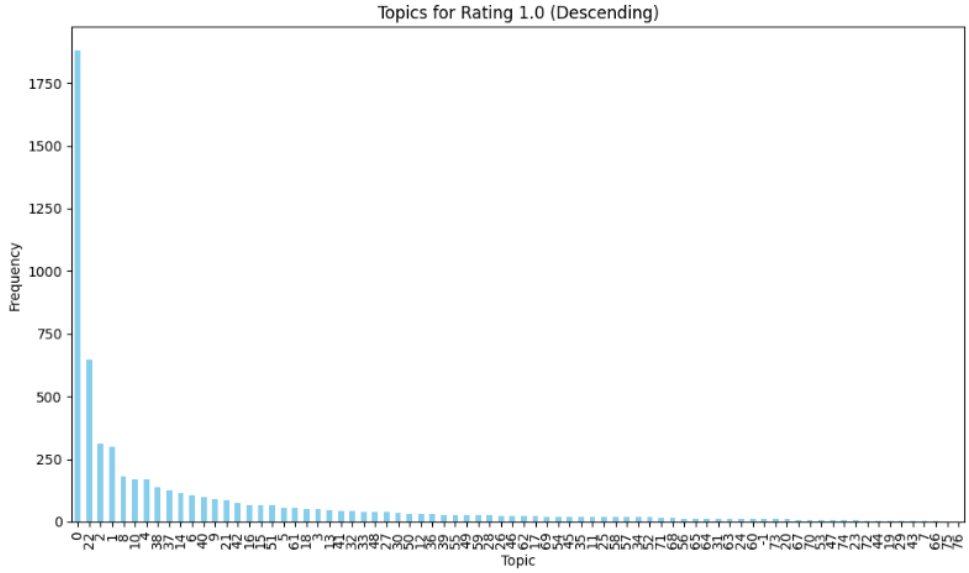
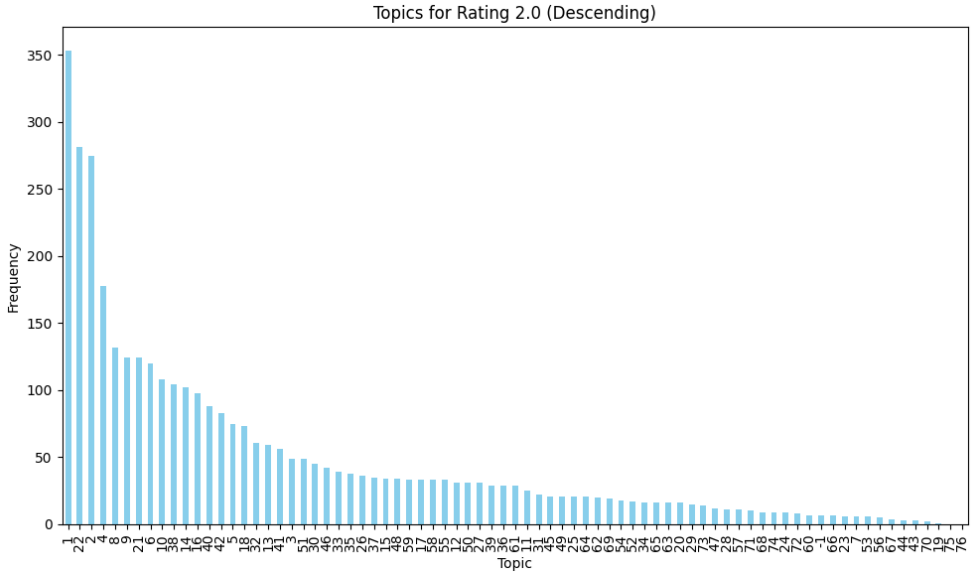


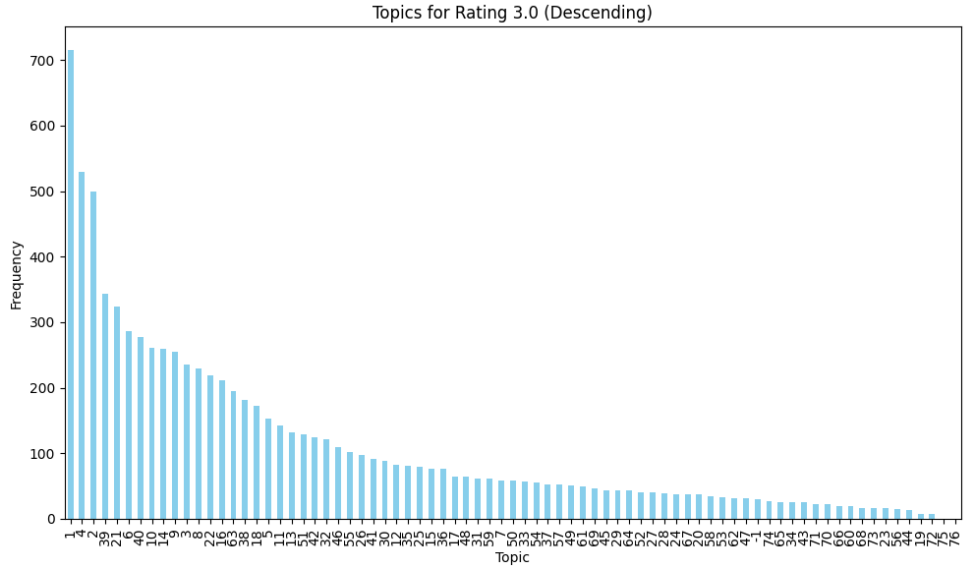
Figure 46: Topic distribution across all product ratings in the final BERTopic model

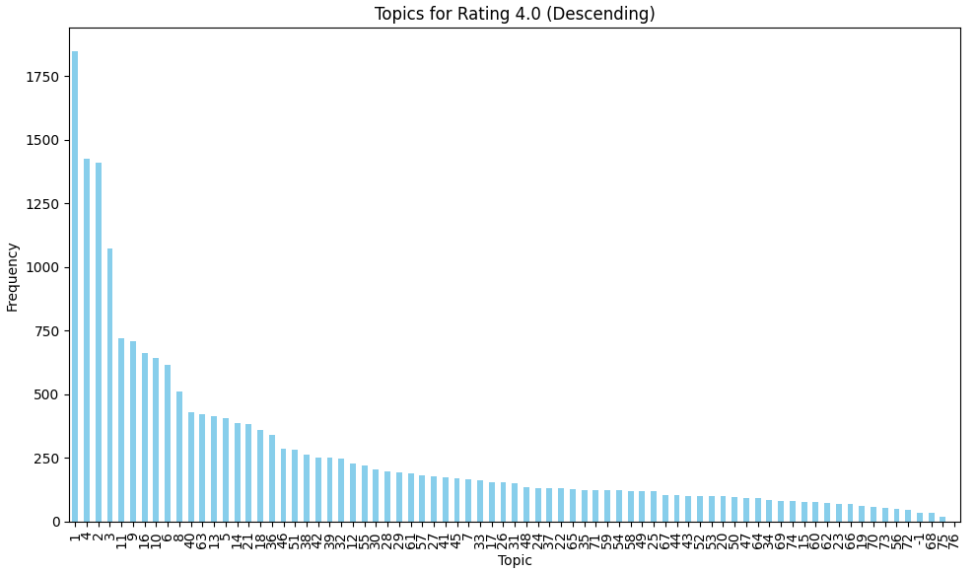
The bar charts in Figure 48 represent the distributions of topics across different rating levels, specifically for ratings 1.0 through 5.0. Each bar chart shows the frequency of topics associated with each rating, sorted in descending order. According to Figure 48, one of the most common topics associated with low ratings 1.0 and 2.0 is topic 22 (returned didn, didn work). This suggests that functionality issues or product defects are a common thread in negative customer experiences. This shows that bad product quality has a high impact on negative customer satisfaction, which suggests a needed improvement in product quality. Conversely, in the higher rating categories of 4.0 and 5.0, topic 3 (good product, buen producto) and topic 4 (great price, good price) are among the most frequent topics. This indicates that perceptions of value for money and good product quality are strong contributing factors to positive customer reviews. The appearance of "buen producto" alongside English terms indicates that this positive sentiment cuts across different language groups within the customer base.

Expanding on this analysis, businesses can explore the content of these topics for more insights. For example, the content of reviews associated with topic 22 (returned didn, didn work) can reveal factors influencing the negative sentiment, which could lead to product improvements that reduce returns and increase customer satisfaction. Meanwhile, identifying positive attributes that drive topic 3 (good product, buen producto) and topic 4 (great price, good price) could further enhance customer satisfaction with product quality and value. Furthermore, tracking these topics over time could reveal trends in customer priorities and emerging issues. For instance, if the frequency of mentions for "returned didn't, didn't work" begins to rise, it could signal an emerging issue in product quality that needs immediate attention.




Topics for Rating 2.0 (Descending)
Frequency
Topic


Topics for Rating 3.0 (Descending)
Frequency
Topic


Topics for Rating 4.0 (Descending)
Frequency
Topic

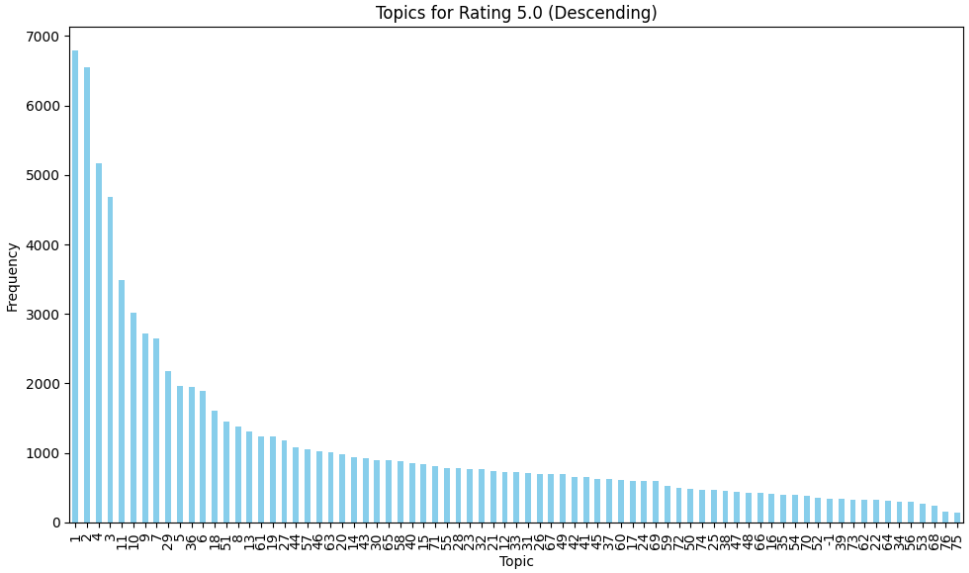


Figure 47: Topic distribution across each product ratings

Moreover, given any product, one can easily identify the topics associated with that product. For example, among the products with a 1.0 rating linked to topic 22 (returned didn, didn work) such as B0027V760M and B00ACGMOA6, there is a wider range of topics in the feedback. For product B0027V760M, while some customers noted issues with returns or malfunctions, the majority of feedback was positive. The topic distribution in Figure 49 shows that topic 0 (instrument, guitars), topic 4 (great price, good price), topic 2 (nice cable, good cable), and topic 11 (works great, worked great) are the most dominant. This suggests that most customers express satisfaction with the product's quality or usability. Similarly, product B00ACGMOA6 has mostly good evaluations for its cables, guitar accessories, functionality, overall product quality, and the Yamaha brand, despite some reported faults, as shown in Figure 50. These observations imply that issues leading to a product's return may represent a smaller fraction of the customer feedback since most reviews indicate a positive experience with the product. The topic distribution for each product not only reflects the customer's immediate response to product features but also serves as a diagnostic tool to pinpoint areas for product development and enhance customer experience.

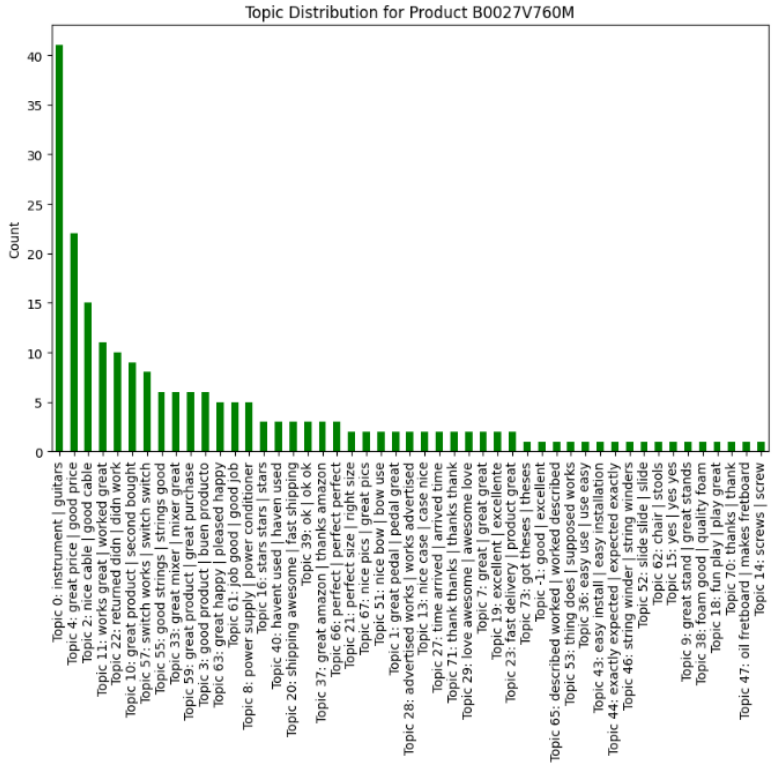


Figure 48: Topic distribution for product B0027V760M

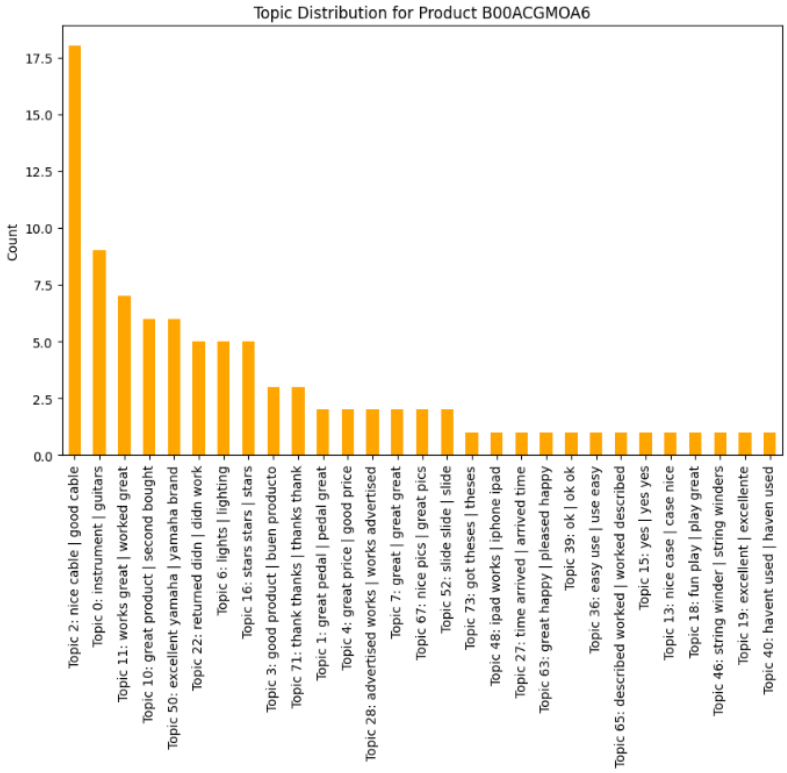


Figure 49: Topic distribution for product B00ACGMOA6

Unlike the final LDA model, where each review is a mixture of all eleven topics, BERTopic model identifies only the most relevant topics for each review. This specificity is applicable to

both long and short-text reviews. This means that for any given review, BERTopic model can pinpoint the topics that are actually mentioned or implied in the text, rather than providing a probability distribution across all topics. For example, the review “G'daughter received this for Christmas present last year and plays if often.” has topic 75 (son loves, loves son) and topic 24 (gave gift, given gift) as shown in Figure 51.

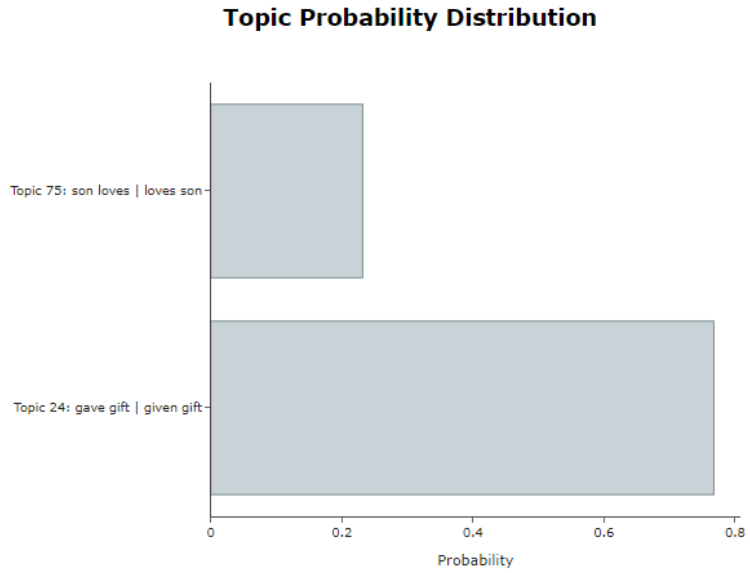


Figure 50: Topic distribution for third review in the dataset

4.5 Discussion

After presenting the results for both LDA and BERTopic, this section shall evaluate and discuss the outcomes of both methods. In general, both techniques provide valuable insights that can significantly contribute to business intelligence and decision-making. The evaluation is based on several criteria, with a primary focus on coherence scores and the interpretation of topics by humans. Depending on the specific use case and business requirements, other considerations such as computation time and data preprocessing procedure serve as supplementary factors in determining the most suitable method for application.

In terms of coherence score, the final LDA model marginally outperforms the final BERTopic model. However, in terms of assessment of topic quality, BERTopic model produces more meaningful and comprehensible topics that were easier to interpret compared to LDA model. This stems from BERTopic’s capability to identify a greater number of topics than LDA. These topics can be organized in a hierarchical order, featuring the main topics and their associated sub-topics. This hierarchical structure enables users to explore these topics in-depth, revealing the different components that make up a major topic. On the contrary, even though the LDA approach presents a clear snapshot of the general theme of the unstructured reviews, it lacks the level of detail offered by BERTopic. When needing a more thorough examination of a specific topic, users of LDA may

need to look through the raw review texts or perform additional analytics to gain additional insights.

In terms of computation time, BERTopic demonstrates an advantage over LDA. The computation time for LDA increases proportionally with the number of topics, which means that models with a higher number of topics may take longer to process. At first glance, LDA appears to be more efficient with an average training time of 25 minutes for each model compared to BERTopic's average training time of more than two hours per model. However, BERTopic can automatically decide the optimal number of topics during its training phase. Whereas LDA lacks this automation ability and requires multiple iterations to experiment with a range of different numbers of topics, alpha and beta values. In reality, it takes a total of 48 hours to exhaustively iterate through all possible combinations of topic numbers and alpha values to find the final LDA model. This renders LDA considerably more time-consuming and resource-intensive than BERTopic, especially when fine-tuning the model for optimal results.

In terms of data preparation, BERTopic is known for its minimal to nonexistent requirement for data preprocessing. On the other hand, LDA requires meticulous cleaning of the data, including the removal of punctuations, numbers, non-alphabetic characters, and stop words. Thereafter, the texts need to be normalized, tokenized, and lemmatized. The method further requires the creation of BoW. In comparison, BERTopic is more efficient than LDA because there is less labor-intensive preparation required. Table 8 provides a detailed comparison between LDA and BERTopic.

Criteria	LDA	BERTopic
Data preprocessing	Requires thorough and meticulous preprocessing steps	Minimal to no data preprocessing required
Number of topics	Numbers of topics must be determined before building model	Automatically determines the number of topics during model creation, but tend to generate too many topics with the default settings. Needs fine tuning to reduce the number of topics and create more coherent topics
Topic relationship per document	Mixes topics within documents	Allocates a single topic to each document
Topic representation	Employs a BoW scheme, overlooking semantics	Uses semantic embeddings to enhance topic significance

Criteria	LDA	BERTopic
Optimal number of topics	The process to ascertain the best topic number is intricate	Employs hierarchical topic reduction for optimal numbers of topics
Outliers	No outliers	Produces outliers as the result of using HDBSCAN, which can be refined post-training
Document length	Ignores length of documents	Prefers shorter documents due to token limits in embedding models
Dataset size	Accommodates any dataset size	Performs better with larger datasets; less effective with smaller collections
Speed and resource usage	Training duration and resource demand scale with topic numbers. Can have longer training time and inexpensive computational resources (CPU)	Shorter training time with 'paraphrase-multilingual-MiniLM-L12-v2' embedding model. Other embedding models might have longer training time and potentially expensive computational resources (GPU)
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.

Table 8: Comparison between LDA and BERTopic models

Although the LDA model has a slightly higher coherence score, BERTopic is a better choice for this study due to its advantages in terms of topic comprehensibility, computational time, and ease of data preparation. However, it is important to note that each method has its unique strengths and the ultimate decision of which one to use should be based on the particular needs and limitations of the business. To ensure that the chosen method aligns with the business goals, a detailed analysis of the review content is advisable, along with consultation with domain experts and key

stakeholders. Additionally, while topic modeling is an effective tool to organize and simplify unstructured data, the true extraction of insights requires an in-depth examination of the content within reviews.

This study encounters some limitations due to the unavailability of advanced computational resources and domain expertise in the field of musical instruments. The experiment relies solely on the processing capacity of a local CPU. The lack of access to Graphics Processing Unit (GPU) resources extends the training duration for both models. Advanced topic modeling techniques using deep learning like BERTopic typically leverage the processing power of GPU to significantly expedite the processing time for large datasets. Without GPU, the scope of the study is limited to selecting a subset of the Amazon Reviews dataset in order to manage the computational load. This, in turn, affects the study's ability to scale and potentially compromised the depth of analysis due to the smaller data sample. Moreover, the limitation in computational power also restricts rapid iteration and model optimization.

The lack of domain expertise in musical instruments is another limitation of this study. While topic modeling algorithms are proficient at detecting patterns and clustering similar terms, the true meaning and significance of these terms can be interpreted correctly only with a deep understanding of the subject matter, particularly in specialized fields such as musical instruments. For instance, the topic modeling algorithm might group terms such as “reed,” “mouthpiece,” and “valve” together based on their co-occurrence in text, but only a domain expert would recognize these as specifically pertaining to wind instruments. Similarly, “hammer,” “keys,” and “pedal” might be common words across various contexts, but within the realm of musical instruments, they are distinctly associated with keyboards. Domain experts can accurately identify the context in which these terms are used, distinguishing between their general and industry-specific meanings. This interpretation is critical because it affects how the topics are labeled and understood.

Moreover, specific issues that are intrinsic to musical instruments might be misunderstood or overlooked by someone without the requisite background knowledge. For example, the word “action” could be related to the physical mechanism of a piano or guitar rather than implying movement or use. Similarly, “bridge” might refer to a guitar part rather than a structure. Additionally, certain phrases might signify common problems unique to the field, such as “fret buzz” in guitars or “pad leakage” in woodwind instruments, which could be identified as significant topics by an expert. These are not just words but represent potential quality issues that could impact customer satisfaction and product design. Recognizing and prioritizing such topics could be invaluable for businesses for quality control, innovation, and customer service enhancement.

4.6 Chapter Summary

This chapter presents an experiment on the Amazon Reviews dataset using LDA and BERTopic that implements the theoretical frameworks established in chapter three. The findings are evaluated, and the study's limitations are discussed. The key points are as follows:

- Data exploration reveals that the data spans from 2003 to 2018 with the majority of the data from 2013 to 2017. Most of the reviews receive 4.0 and 5.0 ratings. Most of the length of the text is short text (< 200 words) with a few exceptions of very long text. There are 48 reviews with missing content, 18,571 duplicates, and 23,015 reviews from unverified users in the dataset. 7.5% of the dataset is comprised of 17,416 reviews in 30 languages other than English.
- Data preprocessing has two steps: data cleaning and text preprocessing. Data cleaning phase removes 32,452 entries, including review with missing content, duplicates, and reviews from unverified users. The clean data has 198,940 reviews in total. This data can now be fed into BERTopic for topic modeling development. For LDA, it requires additional text preprocessing steps including tokenization, stop word removal, lemmatization, and corpus building.
- LDA models are created using different combinations of numbers of topics and alpha values. The number of topics is selected from 2 to 20, while alpha values are chosen from auto, 0.01, 0.1, and 1. During the evaluation process, coherence scores are compared, PyLDAvis visualizations are examined, and frequent words associated with each topic are analyzed. The final LDA model has 11 topics with a coherence score of 0.530945086.
- The BERTopic model was first created using the default settings and later improved by adjusting the parameters to support multiple languages and automatically merge topics. After that, other hyperparameters were fine-tuned to produce the final BERTopic model. This model identified 75 topics (excluding outliers) and achieved an overall coherence score of 0.43627.
- After selecting the final LDA and BERTopic models, further analysis provide valuable insights for businesses. This includes identifying the most and least discussed topics, tracking the shift in topic trends over time, understanding how topics are represented across different rating levels, identifying which topics are linked to the highest and lowest customer ratings, and determining which products are associated with these topics. For any given product, such analysis can reveal which topics are being most frequently discussed by customers.
- Both LDA and BERTopic are good models, but BERTopic is the preferred choice for this study. Although the LDA model has a slightly higher coherence score, BERTopic

has several advantages in terms of topic comprehensibility, computational efficiency, and simplicity of data preparation. A detailed comparison of the two methods is presented in Table 8 for reference purposes.

- Topic modeling is a powerful tool for simplifying and structuring unstructured data analysis. To ensure that the results of the modeling method align with strategic business objectives, it is recommended to thoroughly examine the content of the review. Additionally, consulting domain experts and key stakeholders is essential to ensure that the insights are both accurate and applicable.
- This study encountered some limitations, notably the lack of advanced computational resources like GPUs, which hinder the speed and scale of model training. Furthermore, the absence of domain expertise in musical instruments may have limited a complete understand of the identified topics.

5 Conclusion

The main objective of this study is to explore the application of topic modeling to cultivate actionable business insights from unstructured data. The process involves a comprehensive review of recent literature on various topic modeling methods, and LDA and BERTopic are selected for an empirical study on customer reviews of musical instruments on Amazon. Guided by three research questions, this study seeks to understand the strengths, limitations, and comparative advantages of these two techniques. Some conclusions are derived from revisiting the research questions.

1. **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?**

   The results of the experiment show that both final LDA and BERTopic models are proficient in identifying general topics within the dataset. While LDA offers users a broad overview with fewer topics, BERTopic produces more topics that can be ranked in hierarchy order. This granular insight offered by BERTopic can be quite beneficial for businesses to delve deeper into each topic, giving them the flexibility to perform more in-depth analysis. However, it is important to note that this experiment was conducted in a way that each review was classified into a single topic with the highest probability in order to easily compare the results of both models. That may result in some potential information loss from reviews that cover multiple topics.

2. **Do the identified topics offer any valuable insights for businesses?**

   Both LDA and BERTopic models are able to provide many valuable insights for business analytics and decision-making processes. These insights range from identifying trending topics over time, analyzing whether a topic represents a positive aspect that should be maintained or highlights an issue that needs improvement, identifying products associated with each topic, identifying top products that are linked to each topic, or pinpointing topics associated with specific products. Such insights are invaluable for marketing, product development, and quality assurance teams. They can be used to showcase successful products in promotional campaigns, enhance the products that are well-received by customers, or address issues that are mentioned in negative feedback. This aids businesses in enhancing customer satisfaction and more closely aligning with consumer expectations.

3. **Between the two topic modeling methods, which method performs better on the Amazon Reviews dataset?**

   Although both LDA and BERTopic perform quite well, BERTopic has a slight edge over LDA in several key aspects, including computational efficiency, resource utilization, coherence scores, and overall topic quality, in this particular use case of Amazon customer reviews. Nevertheless, this preference should be taken with caution, as it may not generalize across all use cases. LDA holds several advantages over BERTopic in other contexts as details are described in Table 8. Therefore, it is advisable to chose the best methods that is suitable for a particular use case.

The process of topic modeling is fundamentally exploratory, aimed at gaining initial insights from a large amount of unstructured data. It is by no means a perfect solution but rather a starting point for deeper analysis. Human inspection of topic quality and further analysis like sentiment analysis are highly recommended so that they can give a well-informed picture and support decision-making. Additionally, the interpretation of topic modeling outcomes is inherently subjective, varying significantly across different people. Analysts and stakeholders could potentially draw different conclusions from the same data. This subjectivity necessitates a balanced approach, incorporating diverse perspectives to ensure that the insights align with business goals.

Future research endeavors could benefit from incorporating cutting-edge language models such as OpenAI's GPT-3.5 Turbo for the task of automatic topic labeling in BERTopic. This advanced artificial intelligence could potentially streamline the process by generating intuitive and coherent labels for groups of words associated with each topic, thereby enhancing the interpretability of the topic models. Utilizing such sophisticated artificial intelligence technology would allow businesses to quickly and efficiently understand the underlying themes within large datasets, facilitating a more agile response to market trends and customer feedback. This approach would not only save valuable time and resources but also potentially increase the accuracy of topic identification and labeling, provided that the necessary computational resources are available to support its implementation.

In conclusion, this thesis presents the practical applications and comparison of LDA and BERTopic in extracting meaningful information from Amazon customer reviews. It also highlights the critical need for careful consideration, further analysis, and collaborative interpretation of the results to drive informed business decisions.

List of literature

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, *3*, 42. https://doi.org/10.3389/frai.2020.00042

An, Y., Oh, H., & Lee, J. (2023). Marketing Insights from Reviews Using Topic Modeling with BERTopic and Deep Clustering Network. *Applied Sciences*, *13*(16), 9443. https://doi.org/10.3390/app13169443

Anantharaman, A., Jadiya, A., Siri, C. T. S., Adikar, B. N., & Mohan, B. (2019). Performance Evaluation of Topic Modeling Algorithms for Text Classification. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 704–708). IEEE. https://doi.org/10.1109/ICOEI.2019.8862599

Anees, A. F., Shaikh, A., Shaikh, A., & Shaikh, S. (2020). Survey paper on sentiment analysis: Techniques and challenges. https://easychair.org/publications/preprint_download/sc2h

Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0255-7

Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, *44*(9), 875–886. https://doi.org/10.1080/1206212x.2021.1974663

Bicalho, P., Pita, M., Pedrosa, G., Lacerda, A., & Pappa, G. L. (2017). A general framework to expand short text for topic modeling. *Information Sciences*, *393*, 66–81. https://doi.org/10.1016/j.ins.2017.02.007

Bisgin, H., Liu, Z., Fang, H., Xu, X., & Tong, W. (2011). Mining FDA drug labels using an unsupervised learning technique--topic modeling. *BMC Bioinformatics*, *12 Suppl 10*(Suppl 10), S11. https://doi.org/10.1186/1471-2105-12-S10-S11

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In W. W. Cohen & A. Moore (Eds.), *Proceedings: [June 25-29, 2006, Pittsburgh, PA, USA]* (pp. 113–120). ACM. https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Chakkarwar, V., & Tamane, S. C. (2020). Quick Insight of Research Literature Using Topic Modeling. In Y.-D. Zhang, J. K. Mandal, C. So-In, & N. V. Thakur (Eds.), *Smart Innovation, Systems and Technologies: Vol. 165. Smart Trends in Computing and Communications: Proceedings of SmartCom 2019* (1st ed. 2020, Vol. 165, pp. 189–197). Springer Singapore; Imprint: Springer. https://doi.org/10.1007/978-981-15-0077-0_20

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, 1–13. https://doi.org/10.1016/j.knosys.2018.08.011

Chen, Y., Rabbani, R. M., Gupta, A., & Zaki, M. J. (2017). Comparative text analytics via topic modeling in banking. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). IEEE. https://doi.org/10.1109/SSCI.2017.8280945

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10s), 1–35. https://doi.org/10.1145/3507900

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? — An empirical investigation of panel data. *Decision Support Systems*, *45*(4), 1007–1016. https://doi.org/10.1016/j.dss.2008.04.001

Egger, R., & Yu, J. (2021). Identifying hidden semantic structures in Instagram data: a topic modelling comparison. *Tourism Review.* Advance online publication. https://doi.org/10.1108/TR-05-2021-0244

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

Fan, A., Doshi-Velez, F., & Miratrix, L. (2017, January 12). Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling. http://arxiv.org/pdf/1701.03227v3

García-Méndez, S., Arriba-Pérez, F. de, Barros-Vila, A., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of relevant information, predictions and

forecasts in financial news through topic modelling with Latent Dirichlet Allocation. *Applied Intelligence*, *53*(16), 19610–19628. https://doi.org/10.1007/s10489-023-04452-4

Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4430182

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/pdf/2203.05794.pdf

Gu, S., Ślusarczyk, B., Hajizada, S., Kovalyova, I., & Sakhbieva, A. (2021). Impact of the COVID-19 Pandemic on Online Consumer Purchasing Behavior. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(6), 2263–2281. https://doi.org/10.3390/jtaer16060125

Gui, L., Leng, J., Pergola, G., Zhou, Y., Xu, R., & He, Y. (2019). Neural Topic Model with Reinforcement Learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 3476–3481). IEEE. https://doi.org/10.18653/v1/D19-1350

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*(7), 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS One*, *15*(5), e0232525. https://doi.org/10.1371/journal.pone.0232525

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*, *13*(2), 586–632. https://doi.org/10.5465/annals.2017.0099

Haque, T. U., Saber, N. N., & Shah, F. M. (2018). Sentiment analysis on large scale Amazon product reviews. In *ICIRD 2018: 2018 IEEE International Conference on Innovative Research and Development: May 11, 2018, AIT Conference Center Bangkok, Thailand* (pp. 1–6). IEEE. https://doi.org/10.1109/ICIRD.2018.8376299

Heap, B., Bain, M., Wobcke, W., Krzywicki, A., & Schmeidl, S. (2017, September 18). *Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems*. http://arxiv.org/pdf/1709.05778.pdf

Hofmann, T. (1999). Probabilistic latent semantic indexing. In F. Gey, M. Hearst, & R. Tong (Eds.), *Proceedings of the 22nd annual international ACM SIGIR conference on Research*

and development in information retrieval (pp. 50–57). ACM. https://doi.org/10.1145/312624.312649

Islam, T. (2019). Yoga-Veganism: Correlation Mining of Twitter Health Data. https://doi.org/10.13140/RG.2.2.10252.16009

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2017, November 12). *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey*. http://arxiv.org/pdf/1711.04305.pdf

Jiang, N., Crooks, A. T., Kavak, H., & Wang, W. (2023). Leveraging newspapers to understand urban issues: A longitudinal analysis of urban shrinkage in Detroit. *Environment and Planning B: Urban Analytics and City Science*, Article 23998083231204695. Advance online publication. https://doi.org/10.1177/23998083231204695

Juan, L., Wang, Y., Jiang, J., Yang, Q., Wang, G., & Wang, Y. (2020). Evaluating individual genome similarity with a topic model. *Bioinformatics (Oxford, England)*, *36*(18), 4757–4764. https://doi.org/10.1093/bioinformatics/btaa583

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Krishnan, A. (2023, August 19). Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis. https://doi.org/10.48550/arXiv.2308.11520

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 302–308). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2050

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic Modeling for Short Texts with Auxiliary Word Embeddings. In R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, & J. Zobel (Eds.), *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 165–174). ACM. https://doi.org/10.1145/2911451.2911499

Liang, J., Liu, P., Tan, J., & Bai, S. (2014). Sentiment Classification Based on AS-LDA Model. *Procedia Computer Science*, *31*, 511–516. https://doi.org/10.1016/j.procs.2014.05.296

Liu, H., & Du, F. (2023). Research on E-Commerce Platforms' Return Policies Considering Consumers Abusing Return Policies. *Sustainability*, *15*(18), 13938. https://doi.org/10.3390/su151813938

Luo, J., Nan, G., Li, D., & Tan, Y. (2023). AI-Generated Review Detection. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.4610727

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2-3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

May, C., Cotterell, R., & van Durme, B. (2016, August 13). An Analysis of Lemmatization on Topic Models of Morphologically Rich Language. http://arxiv.org/pdf/1608.03995.pdf

Mazarura, J., & Waal, A. de (2016). A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech): 30 November - 2 December, Stellenbosch, South Africa* (pp. 1–6). IEEE. https://doi.org/10.1109/RoboMech.2016.7813155

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. http://arxiv.org/pdf/1802.03426v3

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013, October 17). Distributed Representations of Words and Phrases and their Compositionality. http://arxiv.org/pdf/1310.4546.pdf

Miyajiwala, A., Ladkat, A., Jagadale, S., & Joshi, R. (2022). On Sensitivity of Deep Learning Based Text Classification Algorithms to Practical Input Perturbations. In K. Arai (Ed.), *Lecture notes in networks and systems: volume 506-508, Intelligent computing: Proceedings of the 2022 Computing Conference* (pp. 613–626). Springer. https://doi.org/10.1007/978-3-031-10464-0_42

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec.

Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software*, *3*(23), 655. https://doi.org/10.21105/joss.00655

Müller, T., Cotterell, R., Fraser, A., & Schütze, H. Joint Lemmatization and Morphological Tagging with Lemming. In (pp. 2268–2274). https://doi.org/10.18653/v1/d15-1272

Nagarhalli, T. P., Vaze, V., & Rana, N. K. (Eds.) (2021). Impact of Machine Learning in Natural Language Processing: A Review.

Nayak, A. S., & Kanive, A. P. (2016). Survey on Pre-Processing Techniques for Text Mining. *International Journal of Engineering and Computer Science.* Advance online publication. https://doi.org/10.18535/ijecs/v5i6.25

Ni, J., Li, J., & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1018

Nigam, K., Mccallum, A. K., Thrun, S., & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, *39*(2/3), 103–134. https://doi.org/10.1023/A:1007692713085

Nogoev, A., Yazdanifard, R., Shahriar, M., Behrang, S., & Meera, M. (2011). The Evolution and Development of E-Commerce Market and E-Cash. In Y. Xie (Ed.), *International Conference on Measurement and Control Engineering 2nd (ICMCE 2011)* (pp. 245–252). ASME Press. https://doi.org/10.1115/1.859858.paper35

Oelke, D., Strobelt, H., Rohrdantz, C., Gurevych, I., & Deussen, O. (2014). Comparative Exploration of Document Collections: a Visual Analytics Approach. *Computer Graphics Forum*, *33*(3), 201–210. https://doi.org/10.1111/cgf.12376

Patra, A., & Singh, D. (2013). Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method. *International Journal of Computer Applications*, *68*(17), 37–41. https://doi.org/10.5120/11674-7301

Pooja, K., & Upadhyaya, P. (2022). What makes an online review credible? A systematic review of the literature and future research directions. *Management Review Quarterly.* Advance online publication. https://doi.org/10.1007/s11301-022-00312-6

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and Sparse Text Topic Modeling via Self-Aggregation. In (pp. 2270–2276). AAAI Press/International Joint Conferences on Artificial Intelligence. https://scholars.cityu.edu.hk/en/publications/short-and-sparse-text-topic-modeling-via-selfaggregation(f7f973ab-812b-4fae-ad8f-dced13195b08).html

Ray, S. K., Ahmad, A., & Kumar, C. A. (2019). Review and Implementation of Topic Modeling in Hindi. *Applied Artificial Intelligence*, *33*(11), 979–1007. https://doi.org/10.1080/08839514.2019.1661576

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. https://doi.org/10.48550/arXiv.1403.6397

Shehu, E., Papies, D., & Neslin, S. A. (2020). Free Shipping Promotions and Product Returns. *Journal of Marketing Research*, *57*(4), 640–658. https://doi.org/10.1177/0022243720921812

Stanisz, T., Drożdż, S., & Kwapień, J. (2024). Complex systems approach to natural language. *Physics Reports*, *1053*, 1–84. https://doi.org/10.1016/j.physrep.2023.12.002

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *EMNLP-CoNLL '12, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952–961). Association for Computational Linguistics.

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In J. (. International Conference on Data Science and Advanced Analytics Tokyo (Ed.), *DSAA 2017 : proceedings : 2017 International Conference on Data Science and Advanced Analytics : Tokyo, Japan, 19-21 October 2017* (pp. 165–174). IEEE. https://doi.org/10.1109/DSAA.2017.61

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

Thompson, L., & Mimno, D. (2020, October 23). Topic Modeling with Contextualized Word Representation Clusters. http://arxiv.org/pdf/2010.12626.pdf

Wang, X., & Mccallum, A. (2006). Topics over time. In L. Ungar, M. Craven, D. Gunopulos, & T. Eliassi-Rad (Eds.), *KDD-2006: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining August 20-23, 2006, Philadelphia, PA, USA* (pp. 424–433). ACM Press. https://doi.org/10.1145/1150402.1150450

Xie, P., & Xing, E. P. (2013, September 26). Integrating Document Clustering and Topic Modeling. http://arxiv.org/pdf/1309.6874v1

Xu, A., Qi, T., & Dong, X. (2020). Analysis of the Douban online review of the MCU: based on LDA topic model. *Journal of Physics: Conference Series*, *1437*(1), 12102. https://doi.org/10.1088/1742-6596/1437/1/012102

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In D. Schwabe, V. Almeida, H. Glaser, R. Baeza-Yates, & S. Moon (Eds.), *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445–1456). ACM. https://doi.org/10.1145/2488388.2488514

Yin, H., Song, X., Yang, S., & Li, J. (2022). Sentiment analysis and topic modeling for COVID-19 vaccine discussions. *World Wide Web*, *25*(3), 1067–1083. https://doi.org/10.1007/s11280-022-01029-y

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In S. Macskassy, C. Perlich, J. Leskovec, W. Wang, & R. Ghani (Eds.), *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233–242). ACM. https://doi.org/10.1145/2623330.2623715

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16 Suppl 13*(Suppl 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593