Identifying Customer Preferences and Marketing Insights with BERTopic from E-commerce Reviews

CONTENTS

CHAPTER I : INTRODUCTION

1.1 Background and Significance

In the digital era, online platforms have become a vital medium for individuals to post and share reviews on various products and services. These reviews offer invaluable insights to vendors and manufacturers, reflecting the consumers' perspectives on a business's offerings. The information contained within these reviews has become a crucial resource to gain insights into consumer sentiments and preferences and can significantly enhance companies' marketing strategies. However, due to the qualitative and unstructured format of reviewer prose, the challenge lies in analyzing and extracting meaningful insights, which is a formidable task given the complexity and volume of the data involved. The introduction of natural language processing models provides a great marketing asset to analyze this rich available data.

The performance of text analysis models heavily depends on the context of the reviews, the terminology used, and the domain to which the reviews belong. Traditional models and methods, which rely on labeled data from other categories, often fail to achieve high accuracy due to the specific and varied nature of review content. A fundamental requirement for such analysis is the availability of a high-quality labeled dataset which is notably labor-intensive and costly, particularly when the dataset required for comprehensive analysis is large. The challenge is further compounded in scenarios where there is a scarcity of labeled training data.

Machine learning emerges as a potent solution to these challenges, offering a way to circumvent the need for labeled text data, especially for topic modeling, a machine learning technique designed to automatically identify the themes or topics present within a large corpus of text. In content analysis, topic modeling can uncover hidden thematic structures in text data, enabling the analysis of customer reviews without the prerequisite of a labeled dataset.

Among the various models employed for topic modeling, Latent Dirichlet Allocation (LDA) has been widely used in marketing to analyze consumer-related topics across different domains, such as food distribution (Jiyoon, K.2021), food tourism (Heekyung, N 2020) and airline services in Korea (Lee, B.-J. 2021) to

analyze consumer opinions. Recent studies have explored advanced techniques such as BERTopic, which utilizes clustering algorithms based on review data to analyze consumer intent and extract relationships between products, offering new insights and methodologies for accurate topic modeling.

The market for ergonomic chairs has witnessed significant growth, with forecasts projecting a compound annual growth rate (CAGR) of 7% from 2022 to 2031, culminating in a global market value of USD 16.88 billion by 2031. North America, in particular, has emerged as a leading market, holding a 31.6% share in 2022. Dominated by a few large-scale players, the industry has seen major initiatives aimed at launching various types of ergonomic chairs to meet the rising consumer demand. These companies have also focused on leveraging technology and materials advancements, diversifying service offerings, and enhancing their online presence as core business strategies to maintain and expand their market share.

This research seeks to address the insufficiency of labeled training datasets for ergonomic chair reviews by employing machine learning and topic modeling techniques. The study begins with the collection of ergonomic chair review texts, followed by a series of preprocessing tasks including text deduplication, data cleansing, text segmentation (tokenization), part-of-speech tagging, and the removal of stop words. These preparatory steps ensure the data is primed for in-depth analysis. The processed text data then undergoes topic modeling to identify and extract prevalent themes. Visualizations generated from the text data offer a graphical representation of the most frequent and significant terms related to ergonomic chairs.

With utilization of models such as BERTopic, the study aims to categorize the review data into specific topics, extracting the advantages and disadvantages associated with each and uncovering key insights into consumer needs, opinions, complaints, and emotional expressions. Through this approach, the research aspires to provide marketing researchers with up-to-date information on machine learning methods in marketing science and bridge the gap in customer preference analysis within the specific domain of ergonomics chair reviews.

1.2 Literature review

Simple definition of ergonomics involves the adoption of appropriate postures and equipment for individuals engaged in sedentary or repetitive tasks over extended periods. Implementing ergonomic principles within office environments is pivotal for safeguarding employees' physical health and mental well-being, especially for those confined to a single posture for long durations. Proper ergonomic practices are essential for maintaining spinal health, preventing conditions such as carpal tunnel syndrome, and fostering healthier postural habits.

The escalating demand for ergonomic chairs has been prominently observed across various industrial sectors, manufacturing entities, educational institutions, and organizations in recent years, propelling the expansion of the ergonomic chair market globally. This surge in demand is attributed to the chairs' beneficial attributes, such as adjustability, stability, and ease of use, catering primarily to settings where prolonged sitting is requisite. Meyer and Fourie (2015) elucidate that ergonomics, as a discipline, endeavors to optimize the congruence between individuals and their work settings, applicable across diverse environments. The nature of office tasks has transformed, necessitating a highly collaborative and interactive work setting, thus shifting from a conventional "office ergonomic" paradigm, focused on engineering and cognitive ergonomics, towards a more "holistic" approach. This comprehensive perspective integrates the physical and mental facets of ergonomics with the dynamics of social interaction within a collective workspace.

Furthermore, the significance of digital marketing within the industrial sector is on the rise, notably in fields characterized by intricate sales processes, as highlighted by Järvinen (2015). Farman (2014) delineates seven crucial digital marketing strategies for the furniture business, which include understanding the market landscape and positioning, crafting a website layout that adapts seamlessly across various devices, enhancing user experience, setting up a virtual storefront and e-commerce payment solutions to facilitate online transactions, enhancing online visibility, engaging content tailored to target audience's needs and promoting specific furniture designs or discounts through targeted display and search advertising. Digital marketing serves as an electronic communication by marketers to advocate for their products and services in the market (Warbung et la., 2023). Precisely, it is

defined as the process of promoting and acquiring information, products, and services via computer networks or the internet (Shankar, 2021). Ultimately, digital marketing empowers marketers to extend their product outreach to users through diverse channels, encompassing email marketing, online advertising, social media marketing, mobile marketing, and more (Mandagi & Aseng, 2021; Tatemba & Rantung, 2021; Komaling & Taliwongso, 2023).

The recent study by Munir (2023) looked into 500 expert articles and studies on digital marketing from 2018 to 2023. It points out that nowadays, digital marketing mostly uses the Internet and social media. However, it's expected to shift towards understanding people's emotions and how they behave as consumers. There's a growing interest in using big data analysis, artificial intelligence (AI), and machine learning to make marketing efforts more effective and to spot potential customers.

Customer reviews are seen as vast collections of text that are beyond traditional ways of analysis. These digital footprints of human opinions and actions, along with the rapid increase in computer power, have led to new marketing studies that use big data and text-mining algorithms (Lucini et al. 2020; Ordenes et al. 2014; Chung et al. 2022; Jia 2018; Agrawal and Mittal 2022). This new wave of research relies on web scraping, data mining, and text-mining algorithms to pull meaningful insights from text data (Baka 2016). While content analysis has been popular, these studies now use Natural Language Processing (NLP), machine learning, and text-mining techniques for a more effective way to collect, process, analyze, visualize, and make sense of customer reviews (Elragal and Klischewski 2017).

As online reviews keep stacking up on various websites, finding useful information gets tougher. Topic modeling has become quite popular (Albalawi et al. 2020) for its ability to uncover these patterns (Luo et al. 2020; Egger 2022). One widely used method is Latent Dirichlet Allocation (LDA) developed by Blei et al. (2003), although it has faced some criticism, mainly regarding data preprocessing challenges (Zhou et al. 2017). To overcome these issues, newer techniques like Top2Vec, Sentence-BERT (SBERT), and BERTopic have been developed, offering easier implementation and lighter data preprocessing (Zhou et al. 2017).

In the last few years, as NLP and machine learning have advanced, topic modeling has also found its place in marketing research. Büschken and Allenby

(2016) introduced a new sentence-based topic model that offers a more detailed topic structure. Other studies have looked into customer reviews in tourism and travel, identifying key aspects of customer service (Guo et al. 2017) or visitor experiences in theme parks (Luo et al. 2020). Zhang et al. (2021a, b) used LDA and sentiment analysis on thousands of Airbnb reviews, while Heng et al. (2018) applied deep learning to analyze grocery product reviews on Amazon. Hendry et al. (2021) and Filieri et al. (2022) used BERTopic, Top2Vec, and LDA to analyze customer interactions with e-commerce chatbots and reviews on travel sites, respectively, to understand customer sentiments towards service robots.

In the realm of marketing strategies for ergonomic chairs, scholarly attention has predominantly been focused on empirical research which reveals a lag in the widespread integration of the term "ergonomics" within marketing strategies, particularly within the furniture industry. The numerous literature concerning marketing strategies in the furniture sector has prioritized aesthetic attributes over ergonomic considerations. Ismail and Rahman (2017) have identified a notable deficiency in the representation of "ergonomics" on furniture industry websites, advocating for future marketing strategies to emphasize ergonomic features prominently on company platforms to enhance public health awareness and to effectively utilize digital marketing tools.

Furthermore, the academic discourse surrounding online reviews encompasses a broad spectrum of investigations, including the usefulness of online reviews, consumer purchase intentions, empirical studies on online reviews, and sentiment analysis of such reviews, along with a comparative analysis of different models utilized in these methodologies. This diversity highlights the imperative for researchers employing topic modeling to possess a nuanced understanding of each algorithm's capabilities and potential strengths, which are intricately linked to their respective data properties.

This current study diverges from a detailed exposition on the computational foundations of topic modeling and its associated techniques. Instead, it pivots towards an exploration of the methodological processes and the semantic outcomes derived from employing various algorithms within the context of analyzing customer reviews in the ergonomic chair market. By undertaking this approach, the research

will shed light on a deeper and more nuanced understanding of how topic modeling can be leveraged to decipher consumer sentiments and preferences within this niche market.

1.3 Research Goals and Questions

This thesis has the following research goals. Firstly, the primary goal is to create a robust framework that integrates the novel NPFST approach with established methods such as Word2Vec, HDBSCAN, MMR, and UMAP, enhancing the BERTopic model's ability to extract and analyze marketing-related information from review data. This framework aims to advance the field of topic modeling by providing deeper insights into consumer preferences and market trends. Moreover, this research aims to showcase how topic modeling, particularly the Bertopic model, can be directly or indirectly employed to address a variety of marketing problems. By capturing the nuances and related topics within review data, the study seeks to offer a detailed understanding of consumer behavior towards ergonomic chairs and potentially other products. Finally, by training and fine-tuning the model specifically for ergonomic chair review datasets, the thesis intends to introduce a methodological framework that can be adopted by other researchers and companies. In summary, the study tries to answer the following research questions:

1. How does the integrated framework enhance the capabilities of the BERTopic model in extracting relevant marketing information from review data?
2. In what ways can topic modeling be utilized to solve marketing problems, and how does it contribute to a better understanding of consumer preferences and market trends?
3. What are the methodological implications of the proposed framework for future research in topic modeling and marketing analysis?
4. How do the insights derived from the topic modeling of ergonomic chair reviews inform product development and marketing strategy formulation?

1.4 Contributions of the Research

Initially, this dissertation introduces a novel framework designed to extract and scrutinize marketing-related insights from review data. This framework integrates a

variety of models, including the newly introduced NPFST, alongside established techniques such as Word2Vec, HDBSCAN, MMR, and UMAP, within the foundational architecture of BERTopic.

Furthermore, the study offers a substantial advancement in the domain of topic modeling, demonstrating its applicability in directly or indirectly tackling diverse challenges within marketing. Particularly, through the adept identification of characteristics and themes in review data, topic modeling, with an emphasis on the BERTopic model, affords a profound insight into consumer inclinations and prevailing market dynamics.

This dissertation delivers an exhaustive analysis of consumer preferences concerning ergonomic chairs. Despite the model being specifically tailored and refined for ergonomic chair datasets, the work proposes a methodological framework extendable to other research endeavors and corporate applications, assessing the effectiveness of topic modeling in analyzing customer feedback. This gleaned intelligence stands as a critical resource for product innovation and the crafting of marketing tactics. Employing this methodology allows businesses to gain a more precise comprehension of consumer sentiments and preferences, thereby informing enhancements to products and services in addition to informing the creation of marketing strategies.

1.5 Organization of Thesis

Chapter I introduces the foundational context and relevance of this thesis, outlining a review of existing literature related to the analysis of customer reviews and topic modeling. It also details the contributions and structure of this thesis. Chapter II delineates the core methodologies employed in this study, encompassing models and algorithms for data preprocessing and topic generation, alongside other pertinent methods in text analysis. Chapter III delves into the criteria for selecting the ergonomics chair review dataset and describes the procedures for cleaning and preparing the dataset for analysis. Chapter IV details the design of the system architecture and the procedures for training Bertopic models. This includes the selection of embedding models, the adjustment of model parameters, and the comprehensive experiments conducted to identify the best configurations for sentiment classification and topic discovery. Chapter V discusses the principal topics

and sentiments revealed in the reviews, examining the implications of these insights for comprehending customer challenges, preferences, and areas of interest. Chapter VI outlines prospective research directions and concludes the thesis.

CHAPTER II : METHODOLOGY

2.1 NLTK (Natural Language Toolkit)

The Natural Language Toolkit (NLTK) is an essential component of the text preprocessing phase in this research, offering a comprehensive suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the English language. Utilized primarily for preparing the ergonomic chair customer reviews for deeper analysis, this research topic harnesses NLTK's ability to facilitate a series of foundational preprocessing tasks that are crucial for refining the dataset and enhancing the quality of subsequent analyses.One of the primary functionalities employed from NLTK is "tokenization", which involves breaking down the text into individual words or sentences. This process is vital as it transforms the raw text into a more analyzable format, enabling the identification of patterns or specific features within the data. Tokenization serves as the first step in organizing the unstructured text data into a structured form.Following tokenization, stop words removal is applied to filter out common words (such as "the", "is", "in", etc.) that appear frequently across texts but hold minimal individual significance. This step is crucial for reducing the dataset's noise, focusing the analysis on words that carry more meaning and are likely to contribute to understanding customer sentiments and preferences.

2.2 Bigram

The research applies a phrase detection algorithm developed by Mikolov and colleagues. This method assesses the frequency of word pairs appearing together compared to their individual occurrences in different contexts. It employs a straightforward statistical approach that calculates a score for each bigram by considering the count of two words appearing together versus their separate appearances in the sentence collection. A discounting coefficient, δ, is introduced to mitigate the overrepresentation of phrases formed from rare words, with δ set to 1 to prevent assigning scores above 0 to phrases seen less than twice.

The total number of tokens in the patent database is denoted by N, calculated as the sum of all token counts in the dataset. Bigrams exceeding a specific threshold score (Tphrase) are recognized as phrases and concatenated with an underscore, treating them as singular terms within the corpus. The algorithm is executed twice on

the preprocessed text, initially identifying bigrams with a higher Tphrase threshold, and subsequently, n-grams up to four words by applying a lower Tphrase. This methodical reduction in the threshold value allows for the discovery of more frequently occurring phrases in the first iteration, such as "lumbar support," and less common phrases in the subsequent iteration, like "lumbar support back."

2.3 Part-of-Speech Tagging

Another significant functionality utilized is “Part-of-Speech (POS) tagging”, which assigns tags to each word in the text, indicating their grammatical roles (such as nouns, verbs, adjectives, etc.). POS tagging is particularly important in this research as it enables the identification of specific word types that are more relevant to customer opinions, such as adjectives and nouns, which often carry significant sentiment or describe key aspects of the products. Specifically, the process involves the application of NLTK's Python pos_tag function(Figure 2-2) to a series of text data–in this case, customer reviews– thereby assigning a corresponding part-of-speech tag to each word in the series. These tags provide essential linguistic information about the grammatical role and category of each word. For example, when the word “sitting” is tagged as a VERB, it would be regularized as “sit” while it would be regularized as “sitting” when it is tagged as a NOUN.

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

Figure 2-2: Universal part-of-speech tag

The integration of NLTK's functionalities into the preprocessing workflow significantly enhances the dataset's readiness for detailed NLP analysis. By efficiently tokenizing the text, removing stop words, and applying POS tagging, the research benefits from a cleaner, more focused corpus that is primed for extracting

meaningful insights into customer experiences and opinions regarding ergonomic chairs. This careful preparation of the text data lays a solid foundation for the application of more advanced analytical techniques, ensuring that the findings derived from the study are both reliable and insightful.

2.4. Vectorization and Embedding Models

2.4.1. CountVectorizer Model

The CountVectorizer model is a text vectorization tool that converts text data into a numerical format, enabling machine learning algorithms to process natural language. By counting the occurrence of each word or term in the dataset, CountVectorizer generates a sparse matrix of term/token frequencies. For example(Figure 2-3): The text of ‘Hello my name is james' , ‘this is my python notebook’ will be generated to integer assignments as below:

	hello	is	james	my	name	notebook	python	this
0	1	1	1	1	1	0	0	0
1	0	1	0	1	0	1	1	1

Figure 2-3: sparse matrix of term/token frequencies

This model is fundamental in preprocessing steps for NLP tasks as it transforms raw text into a format that can be easily analyzed and compared. The simplicity and effectiveness of CountVectorizer make it a standard choice for feature extraction in text classification and topic modeling tasks. Moreover, due to the nature of vocabulary redundancy in reviews, makes CounterVectorizer a necessity in preprocessing.

2.4.2 Text Analysis and Word Embedding

Gensim is a robust and versatile library designed for unsupervised text analysis on raw text and plays a crucial role in the methodology of this research. Word2Vec, specifically in the package, is useful for generating word embeddings, a technique pivotal for understanding natural language data. Word2Vec as used in this research, excels in translating words into high-dimensional vector spaces, effectively

capturing the semantic meanings and relationships between words based on their contextual usage in the text. By converting words into vectors, it allows for the quantitative assessment of word similarities, thereby facilitating tasks such as similarity assessment, clustering, and feature extraction from customer review data. This capability is instrumental in identifying patterns and themes within the feedback on ergonomic chairs, enabling a nuanced understanding of customer sentiments and preferences. For instance, through similarity assessments, Word2Vec helps in grouping synonymous terms or related product features, offering humanly understandable insights that may be critical for product development and marketing strategies.

2.3. Phrase Detection and Extraction

2.3.1 Gensim

Gensim is further leveraged in this study for its phrase detection capabilities. Unlike the direct approach of NPFST (explained below), which focuses on extracting contiguous word sequences, Gensim employs a statistical model to identify meaningful bigrams or trigrams. This model assesses the co-occurrence patterns of words within the corpus, considering not just immediate adjacency but also the significance of word pairs or triplets within broader textual contexts.Gensim's phrase detection is particularly valuable for distilling complex customer feedback into actionable insights. By identifying statistically significant phrases, it aids in uncovering not only explicit but also implicit expressions of customer opinions and experiences. This contrasts with NPFST's method, offering a complementary perspective that enriches the analysis. For example, while NPFST might capture direct mentions of "lumbar support" as a key feature discussed in reviews, Gensim's phrase detection could reveal related phrases like "back pain relief" or "comfortable seating" based on their contextual significance, providing a broader view of customer concerns and preferences.This comprehensive approach not only enhances the understanding of semantic relationships between words but also extends the capability to extract and interpret meaningful patterns and themes from the text data, thereby supporting a more informed and nuanced analysis of customer feedback.

2.3.2. NPFST (Novel Phrase Finding and Segmentation Tool)

The Novel Phrase Finding and Segmentation Tool (NPFST) stands out for its ability to precisely identify and extract contiguous word sequences directly from corpus and capturing the original and explicit phrases mentioned in customer reviews. This specificity in phrase detection is critical, because it allows for the retention of the nuanced and often colloquial expressions used by customers to describe their experiences and opinions about ergonomic chairs.Gensim employs a statistical model to identify significant word co-occurrences, based on their distribution within the dataset, while, NPFST, on the other hand, takes the linear sequence of words as they appear in the larger text. By relying on the explicit sequential order of words, NPFST is adept at capturing phrases that are bound by phrase structure, such as "back support" or "armrest comfort." This method ensures that the phrases extracted are faithful representations of the language used in the reviews, maintaining the context and specificity intended by the customers.

The direct extraction method employed by NPFST is exceptionally effective in identifying phrases that customers use to express critical aspects of their experiences, such as "easy to assemble" or "lumbar support." These phrases are often central to understanding customer satisfaction and product features, making their accurate representation vital. The clarity and specificity of phrases preserved by NPFST also facilitate a more straightforward interpretation of customer sentiments and preferences, which is essential for drawing actionable insights from the data.

2.3. Clustering and Dimensionality Reduction

2.3.1. UMAP (Uniform Manifold Approximation and Projection)

UMAP (Uniform Manifold Approximation and Projection) is a novel dimensionality reduction technique that efficiently maps high-dimensional data into a lower-dimensional space, facilitating visualization and subsequent analysis while preserving the intrinsic structure of the data. Unlike traditional methods such as PCA (Principal Component Analysis), UMAP adopts a more sophisticated approach based on topological data analysis, making it particularly effective at uncovering hidden patterns and structures within complex datasets. UMAP assumes a uniform distribution on the Riemannian manifold and then projects it into a lower-dimensional

space in a way that maintains the significant relationships between data points as much as possible. This is achieved through a combination of local and global optimizations that ensure both local neighborhood structures and the broader topological features of the data are preserved, an advantage over similar reduction techniques such as t-SNE. In this research, UMAP is utilized as a critical step before clustering with HDBSCAN, as it significantly reduces the inconvenience of dimensionality by transforming the high-dimensional feature space into a more manageable dataset. This preprocessing step enhances the effectiveness of the HDBSCAN algorithm, enabling it to identify clusters more accurately and efficiently in the reduced space. The integration of UMAP into this clustering workflow emphasizes the commitment to leverage cutting-edge methodologies to achieve deeper insights into the data, particularly in scenarios where traditional clustering techniques might struggle due to high dimensionality.

2.3.2. HDBSCAN

HDBSCAN, an advanced variant of the traditional DBSCAN algorithm, enhances the method of grouping data by density and pinpointing outliers through an unsupervised learning technique. This method employs a granular, bottom-up strategy for developing a hierarchical cluster framework, progressively amalgamating or dividing data points according to their density-based similarity. The core principle of hierarchical clustering involves initiating with individual data points as separate clusters, and then, through an iterative process of merging or dividing, forming a hierarchy until a predefined criterion is satisfied. This process culminates in a hierarchical arrangement, typically depicted through a dendrogram, illustrating the linkage among clusters at varying levels of closeness.

In this research, HDBSCAN stands out among several hierarchical clustering techniques for its unique approach. It assesses the mutual reachability among data point pairs and constructs a minimum spanning tree from these assessments, which is then converted into a hierarchical clustering structure. This methodology is particularly effective in identifying clusters of diverse shapes and sizes, offering a significant improvement over traditional methods like K-means clustering. A critical consideration in clustering analysis is the data's dimensionality; high-dimensional data can hinder effective clustering due to the "curse of dimensionality." Hence,

integrating dimensionality reduction techniques, such as UMAP, into the clustering workflow is advantageous for mitigating these challenges.

2.4. Topic Modeling and Diversity

2.4.1. MMR (Maximal Marginal Relevance)

The Maximal Marginal Relevance (MMR) technique is a sophisticated approach designed to optimize the balance between relevance and diversity within topic modeling tasks. In the context of this research, MMR is employed to ensure that the extracted topics are not only related to the core subject matter but also represent a broad spectrum of customer opinions. The study aims to mitigate the common issue of topic redundancy, where similar or overlapping topics might dominate the analysis, thereby obscuring less frequent but equally valuable insights. MMR operates by iteratively selecting items that maximize a combined score of relevance to the query (or the main topic of interest) and marginality, which is the dissimilarity to items already selected. This dual consideration allows MMR to identify topics that are both highly relevant to the primary focus of the research and distinct from each other, ensuring a comprehensive exploration of the dataset. For example, when applied to ergonomic chair reviews, MMR helps in distinguishing topics related to comfort, durability, design, and affordability, while preventing the dominance of any single aspect that could skew the overall understanding of customer feedback. This methodology enhances the ability to derive actionable insights from the reviews.

2.4.2 BERT (Bidirectional Encoder Representations from Transformers)

The BERT model, introduced by Google AI, is a groundbreaking pre-trained language model leveraging deep learning techniques. As a variant of neural network-based language models, BERT is capable of being trained on extensive collections of unlabelled textual data and subsequently applied to a wide range of natural language processing (NLP) tasks such as text classification, annotation, and question answering. The primary benefit of pre-trained language models like BERT is their adaptability to various NLP tasks through fine-tuning, eliminating the necessity to rebuild the model from the ground up for each new application. BERT's architecture derives from the Encoder component of the Transformer model, which enables it to function as a bidirectional language model. This bidirectional capability

is evident in its Masked LM (Language Model) pre-training approach, where the model receives words simultaneously and utilizes the surrounding contextual information to predict the identities of masked elements.

Processing text data with the BERT model involves two main steps: pre-training and fine-tuning. During pre-training, the model undergoes self-supervised training with a large amount of unlabelled text data to learn textual features and extract deeper vector representations, forming the pre-trained model. In the fine-tuning phase, converged model parameters from the pre-training process are used as input for a start model. The model is then trained further with manually annotated text datasets for specific tasks, enhancing model fitting and convergence. Through these steps, the BERT model can effectively complete natural language processing tasks.

2.4.3 SentenceBERT

Sentence-BERT, or SBERT, represents a modification of the traditional BERT (Bidirectional Encoder Representations from Transformers) framework, incorporating Siamese and triplet network architectures to produce semantically rich sentence embeddings for natural language processing (NLP) tasks. In practice, SBERT serves as a foundational step in clustering documents by topic and discerning the underlying themes within these groups. However, this process often necessitates a preliminary reduction in the dimensionality of document embeddings to mitigate the challenges clustering algorithms face in high-dimensional settings.Following the establishment of SBERT's framework, the discussion shifts towards BERTopic, a technique that builds upon sentence embeddings to perform topic modeling. BERTopic capitalizes on the embeddings generated by SBERT, employing dimensionality reduction and clustering algorithms to group documents into topics. This method allows for the extraction of thematic structures from large text corpora, leveraging the semantic embeddings provided by SBERT to uncover and organize underlying topics within the data. By integrating SBERT's embeddings with topic modeling, BERTopic offers a sophisticated approach to understanding and categorizing the content of extensive document collections, highlighting the interconnected roles of these advanced NLP methodologies in extracting meaningful insights from textual data.

2.4.4. BERTopic

BERTopic employs BERT embeddings alongside class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) to facilitate topic modeling, effectively organizing a collection of documents into distinct topic groups. This methodology enhances the traditional Term Frequency-Inverse Document Frequency (TF-IDF) by incorporating the context in which words appear within documents. It evaluates a word's significance by considering its frequency within a particular document and contrasting this with its distribution across a broader document set. The foundational principle of BERTopic lies in its ability to detect and cluster documents sharing similar semantic themes by embedding them into a vector space for straightforward comparison. Utilizing the SBERT framework, BERTopic achieves superior performance in tasks involving sentence embeddings.

Owing to the high-dimensional nature of BERT embeddings, clustering them directly poses computational challenges. To address this, the technique leverages UMAP for dimensionality reduction, making it adaptable to various language models, irrespective of their dimensional disparities (McInnes et al. 2018). Following dimension reduction, HDBSCAN is applied for clustering the reduced embeddings (McInnes et al. 2017). The process concludes with the generation of topic representations through c-TF-IDF which offers a more targeted approach to evaluating word significance within the realm of topic modeling, allowing for a more nuanced analysis (Grootendorst 2020a).

CHAPTER III : DATA COLLECTION

3.1 Data retrieval

The study's dataset was sourced from customer reviews of ergonomic chair products, which were posted on Amazon, Wayfair, and the brands' own official platforms, all predominant in the U.S. market with timeframe spanned from March 2006 to December 2023. (Figure 3-1)

Amazon and Wayfair are recognized as two of the leading e-commerce platforms in the United States, these platforms provide a comprehensive array of review data, encompassing both customer and expert opinions, star ratings, overall ratings, and an analytical breakdown of the products' quality, usability, and value. This includes a detailed list of advantages and disadvantages for each product. The process of data acquisition was carried out in December 2023. The reviews, which were publicly accessible, were extracted using the Pandas (McKinney 2011) and BeautifulSoup (Richardson 2019) Python libraries, facilitating the scraping of this data.

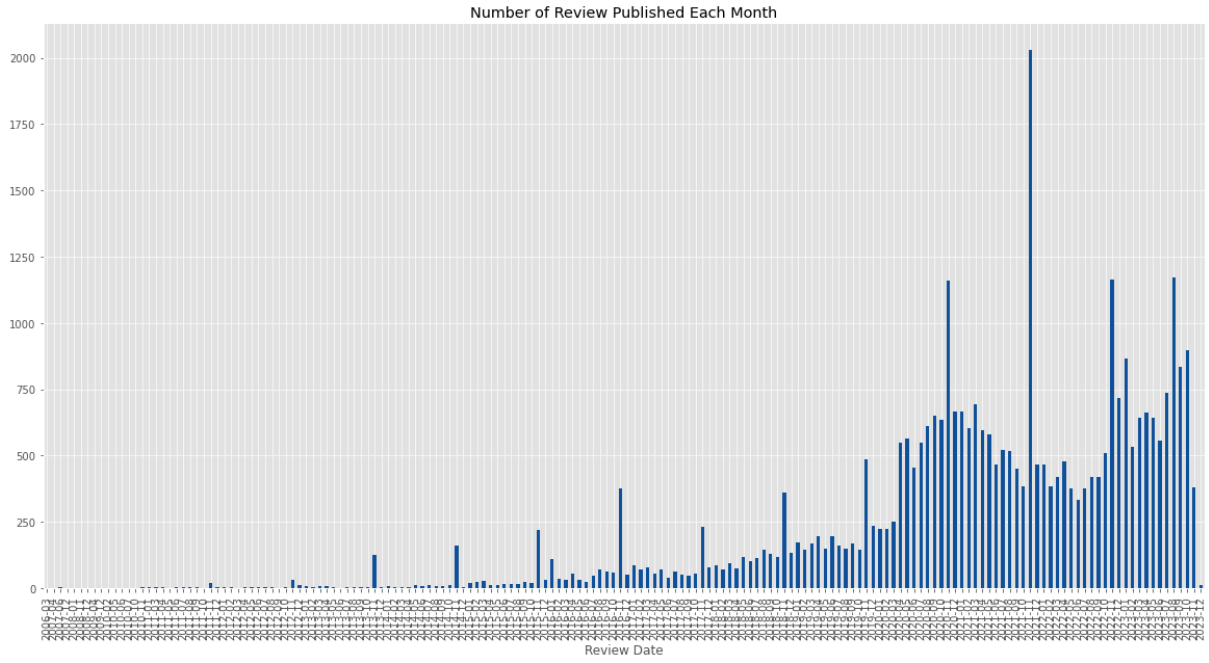


Figure 3-1 Distribution of Review Data According to Monthly Publication Timelines

The original dataset consisted of 51,225 customer evaluations. To ensure

document uniformity, only substantial reviews from U.S. customers were selected; thus, reviews in languages other than English, those that were blank, or contained only a single word were excluded. Out of the 51,225 evaluations, 6,596 were identified as duplicates, and half of these were removed to eliminate redundancy, utilizing the "Remove Duplicates" function in an Excel spreadsheet. After the process of cleaning and removing duplicates, the dataset was refined to include 35,240 customer evaluations (referenced in Figure 3-2). Wayfair contributed the largest portion of usable customer data based on the predefined criteria, with Amazon and their official websites combined contributing to approximately 65% of the total dataset prepared for analysis.

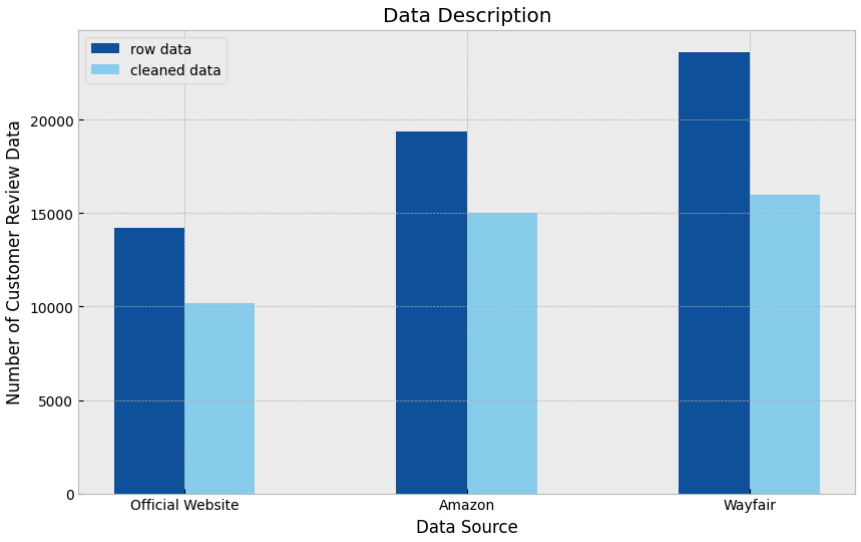
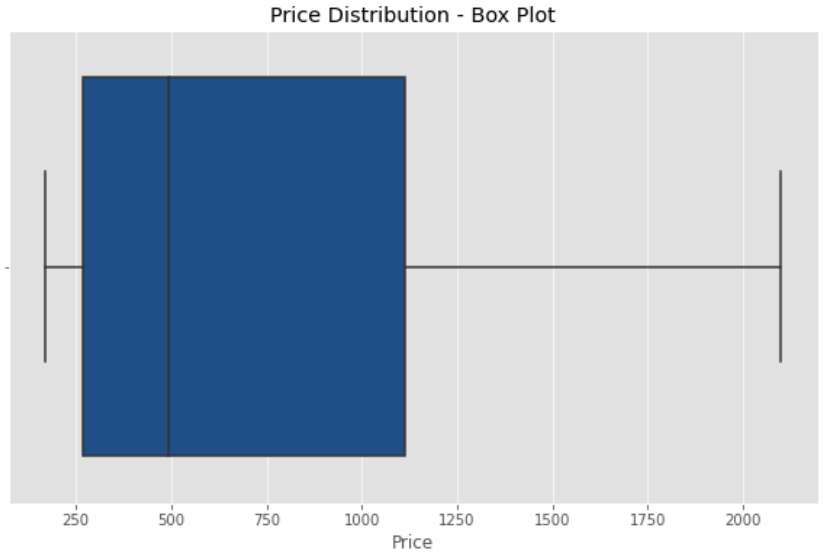


Figure 3-2 Distribution of collected review data by data source

The dataset for the ergonomic chair study showcases a varied price range, divided into four equal categories, each constituting 25% of the total product array(Figure 3-3). The initial category features chairs priced between 250 and 400 USD, offering budget-friendly choices. The subsequent quarter of the products falls into the mid-range price bracket, ranging from 400 to 550 USD, which likely reflects an enhancement in features or brand prestige. The third quartile encompasses chairs priced from 550 to 1000 USD, suggesting a shift towards premium products, possibly equipped with advanced ergonomic attributes and higher quality materials. The final category includes chairs with prices exceeding 1000 USD, belonging to the luxury segment. These chairs are presumed to provide exceptional comfort and

state-of-the-art ergonomic designs, along with possible custom or luxury features. This price distribution offers a comprehensive insight into the ergonomic chair market, addressing a broad spectrum of consumer preferences and financial capacities.

Additionally, the data description indicates a pronounced lean towards high customer satisfaction in the ergonomic chair sector, with an impressive 65.8% of the products receiving a 5-star rating, the highest achievable in a 1-5 star rating scale. This substantial figure points to a strong consumer approval and contentment with these products, signifying their effectiveness in delivering ergonomic support and comfort. Conversely, there exists a noticeable portion of dissatisfaction, as about 10% of the chairs received a 1-star rating. This contrast underlines the presence of products that may not meet consumer expectations or suffer from issues in quality, comfort, or functionality. Such a divergence in ratings highlights the range of consumer experiences and perceptions in the ergonomic chair market, reflecting both high satisfaction and areas of potential improvement. The original review data are presented in the table 3-4.



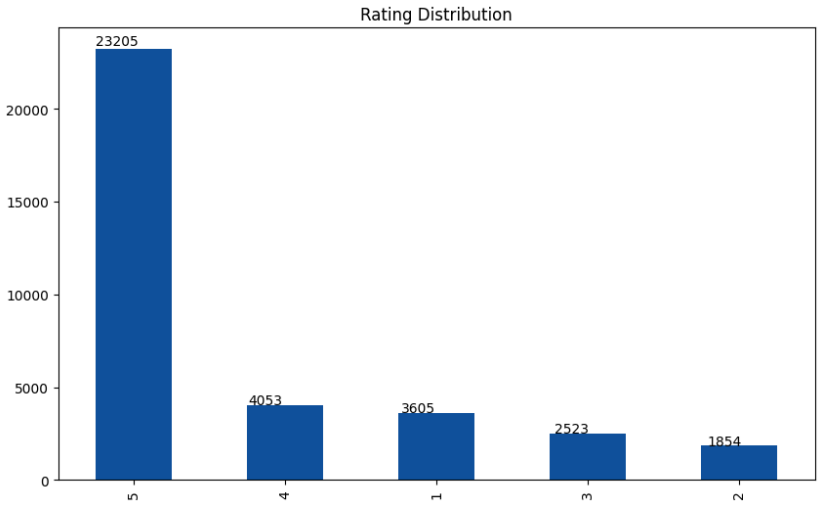


Figure 3- 3 Basic information about reviews dataset. Above: Distribution of collected review data price. Below: Distribution of collected review data star ratings (1-5)

index	Review Date	country	Rate	Price	Review
26701	2021-11-30	United States	4	735.0	Great for back supportMy back feels better but the seat isn't that comfy after 2 months of use.
33449	2022-03-29	United States	5	259.0	I'd give it 6-Stars if I could.I've used it everyday for over a year. I'm sitting in it now as I write this review. It is still comfortable and performing as advertised. Best thing I have bought off of the Zon for quite a while. I'd give it 6-satrs if I could
3499	2021-06-22	United States	3	229.99	So far, okay So far this chair isn't exactly what I was expecting. I love the adjustable armrests, but for me the lumbar support is too low (even adjusted on the highest setting) and pretty hard / uncomfortable on my back. Going to order an extra lumbar support pillow and hopefully that'll make it more comfortable. Overall not bad, but not the best.
4324	2021-12-03	United States	4	189.99	Great chair except for the armrests Ergonomically and seat-comfort wise I really like this chair. What I don't like is that the armrests are cupped/concave and not adjustable (they adjust vertically but the horizontal angle is fixed). If I'm sitting with my arms resting on the armrests perpendicular to my body it's ok, but when I bend my arms in to a typing position the armrests are uncomfortable. If the armrest were horizontally adjustable this chair would be nearly perfect.
7757	2023-04-09	United States	5	319.99	Love this chair! This chair is so comfortable! I love it. I was afraid it would stain easily and so far I've had no issues. I did buy rollerblade wheels for it as I feel the wheels on it didn't work great on the carpet in my office. I'm 5'4" and it's perfect for me.
8362	2023-06-29	United States	5	249.5	Outstanding Office Chair! Chair arrived quickly & was easy to assemble. Comfort is perfect for a tall person (6'4").Customer Service (Cayla) has been great in providing timely responses to product inquiries.Highly recommend!

Table 3-4 Crawls some of the comment texts

3.2 Data Preprocessing

3.2.1 Removal of Stop Words

After removing punctuations and removing words smaller than 3 letters , the detected phrases are processed one more time to split the known stopwords from the NLTK stopwords lists and punctuation marks. Beyond relying solely on NLTK's standard stopword list, this thesis utilizes custom stopwords and enables a more domain-specific analysis tailored to the unique characteristics of the dataset, which consists of user-generated product reviews for ergonomic chairs.

The dataset exhibits recurrent mentions of terms such as 'chair,' 'product,' and 'furniture,' which, while intrinsic to the context, do not contribute substantially to the analysis, extraneous terms like 'https,' 'buy,' 'purchase,' 'order,' 'for,' 'account which

are common in everyday language but may not hold substantial relevance within the context of sentiment analysis or feature assessment of ergonomic chairs. The inclusion of brand names like 'Steelcase,' 'Herman Miller,' and 'wayfair' is commonplace in these reviews. While these brands undoubtedly hold significance, their prominence may bias the analysis when attempting to gauge the objective sentiments and chair features. By eliminating these terms from the analysis, the experiment promotes greater precision and discernment in the interpretation of the reviews and ensures a more impartial evaluation of the user feedback.

Through multiple iterations and updates of the word corpus, the final segmentation and stop word removal results are presented in the Table 3-5.

index	Review Date	Review	Remove Stopword	Unigram Word	Review word count
26701	2021-11-30	Great for back supportMy back feels better but the seat isn't that comfy after 2 months of use.	back supportmy back feel better comfy 2 month use .	['back', 'supportmy', 'back', 'feel', 'comfy', 'month']	18
33449	2022-03-29	I'd give it 6-Stars if I could.I've used it everyday for over a year. I'm sitting in it now as I write this review. It is still comfortable and performing as advertised. Best thing I have bought off of the Zon for quite a while. I'd give it 6-satrs if I could	'd give 6-stars could.i used everyday year . 'm sitting write . still comfortable performing advertised . best bought zon quite . 'd give 6-satrs could	['give', '6-stars', 'could.i', 'use', 'everyday', 'year', 'sit', 'write', 'still', 'comfortable', 'perform', 'advertised', 'best', 'quite', 'give', '6-satrs', 'could']	52
3499	2021-06-22	So far, okay So far this chair isn't exactly what I was expecting. I love the adjustable armrests, but for me the lumbar support is too low (even adjusted on the highest setting) and pretty hard / uncomfortable on my back. Going to order an extra lumbar support pillow and hopefully that'll make it more comfortable. Overall not bad, but not the best.	far , okay far exactly expecting . love adjustable armrest , lumbar support low (even adjusted highest setting) pretty hard / uncomfortable back . going extra lumbar support pillow hopefully 'll make comfortable . overall bad , best .	['okay', 'exactly', 'expect', 'love', 'adjustable', 'armrest', 'lumbar', 'support', 'even', 'adjust', 'highest', 'set', 'pretty', 'hard', 'uncomfortable', 'back', 'go', 'extra', 'lumbar', 'support', 'pillow', 'hopefully', 'make', 'comfortable', 'overall', 'best']	63
4324	2021-12-03	Great chair except for the armrests Ergonomically and seat-comfort wise I really like this chair. What I don't like is that the armrests are cupped/concave and not adjustable (they adjust vertically but the horizontal angle is fixed). If I'm sitting with my arms resting on the armrests perpendicular to my body it's ok, but when I bend my arms in to a typing position the armrests are uncomfortable. If the armrest were horizontally adjustable this chair would be nearly perfect.	except armrest ergonomically seat-comfort wise really . armrest cupped/concave adjustable (adjust vertically horizontal angle fixed) . 'm sitting arm resting armrest perpendicular body ok , bend arm typing position armrest uncomfortable . armrest horizontally adjustable nearly perfect .	['except', 'armrest', 'ergonomically', 'seat-comfort', 'wise', 'really', 'armrest', 'cupped/concave', 'adjustable', 'adjust', 'vertically', 'horizontal', 'angle', 'fix', 'sit', 'arm', 'rest', 'armrest', 'perpendicular', 'body', 'bend', 'arm', 'type', 'position', 'armrest', 'uncomfortable', 'armrest', 'horizontally', 'adjustable', 'nearly', 'perfect']	80
7757	2023-04-09	Love this chair! This chair is so comfortable! I love it. I was afraid it would stain easily and so far I've had no issues. I did buy rollerblade wheels for it as I feel the wheels on it didn't work great on the carpet in my office. I'm 5'4" and it's perfect for me.	love ! comfortable ! love . afraid stain easily far ' issue . rollerblade wheel feel wheel ' work carpet office . ' 5 ' 4 " ' perfect .	['love', 'comfortable', 'love', 'afraid', 'stain', 'easily', 'issue', 'rollerblade', 'wheel', 'feel', 'wheel', 'work', 'carpet', 'office', 'perfect']	55
8362	2023-06-29	Outstanding Office Chair! Chair arrived quickly & was easy to assemble. Comfort is perfect for a tall person (6'4").Customer Service (Cayla) has been great in providing timely responses to product inquiries.Highly recommend!	outstanding office ! arrived quickly & easy assemble . comfort perfect tall person (6 ' 4 ") .customer service (cayla) providing timely response inquiries.highly recommend !	['outstanding', 'office', 'arrive', 'quickly', 'easy', 'assemble', 'comfort', 'perfect', 'tall', 'person', '.customer', 'service', 'cayla', 'provide', 'timely', 'response', 'inquiries.highly', 'recommend']	32

Table 3-5 review data word cut, remove stop word result. 'Review': original review content. 'Remove Stopword': remove stopwords from content. 'Unigram Word': extract unigram words from review.'Review word count': number of words in the original review.

3.2.2 Part-of-Speech Tagging

After gaining a unigram bag of words, the focus is on identifying and isolating relevant bigrams within the text. To achieve this, a filtering step is implemented, wherein only those words that are categorized as nouns (both singular and plural, denoted as "NN" and "NNS," respectively) and adjectives (represented as "JJ") are retained. The rationale behind this filtering is to capture adjective-noun combinations and noun-noun pairs, which are often the meaningful descriptive attributes consumers used to describe product features and purchasing experience, the comparison of phrase generated after part-of-speech tagging and original review are presented in the Table 3-6.

index	Review Date	Review	After POS Tag
26701	2021-11-30	Great for back supportMy back feels better but the seat isn't that comfy after 2 months of use.	['supportmy', 'feel', 'comfy', 'month']
33449	2022-03-29	I'd give it 6-Stars if I could.I've used it everyday for over a year. I'm sitting in it now as I write this review. It is still comfortable and performing as advertised. Best thing I have bought off of the Zon for quite a while. I'd give it 6-satrs if I could	['6-stars', 'could.i', 'use', 'everyday', 'year', 'sit', 'comfortable', 'perform', 'give', '6-satrs']
3499	2021-06-22	So far, okay So far this chair isn't exactly what I was expecting. I love the adjustable armrests, but for me the lumbar support is too low (even adjusted on the highest setting) and pretty hard / uncomfortable on my back. Going to order an extra lumbar support pillow and hopefully that'll make it more comfortable. Overall not bad, but not the best.	['adjustable', 'armrest', 'lumbar', 'support', 'hard', 'uncomfortable', 'extra', 'lumbar', 'support', 'comfortable', 'overall']
4324	2021-12-03	Great chair except for the armrests Ergonomically and seat-comfort wise I really like this chair. What I don't like is that the armrests are cupped/concave and not adjustable (they adjust vertically but the horizontal angle is fixed). If I'm sitting with my arms resting on the armrests perpendicular to my body it's ok, but when I bend my arms in to a typing position the armrests are uncomfortable. If the armrest were horizontally adjustable this chair would be nearly perfect.	['seat-comfort', 'wise', 'armrest', 'cupped/concave', 'adjustable', 'adjust', 'horizontal', 'angle', 'fix', 'sit', 'arm', 'rest', 'perpendicular', 'body', 'arm', 'type', 'position', 'uncomfortable', 'adjustable', 'perfect']
7757	2023-04-09	Love this chair! This chair is so comfortable! I love it. I was afraid it would stain easily and so far I've had no issues. I did buy rollerblade wheels for it as I feel the wheels on it didn't work great on the carpet in my office. I'm 5'4" and it's perfect for me.	['love', 'comfortable', 'love', 'afraid', 'rollerblade', 'wheel', 'wheel', 'work', 'carpet', 'office', 'perfect']
8362	2023-06-29	Outstanding Office Chair! Chair arrived quickly & was easy to assemble. Comfort is perfect for a tall person (6'4").Customer Service (Cayla) has been great in providing timely responses to product inquiries.Highly recommend!	['outstanding', 'office', 'arrive', 'easy', 'assemble', 'comfort', 'tall', 'person', 'service', 'cayla', 'response', 'inquiries.highly', 'recommend']

Table 3-6: review words of adjective and noun

3.3 Feature extraction

3.3.1 Gensim for Phrase Detection

Utilizing the Gensim library, a pivotal step in the text preprocessing pipeline involves the sophisticated detection and formation of bigrams from the dataset composed of nouns and adjectives extracted from ergonomic chair customer reviews.

```
docs= list(df['n_adj'])

phrases = gensim.models.Phrases(docs, min_count=10, threshold=20)

bigram_model = gensim.models.phrases.Phraser(phrases)
```

The code snippet employs Gensim's Phrases model, configured with specific parameters such as `min_count=10` and `threshold=20`, to identify and solidify meaningful bigram phrases within the corpus. The `min_count` parameter plays a critical role by stipulating that a word pair must appear together in the dataset at least ten times to be considered a valid phrase, effectively filtering out rare and potentially irrelevant co-occurrences. This ensures that the identified bigrams are not only statistically significant but also relevant and common within the context of the reviews. The `threshold` parameter further refines this process by setting a high bar for what constitutes a bigram, based on the strength of association between the word pair, thus favoring those combinations that occur more frequently together than by chance.

This Gensim's phrase detection results(Figure 3-7) in the extraction of bigrams that capture key concepts and attributes relevant to customer opinions and experiences, such as "customer_service" and "head_rest". By focusing on noun-adjective and noun-noun pairs, the approach enriches the analytical dataset with phrases that are descriptive and informative, laying a robust foundation for

deeper semantic analysis.

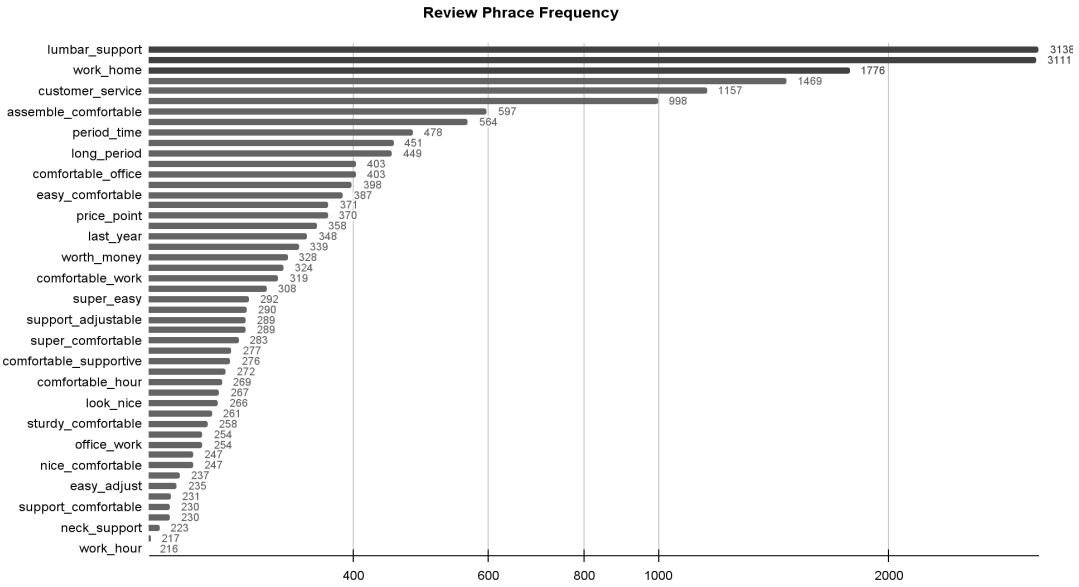


Figure 3-7 review phrase frequency by Gensim

3.3.2 Word2Vec Enhancement

Following the identification of bigrams, the research leverages Word2Vec, by treating each bigram as a single unit within the Word2Vec framework, the model is able to encapsulate the unique context in which these word pairs are used, thereby enriching the semantic relationships captured in the embeddings.

```
w2vmodel = Word2Vec(bigram_model[docs], sg=1, hs= 1, seed=33)

bigram_counter = Counter()
```

The code snippet initializes a Word2Vec model from the Gensim library, specifically designed to process bigrams extracted from a corpus of documents (`docs`). The model is configured with the following parameters: `sg=1` to use the skip-gram architecture, which is effective for handling infrequent words; `hs=1` to employ hierarchical softmax for training the model; and `seed=33` to ensure reproducibility of results.

This setup trains the Word2Vec model on the bigrams identified by the Gensim Phraser (`bigram_model`), which are expected to encapsulate more meaningful and contextually relevant phrases from the original text data.

The culmination of this process is evident in the final results, where bigrams such as "customer_service", "head_rest", and "long_period" emerge as key phrases, reflecting significant aspects discussed in customer reviews. These bigrams, now enriched with contextual and semantic depth through Word2Vec embeddings, offer valuable insights into customer sentiments and preferences, underpinning the research's analytical objectives. The frequency of these phrases, as indicated by the final counts, highlights their importance and prevalence in the dataset, affirming the effectiveness of the combined Gensim and Word2Vec methodology in extracting meaningful and relevant themes from the customer reviews. Upon training, the model is able to capture nuanced semantic relationships between words and phrases, resulting in a rich vector representation for each bigram.

This representation aids in various NLP tasks such as similarity assessment, topic modeling, and sentiment analysis, by providing a deeper understanding of the context and semantics encoded in the language used within the customer reviews.The "Review Phrase Frequency - Word2Vec" listed below (Figure 3-8)represents the most frequent bigrams identified in the dataset, along with their corresponding frequencies. These bigrams, such as "customer_service", "head_rest", and "long_period", highlight the key themes and topics discussed in the ergonomic chair reviews, offering valuable insights into customer preferences, experiences, and concerns.

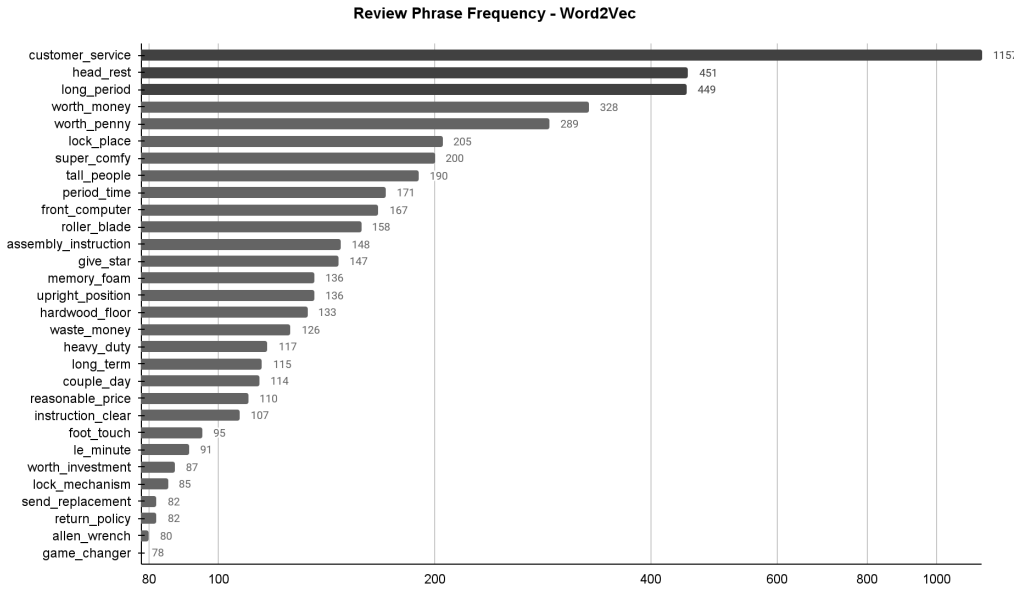


Figure 3-8 review phrase frequency by Word2Vec

3.3.3 NPFST

This thesis incorporates the Novel Phrase Finding and Segmentation Tool (NPFST) alongside Gensim for the extraction of bigrams from datasets comprising ergonomic chair reviews(Figure 3-9). This comparative analysis illuminates the methodological divergences and resultant disparities between NPFST and Gensim, substantiating the decision to prioritize NPFST for in-depth evaluation within the study. NPFST distinguishes itself through a direct approach to phrase detection, scanning for consecutive word sequences delineated by spaces, thereby preserving the phrases in their authentic lexical form. This straightforward methodology is particularly adept at identifying overt, textually manifest phrases, facilitating the capture of directly expressed terms without alteration. Conversely, Gensim employs a more complex statistical model that identifies significant word co-occurrences within a given range, capturing not only adjacent word pairs but also non-contiguous but relevant bigrams based on statistical significance. This process effectively consolidates identified phrases into singular tokens, streamlining the text for further analysis by condensing phrase representations.

The contrast in methodologies between NPFST and Gensim—where the former focuses on the preservation of original textual structures and the latter on a statistical representation of textual relationships—highlights the inherent trade-offs between straightforward lexical matching and nuanced statistical inference. The preference for NPFST in this research is underpinned by its ability to directly extract and maintain the integrity of explicit phrases from the text, offering a clear and unmodified reflection of the data's original linguistic patterns. This choice is validated through a methodical comparison, showcasing each tool's unique strengths and applications in processing complex text data.

While NPFST tends to extract several repeated topic from the same paragraph, (e.g: 'good imitation', 'good imitation of restless', 'good imitation of restless legs', 'good imitation of restless legs syndrome')after removing the repeating phrase, the results table showcases the efficiency of both methodologies in extracting meaningful phrases from ergonomic chair review data. The Graph 3-9 presents a visual analysis of phrase frequency within a given dataset, utilizing a coordinated system where the Y-axis represents the relative frequency of each phrase (calculated

as the frequency of each phrase divided by the total number of phrases), and the X-axis indicates the absolute frequency of each phrase. The graph distinguishes between topics using both color and node size as indicators of frequency significance. Here, cooler colors (such as blue) and larger node sizes signify phrases with higher scores, indicating a greater frequency of occurrence within the dataset. Conversely, warmer colors (such as red) and smaller node sizes are used to denote phrases with lower scores, reflecting a less frequent appearance. This visual methodology allows for an immediate, intuitive understanding of the distribution and significance of phrases across different topics, highlighting the most prevalent themes by their prominence in terms of color and scale.

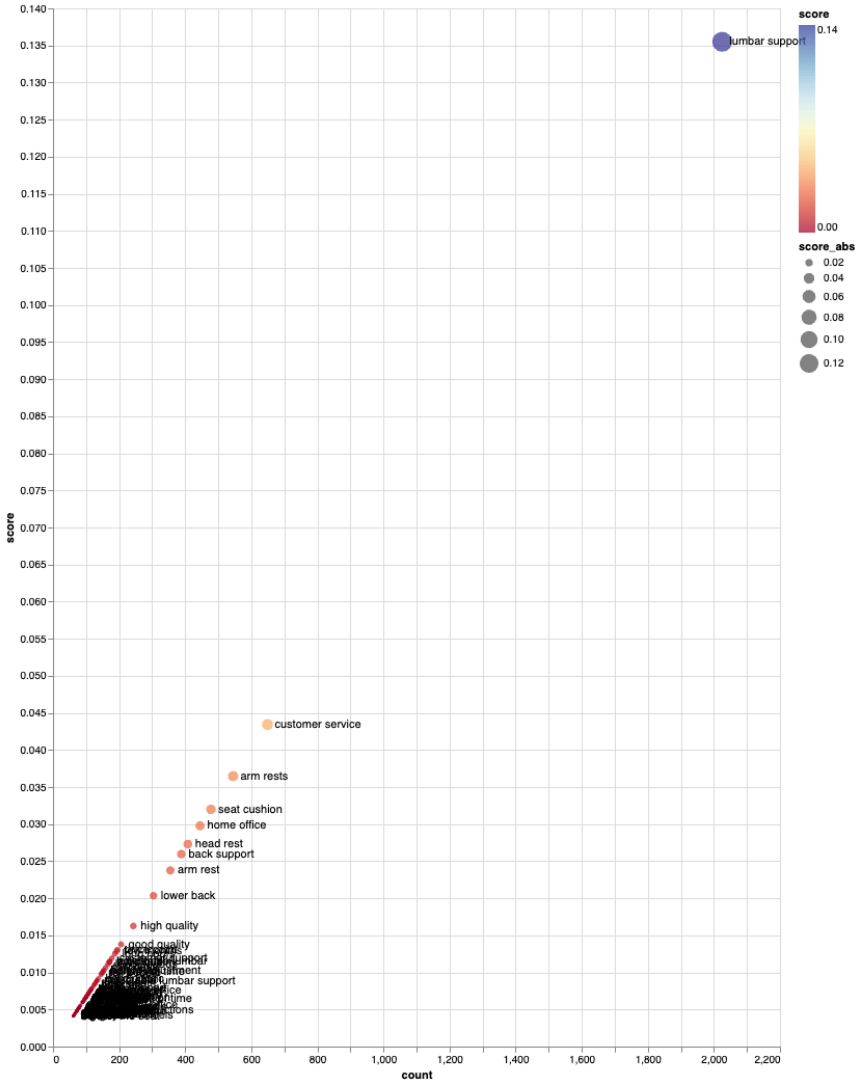


Figure 3-9 review phrase frequency by NPFST(y-axis: score = phrase frequency / total phrase counts; x-axis: count: phrase frequency)

The outputs differ not only in format but also in the types of phrases emphasized by each technique. For instance, NPFST highlights terms like "lumbar support" and "office chair" with their frequencies preserved in their natural, unaltered form. Conversely, Gensim, through its merged token format, identifies statistically significant bigrams such as "lumbar_support" and "easy_assemble," which may indicate deeper contextual associations beyond mere adjacency.

The decision to utilize NPFSTresults for further analysis in this thesis hinges on it’s clarity and Intuitiveness, the output format of NPFST, which maintains the original phrase boundaries, this is particularly beneficial for analyses where the explicit expression and surface-level readability of phrases are paramount. Given the focus of this thesis on exploring explicit expressions and user sentiments within ergonomic chair reviews, the direct and unmodified representation of bigrams by NPFST aligns more closely with the analytical goals.

CHAPTER IV : MODEL SETUP

4.1 Embedding Models and parameter selection

In the course of the study, the researcher implemented several modifications and customized some embedding models to the default settings of the BERTopic model, enhancing its efficacy and ensuring the originality of the output.

```
topic_model_5 = BERTopic(
 embedding_model=SentenceTransformer("sentence-transformers/all-mpnet-base-v2")
 umap_model=UMAP(n_neighbors=15,
                 n_components=5,
                 min_dist=0.0,
                 metric='cosine',
                 random_state=42)
 hdbscan_model=HDBSCAN(min_cluster_size=50,
                 metric='euclidean',
                 cluster_selection_method='eom',
                 prediction_data=True)
 representation_model=MaximalMarginalRelevance(diversity=0.3)
 top_n_words=50,
 verbose=True,
 nr_topics=50
)
embeddings = sentence_model.encode(text, show_progress_bar=True)
```

4.1.2 Embeded Douments

This study used a hugging face open source sentence transformer pretrained model, particularly the "all-mpnet-base-v2" version, which represents a significant advancement in generating sentence embeddings. It maps sentences & paragraphs

to a 768 dimensional dense vector space and can be used for tasks like clustering or semantic search, making it an ideal selection for sentence-transformation tasks especially for customer review data within the BERTopic framework.

When the method sentence_model.encode(*text, show_progress_bar=True*) is invoked, each piece of text from the dataset is processed through this model. The 'text' strings here used NPFST noun phrases. NPFST focuses on the most relevant aspects of the reviews for topic analysis, and reduces the noise in the data by omitting less informative parts of the text, such as filler words or off-topic comments. Furthermore, it improves the computational efficiency of analysis, especially with large datasets with over 35,000 data points, Extracting phrases before encoding reduces the volume of text processed by the sentence transformer, aiding in scalability.

4.1.3 Reduce dimensionality and clustering embeddings

The tailored selection of UMAP and HDBSCAN parameters for analyzing ergonomics chair reviews demonstrates a strategic approach to capturing both detailed and broad trends in customer feedback. By balancing local and global data structures through UMAP's `n_neighbors` and emphasizing thematic similarities with cosine similarity as a metric. This precision, combined with the dimensionality reduction to five dimensions (`n_components=5`), simplifies the data while preserving critical information about customer satisfaction areas. Five dimensions were finalized through suggestion from source code and also experimentation. The minimal distance setting (`min_dist=0.0`) further sharpens the focus on subtle differences in feedback, enhancing the granularity of the analysis.

On the clustering side, HDBSCAN's parameters are chosen to complement UMAP's data preparation by focusing on significant patterns through `min_cluster_size` and `metric='euclidean'`, ensuring that the clustering is both meaningful and interpretable. Together, these settings forge a robust framework for identifying key insights from customer reviews, guiding product enhancements and strategic decisions with a comprehensive understanding of consumer preferences and sentiments, interpretable solutions tailored to the complexities of customer feedback analysis.

4.1.5 Representation Model

The `diversity` parameter in MMR controls the balance between relevance and diversity. A value of 0 would make the selection purely based on relevance, ignoring diversity, leading to a potentially narrow set of topics or documents closely resembling each other. Conversely, a higher value places more emphasis on diversity.The study set a diversity value of 0.3, the goal is to ensure the topics or summaries generated cover different aspects of the ergonomics chair reviews, from comfort and design to durability and customer service, providing a holistic view of customer sentiment.

4.1.6 Model Fitting

```
topics, probs = topic_model_5.fit_transform(bigram, embeddings) # fit of topic model

topic_df = topic_model_5.get_document_info(bigram)

topic_df.Topic.value_counts()
```

After instantiating the BERTopic model, it was applied to the preprocessed text data stored in the "bigram" column of the data frame which is the pre-processed bigram data by Word2Vec extracted from review data. Upon configuration, the fit_transform method was invoked on the pre-processed bigram tokens and their corresponding embeddings. This method not only trained ‘topic_model_5’ but also transformed the data to yield an array of topic identifiers and their probabilities, thus quantifying the strength of association between documents and their inferred topics. The succeeding command, topic_model_5.get_document_info(bigram), retrieved detailed information about each document's topic allocation, encapsulated within topic_df. An examination of the frequency of topics, through topic_df.Topic.value_counts(), provided insights into the distribution and prevalence of topics across the corpus.

The topic model's efficacy was gauged by the output of get_topic_info() method. This function furnished a comprehensive view of the topics, including their size and coherence. The coherence score, in particular, evaluated the semantic similarity between the high-scoring words within each topic. Choosing to use bigram data by Word2Vec as input for the BERTopic model, instead of other phrase models generated by Gensim or the original unprocessed review data, ensures analysis

capture more contextual information than single words, providing insights into specific features or aspects of the chairs that customers discuss (e.g., "lumbar support," "adjustable height")

4.2 Result and Discussion of Trained Model

4.2.1 Topic Representation

This detailed analysis of the BERTopic model's application to the dataset not only underscored the model's capability to unveil thematic structures within the text but also demonstrated the intricate interplay between the choice of hyperparameters and the quality of the resulting topics. The top 12 topics, as delineated in Table 4-1, presented an empirical foundation for subsequent qualitative analyses, paving the way for a deeper understanding of the corpus's thematic composition.

Topic	Name	Representation	Representative_Docs	Top_n_words
-1	-1_support_comfortable_hour_lumbar	['support', 'comfortable', 'hour', 'lumbar', '...	["['uncomfortable', 'lumbar', 'support', 'hop'...	support - comfortable - hour - lumbar - desk -...
0	0_comfortable_assemble_sturdy_stylish	['comfortable', 'assemble', 'sturdy', 'stylish...	["['easy', 'assemble', 'comfortable']", "['lov...	comfortable - assemble - sturdy - stylish - ad...
1	1_lumbar_support_sit_comfortable	['lumbar', 'support', 'sit', 'comfortable', 'b...	["['lumbar', 'support', 'lumbar', 'support', '...	lumbar - support - sit - comfortable - back - ...
2	2_support_lumbar_adjustable_mesh	['support', 'lumbar', 'adjustable', 'mesh', 'c...	["['embody', 'investment', 'pro', 'disappear',...	support - lumbar - adjustable - mesh - cushion...
3	3_height_size_comfortable_petite	['height', 'size', 'comfortable', 'petite', 'l...	["['comfortable', 'tall', 'person', 'love']", ...	height - size - comfortable - petite - leg - f...
4	4_wheel_caster_carpet_roller	['wheel', 'caster', 'carpet', 'roller', 'hardw...	["['price', 'assemble', 'perfect', 'short', 'f...	wheel - caster - carpet - roller - hardwood_fl...
5	5_home_comfortable_assemble_quality	['home', 'comfortable', 'assemble', 'quality',...	["['home', 'office']", "['home', 'office']", "...	home - comfortable - assemble - quality - stur...
6	6_customer_service_customer_service_company	['customer_service', 'customer', 'service', 'c...	["['customer_service', 'easy', 'assemble']", "...	customer_service - customer - service - compan...
7	7_quality_material_value_sturdy	['quality', 'material', 'value', 'sturdy', 'co...	["['quality', 'price']", "['quality', 'price']...	quality - material - value - sturdy - construc...
8	8_adjustable_comfortable_armrest_assemble	['adjustable', 'comfortable', 'armrest', 'asse...	["['rest']", "['comfortable', 'love', 'rest', ...	adjustable - comfortable - armrest - assemble ...
9	9_cushion_pillow_head_rest_support	['cushion', 'pillow', 'head_rest', 'support', ...	["['cushion']", "['cushion']", "['cushion', 'c...	cushion - pillow - head_rest - support - comfo...
10	10_color_comfortable_blue_color_bright	['color', 'comfortable', 'blue_color', 'bright...	["['nice', 'color', 'comfortable', 'happy']", ...	color - comfortable - blue_color - bright - de...

Table 4-1 Topic Representation

The table elaborates on the intricate topology of topics harvested by the BERTopic algorithm, showcasing a meticulous array of keywords that encapsulate the essence and thematic structure of each topic. The biggest group is Topic -1 , which corresponds to outliers, topic 0, captured under the moniker 0_comfortable_assemble_sturdy_stylish, is articulated through keywords such as 'comfortable', 'assemble', 'sturdy', and 'stylish'. These terms collectively underscore a dialogue revolving around the ease of assembly, robustness, and aesthetic appeal of furniture, implicating a significant focus on consumer satisfaction and the practical aesthetics of home or office decor. It encompasses a melange of otherwise distinct categories. This discourse extends an invitation to explore the intersectionality of functionality, style, and user experience in product design and utility.

The representation of Topic 1, identified as 1_lumbar_support_sit_comfortable, through keywords like 'lumbar', 'support', 'sit', 'comfortable', magnifies the concentration on ergonomic design and its pivotal role in enhancing sitting comfort and spinal health. This narrative seamlessly integrates with the broader conversation on workplace wellness and the imperative of incorporating ergonomic principles in seating solutions to mitigate discomfort and foster wellbeing.

Further dissecting the thematic fabric, Topic 6, labeled as 6_customer_service_customer_service_company, with keywords like 'customer_service', 'customer', 'service', 'company', delves into the realm of customer service excellence and the corporate ethos of customer-centricity. This exploration illuminates the crucial aspect of customer interactions and the pivotal role of service quality in shaping customer perceptions and loyalty, thereby weaving a narrative that underscores the symbiotic relationship between service excellence and business success.

The BERTopic model's adeptness at distilling these topics into cohesive and interpretable themes offers a panoramic view of the dataset's thematic landscape, enabling an agile and nuanced exploration of the data. By demystifying the thematic essence through a distilled lexicon of keywords, the model empowers researchers and analysts to navigate the complexities of the dataset with enhanced clarity and insight. This methodical unraveling of topics facilitates a granular understanding of the data, fostering a rich dialogue around the themes and enabling stakeholders to engage with the data in a more informed and meaningful manner.

4.2.2 Topic Word Scores

In the comprehensive analysis of the dataset, the BERTopic algorithm was employed to discern and quantify the underlying themes present in the text corpus.

The results, as depicted in Figure 4-2, elucidate the top 20 topics' word scores, emphasizing the highest scoring terms within each identified topic, thereby providing an insightful measure of term relevance. For instance, Topic 2, marked by pivotal terms such as 'lumbar', 'support', 'mesh', and 'adjustable', achieved a notable topic word score, underpinning the terms' significance and their strong correlation to the topic.

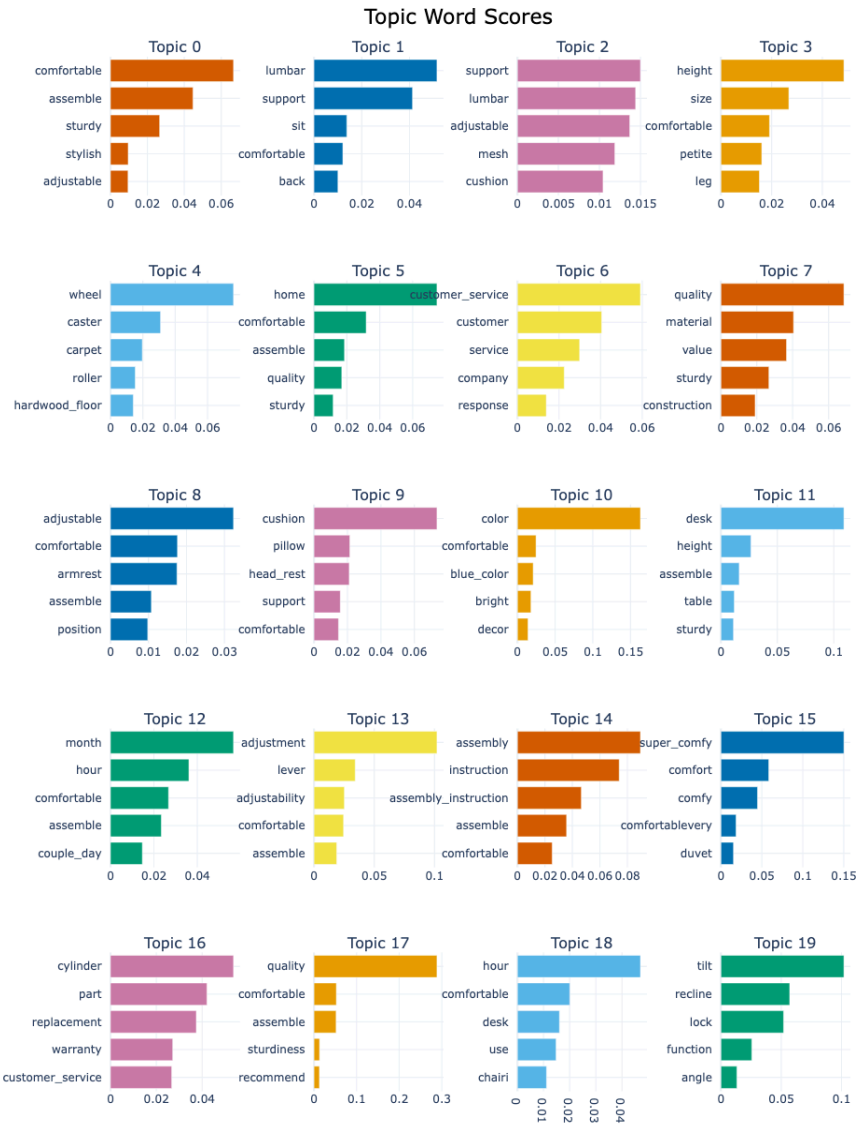


Figure 4-2 Topic Word Score

Topic 1, which encapsulates the critical components of ergonomic support, as indicated by keywords like 'lumbar', 'support', and 'comfortable'. Topic 4, with a focus on mobility, is exemplified by terms such as ‘wheel’, 'caster', and 'carpet', signaling an emphasis on comfort in the sense of portability of the user in a defined work space. Subsequent topics, such as Topic 6's 'customer_service', reflect on the integral aspects of consumer relations and service quality.

The descending order of word scores within each topic unveils the relative weight and significance of individual terms, providing a gradation of relevance that further clarifies the thematic structure. Terms with lower scores, while still relevant,

may not define the topic's core as strongly. A summary of the topics alongside their top n words offers a quick reference and a thematic synopsis. Each entry in the table provides a distilled representation of the topic, an array of representative documents, and a selection of key terms, thereby encapsulating the thematic essence succinctly.

4.2.3 Hierarchical Structures

The hierarchical clustering dendrogram, as depicted in the accompanying figure(Figure 4-3), serves as a testament to the meticulous stratification of the dataset's thematic elements. This visual representation elucidates the hierarchical relationships and proximities between the topics extracted by the BERTopic algorithm. Each branch of the dendrogram represents a cluster of terms that coalesce around a central theme, signifying a confluence of conceptual similarity.

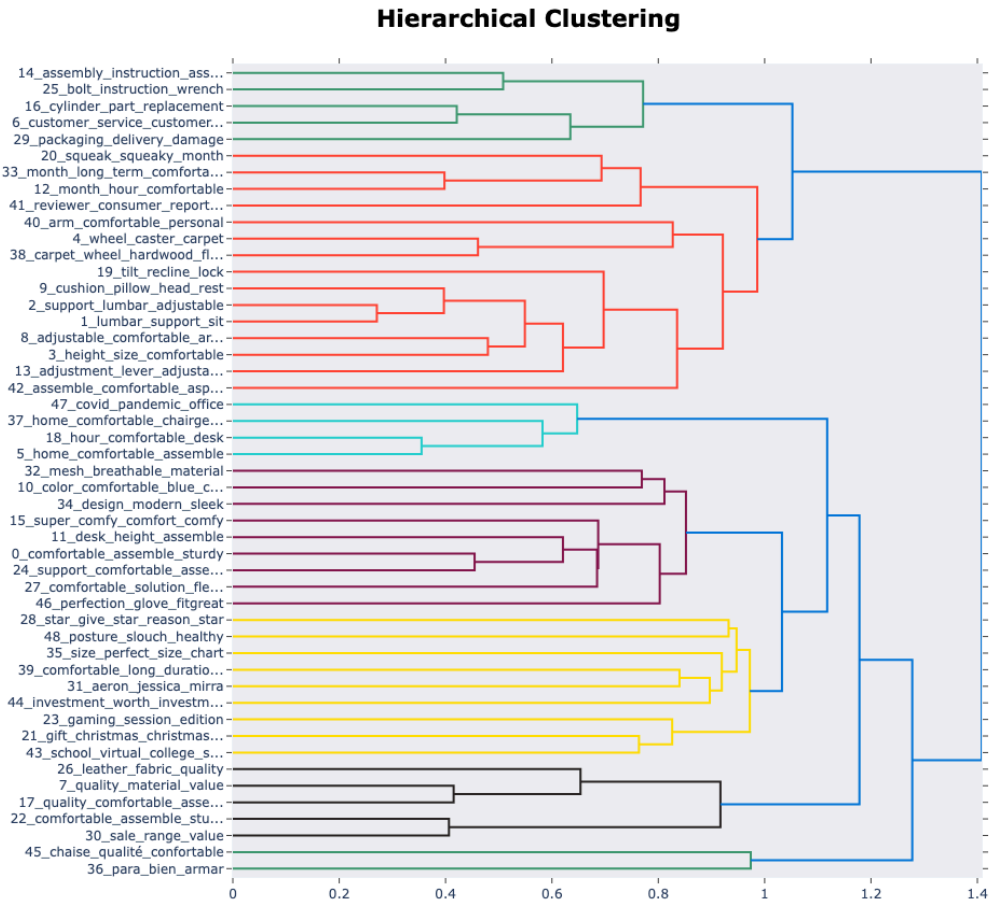


Figure 4-3 Topic Representation

The clustering process begins with each topic as a standalone entity and progressively merges them based on their similarity, which is commonly measured by distance metrics. The vertical lines represent the topics, while the horizontal lines link the topics, indicating the point at which they are considered sufficiently similar to be joined into a single cluster. The length of the horizontal lines, or dendrites, correlates with the distance or dissimilarity between clusters: shorter lines imply greater similarity. The clusters colored in blue and red suggest a strong thematic linkage, possibly around core aspects such as 'comfortable', 'assemble', and 'quality'. These clusters might represent an overarching theme of ergonomic or quality-centric product design. The dendrogram allows for the identification of such broad themes and provides a collective understanding of the dataset's granular thematic structure.

The color-coding within the dendrogram provides additional layers of interpretation. Each color may represent a different method of linkage used in the clustering process, such as single, complete, or average linkage. These methods vary based on the strategy of defining the distance between clusters, which could be based on the closest points, farthest points, or average distances, respectively. Clusters with a yellow undertone, conversely, might symbolize a more diverse but interrelated set of topics, perhaps converging around user experience, functionality, and aesthetic design, as suggested by terms like 'mesh', 'breathable', 'comfortable', and 'blue_color'. Such diversity within proximity can signal intersecting sub-themes that warrant further exploration.

While the hierarchical clustering dendrogram provides a detailed overview of topic relationships, its utility is not without limitations. One prominent issue is the potential for misinterpretation of groupings, as some may not intuitively align with the underlying thematic content. The clustering algorithm relies heavily on distance metrics, which can sometimes lead to topics being grouped based on numerical proximity rather than conceptual coherence. This is especially true when the algorithm encounters high-dimensional data or abstract themes that are not easily quantifiable.

For instance, within the dendrogram, clusters might encompass a wide array of terms with a loose conceptual connection, such as those involving 'comfortable', 'mesh', and 'blue_color'. These could potentially be conflated under a broad category

of product design, but may in fact represent distinct sub-themes like material properties and aesthetic preferences, which are not necessarily related. Furthermore, the 'one-size-fits-all' nature of linkage methods could force disparate topics into a single cluster, especially when considering the most dissimilar pair (complete linkage) or when averaging distances (average linkage), potentially obscuring meaningful distinctions.

Additionally, the dendrogram's two-dimensional representation of potentially multi-dimensional relationships can oversimplify the data, leading to a loss of nuance and obscuring complex inter-topic relationships. Lastly, the algorithm's sensitivity to outliers can result in single-topic clusters that appear to be significantly disconnected from the main body of the dendrogram. These outliers such as topic 45 and 36 may represent unique or rare topics that do not easily merge with others, which could be informative or misleading, depending on the context of the analysis.

In summary, while hierarchical clustering dendrograms serve as powerful tools for visualizing topic relationships, they require careful interpretation and validation to ensure that the clusters represent meaningful and coherent themes. Researchers must remain critical of the algorithm's outputs, complementing them with domain knowledge and additional qualitative analysis to fully leverage the insights provided by the data.

4.2.4 Intertopic Distance Map

Similar to LDA and the accompanying LDAvis library, the Bertopic library offers an interactive interface displaying each topic along with its associated words and scores(Figure 4-4). Specifically, it provides a spatial representation of topics within a two-dimensional plane, where the distance between points reflects the dissimilarity between topics. Topics that are closely positioned or overlapped indicate thematic overlap, while those further apart suggest distinct subject areas. By zooming in and out the map, researchers can take a closer look into each topic's top words and size. This map is particularly useful for identifying isolated topics or those that form dense clusters, indicating areas of concentrated discourse or potential interdisciplinary intersections.

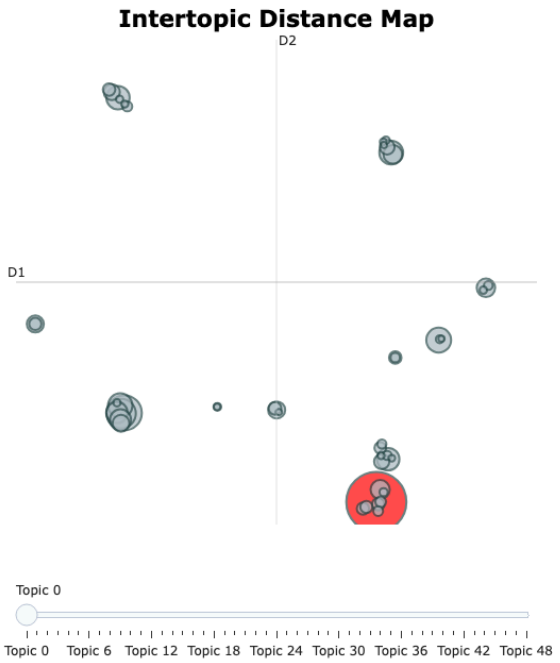


Figure 4-4 Intertopic Distance Map

Hierarchical and interatomic distance maps collectively enable a multifaceted analysis of the dataset, offering insights into both the macro-structure and the micro-structure of the thematic landscape, these charts serve as a comprehensive tool for academic research.

4.3 Topic Label Construction

The construction of topic labels is a critical feature engineering step that plays a pivotal role in the interpretability and usability of machine learning models, particularly in natural language processing (NLP) tasks. In the context of the BERTopic model, which is designed to uncover latent thematic structures within a corpus, the assignment of descriptive labels to each topic is a fundamental post-processing step that transforms abstract topic clusters into interpretable semantic categories. For this study, the topic labels were meticulously crafted to encapsulate the essence of each topic cluster. The table below presents a subset of these labels:

index	Topic Name	Topic Label
0	-1_support_comfortable_hour_lumbar	Lumbar Support and Comfort Over Hours
1	6_customer_service_customer_service_company	Customer Service and Company Interaction
2	2_support_lumbar_adjustable_mesh	Adjustable Lumbar Support and Mesh Design
3	16_cylinder_part_replacement_warranty	Cylinder Parts and Replacement Warranty
4	11_desk_height_assemble_table	Assembly Ease and Desk Height Compatibility
5	1_lumbar_support_sit_comfortable	Lumbar Support Comfort
6	22_comfortable_assemble_sturdy_ajustments	Comfort, Sturdiness, and Adjustment Ease
7	9_cushion_pillow_head_rest_support	Cushioning and Headrest Support
8	3_height_size_comfortable_petite	Comfort for Petite Sizes and Height Adjustment
9	32_mesh_breathable_material_fabric	Breathable Mesh Material
10	30_sale_range_value_dollar	Sale Range and Value for Money
11	25_bolt_instruction_wrench_assembly	Assembly Instructions and Tools
12	4_wheel_caster_carpet_roller	Wheel and Caster Performance on Carpets
13	26_leather_fabric_quality_color	Leather Quality, Fabric, and Color Options
14	13_adjustment_lever_adjustability_comfortable	Lever Adjustments and Comfort
15	42_assemble_comfortable_aspect_quality	Assembly, Comfort, and Quality Aspects
16	5_home_comfortable_assemble_quality	Home Chair Comfort and Quality
17	23_gaming_session_edition_comfortable	Gaming Edition Comfort
18	18_hour_comfortable_desk_use	Comfort for Long Desk Hours
19	17_quality_comfortable_assemble_sturdiness	Quality, Comfort, Assembly, and Sturdiness
20	7_quality_material_value_sturdy	Material Quality and Sturdiness
21	12_month_hour_comfortable_assemble	Monthly Comfort and Assembly
22	31_aeron_jessica_mirra_redesign	Aeron and Mirra Chair Redesigns
23	29_packaging_delivery_damage_assemble	Packaging, Delivery, and Damage Issues
24	24_support_comfortable_assemble_recommend	Support, Comfort, and Recommendations
25	15_super_comfy_comfort_comfy_comfortablevery	Superior Comfort and Comfiness
26	28_star_give_star_reason_star_love	Star Ratings and Reasons for Love
27	8_adjustable_comfortable_armrest_assemble	Adjustable Armrests and Comfort
28	0_comfortable_assemble_sturdy_stylish	Comfort, Style, and Sturdiness
29	14_assembly_instruction_assembly_instruction_assemble	Assembly Instructions and Process
30	19_tilt_recline_lock_function	Tilt, Recline, and Lock Features
31	20_squeak_squeaky_month_sound	Squeakiness and Noise Issues
32	40_arm_comfortable_personal_wheel	Arm Comfort and Personalization
33	35_size_perfect_size_chart_comfortable	Size Fit and Comfort
34	41_reviewer_consumer_report_read_assemble	Consumer Reviews and Reports
35	44_investment_worth_investment_investmentthe_notice	Investment Worth and Noticeability
36	48_posture_slouch_healthy_comfortable	Posture Support and Health
37	10_color_comfortable_blue_color_bright	Color Options and Comfort
38	21_gift_christmas_christmas_gift_christmas_present	Ideal as Christmas Gifts
39	46_perfection_glove_fitgreat_comfortable	Perfect Fit and Comfort
40	38_carpet_wheel_hardwood_floor_roller	Performance on Carpet and Hardwood
41	27_comfortable_solution_flexible_work	Comfortable Solutions for Flexible Work
42	34_design_modern_sleek_superior	Modern and Sleek Design
43	39_comfortable_long_duration_stretch_composer	Long Duration Comfort and Stretch
44	33_month_long_term_comfortable_daily	Long-term Daily Comfort
45	36_para_bien_armar_excelente	Excellent Assembly Experience
46	47_covid_pandemic_office_coronavirus	COVID Pandemic and Office Adaptations
47	43_school_virtual_college_student_online_class	Suitability for School and Online Classes
48	37_home_comfortable_chairgetting_pricework	Home Comfort and Price-to-Work Ratio
49	45_chaise_qualité_confortable_skin	Quality and Skin Comfort

Table 4-7 Topic Label Construction

Each label conveys the central theme of the topic it represents, transforming the numerical topic identifier into a more tangible and comprehensible concept. For instance, the label "Lumbar Support and Comfort Over Hours" succinctly captures the topic's focus on ergonomic design and long-term comfort, while "Customer Service and Company Interaction" reflects a topic centered around the dynamics of customer engagement and company responses. Other topics capture brand recognition such as Topic 31 with key words in “Mirra” and “Aeron”. The creation of these labels involved analyzing the most representative terms within each topic, considering the context of their occurrence, and synthesizing a phrase that accurately reflects the topic's content. This process enhances the communicative efficacy of the model's outputs, allowing stakeholders to engage with the findings intuitively.

In concluding this analysis, the examination of hierarchical diagrams and intertopic distance has led to the generation of the final topics presented in a manner that resonates with human understanding (Table 4-5).

Topics From Aggregate Data	
Adjustable Lumbar Support and Mesh Design	Aeron and Mirra Chair Redesigns
Arm Comfort and Personalization	Assembly and Instructions
Assembly, Comfort, and Quality Aspects	Color Options and Comfort
Comfort and Adjustability	Comfort for Petite Sizes and Height Adjustment
Comfortable Solutions for Flexible Work	Cushioning and Headrest Support
Customer Service and Experience	Cylinder Parts and Replacement Warranty
Home Chair Comfort and Quality	Long Duration Comfort and Stretch
Long-term Daily Comfort	Lumbar Support and Comfort
Material and Design Quality	Monthly Comfort and Assembly
Size Fit and Comfort	Squeakiness and Noise Issues
Superior Comfort and Comfiness	Tilt, Recline, and Lock Features
Use Case Specific Comfort	Wheel and Surface Compatibility

Table 4-5 Organized Final Topics from BERTopic

In the realm of office ergonomics and consumer experience, each topic label has been meticulously crafted to convey the essence of the discussion it encapsulates. For instance, "Adjustable Lumbar Support and Mesh Design" evolves beyond a mere identifier, painting a vivid picture of ergonomic innovation and

breathable material application for enhanced comfort during extended use, the label "Customer Service and Experience" delves into the realm of consumer interactions, shedding light on the pivotal role of customer engagement and the responsiveness of corporate support structures. The prominence of brand-centric discussions is evident in topics with keywords like "Aeron and Mirra Chair Redesigns," where the confluence of legacy and innovation is explored.

The labeling process is a thoughtful curation of terms that are most emblematic of their respective topics, ensuring that the semantic core of each discussion is distilled into a succinct, relatable phrase. This not only aids in the communicative clarity of the model's outputs but also fosters an intuitive connection with stakeholders, enhancing the accessibility of complex data. With the topic labels set, a deeper narrative unfolds. The "Functional Features and Adjustability" cluster underscores the critical intersection of user adaptability and comfort, illustrating how functional design is pivotal to user satisfaction. This narrative extends to the discourse on product usage, where the synergy between flexible work solutions, individual customization, and the pursuit of unparalleled comfort is examined.

Furthermore, "Product Maintenance and Longevity" comes into sharp focus, emphasizing durable product design as evidenced by concerns over noise issues and part robustness, thus highlighting the importance of longevity in product design. And "Market Value and Purchasing Factors" distill the quintessence of consumer decision-making, merging subjective comfort with objective economic considerations, thereby guiding purchasing decisions.

"Cultural and Seasonal Considerations" reveal the thematic significance of social currents, as seen in the nuances of seasonal gifting and pandemic-driven product adaptations, underscoring the socio-cultural adaptability of product relevance.

Lastly, the "Visual Design and Preferences" theme emerges, illustrating consumer aesthetic preferences and the spirit of contemporary design, which collectively inform the visual and tactile narratives of product offerings.

CHAPTER V: SUPPLEMENTARY ANALYSIS

5.1 Word/Phrase Analysis

After preprocessing the textual data in the ergonomics chair reviews, the study employed Python's word cloud library' to visualize the segmentation of the review text. Words that appeared more frequently in the text comments would create a striking visual impact on the word cloud. In the graph, the 100 most frequently addressed words were selected for word cloud generation.

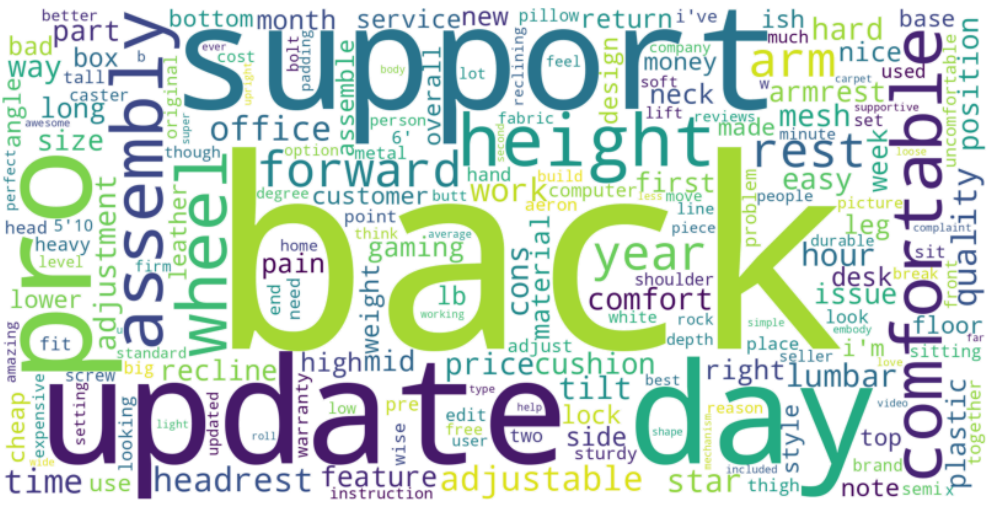


Figure 5-1: word cloud diagram

Crucial findings from the analysis of keywords suggest a pronounced preference for ergonomic features in products, as evidenced by the frequent references to "back," "support," "lumbar," and "comfortable." Such terms underscore a significant consumer demand for products designed to provide substantial support and comfort, especially for individuals suffering from back pain or those aiming to avert it. The focus on "lumbar support" and "comfortable" seating underscores the importance of furniture that accommodates prolonged use without causing discomfort.

Additionally, the analysis highlights the value placed on adjustability and material quality. Keywords like "adjustable," "quality," "mesh," and "armrests" indicate a consumer preference for customizable furniture that caters to individual body shapes and preferences. The specific mention of "mesh" materials points to a preference for breathable fabrics, enhancing comfort over extended periods. The

inference of high-quality materials and construction speaks to a demand for products that are durable and provide consistent support.

Moreover, the emphasis on terms such as "price," "worth," "easy to assemble," and "durable" reveals a consumer trend towards seeking products that deliver excellent value for money. The focus on ease of assembly and durability indicates a preference for practical, durable solutions in office furniture, highlighting an interest in making long-term, cost-effective investments.

To further explore the customer preference, the study compared the key phrases extracted from techniques of NPFST, Gensim, and Word2Vec. The analysis leverages phrase detection to capture not just individual words but also the context in which they are used, enhancing the depth and accuracy of the insights gained.The table below (Table 5-2) displays the results of themes by descending frequencies:

1	NPFST	Frequency	Gensim	Frequency	Word2Vec	Frequency
2	lumbar support	2025	lumbar_support	3138	customer_service	1157
3	office chair	1591	easy_assemble	3111	head_rest	451
4	great chair	903	work_home	1776	long_period	449
5	customer service	649	home_office	1469	worth_money	328
6	herman miller	546	customer_service	1157	worth_penny	289
7	arm rests	545	comfortable_easy	998	lock_place	205
8	comfortable chair	511	assemble_comfortable	597	super_comfy	200
9	seat cushion	478	high_quality	564	tall_people	190
10	good chair	460	period_time	478	period_time	171
11	home office	445	head_rest	451	front_computer	167
12	head rest	408	long_period	449	roller_blade	158
13	office chairs	390	worth_price	403	assembly_instruction	148
14	back support	388	comfortable_office	403	give_star	147
15	arm rest	355	long_time	398	memory_foam	136
16	lower back	304	easy_comfortable	387	upright_position	136
17	back of the chair	286	comfortable_support	371	hardwood_floor	133
18	new chair	280	price_point	370	waste_money	126
19	best chair	262	comfortable_adjustable	358	heavy_duty	117
20	nice chair	257	last_year	348	long_term	115
21	desk chair	254	office_comfortable	339	couple_day	114
22	high quality	243	worth_money	328	reasonable_price	110
23	ergonomic chair	241	comfortable_look	324	instruction_clear	107
24	other chairs	236	comfortable_work	319	foot_touch	95
25	old chair	225	comfortable_love	308	le_minute	91
26	good quality	206	super_easy	292	worth_investment	87
27	price point	195	adjustable_lumbar	290	lock_mechanism	85
28	few months	192	support_adjustable	289	send_replacement	82
29	long hours	187	worth_penny	289	return_policy	82
30	quality chair	186	super_comfortable	283	allen_wrench	80
31	customer support	178	comfortable_sturdy	277	game_changer	78

Table 5-2 : phrase extraction comparison

From the table, it’s clear that high mentions of phrases like "lumbar support" (NPFST) and "lumbar_support" (Gensim) underscore the critical importance of ergonomic features in customer satisfaction, aligning with the prominent mentions of "arm rests" and "seat cushion" in NPFST results. Gensim's detection of "easy_assemble" and "assemble_comfortable" as highly frequent phrases suggests that customers highly value straightforward assembly processes. Highlighting easy assembly in product descriptions and marketing materials could address this consumer priority effectively.

The prominence of "customer service" in Word2Vec results, contrasted with its lower frequency in NPFST, indicates a nuanced context around service experiences that Word2Vec captures effectively. Marketing communications should emphasize the brand's commitment to superior customer service, potentially turning a common point of customer anxiety into a strong selling proposition.

The mention of "high quality" across methodologies points to the importance of quality assurance in customer decision-making. Marketing efforts should focus on the durability and quality of materials, supported by warranties or guarantees, to reassure potential buyers.Analysis reveals specific product features and concerns, such as "head rest," "long_period" of use, and "worth_money." Creating content that addresses these topics, from blog posts to buying guides, can improve search engine optimization (SEO) and engagement by aligning with customer interests.

Phrases with high frequencies across models, such as "ergonomic chair," "office chair," and nuanced terms like "super_comfy" and "tall_people," should be incorporated into website content, product descriptions, and blog posts to improve search engine visibility and attract organic traffic. Phrases like "great chair" and "comfortable chair" reflect positive customer experiences. Marketing materials can incorporate such testimonials to provide social proof, enhancing trust and confidence among prospective customers.

The analysis also surfaces areas for improvement, such as "assembly_instruction" and "customer_service." Addressing these concerns publicly and showing commitment to enhancement can improve brand perception.The mention of specific brands, such as "herman miller," indicates comparative shopping behavior. Marketing strategies should consider direct comparisons or highlight unique

selling propositions that differentiate the brand from competitors.The data suggests varying priorities, such as comfort ("super_comfy"), assembly ease ("easy_assemble"), and suitability for long hours of use ("long_period"). Marketing campaigns could be tailored to address these distinct customer segments, offering personalized recommendations based on specific needs or concerns.

In conclusion, the analytical methods applied to the customer review data not only reveal the multifaceted aspects of consumer feedback but also provide actionable insights for refining marketing strategies. By focusing on product features, customer service, and quality assurance, and leveraging these insights for targeted content marketing, brands can enhance their competitive edge and foster stronger customer relationships.

5.1 Decoding Customer Satisfaction

Figure 5-3 offers a bar chart representation of the average review rates by topic, quantifying customer satisfaction or the perceived quality of different product features. Each bar represents a topic, and its height corresponds to the average review score, which is presumably on a 1 to 5 scale, with 5 being the highest level of satisfaction.

This chart provides a clear, at-a-glance summary of which topics are associated with higher or lower customer satisfaction. Topics with taller bars, such as “Home Chair Comfort and Quality ”, “Superior Comfort and Comfiness”, “Use Case Specific Comfort" suggest areas where customers feel their expectations are being met or exceeded. Conversely, shorter bars, potentially for topics like “Cylinder Parts and Replacement Warranty”, “Long-term Daily Comfort", "Squeakiness and Noise Issues," may highlight areas in need of attention from manufacturers.

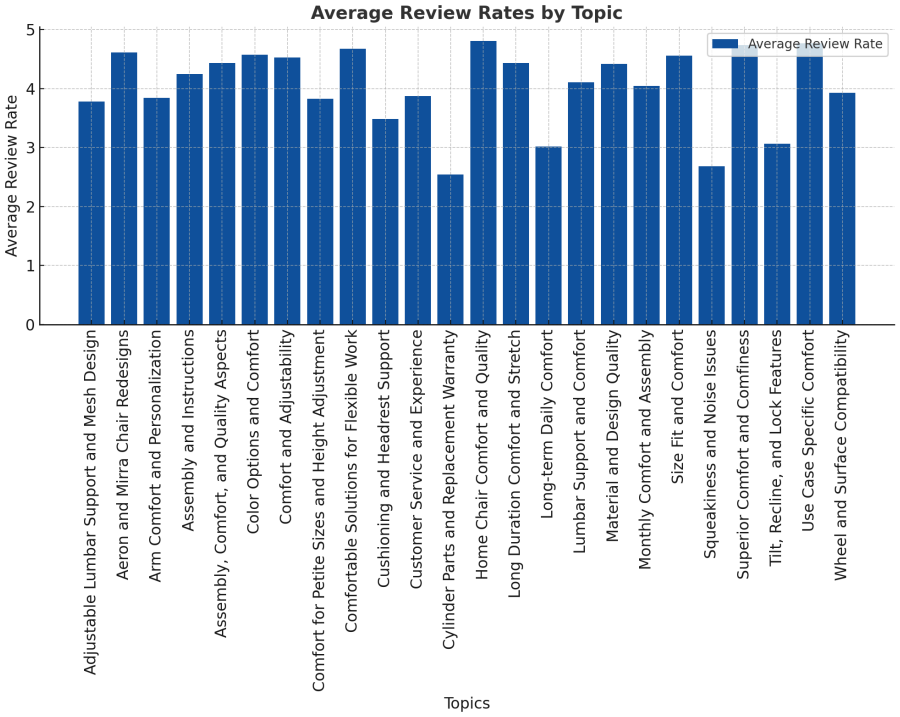


Figure 5-3 Average Review Rates by Topic

In light of the analysis of average review ratings, manufacturers might consider a series of enhancements to improve product satisfaction. Firstly, an upgrade in the design and warranty of the pneumatic cylinder is suggested, utilizing higher-grade materials to enhance its longevity and functionality. Offering an extended warranty for these components could reassure customers about the product's durability and quality. Additionally, making the cylinder easier to service, enabling replacement without the need for specialized tools, would be advantageous.

For sustained daily comfort, chairs should incorporate adjustable lumbar support to accommodate various body shapes and comfort preferences. The use of memory foam or equivalent advanced materials for the seating and backrest can significantly increase comfort during prolonged usage.

Addressing issues of squeakiness and noise is also crucial. Stringent quality control measures should be implemented to ensure all moving parts are well-lubricated and fit securely, reducing noise. Introducing noise-dampening materials at critical joints and connections can further minimize sound emissions. Providing customers with a maintenance guide for regular upkeep, such as lubrication and tightening of components, would also be beneficial.

Lastly, comprehensive ergonomic improvements are recommended. Chairs should offer a full spectrum of adjustments, including seat height, backrest tilt, and armrest positioning, to suit various body sizes and desk configurations. Incorporating a synchro-tilt mechanism that coordinates the chair's movement can enhance posture and overall comfort, rounding off a holistic approach to ergonomic office furniture design.

The chart also enables a comparative analysis across different aspects of the product experience. If "Material and Design Quality" scores highly but "Assembly and Instructions" does not, it could suggest that while the product is well-received once set up, the assembly process poses challenges to customers, potentially impacting overall satisfaction.

5.2 Review Patterns Over Time

5.2.1 Monthly Topic Frequency

The visualization of data through graphical means is a critical aspect of data interpretation in research, linking the generated topics with time, the researcher analyzed the frequency of different topics at different points in time, and generated the following charts, which provides a vivid illustration of trends, patterns, and anomalies within a dataset.

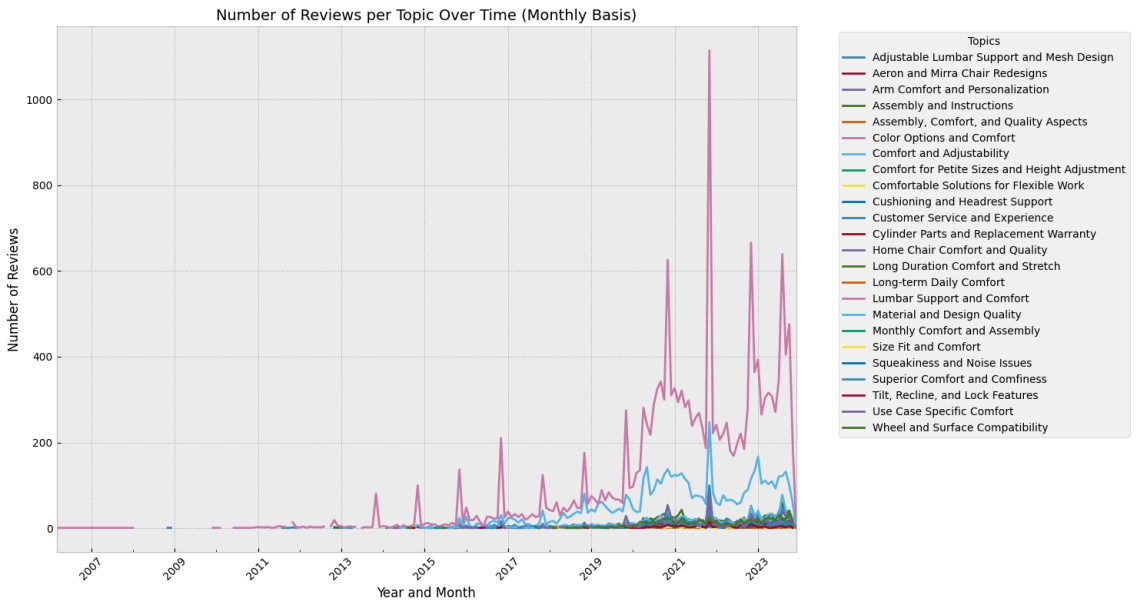
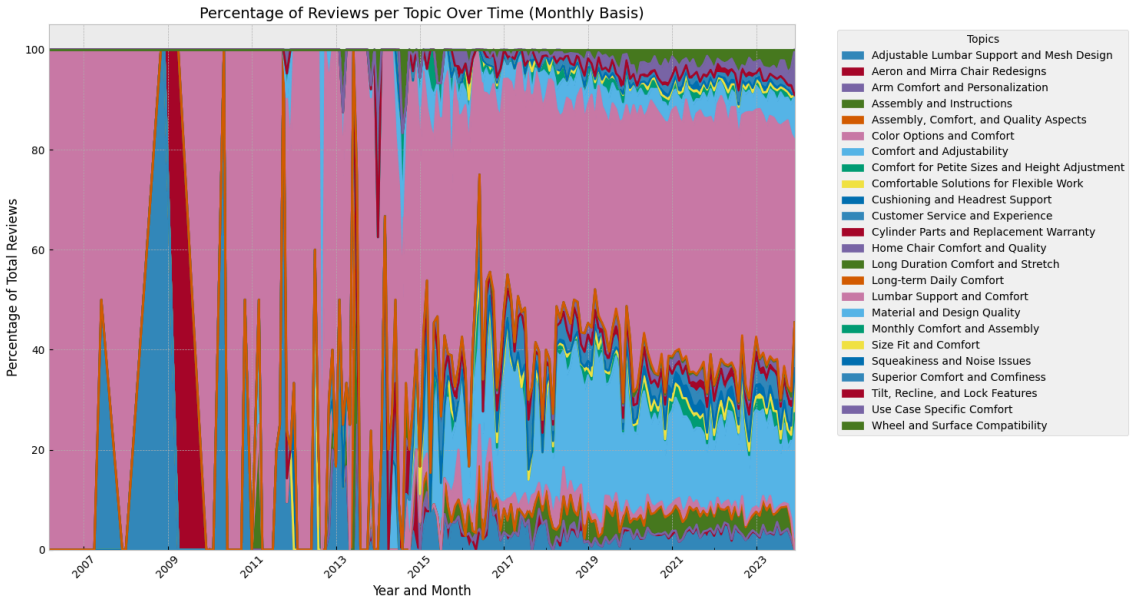


Figure 5-4 Review Patterns Over Time



(up:Number of Reviews per Topic Over Time (Monthly Basis) ; down: Percentage of Reviews per Topic Over Time (Monthly Basis))

On the upper figure 5-4 , The x-axis represents time, spanning several years, while the y-axis represents the number of reviews. This chart can provide insights into trends over time, highlighting when certain features of ergonomic chairs became hot topics among customers and when they declined in discussion.

There are significant spikes in reviews for certain topics at specific times, indicating that these features became particularly relevant or problematic.Some topics have consistent review counts over time, suggesting steady interest or ongoing issues.The overall volume of reviews increases over time, indicating either a growth in the customer base, increased engagement with the review process, or perhaps a combination of both.

Downside figure presents a stacked area chart that traces the evolution of customer feedback on different product-related topics over time. Similar to the first plot but this time pivoting the “percentage ”values for plotting to visualize the percentage of total monthly reviews that each topic represents over time. Each colored section of the chart corresponds to a specific topic, with its width at any point proportional to the percentage of total reviews discussing that topic in a given month. This type of visualization is particularly effective at demonstrating how the relative importance or popularity of topics changes over time.

The trend analysis(Figure 5-5), focusing on the period from 2016 onwards and excluding "Lumbar Support and Comfort," visualizes the evolution of customer feedback for selected topics: "Comfort and Adjustability," "Material and Design Quality," and "Assembly and Instructions."

	YearMonth	Adjustable Lumbar Support and Mesh Design	Aeron and Mirra Chair Redesigns	Arm Comfort and Personalization	Assembly and Instructions	Assembly, Comfort, and Quality Aspects	Color Options and Comfort	Comfort and Adjustability	Comfort for Petite Sizes and Height Adjustment	Comfortable Solutions for Flexible Work	...	Long-term Daily Comfort	Lumbar Support and Comfort	Material and Design Quality	Monthly Comfort and Assembly	Size Fit and Comfort	Squeakiness and Noise Issues	Superior Comfort and Comfiness	Tilt, Recline, and Lock Features	Use Case Specific Comfort	Wheel and Surface Compatibility
74	2016-01	2.777778	NaN	0.925926	3.703704	NaN	1.851852	25.000000	1.851852	NaN	...	NaN	44.444444	6.481481	NaN	NaN	NaN	1.851852	NaN	4.629630	NaN
75	2016-02	2.777778	NaN	NaN	NaN	NaN	5.555556	16.666667	2.777778	NaN	...	NaN	52.777778	2.777778	5.555556	NaN	NaN	NaN	NaN	2.777778	NaN
76	2016-03	NaN	3.333333	NaN	6.666667	NaN	NaN	3.333333	3.333333	NaN	...	NaN	63.333333	6.666667	3.333333	NaN	3.333333	NaN	NaN	3.333333	3.333333
77	2016-04	1.886792	NaN	NaN	1.886792	NaN	1.886792	20.754717	7.547170	NaN	...	NaN	54.716981	3.773585	NaN	NaN	NaN	NaN	NaN	1.886792	NaN
78	2016-05	NaN	NaN	3.333333	NaN	NaN	6.666667	30.000000	NaN	6.666667	...	NaN	46.666667	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
165	2023-08	3.074295	NaN	0.170794	5.038429	0.170794	0.512383	10.418446	1.964133	0.853971	...	0.426985	54.568745	6.660974	0.683177	0.341588	0.512383	1.366354	0.341588	3.159693	2.134927
166	2023-09	3.721489	NaN	NaN	4.681873	0.120048	0.840336	15.846339	2.641056	0.360144	...	0.360144	48.619448	4.921969	1.560624	0.240096	0.600240	1.080432	0.240096	3.841537	2.160864
167	2023-10	4.120267	NaN	0.111359	4.231626	0.222717	0.445434	11.247216	3.118040	0.556793	...	0.445434	53.006682	4.565702	0.890869	0.222717	0.556793	0.334076	0.556793	2.449889	4.565702
168	2023-11	2.374670	NaN	NaN	2.902375	NaN	1.055409	15.567282	0.263852	0.791557	...	1.055409	51.451187	6.068602	1.319261	NaN	0.791557	1.055409	0.527704	3.430079	3.693931
169	2023-12	NaN	NaN	NaN	NaN	NaN	NaN	27.272727	NaN	9.090909	...	NaN	36.363636	9.090909	NaN	NaN	NaN	NaN	NaN	9.090909	NaN

96 rows × 25 columns

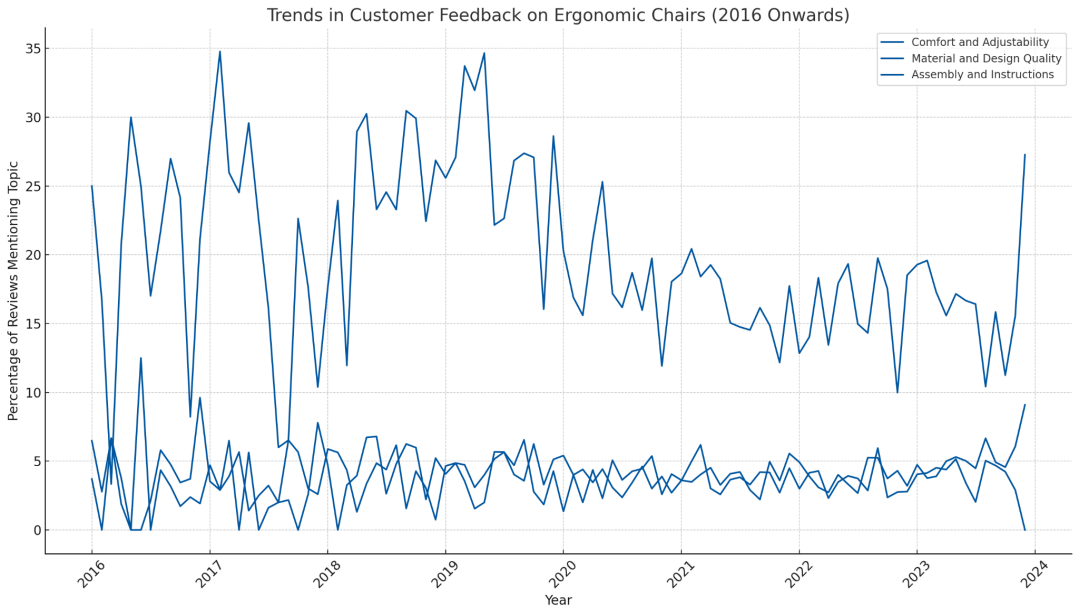


Figure 5-5 Trends in Customer Feedback (2016 onwards)

(up: table represents topic percentage; down:chart outline 3 most frequent topics)

Starting with "Comfort and Adjustability," there is a significant variance in the frequency of feedback over the years. This suggests that consumers' expectations and experiences with the ergonomic aspects of chairs are fluctuating, possibly due to changes in chair design, evolving ergonomic standards, or differing consumer needs. This is a critical area where consumer feedback is highly variable, indicating that ergonomic chair manufacturers need to continually innovate and respond to the changing demands and feedback from customers to maintain satisfaction.

The topic of "Material and Design Quality" displays a trend that could indicate a steady interest over time with some peaks and troughs. These variations could correlate with product releases, changes in material quality, or public discourse about sustainable materials and durability. Consumers maintain a consistent interest in the quality of materials and design, but the fluctuations suggest that certain events or changes in the market can greatly influence consumer satisfaction. Manufacturers should monitor these trends closely and strive for high-quality, consistent product experiences.

Lastly, the "Assembly and Instructions" topic shows a less volatile trend but with notable spikes. These could reflect times when either new products were introduced with different assembly requirements, or when there was an increase in feedback related to the clarity and user-friendliness of the assembly instructions provided. While generally less discussed, spikes in feedback indicate that there are occasional but significant concerns with the assembly process of ergonomic chairs. This implies that there might be opportunities for improvement in the clarity and simplicity of instructions or in the design of the chairs for easier assembly.

Manufacturers and designers of ergonomic chairs should closely analyze these trends and the underlying factors that contribute to them to address potential issues and improve the overall customer experience. This could involve investing in better ergonomic designs, ensuring high-quality materials are used, or simplifying the assembly process and improving instructions to enhance customer satisfaction and loyalty.

5.2.2 Topics Distribution

The plot below(Figure 5-6) provides a Distribution of Feedback Across Topics for the last full year in the dataset, using a pie chart to visually represent the share of customer feedback dedicated to various topics. This visualization helps identify the most dominant concerns among customers, highlighting the areas that received the most attention in reviews.

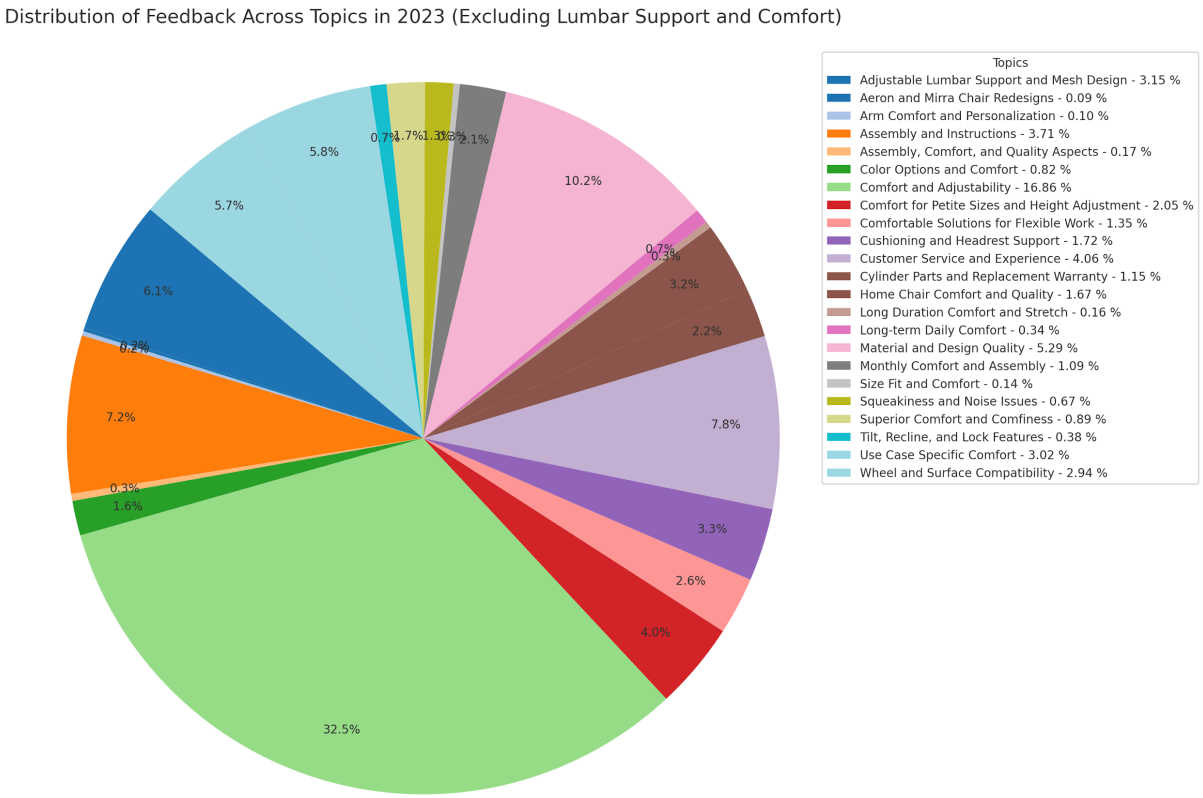


Figure 5-6 Distrubution of Feedback Across Topics in 2023

The "Lumbar Support and Comfort" topic represents 46.2% of the total topic numbers. Excluding the dominant topic ensures that each segment of the pie is clearly labeled, making it easier to identify the topics represented in the chart. The legend, positioned outside the plot area to the upper right, allows for an unobstructed view of the data while providing detailed information about the topics included in the analysis.

The pie chart identifies which aspects of ergonomic chairs are most frequently mentioned in customer reviews. These areas, such as "Comfort and Adjustability" and "Material and Design Quality," should be the primary focus in marketing materials, emphasizing how your products excel in these regards. The distribution highlights consumer priorities when choosing ergonomic chairs. For instance, if topics related to comfort and adjustability dominate the feedback, it indicates a significant market demand for ergonomic chairs that offer superior comfort and customization.

However, less dominant topics in the feedback distribution may represent niche areas that are underserved by current market offerings. These could be opportunities for differentiation, allowing brands to stand out by addressing these specific consumer needs. The distribution also points to areas where there might be room for product improvements. Integrating this feedback into product development can enhance product offerings and customer satisfaction.

5.2.2 Seasonal Analysis

To investigate repeating patterns over time, the researcher aggregated reviews by month to discern any cyclical trends or seasonal behaviors in consumer feedback.This step represents the average frequency of topics appearing in the same month across different years.Such aggregation allows for the smoothing out of anomalies and highlights consistent trends across years, enabling a focused examination of how consumer sentiments or topics of interest fluctuate throughout the year. The processed data frame is shown as table below(Table 5-7).

Merged_Topics	Adjustable Lumbar Support and Mesh Design	Aeron and Mirra Chair Redesigns	Arm Comfort and Personalization	Assembly and Instructions	Assembly, Comfort, and Quality Aspects	Color Options and Comfort	Comfort and Adjustability	Comfort for Petite Sizes and Height Adjustment	Comfortable Solutions for Flexible Work	Cushioning and Headrest Support	...	Long-term Daily Comfort	Lumbar Support and Comfort	Material and Design Quality	Monthly Comfort and Assembly	Size Fit and Comfort	Squeakiness and Noise Issues	Superior Comfort and Confiness	Tilt, Recline, and Lock Features	Use Case Specific Comfort	Wheel and Surface Compatibility
Month																					
1	4.814726	0.346420	0.555386	4.055384	0.493095	1.442128	19.229802	1.896200	0.658888	5.861519	...	0.248611	60.570320	5.396839	2.203988	0.397764	0.497264	1.289516	0.638914	5.398162	4.806904
2	8.537777	1.995142	0.188324	4.377956	NaN	2.655703	20.233819	2.182949	0.604089	1.696673	...	0.618811	60.502370	3.754528	1.792444	0.332226	0.977663	1.077577	0.939295	1.636643	1.992944
3	11.641568	1.744548	1.219527	6.689435	0.274811	1.762862	19.270893	2.506120	0.663400	1.528755	...	0.333099	55.309248	5.416777	1.364101	0.642509	1.194582	0.698091	0.745657	2.635439	3.966995
4	4.605017	0.420168	0.195951	4.436571	0.257371	3.120363	21.946359	2.955398	0.524910	1.842943	...	0.375610	64.488222	4.429679	2.289906	0.502555	0.880642	1.405263	0.647466	2.035487	2.320086
5	14.175014	NaN	0.953834	3.500276	0.385223	2.581805	24.044467	3.627794	1.726918	1.728802	...	0.280401	57.798762	4.778995	2.632697	0.425805	0.799633	1.371261	0.703121	1.860975	2.336589
6	4.306355	0.358423	0.470577	5.582711	25.325999	3.106275	20.453565	2.636928	0.351686	2.317912	...	0.329432	56.085873	5.828541	1.492245	0.313224	0.732975	1.202725	0.596675	2.136181	2.797307
7	9.013539	1.612903	0.163628	6.238203	0.722896	2.501987	16.593760	5.604673	0.711867	1.600180	...	0.409853	55.443355	5.061872	2.597228	0.793262	0.666593	2.594201	0.621374	2.311183	3.069776
8	5.606292	0.163934	0.625012	4.423590	0.167364	2.413968	16.679980	2.299734	0.702169	4.144426	...	0.255845	58.914769	5.868916	3.472940	0.461492	0.818273	1.413673	0.916574	2.034039	4.074465
9	7.093724	4.347826	1.152340	4.605196	0.454043	3.461033	19.608998	2.385904	0.352524	1.869233	...	0.396296	54.340909	13.808582	1.789274	0.869549	0.972786	1.209214	0.540853	2.325253	3.568152
10	12.992408	1.022385	0.176811	3.822303	0.192515	2.499410	18.828506	3.095901	0.547415	2.151975	...	0.887284	61.152562	6.077015	1.744764	0.190348	0.752776	1.674661	1.077291	2.091451	3.246325
11	13.203371	1.922483	0.602449	3.435550	0.805245	1.507283	11.682703	1.718046	0.630321	1.700019	...	0.462890	57.219710	3.597795	1.510543	0.608545	0.462142	0.968502	0.719349	3.342393	2.293265
12	15.331096	10.208914	0.305082	3.800715	0.217901	3.127872	19.861817	2.431168	2.040371	1.558675	...	0.220014	60.282353	7.164959	1.824690	0.560120	0.742853	1.331112	1.169227	3.263551	2.251834

Table 5-7 Review Seasonal Analysis

The data is depicted in a stacked area chart(Figure 5-8) , where the cumulative topics represent 100% of the reviews for each month. Different colors are employed to distinguish between topics, and the varying thickness of each color segment illustrates the relative significance of each topic over time. Although outliers have been excluded, there are still some anomalous values. For example, the orange line representing the topic "Assembly, Comfort, and Quality Aspects" shows a sharp increase in the number of comments in June, which may be attributed to a lack of diversity in the sample sources. However, on a general trend, September, October, and March are periods with high volumes of customer comments.

The chart's detailed time frame enables the observation of both short-term

variations and long-term patterns. Seasonal fluctuations, such as an uptick in mentions of "Arm Comfort and Personalization" during periods traditionally associated with heightened sales, could signal market reactions to promotional endeavors or shifts in consumer engagement patterns.

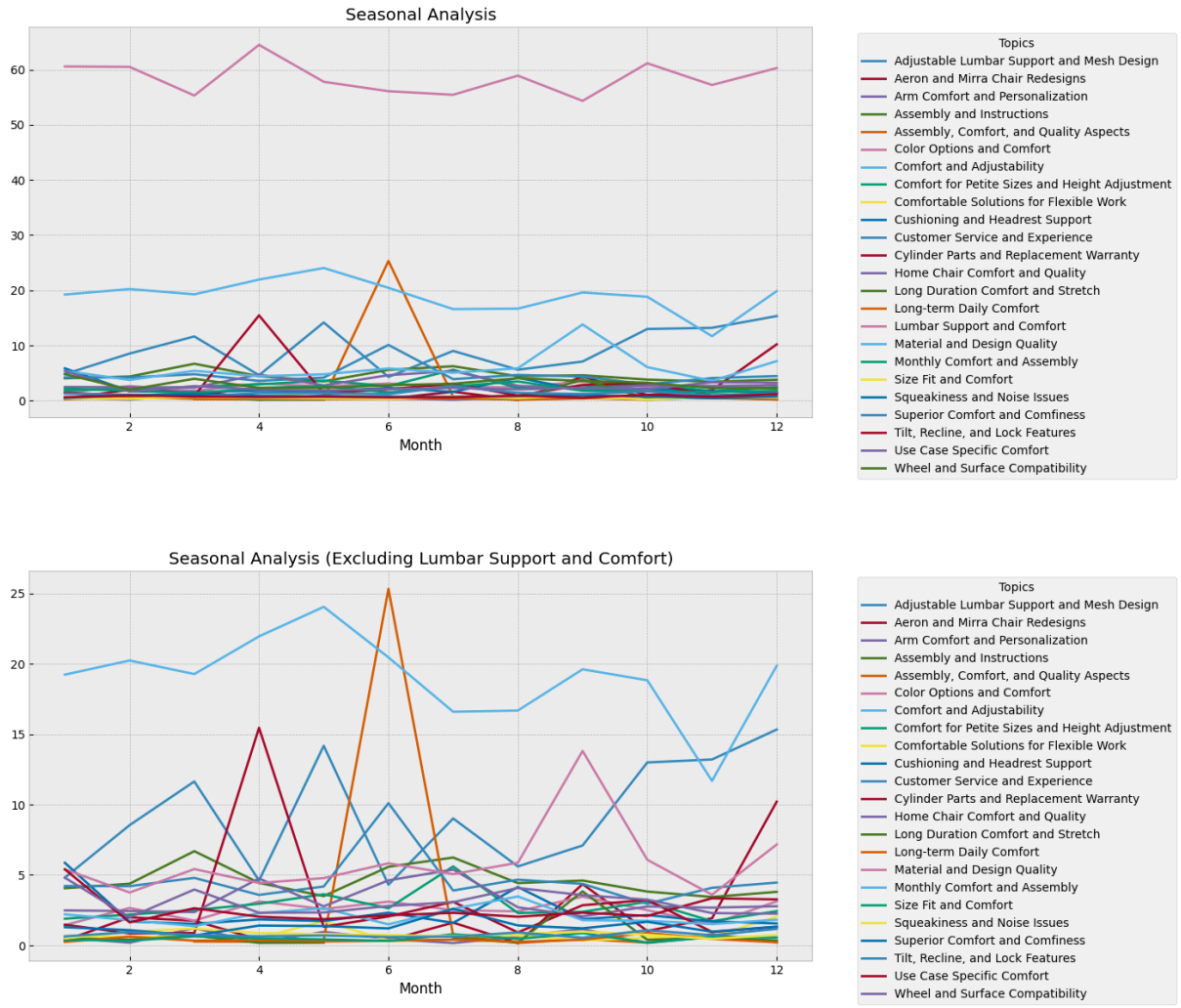


Figure 5-8 Review Seasonal Analysis (down: exclude outlier: Lumbar Support and Comfort)

Moreover, a diminishing frequency of discussions about "Squeakiness and Noise Issues" over time could suggest advancements in product quality or the successful resolution of common complaints. The analysis reveals that certain topics garner more attention during specific times of the year, providing valuable insights for tailoring marketing strategies. For instance, marketing initiatives could be optimized to coincide with periods of increased interest in features like "Comfort for Petite Sizes and Height Adjustment."

5.2.3 Trending Recognition

In conducting trend analysis, the researcher employs a moving average to smooth out short-term fluctuations and highlight longer-term trends within the dataset(Figure 5-9). The data frame represents the moving average of different topics over 6 months(from June 2023 to December 2023), calculates the difference between each period's moving average and its predecessor with the latest 6 periods, then takes the average of these differences for each topic. This average change over the period helps identify the direction and the magnitude of the trend for each topic.

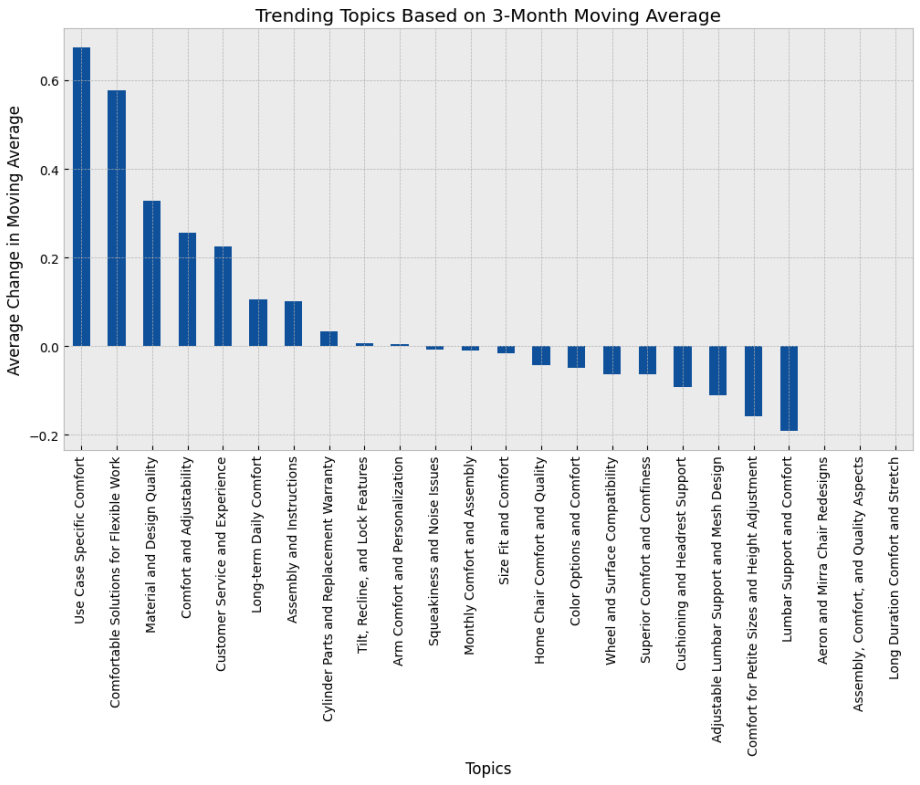


Figure 5-9 Trending Topics Based on 3-Month Moving Average

The 'Use Case Specific Comfort' tops the chart with the highest positive value, indicating a significant uptick in consumer discussions or positive sentiment around this topic. This suggests that products designed with a focus on specific use cases are resonating well with consumers. The rise in this sentiment could be attributed to an increased consumer awareness of personal well-being, or it might reflect a market response to recent product launches that have successfully targeted niche user needs.

Following closely is 'Comfortable Solutions for Flexible Work', which has seen a substantial rise, likely influenced by the growing trend of remote work and home offices. This increase suggests that consumers are actively seeking out products that cater to the flexibility required by changing work environments. A potential driver for this sentiment could be the COVID-19 pandemic, which forced many individuals to adapt their living spaces into functional work areas, thus prioritizing comfort and flexibility in product design.

'Material and Design Quality' also shows a notable positive trend. This positive sentiment growth might be indicative of consumers placing greater value on the build quality and aesthetic appeal of products. In a market saturated with options, it is plausible that consumers are becoming more discerning, preferring products that offer durability and a pleasing design. This shift could be driven by a consumer base that is increasingly willing to invest in higher-quality goods that promise longevity and a timeless appeal.

Conversely, topics such as 'Lumbar Support and Comfort' and 'Comfort for Petite Sizes and Height Adjustment' are trending negatively. This downward trend could suggest either a market saturation for these features or a possible indication of consumers finding the current offerings inadequate. It might also reflect a shift in marketing focus away from these features, leading to a decline in consumer discussions or sentiment.

The presence of 'NaN' (not a number) values for topics like 'Aeron and Mirra Chair Redesigns', 'Assembly, Comfort, and Quality Aspects', and 'Long Duration Comfort and Stretch' suggests a lack of data or insufficient mentions to calculate a meaningful trend. This absence of data could signal a need for further investigation into these topics. Are these areas less critical to consumers, or are they simply underrepresented in the data collected? It may also point to a gap in the market where consumers have not yet fully realized the importance or value of these aspects.

5.3 Ergonomics Chair Marketing Strategy Suggestion

Drawing on the insights from topic analysis, a series of marketing strategies emerge, aimed at bolstering brand image and engaging customers more effectively.

At the heart of these strategies lies the emphasis on outstanding customer service. The frequent references to 'customer_service' highlight the critical nature of incorporating a responsive support system, straightforward replacement procedures, and clear return policies as central components of a marketing strategy. Such a focus has gained increased prominence in discussions about customer service since 2015, representing a significant portion of consumer feedback and pointing towards a crucial direction for marketing narratives.

The analysis also reveals a marked interest in ergonomic features among consumers, with terms like 'head_rest', 'long_period', and 'tall_people' indicating a preference for products designed for varied physical needs and extended usage. Marketing messages should thus highlight the ergonomic design of products, focusing on features like 'memory_foam' for added comfort and compatibility with 'hardwood_floor', showcasing the brand's dedication to satisfying a broad spectrum of consumer demands.

Another pivotal insight pertains to the articulation of a compelling value proposition. Mentions of 'worth_money', 'worth_penny', and 'reasonable_price' point to a consumer focus on value for money. Marketing initiatives need to communicate the long-term advantages and durability of the products, positioning them as investments in personal well-being and environmental sustainability, thus addressing concerns over cost-effectiveness.

The importance of product usability and straightforward assembly is underscored by references to 'assembly_instruction', indicating a consumer preference for easy installation. This presents an opportunity for marketing to simplify the assembly process, potentially through engaging instructional content, thereby enriching the initial product interaction.

Incorporating customer testimonials and positive reviews, as suggested by 'give_star', plays a crucial role in influencing purchase decisions. Embedding these elements into marketing campaigns can enhance credibility and foster trust, utilizing social proof to attract prospective customers.

Directly addressing customer concerns, especially regarding perceived value as indicated by 'waste_money', can bolster confidence in the brand. Marketing

communications that elaborate on product specifications, warranty details, and after-sales support can alleviate concerns and highlight the brand's dedication to customer satisfaction.

Insights into specific product features, such as 'roller_blade', 'upright_position', and 'lock_mechanism', offer valuable guidance for both product development and marketing strategies. Utilizing this feedback to guide product innovation and marketing messages ensures alignment with consumer experiences and enables targeted promotions or offers on highly valued product attributes.

Opportunities for seasonal and situational marketing, derived from terms like 'couple_day' and 'front_computer', suggest a refined approach to campaign execution. Customizing marketing efforts to align with seasonal patterns and specific user contexts can increase relevance and engagement.

Moreover, highlighting after-sales service, particularly the ease of securing replacements and the provision of tools like 'allen_wrench', can distinguish the brand in a competitive marketplace. Portraying a hassle-free post-purchase experience in marketing collateral can further elevate the brand's perception.

In conclusion, leveraging 'game changer' features that deeply resonate with consumers can significantly differentiate the brand. Spotlighting these innovative elements in marketing endeavors can captivate consumer interest and underscore the brand's unique value proposition, distinguishing it within a competitive arena.

CHAPTER VI : FUTURE SCOPE AND CONCLUSION

5.1 Conclusions

In conclusion, this thesis introduces a Bertopic Modeling-based approach to analyzing customer preferences, adeptly navigating the complexities of extracting valuable insights from consumer feedback for enhanced marketing strategy formulation. Utilizing online customer reviews, the approach trains localized topic analysis models on specific categories, employing the BERTopic model in conjunction with sophisticated clustering and natural language processing techniques. This integrated framework significantly improves the model's capability to derive relevant marketing insights from the vast array of consumer reviews. The careful selection of categories for global model training is pivotal in optimizing model performance. This enhancement fosters a detailed understanding of consumer sentiments, preferences, and expectations, thereby enabling the extraction of precise marketing information crucial for strategic decision-making.

By tailoring and refining the Bertopic model to analyze ergonomic chair review datasets, the approach adeptly pinpoints the relevant topics. Through the phrase isolation and applying the model, the analysis further explores emerging trends, reveals latent consumer needs, and provides insights into brand perception across diverse customer segments. Analyzing predominant topics within consumer dialogue allows businesses to align their offerings with current demands, customize messaging to engage distinct audiences effectively, and anticipate market trends with enhanced precision. This method empowers companies to maintain a competitive edge, catering to consumer needs.

The insights derived from topic modeling present a solid foundation of customer priorities, areas of dissatisfaction, and unmet needs within the ergonomic chair market. Marketing strategies benefit from these insights by crafting compelling narratives that address consumer concerns, emphasize product advantages, and distinguish the brand in a competitive environment. Moreover, recognizing the subtleties in consumer feedback aligns product offerings with market expectations, significantly improving customer satisfaction and loyalty.

This thesis lays the groundwork for incorporating advanced machine learning

algorithms into marketing research, underscoring the significance of data quality, algorithmic transparency, and interpretability. This study not only uncovers prevalent topics and customer preferences but also sets a new benchmark for methodological rigor in the intersection of machine learning and marketing research.

5.2 Limitations

One limitation of the study is the exclusive reliance on online reviews from e-commerce platforms as the primary data source, which could introduce a potential bias. This approach means that sentiments and topics extracted from these reviews might not fully reflect the complete range of consumer opinions. The reviews are sourced solely from customers who have made purchases, omitting insights from potential buyers who have yet to purchase but may hold valuable opinions and preferences regarding ergonomic chairs. The researcher faced constraints in accessing a broader array of online review data due to the unstructured nature of reviews, variations in formats across different websites, and web crawling restrictions. The study's focus on participants from the United States region, means the results may only represent the views and demographics of the U.S. market. This geographic limitation restricts the applicability of the findings to other regions, potentially overlooking diverse consumer perspectives and preferences that vary across different cultural and economic backgrounds.

The following limitations are also acknowledged after a retrospective critique of the research’s methodology from data collection through results analysis:

- Words with contextual meaning (dual meaning) can be extracted differently in the models (example as support for customer support or back support)
- the collection of data was across three sources, may be biased in customer representation and sampling can introduce bias
- fine tuned model performance on a pre-trained model versus an ideal albeit unavailable model trained on reviews dataset specifically
- further adaptations and hyper parameters may improve the results

5.3 Future Scope and Challenges

The Bertopic Model shows great promise in generating topic clusters from a given set of unlabeled data, , employing a variety of clustering algorithms such as NPFST, Word2Vec, HDBSCAN, MMR, and UMAP within the BERTopic framework. This integration offers a unique method for modeling topics from customer reviews. However, the success of these algorithms in effectively and efficiently identifying topics hinges on their compatibility with and optimization for BERTopic. The introduction of the NPFST method, although promising, faces potential integration challenges and the need for performance benchmarking against more conventional techniques. Future investigations could focus on developing metrics for evaluating the effectiveness of these methodologies, alongside comparisons across different methods and embedding models.

Another challenge is training models on large, unstructured datasets. Customer reviews, inherently unstructured and filled with irregularities such as typos and slang, present complexities in text analysis. Transforming these texts into a form amenable to analysis, without altering their original sentiment and meaning, remains a formidable challenge. The quality of the data, including its completeness and representativeness, is crucial for ensuring the reliability of topic modeling results. Future research should explore strategies for managing imbalanced datasets, enhancing data quality for underrepresented categories.

Moreover, while the choice of ergonomic chairs as a research sample provides a focused lens for study, it also limits the generalizability of the findings to other product categories. The insights and model parameters obtained may not directly translate to products or services outside the ergonomic or office furniture domain. Expanding the application of the Bertopic Model to other industries and fields represents a promising direction for future research. Future research directions could expand upon this foundation by investigating alternative data sources, enhancing text preprocessing methods, and applying the framework to various industries. Such explorations would deepen the academic conversation regarding the application and constraints of topic modeling in marketing analysis and offer a methodological blueprint for subsequent research endeavors.

REFERENCES

Wang, Z., Chen, J., Chen, J., et al. (2023). "Identifying interdisciplinary topics and their evolution based on BERTopic". *Scientometrics*

Yazıcı, G., Ozansoy Çadırcı, T. (2023). "Creating meaningful insights from customer reviews: a methodological comparison of topic modeling algorithms and their use in marketing research". *Journal of Marketing Analysis*

Mosteller, J.R., Mathwick, C. (2016). "Online reviewer engagement: A typology based on reviewer motivation". *Journal of Service Research*

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

Alvargonzález, D. (2011). "Multidisciplinarity, interdisciplinarity, transdisciplinarity, and the sciences". *International Studies in the Philosophy of Science*, 25(4), 387–403

Callon, M., Courtial, J.P., Turner, W.A., Bauin, S. (1983). "From translations to problematic networks: An introduction to co-word analysis". *Social Science Information*, 22(2), 191–235

Chen, B., Tsutsui, S., Ding, Y., Ma, F. (2017). "Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval". J*ournal of Informetrics*, 11(4), 1175–1189

Dong, K., Xu, H., Luo, R., Wei, L., Fang, S. (2018). "An integrated method for interdisciplinary topic identification and prediction: A case study on information science and library science". *Scientometrics*, 115, 849–868

Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". arXiv preprint arXiv:2203.05794

Hall, D., Jurafsky, D., Manning, C.D. (2008). "Studying the history of ideas using topic models". In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 363–371

Jiang, L., Zhang, T., Huang, T. (2022). "Empirical research of hot topic recognition

and its evolution path method for scientific and technological literature". *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 26(3), 299–308

Leydesdorff, L., Hellsten, I. (2006). "Measuring the meaning of words in contexts: An automated analysis of controversies about ‘monarch butterflies’, ‘frankenfoods’, and ‘stem cells’". *Scientometrics*, 67(2), 231–258

Leydesdorff, L., Ismael, R. (2011). "Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations". *Journal of Informetrics*, 5(1), 87–100

Leydesdorff, L., Wagner, C.S., Bornmann, L. (2019). "Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient". *Journal of Informetrics*, 13(1), 255–269

Li, M. (2017). "An exploration to visualize the emerging trends of technology foresight based on an improved technique of co-word analysis and relevant literature data of WOS". *Technology Analysis & Strategic Management*, 29(6), 655–671

Li, J. (2014). "The concept and measurement of interdisciplinarity". *Documentation, Information & Knowledge*, 3, 87–93

Aggarwal, C.C., Hinneburg, A., Keim, D.A. (2001). "On the surprising behavior of distance metrics in high dimensional space". *Proceedings of the International Conference on Database Theory*, pp. 420-434. Springer

Allaoui, M., Kherfi, M.L., Cheriet, A. (2020). "Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study". *International Conference on Image and Signal Processing*, pp. 317-325. Springer

Angelov, D. (2020). "Top2vec: Distributed representations of topics". arXiv preprint arXiv:2008.09470

Baturo, A., Dasandi, N., Mikhaylov, S.J. (2017). "Understanding state preferences with text as data: Introducing the UN General Debate Corpus". *Research & Politics*, 4(2), 2053168017712821

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U. (1999). "When is “nearest neighbor” meaningful?". *International Conference on Database Theory*, pp. 217-235. Springer

Bianchi, F., Terragni, S., Hovy, D. (2020a). "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence". arXiv preprint arXiv:2004.03974

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., Fersini, E. (2020b). "Cross-lingual contextualized topic models with zero-shot learning". arXiv preprint arXiv:2004.07737

Blei, D.M., Lafferty, J.D. (2006). "Dynamic topic models". In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113-120

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research*, 3, 993-1022

Bouma, G. (2009). "Normalized (pointwise) mutual information in collocation extraction". *Proceedings of GSCL*, 30, 31-40

Anoop, V.S., Asharaf, S. (2017). "A topic modeling guided approach for semantic knowledge discovery in e-commerce". *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(4), 40–47

Kyeong, J., Jiyoon, K., Lee, J.W., Lee, Y., Lee, S.B. (2021). "Text Mining Analysis of Consumer Perception of Food Distribution Platforms: Focusing on Topic Modeling". *Journal of Foodservice Management*, 24, 71–100

Bumjun, L., Heekyung, N. (2020). "Food tourism market segmentation approach using topic modeling analysis: Focusing on benefits sought". *Korean Journal of Hospitality and Tourism*, 29, 187–204

Cho, M.-K., Lee, B.-J. (2021). "Comparison of service quality of full-service carriers in Korea using topic modeling: Based on reviews from TripAdvisor". *Journal of Hospitality and Tourism Studies*, 23, 152–165

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

Lloyd, S. (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory*, 28(2), 129-137

Campello, R.J.G.B., Moulavi, D., Sander, J. (2013). "Density-Based Clustering Based

on Hierarchical Density Estimates". *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*, vol 7819. Springer, Berlin, Heidelberg

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Vol. 96, 226–231

Xie, J., Girshick, R., Farhadi, A. (2016). "Unsupervised deep embedding for clustering analysis". *International Conference on Machine Learning*, pp. 478-487

Yang, B., Fu, X., Sidiropoulos, N.D., Hong, M. (2017). "Towards k-means-friendly spaces: Simultaneous deep learning and clustering". *International Conference on Machine Learning*, pp. 3861-3870

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research*, 3(Jan), 993–1022

Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". arXiv preprint arXiv:2203.05794

Reimers, N., Gurevych, I. (2019). "SentenceBERT: Sentence embeddings using siamese BERT networks". arXiv preprint arXiv:1908.10084

Blei, D.M., Lafferty, J.D. (2006). "Dynamic topic models". *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113-120

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W. (2021). "Informer: Beyond efficient transformer for long sequence time-series forecasting". In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11106-11115

Wu, H., Xu, J., Wang, J., Long, M. (2021). "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting". *Advances in Neural Information Processing Systems*

Caron, M., Müller, O. (2020). "Hardening Soft Information: A Transformer-Based Approach to Forecasting Stock Return Volatility". *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, pp. 4383-4391

Anoop, V.S., Asharaf, S. (2017). "A topic modeling guided approach for semantic knowledge discovery in e-commerce". *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(4), 40–47

Li, W., Yin, J., Chen, H. (2018). "Supervised Topic Modeling Using Hierarchical Dirichlet Process-Based Inverse Regression: Experiments on E-Commerce Applications". *IEEE Transactions on Knowledge and Data Engineering*, 30(6), 1192-1205

Chen, R., Xu, W. (2017). "The determinants of online customer ratings: a combined domain ontology and topic text analytics approach". *Electronic Commerce Research*, 17, 31–50

Dong, L.Y., Ji, S.J., Zhang, C.J., Zhang, Q., Chiu, D.W., Qiu, L.Q., Li, D. (2018). "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews". *Expert Systems with Applications*, 114, 210-223

Hajek, P., Hikkerova, L., Sahut, J.M. (2023). "Fake review detection in e-Commerce platforms using aspect-based sentiment analysis". *Journal of Business Research*, 167, 114143

Ye, X., Lian, Z., She, B., Kudva, S. (2020). "Spatial and big data analytics of E-market transactions in China". *GeoJournal*, 85, 329-341

Hong, W., Zheng, C., Wu, L., Pu, X. (2019). "Analyzing the relationship between consumer satisfaction and fresh e-commerce logistics service using text mining techniques". *Sustainability*, 11(13), 3570

Wang, Y.D., Jiang, B.T., Ye, X.Y. (2016). "A method for studying the development pattern of urban commercial service facilities based on customer reviews from social media". *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41, 577-578

Kyeong, J., Jiyoon, K., Lee, J.W., Lee, Y., Lee, S.B. (2021). "Text Mining Analysis of Consumer Perception of Food Distribution Platforms: Focusing on Topic Modeling". *Journal of Foodservice Management*, 24, 71–100

Bumjun, L., Heekyung, N. (2020). "Food tourism market segmentation approach

using topic modeling analysis: Focusing on benefits sought". *Korean Journal of Hospitality and Tourism*, 29, 187–204

Soyeon, L., Yeongok, K. (2022). "Analysis of Apartment Interior Trend Using Topic Modeling: Focusing on ‘Today’s House’ Review Data". *Proceedings of the KMIS 2022: 14th International Conference on Knowledge Management and Information Systems*, Valletta, Malta, pp. 141–149

Cho, M.-K., Lee, B.-J. (2021). "Comparison of service quality of full-service carriers in Korea using topic modeling: Based on reviews from TripAdvisor". *Journal of Hospitality and Tourism Studies*, 23, 152–165

Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S.S.K., Dwivedi, S., Raykar, V. (2020). "Attention based multi-modal new product sales time-series forecasting". *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3110-3118

Kudo, T. (2006). "MeCab: Yet another part-of-speech and morphological analyzer". https://sourceforge.net/projects/mecab/

Sentence Transformers. "all-mpnet-base-v2." *Hugging Face,* n.d.*,* https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Lee, J. (2020). "KcBERT: Korean Comments BERT". *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pp. 437-440

Derrick, T.R., Bates, B.T., Dufek, J.S. (1994). "Evaluation of time-series data sets using the Pearson product-moment correlation coefficient". *Medicine and Science in Sports and Exercise*, 26(7), 919-928

Bouma, G. (2009). "Normalized (pointwise) mutual information in collocation extraction". *Proceedings of GSCL*, 30, 31–40

Dieng, A.B., Ruiz, F.J.R., Blei, D.M. (2020). "Topic modeling in embedding spaces". *Transactions of the Association for Computational Linguistics*, 8, 439–453