UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE,
ECONOMICHE E SOCIALI

Corso di Laurea Magistrale in Data Science and Economics

E-commerce Insights: Analyzing Customer Reviews through LDA Topic Modeling and Association Rules

Tesi di Laurea di: Luca Bertoletti
Relatore: Dario Malchiodi
Correlatore: Matteo Zignani

Anno Accademico 2022 / 2023

Contents

Abstract

In today's digital era, the dissemination of information has become incredibly widespread, particularly in product reviews. App stores, initially intended solely for distribution, have transformed into essential platforms where users provide feedback through ratings and reviews, significantly influencing consumer decisions and marketing strategies. However, understanding these reviews can be challenging due to their unstructured nature. This study focuses on extracting insights from user reviews of the prominent Chinese e-commerce app, Temu, using techniques like LDA for topic analysis and association rule mining to understand how different features relate to user ratings. By uncovering key topics and identifying what users like and dislike, this research aims to assist companies in refining their e-commerce strategies, ultimately enhancing user satisfaction and loyalty in the competitive digital marketplace.

1 Introduction

In the era of the internet and social media, the dissemination of information has become increasingly widespread and accessible, profoundly impacting various aspects of society. This transformation extends to product reviews, where numerous platforms now offer consumers several information before making purchasing decisions, as noted by Kim and Chun [26]. In the past decade, app stores like Google Play or Apple AppStore have emerged as crucial sources of user reviews, marking a significant shift in how consumers interact with mobile applications. Originally conceived primarily as distribution platforms, these app stores have evolved into dynamic hubs where users actively engage by providing feedback through ratings and reviews post-download. These feedbacks, encompassing bug reports, feature requests, and overall user experience, offer invaluable insights for app developers and consumers [36]. As highlighted by Wang et al. [46], these reviews often reflect user satisfaction, directly impacting loyalty and influencing marketing strategies.

However, while user reviews offer valuable insights, they also present challenges such as the unstructured nature of feedback and the need for effective analysis to extract meaningful insights, as noted by Al-Hawari et al. [4]. Despite these challenges, the importance of user reviews in app stores cannot be understated, as they provide companies with a direct line of communication with their user base, facilitating continuous improvement and innovation.

The goal of this work is to uncover valuable insights hidden within user reviews, with a particular focus on those from the well-known Chinese e-commerce app, Temu. This study will utilize LDA to analyze pre-processed text reviews and extract relevant topics along with related terms. Subsequently, keywords will be identified based on their frequency and categorized into specific e-commerce features. The reviews will be segmented into two clusters according to their rating (”positive” and ”negative”), and association rule mining will be employed to uncover the relationship between these features and the assigned rating.

Two main questions are addressed:

1. What are the main topics discussed in the user reviews?
2. Which features do users view positively and which negatively?

Exploring the features that capture users’ interest in e-commerce applications holds significant importance for the digital sector. By identifying which features need attention and improvement companies can refine their strategies and meet customer expectations more effectively, ultimately fostering satisfaction and loyalty in the competitive e-commerce market.

This work is organized as follows. Chapter 2 serves as an introduction to the concept of social media mining, delving into the intersection of social media, data mining, and e-commerce. It provides a foundational understanding of how these elements combine and set the stage for the subsequent analysis. In Chapter 3, a comprehensive literature review is conducted on two key techniques utilized in this study: Latent Dirichlet Allocation (LDA) and association rules. This chapter synthesizes existing research and establishes a theoretical framework for the application of these techniques in analyzing user-generated content. Chapter 4 offers a detailed account of the analysis conducted on users' reviews using Python programming language, outlining the methodology and procedures employed. In Chapter 5, the results of the analysis are presented and discussed, focusing on key insights obtained from the data. Finally, Chapter 6 consolidates the findings and offers conclusions drawn from the study, as well as suggestions for future research directions in the field.

2 Introduction to Social Media Mining

The evolution of the World Wide Web [18] has seen several significant transitions, each marking a shift in how people interact with and utilize the internet. The transition from Web 1.0 to Web 2.0 stands out as one of the most notable. Web 1.0, often referred to as the ”read-only” web, was characterized by static web pages and limited user interaction. During this phase, information was primarily consumed passively, with websites serving as digital brochures rather than interactive platforms.

The rise of Web 2.0 marked a paradigm shift towards a more dynamic and participatory web experience. This transition was possible thanks to advancements in technology, such as the widespread adoption of broadband internet, improved web development tools, and the proliferation of Social Media platforms. Web 2.0 ushered in an era of user-generated content, social networking, and interactive web applications. Websites became more collaborative and engaging, empowering users to create, share, and interact with content in real-time.

2.1 Social Media

Social Media, as defined by Kaplan and Haenlein in [24], is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user generated content. Within this broad definition, there are different kinds of Social Media that need to be differentiated more clearly. Table 1 below shows the different types, characteristics, and some examples of Social Media platforms.

Type	**Characteristics**	**Examples**
Social Network	Web-based services that allow individuals and communities to connect with real-world friends and acquaintances online.	Facebook, Twitter, LinkedIn
Media Sharing	Media sharing is used to find and share photographs, live videos, videos and other kinds of media on the web.	YouTube, Instagram, Snapchat
Discussion Forum	Used for finding, sharing, and discussing different kinds of information, opinions, and news.	Reddit, Quora, Yahoo! Answers
Bookmarking and Content Curation	Social bookmarking sites allow users to bookmark Web content for storage, organization, and sharing.	Pinterest
Consumer Reviews	Used to collect and publish user-generated content in the form of subjective commentaries on existing products, services, entertainment, businesses, places, etc. Some of these sites also provide product reviews.	TripAdvisor, Yelp
Blogging	A journal-like website for users, aka bloggers, to contribute textual and multimedia content. Blogs are generally maintained by an individual or by a community.	WordPress, Medium

Table 1: Different types of Social Media platforms

Consumer reviews platforms like TripAdvisor and Yelp serve as valuable resources for individuals seeking subjective opinions and evaluations of various products and services. These platforms exploit user-generated content to inform potential consumers about the quality and experiences associated with specific offerings. Similarly, in the realm of e-commerce, customer reviews play a crucial role in helping purchasing decisions. Just as travelers rely on TripAdvisor to choose hotels or restaurants, online shoppers can rely on customer reviews on e-commerce platforms like Amazon or eBay to understand the satisfaction levels of previous buyers and determine the suitability of a product.

2.2 Data Mining

Data mining [8] is a multidisciplinary field that involves discovering patterns, trends, and insights from large datasets using various techniques from statistics, machine learning, and database systems. Data mining finds applications in diverse domains such as business intelligence, healthcare, finance, marketing, and scientific research, enabling organizations to make data-driven decisions, uncover hidden patterns, and gain competitive advantages [44].

Data mining algorithms [47] are broadly classified into supervised and unsupervised learning algorithms. Supervised learning involves training a model on labeled data, where each input is paired with the correct output. This allows the algorithm to learn the mapping between inputs and outputs, enabling it to make predictions on unseen data. Classic examples of supervised learning algorithms include linear regression for predicting continuous values and decision trees for classification tasks. On the other hand, unsupervised learning involves training on unlabeled data, where the algorithm must find patterns or structure within the data on its own. Clustering algorithms like k-means and hierarchical clustering are common examples of unsupervised learning, used for grouping similar data points together. In the realm of Social Media mining, unsupervised algorithms like LDA and association rules mining hold significant importance. Social Media platforms generate vast amounts of unstructured data in the form of text, images, videos, and interactions among users. LDA, as an unsupervised algorithm, becomes invaluable for topic modeling within this sea of data. By identifying underlying themes or topics across a collection of Social Media posts or comments, LDA aids in understanding the prevalent discussions, trends, and sentiments within the online community. Similarly, association rules mining, performed through the Apriori algorithm, plays a crucial role in uncovering patterns and relationships among various items or concepts mentioned in Social Media posts. This allows for insights into user behavior, preferences, and interactions, which are invaluable for targeted marketing, recommendation systems, and understanding community dynamics. In the broader context of Social Media mining, these algorithms, alongside other techniques like text mining, natural language processing, clustering, and deep learning, form the backbone for extracting valuable insights and understanding user engagement and behavior in online Social Media.

2.3 Social Media Mining

Social Media mining [47] is a specialized application of data mining focused specifically on extracting insights from Social Media platforms. While data mining is a broader concept applicable to any dataset, Social Media mining targets the vast amounts of user-generated content and interactions on platforms. By applying data mining techniques to Social Media data, it is possible to uncover valuable insights about user behaviors, sentiments, relationships, and trends.

Social Media data differ significantly from traditional attribute value data used in classic data mining. They are vast, full of noise, distributed across various platforms, unstructured, and dynamic. These characteristics present significant challenges for data mining tasks, requiring the development of new, efficient techniques and algorithms. Social Media data can be particularly noisy, making it crucial to filter out irrelevant information before analysis. Additionally, the decentralized nature of Social Media means that data are distributed across multiple platforms, complicating efforts to understand information flows. Furthermore, the unstructured nature of Social Media data makes it difficult to draw meaningful insights from diverse sources. Finally, Social Media platforms are dynamic, evolving continuously, further complicating data analysis efforts.

Social Media mining is useful for a multitude of purposes across different domains. In politics and public opinion research, it allows analysts to gauge public sentiment, track political discourse, and assess the impact of policies or events in real-time. Additionally, Social Media mining plays a crucial role in cybersecurity, enabling organizations to detect and prevent cyber threats, identify malicious activities, and protect user privacy. In business, it helps companies understand customer preferences, monitor brand reputation, and spot emerging trends. Specifically in e-commerce, Social Media mining is invaluable for analyzing customer reviews. By using these techniques, businesses can dive into consumer opinions, spot patterns in feedback, and find areas for product improvement. This understanding helps companies tailor their offerings to better meet customer needs, boosting satisfaction and loyalty. This study focuses on using Social Media mining to extract insights from customer reviews in e-commerce, highlighting its importance in understanding and enhancing customer experiences.

However, Social Media mining also comes with several drawbacks and challenges. One major concern is the privacy of individuals, as the mining process often involves analyzing personal data shared on Social Media without explicit consent. This raises ethical and legal questions regarding data protection and user privacy rights. Furthermore, handling the massive amount of Social Media data quickly becomes a challenge for mining algorithms due to the need for advanced infrastructure and computational resources. The noisy and unstructured nature of this data adds another layer of complexity, making it tricky to clean up and analyze properly. This can introduce biases and inaccuracies in the insights extracted from the data. Moreover, the ever-changing nature of Social Media platforms means that keeping the data relevant and staying updated with platform changes and shifts in user behavior is an ongoing challenge. Despite these challenges, Social Media mining continues to evolve as a powerful tool for understanding and harnessing the wealth of information available on Social Media platforms.

2.4 E-commerce

Formally, e-commerce refers to digitally facilitated commercial transactions among organizations and individuals [28] [12]. Each component of this definition is significant: digitally facilitated transactions encompass those mediated by digital technology, predominantly occurring over the Internet, the Web, or mobile devices. Commercial transactions entail the exchange of value, such as money, across organizational or individual boundaries in exchange for goods and services. Understanding the exchange of value is crucial in delineating the scope of e-commerce, as commerce fundamentally hinges upon this exchange. According to Laudon and Traver [28], e-commerce development went through three stages: innovation, consolidation, and reinvention.

Invention (1995 –2000)

The initial stages of online commerce witnessed a remarkable growth. In this era of innovation, e-commerce primarily revolved around the sale of basic retail goods over the Internet, given the limited bandwidth available which constrained the complexity of products offered. Advertising tactics were largely confined to static display ads, complemented by relatively weak search engine capabilities. The online presence of most major corporations, if existent, typically comprised simple static websites showcasing their brands. This phase of e-commerce concluded in 2000 with a significant downturn in stock market valuations, resulting in the disappearance of numerous companies in what became known as the ”dot-com crash.”

Consolidation (2001 – 2006)

The period from 2001 to 2006 was a turning point for e-commerce. It was a time when companies had to rethink their strategies and focus more on making money rather than just using fancy technology. Big companies started using the internet to strengthen their brands instead of creating new ones, while smaller startups faced challenges in getting funding. E-commerce also expanded beyond just selling products online to offering services like travel bookings and financial management, thanks to faster internet and cheaper computers. Marketing strategies got smarter too, with businesses using data to target customers more effectively. And instead of just having websites, companies began to spread their presence across email, Social Media, and online ads to connect with customers in new ways. This period set the stage for a new chapter in e-commerce, where adaptability and customer satisfaction became key drivers of success.

Reinvention (2007 – present)

From 2007 to the present, e-commerce underwent another radical transformation, driven by the rapid evolution of Web 2.0 technologies, widespread mobile device adoption, and the expansion of local goods and services. This period marked a convergence of sociological, technological, and business forces, shaping what is often termed as the ”social, mobile, local” online world. Entertainment content emerged as a significant revenue stream in e-commerce, with mobile devices becoming both entertainment hubs and on-the-go shopping tools. Marketing strategies evolved with the rise of social networks, leveraging powerful data analytics for personalized and targeted campaigns. Companies expanded their online presence beyond static websites to platforms like Facebook, Twitter, and Instagram, aiming to engage consumers with coordinated messages. These social networks, characterized by user-generated content and high interactivity, presented marketers with unprecedented opportunities for targeted advertising. Moreover, the reinvention of e-commerce gave rise to on-demand personal service businesses like Uber, Airbnb, and Instacart, leveraging the mobile platform to tap into unused assets and create lucrative markets.

An e-commerce is distinguished by its eight defining features. These elements present many opportunities for marketing and sales endeavors. With the ability to deliver tailored, interactive, and engaging messages to specific audience segments, e-commerce emerges as a potent tool for effectively reaching and engaging target markets.

- Ubiquity: Present virtually everywhere and at all times.

- Global reach: Facilitates commercial transactions across cultural and national borders more conveniently and affordably compared to traditional commerce.

- Universal standards: Employed uniformly across nations worldwide, contrasting with the varied technologies prevalent in traditional commerce.

- Richness: Empowers online merchants to convey marketing messages in ways not achievable through conventional commerce methods.

- Interactivity: Enables bidirectional communication between merchants and consumers, offering engagement akin to face-to-face interactions but on a larger, global scale.

- Information density: Represents the abundance and quality of information accessible to all market participants. The Internet minimizes costs associated with information gathering, storage, processing, and dissemination while enhancing information currency, accuracy, and timeliness.

- Personalization and customization: Leveraging increased information density, merchants can tailor marketing messages to specific individuals, achieving levels of personalization and customization previously unattainable.

- Social technology: Embraces a many-to-many communication model, allowing millions of users to generate content consumed by equally numerous others. This fosters the formation of extensive social networks and the aggregation of substantial audiences on Social Media platforms.

Various types of e-commerce exist, each with its own distinct characteristics. Typically, e-commerce are classified based on the nature of the market relationship, specifically, who is selling to whom. There are four primary types [28] [12]:

- B2C e-commerce involves businesses selling directly to consumers, representing the most common type encountered by shoppers.

- B2B e-commerce sees businesses selling goods or services to other businesses, constituting the largest sector of e-commerce.

- C2C e-commerce facilitates direct transactions between consumers, where individuals prepare and list products for sale, relying on platform providers for cataloging, search, and transaction processing.

- C2B e-commerce consists in consumers offer products or services to businesses, often through platforms where individuals can pitch their skills, goods, or ideas to potential buyers. This model flips the traditional buyer-seller relationship on its head, empowering consumers to initiate transactions and negotiate terms with businesses.

In this work, the spotlight will be on reviews on B2C e-commerce platforms. In particular, the analysis will prioritize the examination of customer reviews for the popular Chinese e-commerce site, Temu. The next section will provide a concise overview of the chosen e-commerce.

2.4.1 Temu

Temu is an online retailer that launched in the United States in September 2022 [13]. It is operated by the Chinese e-commerce company PDD Holdings [35]. Temu offers steep discounts on a variety of products, mostly shipped directly from Chinese factories. By removing intermediaries, the e-commerce is able to keep incredibly low prices [14]. Moreover, Temu can no doubt attribute its popularity to its strategy of giving free stuff to users who promote the app on their social networks and get friends and family to sign up.

Starting from February 2023, Temu launched in several other countries among which Canada, Australia and New Zealand. In the following months, Temu was launched in Europe reaching France, Germany, Italy, the Netherlands, Spain and the UK. Temu eventually expanded also into the Latin American market.

The app has ranked No. 1 in both Google Play and Apple stores for much of 2023 [13] thanks to the ads campaigns on social networks such as Instagram, TikTok and YouTube.

Despite the high success, the app is not free from criticism. One common issue is the prevalence of consumer complaints, particularly regarding the quality of products and false product advertising [13]. Additionally, data privacy concerns have emerged as a prominent issue. In May 2023, the United States–China Economic and Security Review Commission expressed apprehensions about the potential risks to users' personal data on Temu. These concerns were heightened following the suspension of Pinduoduo, Temu's sister app in China, from Google Play due to the discovery of malware in certain versions [30].

3 Literature Review

The literature review chapter dives into the methods crucial to this study, highlighting their strengths and weaknesses. It examines the Latent Dirichlet Allocation (LDA) algorithm, showcasing its effectiveness in uncovering hidden patterns in data, yet also discussing its limitations in handling complex relationships within text. Furthermore, it explores the diverse domains where LDA has demonstrated utility, based on past research findings. Similarly, the chapter delves into the Apriori algorithm, emphasizing its strength in identifying frequent itemsets but also acknowledging its challenges in scaling to larger datasets. By analyzing existing literature, this chapter provides insight into how these methods have been applied and their significance in various fields, while also acknowledging the areas where improvements or alternative approaches may be needed.

3.1 Latent Dirichlet Allocation (LDA)

Topic modeling [3] is a natural language processing (NLP) technique used to identify topics present in a collection of texts. It is a type of unsupervised machine learning approach that helps discover hidden thematic structures within a large set of documents. The goal of topic modeling is to automatically extract meaningful topics from a corpus without the need for prior labeling or supervision. Topic modeling finds applications in various fields, including information retrieval, document summarization, content recommendation, and understanding the thematic structure of large text datasets. One of the most popular algorithms for topic modeling is Latent Dirichlet Allocation (LDA)[9]. LDA is a widely used generative probabilistic model used for uncovering latent topics within a collection of documents. At its core, LDA assumes that documents are represented as a mixture of topics, and each topic is characterized by a distribution over words. To estimate its parameters the Expectation-Maximization (EM) algorithm [34] [21] can be used. EM relies on discovering the maximum likelihood estimates of parameters when the data model depends on certain latent variables. EM algorithm contains two steps, the E-step (expectation) and the M-step (maximization). To comprehend how LDA works, it's essential to understand its key elements[9]:

- A word is the fundamental unit of text
- A document is a sequence of words
- A corpus is a collection of documents

Here's a description of the steps of the algorithm [9] [21]:

1. Initialization:
    - Choose the number of topics, K, which is a hyperparameter set by the user.
    - Initialize two matrices:
        - ϕ: Topic-term distribution matrix, where each row represents a topic and each column represents a term in the vocabulary. Initialized randomly.
        - θ: Document-topic distribution matrix, where each row represents a document and each column represents a topic. Initialized randomly.
2. Iterative Optimization:
    - E-Step (Expectation Step):
        - For each document d:
            - For each word w in d:
                - Calculate the probability of each topic k given w and the current state of ϕ and θ.
                - Update the count of w assigned to each topic.
    - M-Step (Maximization Step):
        - Update ϕ and θ to maximize the likelihood of the data given the current assignments of topics to words.
            - Update ϕ using the counts of words assigned to topics in the E-step.
            - Update θ using the counts of words assigned to topics in each document in the E-step.
3. Repeat the E-step and M-step until convergence criteria are met (for example, maximum number of iterations reached or small change in likelihood).
4. Output:
    - ϕ: The final topic-term distribution matrix.
    - θ: The final document-topic distribution matrix.
    - Topics represented by the most probable words in each topic.

The underlying intuition is that documents exhibit multiple topics, where a topic is a multinomial distribution over a fixed vocabulary W: LDA considers documents as mixtures of topics and topics as mixtures of words. The goal of LDA is to automatically discover the topics from a collection of documents. The documents of the collection are modeled as mixtures over K topics, each of which is a multinomial distribution over W. Each topic multinomial distribution ϕ_k is generated by a conjugate Dirichlet prior with parameter β, while each document multinomial distribution θ_d is generated by a conjugate Dirichlet prior with parameter α. Thus, the topic proportions for document d are θ_d, and the word distributions for topic k are ϕ_k. In other words, $\theta_{d,k}$ is the probability of topic k occurring in document d. Respectively, $\phi_{k,w}$ is the probability of word w belonging to topic k [43].

The most crucial parameter that LDA needs is the number of topics specified by the user [15]. LDA assumes that a set of documents (i.e., reviews) consists of multiple topics, and each topic consists of multiple words. LDA can find the relationships between words, then allocate them inside the corresponding topics. However, LDA cannot automatically suggest the number of topics within a collection [11]. Hence, there is a need to find the optimal number of topics, then set it as a parameter before executing the LDA analysis.

Four main methods are often cited for determining the optimal number of topics in a collection. These methods are described in Griffiths and Steyvers (2004) [16], Cao et al. (2009) [11], Arun et al. (2010) [5], and Deveaud et al. (2014) [15]. The general idea is to train multiple LDA models with different numbers of topics and then calculate coherence scores [40] [19] for each model. The number of topics that maximizes the coherence score is often considered the optimal choice. This method is used in the Chapter 4 to determine a suitable number of topics in the reviews.

The topics extracted by the algorithm help to illustrate which of the topics are relevant to the customers. Customers usually only write about their most unforgettable, either positive or negative, experiences with the services. Hence, if a topic is not relevant, reviewers would not write about it often. Consequently, the said topic would not be extracted by the algorithm if it is not discussed frequently.

Nevertheless, LDA still has its limitations. First, LDA is sensitive to preprocessing steps such as tokenization, stop word removal, and stemming. Small changes in preprocessing can lead to significantly different topics being generated.

Second, LDA produces results without the awareness of the meaning. Thus, it still requires human interpretation of the topic and for labeling purposes, like Factor Analysis results [29].

Lastly, the computational complexity of LDA can be prohibitive, particularly with large datasets or a high number of topics, necessitating substantial computational resources for practical applications [3][43].

Grimmer and Stewart [17] stressed that all text analysis techniques are not meant to replace humans but instead to augment humans' abilities. This set of limitations are not exclusive to LDA but should apply to most, if not all, automated text analysis techniques of today.
Previous studies have used LDA in several industries.

- Tourism industry: Song et al. (2020) [42] aimed to understand the perceptions and experiences of citizens of an urban park in New York City.
- Health and safety industry: Bahng and Lee (2020) [6] examined patients' concerns regarding hearing loss and Min et al. (2020)[33] analyzed the issues on occupational accidents.
- Sharing economy industry: Kiatkawsin et al. (2020) [25] provided new findings on Airbnb guests' experiences while Sutherland and Kiatkawsin (2020) [43] studied the relevant topics that drive customer satisfaction in the accommodation sector.
- Social network industry: McCallum et al. (2007) [32] combined the LDA and the Author-Topic (AT) [41] model to get the Author-Recipient-Topic (ART) [32] model for Social Network Analysis. The Idea of ART is to learn topic distributions based on the direction-sensitive messages sent between the senders and receivers.
- Cybersecurity industry: Bergholz, A. et al (2008) [7] developed a new statistical model, the latent Class-Topic Model (CLTOM), which is an extension of LDA, that improves phishing detection.

3.2 Associaton Rules Mining

Association rules are a fundamental concept in data mining introduced by Agrawal et al. (1993) [1] to obtain relevant insights from large transactional databases that might not be immediately apparent through traditional analysis methods.
A transactional database can take various forms depending on the nature of the data being collected and analyzed. A shopping basket database is a classic example, where each transaction corresponds to a customer's purchase, and the items within the transaction are the products bought. This type of database is commonly used for market basket analysis to discover associations and patterns among purchased items.
On the other hand, a text database, as in this case, presents a different perspective. Here, each transaction might represent a document or a piece of text, and the items within the transaction are the words or entities extracted from the text. This configuration is valuable in natural language processing and text mining, enabling the identification of relationships and associations between words within textual data. With the popularity of e-commerce, massive transactional databases are available now [27]. Consequently, text databases containing customer reviews, originated from those transactions, have also become increasingly popular. By mining these databases, businesses can gather valuable feedback on specific products or services, guiding decision-making processes such as product or service development and marketing strategies.
An association rule is represented in the form $(X \to Y)$, where X is an itemset that represents the antecedent, and Y an itemset called consequent where $X \cap Y = \emptyset$. The rule means X implies Y. The classical approach to measure the goodness or significance of the association rules involves calculate three key metrics: support, confidence, and lift.

- Support: The support of an association rule is a measure of how frequently the items or variables involved in the rule appear together in the dataset. It is calculated as the proportion of transactions or instances in the dataset that contain all the items in the rule. Higher support values indicate that the rule is more commonly observed in the dataset, suggesting a stronger association between the items.

  $$\text{Support}(X \to Y) = \frac{\text{Number of transactions containing both } X \text{ and } Y}{\text{Total number of transactions}}$$

  Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item.

- Confidence: Confidence quantifies the strength of an association rule. It measures the proportion of transactions containing the antecedent (the items on the left side of the rule) where the consequent (the items on the right side of the rule) also appears. In other words, confidence indicates the likelihood that the presence of the antecedent implies the presence of the consequent. Higher confidence values signify a stronger predictive power of the rule.

$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \to Y)}{\text{Support}(X)}$$

Suppose the confidence of the association rule $(X \to Y)$ is 80$\%$, it means that 80$\%$ of the transactions that contain X also contain Y together.

- Lift: Lift is a measure that compares the strength of an association rule to what would be expected if the antecedent and consequent were independent of each other. Lift values greater than 1 indicate that the presence of the antecedent increases the likelihood of finding the consequent in a transaction, suggesting a positive association. Lift values equal to 1 indicate independence, while values less than 1 suggest a negative association.

$$\text{Lift}(X \to Y) = \frac{\text{Support}(X \to Y)}{\text{Support}(X)\text{Support}(Y)}$$

These formulas provide a quantitative way to evaluate the strength and significance of association rules in a dataset, aiding in the identification of meaningful patterns and relationships.

Association rules have earned significant attention in Text Mining, as highlighted in [31]. This study explores the utilization of fuzzy association rules on textual transactions. Textual transactions are fundamental for applying association rules to text, opening the possibility of applying association rules to text mining problems. Diverse text entities such as reviews or tweets are treated as transactions, in which each word is an item. This approach enables the extraction of valuable relationships and metrics about co-occurrences in large text databases.

The most widely used algorithm to generate association rules is the A-Priori algorithm [2]. The underlying principles of the A-Priori algorithm are rooted in the exploitation of the monotonicity property, which asserts that if an itemset is frequent, then all its subsets are also frequent. This strategic pruning mechanism not only enhances computational efficiency but also enables the algorithm to scale effectively to large transactional datasets. The A-Priori Algorithm alternates between constructing candidate sets and filtering to find those that are truly frequent as shown in Figure 1.

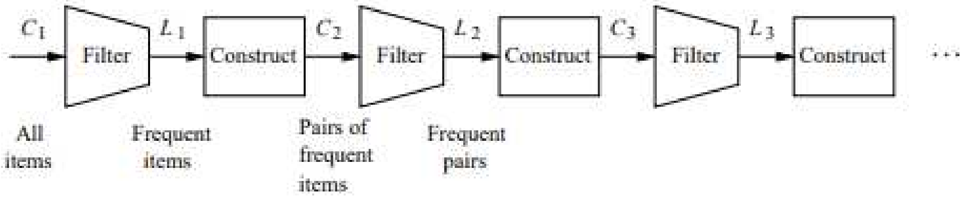


Figure 1: A-Priori algorithm steps

Here a description of the A-Priori algorithm steps [39].

1. Initialize Candidate Itemsets (C1):

    - Start with C1, which contains all singleton itemsets (individual items in the dataset).

2. Generate Frequent Itemsets (L1):

    - Count the occurrences of each item in the dataset.
    - Form L1, the set of frequent items, by selecting items whose counts are at least the support threshold s.

3. Generate Candidate Itemsets (C2):

    - Construct candidate pairs, C2, by forming pairs of items that are both in L1 (frequent items).
    - Test membership in C2 by verifying that both items of each pair are in L1.

4. Generate Frequent Itemsets (L2):

    - Count the occurrences of each candidate pair in the dataset.
    - Determine which candidate pairs appear at least s times to form L2, the set of frequent pairs.

5. Generalize the Process:

    - Continue the process by generating candidate itemsets of size k+1 from frequent itemsets of size k and then finding frequent itemsets of size k+1.
    - For each step, construct candidate itemsets (e.g., triples, quadruples) implicitly based on the frequent itemsets from the previous step.
    - Count the occurrences of candidate itemsets and determine which ones are frequent.

6. Repeat Until Convergence:

    - Repeat the process of generating candidate itemsets and finding frequent itemsets until no more frequent itemsets are found or until reaching the desired itemset size.

7. Find Frequent Itemsets (Lk):

    - For each basket or transaction in the dataset:
        - Look at the items that are in L1 (frequent items).
        - Examine pairs of items to determine if they are in L2 (frequent pairs).
        - Count triples (or larger itemsets) that consist of items from the basket and are candidates in the corresponding candidate set (e.g., C3 for triples).
        - Increment the count of frequent itemsets found in the basket.

8. Output Frequent Itemsets:

    - Return the set of all frequent itemsets found during the iterations of the algorithm.

The Apriori algorithm efficiently identifies frequent itemsets in transaction databases, allowing for the discovery of associations between items. By exploiting the monotonicity property, the algorithm reduces the search space, making it computationally feasible for large datasets. Thanks to this property, the number of possible candidates to evaluate can be reduced. Based on A-Priori algorithm, many new algorithms were designed with some modifications or efficiency improvements [27] [20], [45]].
Association rules are implemented to solve several data mining tasks.

- Summarization

  In their paper, Kacprzyk and Zadrozny (2003) [22] focused on summarizing information from Twitter. The work is based on summarizing threads of conversation on a particular topic, so that the reader can get a reliable and complete idea quickly. A similar proposal came from Phan et al. (2018) [38] that uses association rules to summarize Obama's most important tweets.

- Topic detection:

  In this field, topic modeling techniques such as LDA stands out, but some approximation through association rules can be found. The paper by Cagliero and Fiori (2012) [10] offers a solution through generalized association rules, which provides a compendium of topics used in the social network Twitter. This research uses dynamic association rules that adapt based on the context of the posts and the content of the tweets made by the user. Association rules stand out when being used to obtain knowledge about a specific topic, for example, about cyber bullying as in Zainol et al. 2018 [48]. In this approach, the authors utilize the A-Priori algorithm to uncover specific patterns, aiding in the identification of topics within social networks commonly associated with cyberbullying.

- Sentiment analysis

  In the health field, the paper (Paulose et al. 2018) [37] use association rules mining and natural language processing techniques to mine the social network Twitter for the use of Fentanyl. Following sentiment analysis, they employ association rules to uncover correlations between its usage and other drugs and products posing health risks.

- Collaborative social systems

  User actions in social networks can be used to generate collaborative systems that improve the experience of other users. These systems operate on the assumption that if something proves useful for one user, it's likely to be useful for another user with a similar profile. In [23], Kakulapati and Reddy use the A-Priori algorithm to relate actors and metadata from films to other similar films. It is based on the premise that related films will be a good recommendation for the users.

Association rule mining often poses challenges for end users due to several common issues. One significant problem is the long time it can take for algorithms to produce results. This delay is particularly noticeable when dealing with large datasets or when lowering the frequency thresholds, as the number of association rules generated can quickly become overwhelming. Moreover, as the set of frequent itemsets expands, users are often presented with numerous redundant rules, further complicating the interpretation and usability of the results.
In conclusion, the literature review has provided an overview of the applications and methodologies of both Latent Dirichlet Allocation (LDA) and association rules across various domains. While previous studies have highlighted the individual strengths and limitations of these techniques,there remains a notable gap in research exploring their combined utilization. This presents an opportunity for future analysis to leverage the complementary nature of LDA and association rules. The subsequent chapter in this work will explore the combined application, on a database composed of customer reviews, of these two techniques to address the research questions.

4 Analysis in Python

In this chapter, the focus is on exploring data analysis with Python. It involves several steps to uncover valuable insights. First, the necessary tools are set up by installing and importing required libraries. Then, the data is collected, laying the groundwork for analysis. Subsequently, the data is shaped to facilitate analysis. Exploratory analysis reveals interesting patterns and trends within the data. Before delving deeper, it's essential to clean up the data and prepare it for analysis. Advanced techniques like topic modeling and association rules mining are then applied to gain further insights.

4.1 Install and import libraries

The following libraries were required to provide a toolkit for diverse data analysis, web scraping, natural language processing (NLP) and visualization tasks.

- Pandas and NumPy:

  Pandas is a data manipulation library, and NumPy provides support for large, multi-dimensional arrays and matrices. Together, they facilitate efficient handling, cleaning, and analysis of structured data.

- Google Play Scraper [1]:

  This library is tailored for web scraping the Google Play Store. It provides functionalities like fetching app details, reviews, and other information, making it a valuable tool for extracting data from the Google Play platform.

- Genism:

  Genism is primarily used for topic modeling and document similarity analysis. It provides tools for creating and training topic models, including algorithms like Latent Dirichlet Allocation (LDA). The ”corpora” module helps in building corpora for topic modeling, and the ”Coherence Model” evaluates the coherence of generated topics.

- NLTK (Natural Language Toolkit):

  NLTK is a comprehensive library for natural language processing. In this context, it offers tools for text processing, including tokenization, lemmatization, and stop-words removal. It also provides resources like the WordNet lexical database.

[1] `https://github.com/JoMingyu/google-play-scraper`

- Matplotlib.pyplot and Seaborn:

  Matplotlib is a versatile plotting library, and ”pyplot” provides a MATLAB like interface for creating static plots. Seaborn, built on top of Matplotlib, enhances its aesthetics and adds additional statistical plotting capabilities. Together, they enable the creation of various data visualizations.

- PyLDAvis and PyLDAvis.gensim:

  PyLDAvis is a library for visualizing topic models, and the ”pyLDAvis.gensim” module specifically integrates Gensim’s topic models with PyLDAvis. This combination allows for interactive visualization of topic models, aiding in the interpretation and exploration of complex topics within textual data.

- Efficient Apriori[2]:

  The efficient apriori library is an implementation of the Apriori algorithm, a popular algorithm for association rule mining.

4.2 Data Collection

For this study, customer reviews from the popular Chinese e-commerce named Temu was chosen as the dataset. Reviews from the application were collected from the Google Play Store Website and mined through the Google Play Scraper package in Python. The scraper provides an easier way to scrape data from the Google Play Store by providing APIs without external dependencies.
In particular, the function parameters ‘lang’ and ‘country’ were set equal to ‘en’ and ‘us’ according to the ISO 3166 and ISO 639-1 standards, respectively. Despite the use of these filters, a small set of reviews written in Spanish was scraped and therefore, as shown in the following ‘Data Cleaning and Preprocessing’ section, it was removed.
After scraping the data, these variables were collected:

- Review ID, the unique identifier of the review.
- Username, the name of the user who has written the review.
- User Image, the profile pic of the user.
- Content, the text of the review.
- Score, the score (1 to 5) of the review.
- Thumbs Up Count, number of thumbs up received by the review.

[2] `https://github.com/tommyod/Efficient-Apriori`

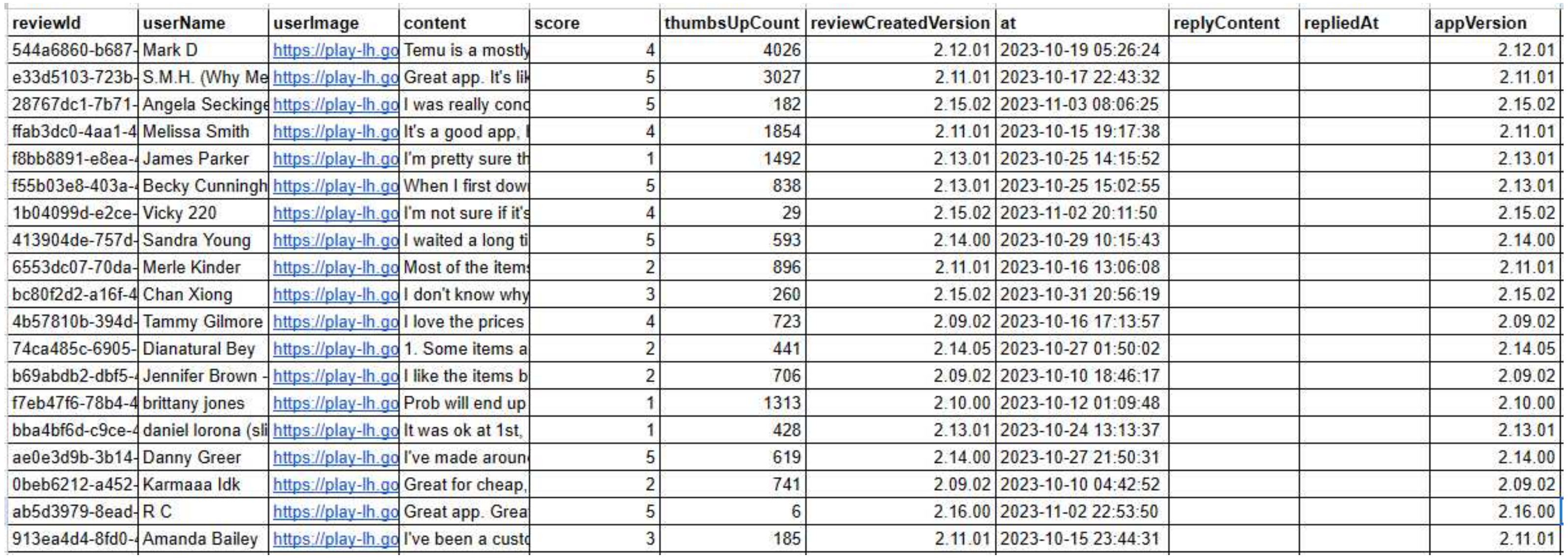
- Review Created Version, app version when the review was written.
- At, date of the review.
- Reply Content, the text written by Temu in response to a review.
- Replied At, date of the Temu's reply.
- App Version, version of the app.

A preview of the dataset is available in the next page in Figure 2.
A summary of the dataset is shown in Table 2:

Application	**Number of Reviews**	**Review Date Period**
Temu	174,554	Sep 2022 – Nov 2023

Table 2: Dataset summary

The review date period goes from the first review available to the first few days of November 2023 (reviews were scraped on 05-11-2023 from Google Play).



reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	appVersion
544a6860-b687-	Mark D	https://play-lh.go	Temu is a mostly	4	4026	2.12.01	2023-10-19 05:26:24			2.12.01
e33d5103-723b-	S.M.H. (Why Me	https://play-lh.go	Great app. It's lik	5	3027	2.11.01	2023-10-17 22:43:32			2.11.01
28767dc1-7b71-	Angela Seckinge	https://play-lh.go	I was really conc	5	182	2.15.02	2023-11-03 08:06:25			2.15.02
ffab3dc0-4aa1-4	Melissa Smith	https://play-lh.go	It's a good app, I	4	1854	2.11.01	2023-10-15 19:17:38			2.11.01
f8bb8891-e8ea-	James Parker	https://play-lh.go	I'm pretty sure th	1	1492	2.13.01	2023-10-25 14:15:52			2.13.01
f55b03e8-403a-	Becky Cunningh	https://play-lh.go	When I first dow	5	838	2.13.01	2023-10-25 15:02:55			2.13.01
1b04099d-e2ce-	Vicky 220	https://play-lh.go	I'm not sure if it's	4	29	2.15.02	2023-11-02 20:11:50			2.15.02
413904de-757d-	Sandra Young	https://play-lh.go	I waited a long ti	5	593	2.14.00	2023-10-29 10:15:43			2.14.00
6553dc07-70da-	Merle Kinder	https://play-lh.go	Most of the items	2	896	2.11.01	2023-10-16 13:06:08			2.11.01
bc80f2d2-a16f-4	Chan Xiong	https://play-lh.go	I don't know why	3	260	2.15.02	2023-10-31 20:56:19			2.15.02
4b57810b-394d-	Tammy Gilmore	https://play-lh.go	I love the prices	4	723	2.09.02	2023-10-16 17:13:57			2.09.02
74ca485c-6905-	Dianatural Bey	https://play-lh.go	1. Some items a	2	441	2.14.05	2023-10-27 01:50:02			2.14.05
b69abdb2-dbf5-	Jennifer Brown -	https://play-lh.go	I like the items b	2	706	2.09.02	2023-10-10 18:46:17			2.09.02
f7eb47f6-78b4-4	brittany jones	https://play-lh.go	Prob will end up	1	1313	2.10.00	2023-10-12 01:09:48			2.10.00
bba4bf6d-c9ce-4	daniel lorona (sli	https://play-lh.go	It was ok at 1st,	1	428	2.13.01	2023-10-24 13:13:37			2.13.01
ae0e3d9b-3b14-	Danny Greer	https://play-lh.go	I've made aroun	5	619	2.14.00	2023-10-27 21:50:31			2.14.00
0beb6212-a452-	Karmaaa Idk	https://play-lh.go	Great for cheap,	2	741	2.09.02	2023-10-10 04:42:52			2.09.02
ab5d3979-8ead-	R C	https://play-lh.go	Great app. Grea	5	6	2.16.00	2023-11-02 22:53:50			2.16.00
913ea4d4-8fd0-	Amanda Bailey	https://play-lh.go	I've been a custo	3	185	2.11.01	2023-10-15 23:44:31			2.11.01

Figure 2: Dataset preview

4.3 Data Manipulation

The data manipulation process is a crucial step involving the transformation and restructuring of raw data to derive meaningful insights. This process is indispensable for obtaining variables in a format conducive to effective data exploration.
In this step, some variables were transformed, and some new variables were added transforming already existing ones. In particular:

- The variable ‘at’ was transformed into datetime format.
- The variable ‘content’ was transformed into string format.
- A new variable ’review_evaluation’ was created by transforming the variable ‘score’. Score values were divided into two groups:
  - Negative, if score value is <= 3
  - Positive, if score value is >= 4

  After labelling all reviews, the results, shown in Table 3, were:

Positive	**Negative**
143,376	31,178

  Table 3: Reviews Evaluation

- A new variable ‘year_month’ was created by transforming the ‘at’ variable.
- A new variable ‘time_slot’ was created by transforming the variable ‘at’. Timestamps in ‘at’ were divided into four slots:
  - Morning, from 6.00 to 11.59
  - Afternoon, from 12.00 to 17.59
  - Evening, from 18.00 to 23.59
  - Night, from 00.00 to 5.59

Therefore, after the data manipulation step, the following variables shown in Table 4 were available:

#	Variable	Non-null count
0	reviewId	174,554
1	userName	174,549
2	userImage	174,554
3	content	174,554
4	score	174,554
5	thumbsUpCount	174,554
6	reviewCreatedVersion	146,614
7	at	174,554
8	replyContent	4,636
9	repliedAt	4,636
10	appVersion	146,614
11	review_evaluation	174,554
12	year_month	174,554
13	time_slot	174,554

Table 4: Non-null count for each variable

4.4 Explorative Data Analysis

Exploratory Data Analysis (EDA) is a crucial phase in the data analysis process that focuses on gaining a deeper understanding of the dataset. It involves employing a variety of statistical and visual techniques to uncover patterns, relationships, and potential outliers within the data.
In this step, both already existing and new variables were exploited to gain useful insights about the data. It was interesting to understand:

- The distribution of the review's score
- The monthly distribution of positive/negative reviews
- The number of reviews per time slot

4.4.1 Reviews Score Distribution

The analysis of reviews scores for the e-commerce platform, shown below in Figure 3, reveals a predominantly positive sentiment, as indicated by the overwhelmingly high number of 5-star ratings, totaling 128,762. Conversely, the relatively lower counts for 1-star (21,384), 2-star (3,968), and 3-star (5,826) ratings suggest that negative experiences or criticisms are comparatively less prevalent. The middle ground is occupied by a notable number of 4-star ratings, totaling 14,614, indicating a substantial degree of contentment among customers. Overall, the distribution of scores emphasizes the overall positive reception of the e-commerce platform, with the majority of users expressing high levels of satisfaction through their 5-star evaluations.

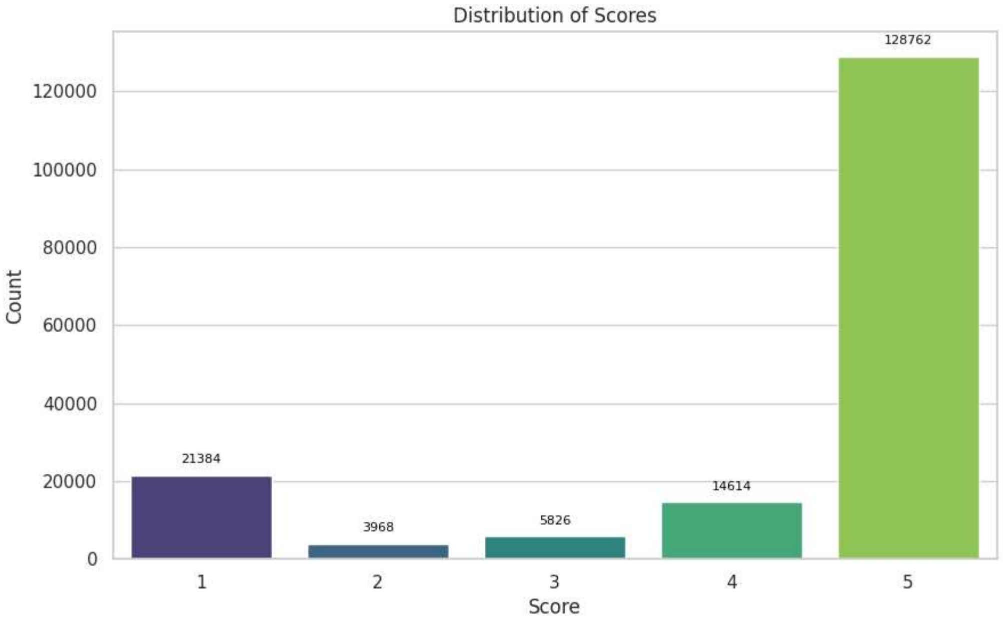


Figure 3: Distribution of scores

4.4.2 Monthly Distribution of Positive and Negative Reviews

The monthly distribution plot of positive and negative reviews, shown below in Figure 4, provides a visual representation of sentiment trends over time. The plot reveals fluctuations in customer sentiment across different months, offering insights into the overall satisfaction levels associated with the services offered by the e-commerce.

Peaks in positive reviews may indicate periods of heightened customer satisfaction, possibly linked to promotions, product launches, or improved service quality. Conversely, spikes in negative reviews might signal issues that need prompt attention, such as product defects, customer service lapses, or marketing missteps.

Analyzing the plot, after the first few months of activity, a significant increase in the number of reviews starting from May 2023 can be observed. Looking at the distribution of positive and negative reviews, the number of positive reviews dominates over the negative reviews in each month.

The monthly distribution of reviews paints a positive picture, with a notable prevalence of positive feedback and an absence of negative peaks. This pattern suggests a consistent and overall positive sentiment among customers. The absence of pronounced fluctuations in negative reviews indicates a sustained satisfaction level, reflecting positively on the overall e-commerce services. This

promising outlook can be leveraged by businesses to reinforce their reputation for delivering a consistently high-quality experience, potentially attracting new customers and fostering loyalty among the existing ones.
This analysis captures the customer feedback over the observed months of e-commerce activity, offering a snapshot of satisfaction trends during this period. However, for a more comprehensive understanding, it is essential to extend the analysis over a more extended timeframe to identify potential seasonality in customer behavior. Long-term observations could reveal recurring patterns, enabling businesses to anticipate and adapt to seasonal fluctuations in customer sentiment, optimizing strategies and resources accordingly.

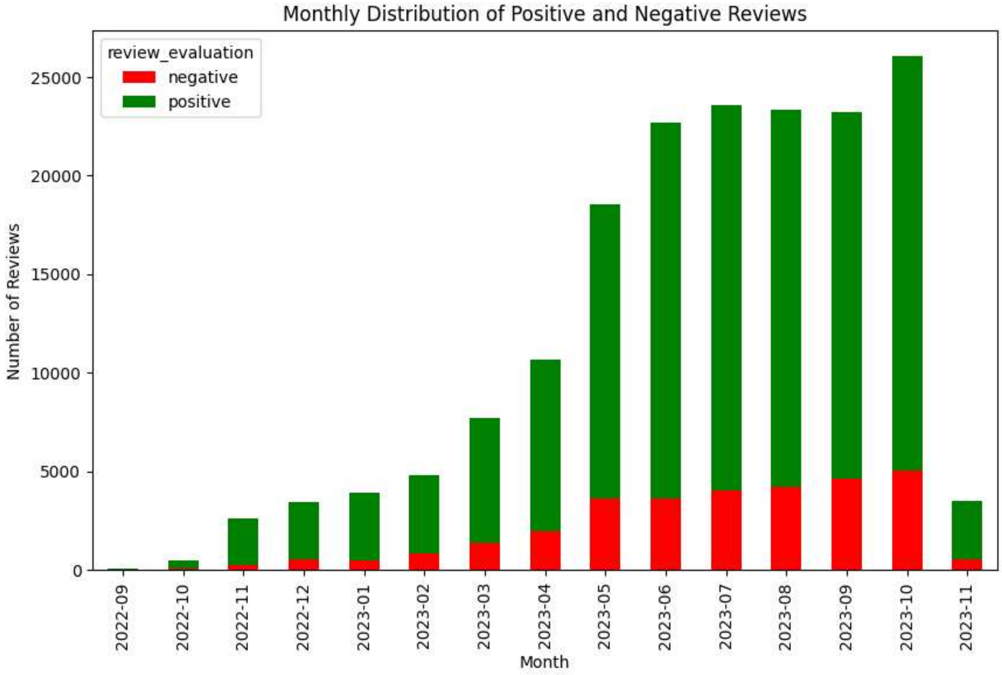


Figure 4: Monthly distribution of positive and negative reviews

4.4.3 Number of reviews per time slot

Understanding the timing of customer reviews within an e-commerce platform holds multiple benefits for effective business management. Firstly, optimizing customer engagement becomes possible by strategically scheduling promotional activities during peak review submission times, enhancing the impact of discounts and product announcements. Efficient allocation of customer support resources is facilitated by identifying high-review periods, ensuring adequate staffing during these crucial hours. For product launches and updates, knowledge of peak user activity enables precise timing, maximizing visibility and adoption rates. Leveraging customer review timings for social media and marketing campaigns enhances visibility and encourages social sharing.

The ability to personalize communication and target users during active periods fosters more meaningful interactions. Identifying recurring review patterns also aids in refining the user experience, addressing concerns at specific times to boost overall satisfaction.

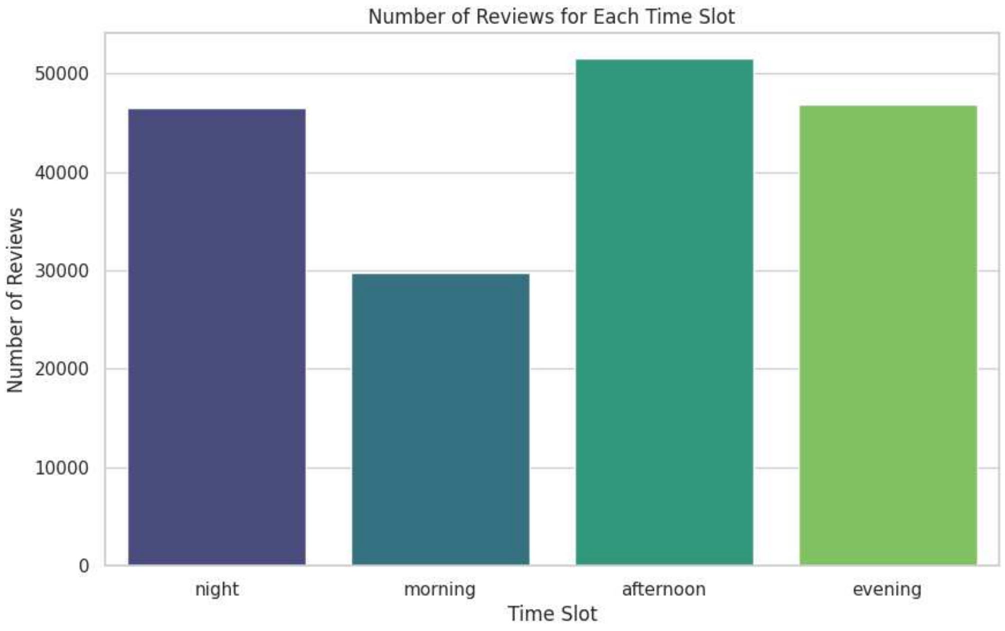
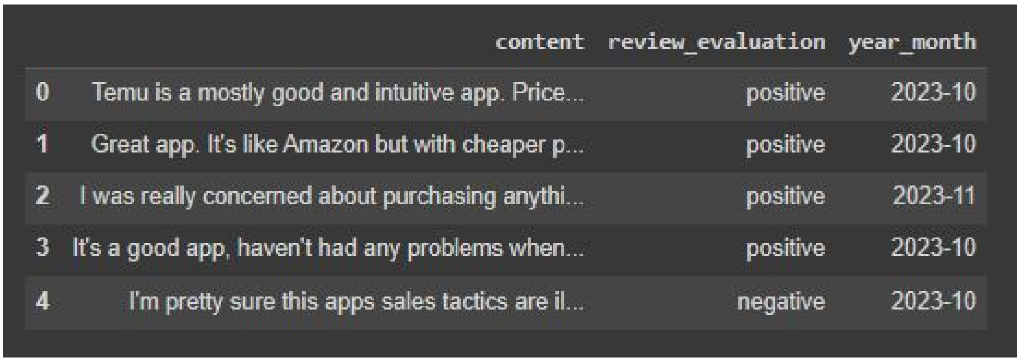


Figure 5: Number of reviews for each time slot

The analysis of review timestamps, shown above in Figure 5, reveals the afternoon dominating as the peak period for customer feedback. This suggests that customers are more actively engaging with the e-commerce platform and expressing their opinions during the afternoon hours. Interestingly, the night and evening time slots exhibit relatively similar review frequencies, indicating sustained engagement throughout the later parts of the day. In contrast, the morning receives the fewest reviews, suggesting a lower level of customer activity during this timeframe. Understanding this chronological pattern can guide strategic decision-making, allowing businesses to focus promotional efforts and customer engagement initiatives during the high-impact afternoon hours. Additionally, allocating resources for customer support and product launches during the more active periods may optimize the overall effectiveness of these endeavors.

4.5 Data Cleaning and Preprocessing

Following the explorative data analysis, certain columns were no more useful and therefore were removed from the dataset. Only three variables were kept as can be seen in Figure 6 below.



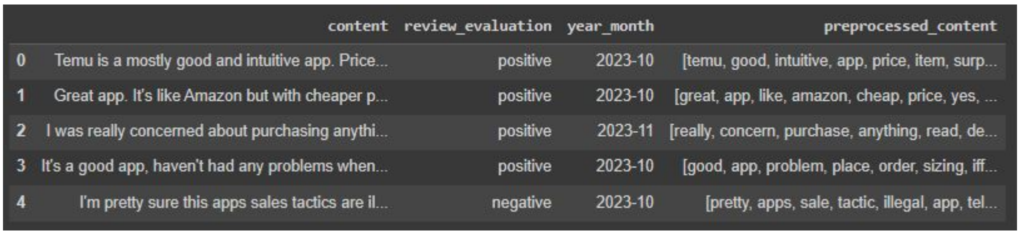
	content	review_evaluation	year_month
0	Temu is a mostly good and intuitive app. Price...	positive	2023-10
1	Great app. It's like Amazon but with cheaper p...	positive	2023-10
2	I was really concerned about purchasing anythi...	positive	2023-11
3	It's a good app, haven't had any problems when...	positive	2023-10
4	I'm pretty sure this apps sales tactics are il...	negative	2023-10

Figure 6: Dataframe

The ’content’ column within the dataset contains raw review text. However, since the goal is to conduct Latent Dirichlet Allocation (LDA), it becomes necessary to preprocess the textual data. This preprocessing step is fundamental to refining the raw text, involving operations such as converting all text to lowercase for uniformity, tokenization to break down the text into meaningful units, and part-of-speech tagging for linguistic insights. Furthermore, the process includes the removal of stop-words, punctuation, and specific words that may introduce noise to the analysis.

The ‘preprocess_text’ function is designed to clean and transform input text for natural language processing (NLP) tasks. First, it converts the entire text to lowercase to ensure uniformity. Next, it tokenizes the text into words and tags each word with its respective part of speech using the Natural Language Toolkit (nltk) library. The function then proceeds to remove punctuation, non-word tokens (since there were emoticons and numbers in the text reviews) and stop words (common words that often do not contribute significant meaning) while lemmatizing the remaining tokens. Additionally, it filters out specific words (the set of Spanish words discussed before) listed in an array, which are deemed irrelevant for the intended analysis. The resulting output is a preprocessed list of meaningful and normalized tokens, ready for further analysis.

After having applied the ‘preprocess_text’ function, the following dataframe in Figure 7 is obtained:



	content	review_evaluation	year_month	preprocessed_content
0	Temu is a mostly good and intuitive app. Price...	positive	2023-10	[temu, good, intuitive, app, price, item, surp...
1	Great app. It's like Amazon but with cheaper p...	positive	2023-10	[great, app, like, amazon, cheap, price, yes, ...
2	I was really concerned about purchasing anythi...	positive	2023-11	[really, concern, purchase, anything, read, de...
3	It's a good app, haven't had any problems when...	positive	2023-10	[good, app, problem, place, order, sizing, iff...
4	I'm pretty sure this apps sales tactics are il...	negative	2023-10	[pretty, apps, sale, tactic, illegal, app, tel...

Figure 7: Dataframe Preprocessed

Below an example of the text that the function receives as input:
[‘Temu is a mostly good and intuitive app. Prices and items are surprisingly good. It’s been great so far with the many orders that I’ve placed and the shipping has always been within the estimates for arrival! However, I am deducting a star because they got rid of the wishlist without explaining why, so the only way to save items is to add them to your cart. That’s disappointing.’]
The output generated is:
[’temu’, ’good’, ’intuitive’, ’app’, ’price’, ’item’, ’surprisingly’, ’good’, ’great’, ’far’, ’many’, ’order’, ’place’, ’shipping’, ’within’, ’estimate’, ’arrival’, ’however’, ’deduct’, ’star’,’get’,’rid’, ’wishlist’, ’without’, ’explain’, ’way’, ’save’, ’item’, ’add’, ’cart’, ’disappointing’]

4.6 Topic Modeling

Next, a function called ‘dictionary_corpus’ has been defined for creating a dictionary and a corpus, which are commonly used components in topic modeling tasks, particularly in the context of Latent Dirichlet Allocation (LDA) to uncover latent topics within the text data.
Below an example of tokenized words and their frequence in the first review:
[[(’add’, 1), (’app’, 1), (’arrival’, 1), (’cart’, 1), (’deduct’, 1), (’disappointing’, 1), (’estimate’, 1), (’explain’, 1), (’far’, 1), (’get’, 1), (’good’, 2), (’great’, 1), (’however’, 1), (’intuitive’, 1), (’item’, 2), (’many’, 1), (’order’, 1), (’place’, 1), (’price’, 1), (’rid’, 1), (’save’, 1), (’shipping’, 1), (’star’, 1), (’surprisingly’, 1), (’temu’, 1), (’way’, 1), (’wishlist’, 1), (’within’, 1), (’without’, 1)]]

4.6.1 Hyperparameter Tuning

Hyperparameter tuning is a crucial step in optimizing the performance of Latent Dirichlet Allocation (LDA). One of the key hyperparameters in LDA is the number of topics, which significantly influences the interpretability and coherence of the resulting topics. Selecting an appropriate number of topics involves a delicate balance, as too few topics may oversimplify the representation of documents, while too many may lead to ambiguity and redundancy. Typically, hyperparameter tuning involves experimenting with different values for the number of topics, such as through coherence scores for varying topic counts. The goal is to identify the number of topics that maximizes coherence values, indicating the most semantically meaningful and coherent topics within the dataset.
Coherence value is a metric used to evaluate the interpretability and quality of topics generated by topic models. It quantifies the degree of semantic similarity among words within the same topic, aiming to measure how well-defined and meaningful the topics are. A higher coherence value suggests that the words within a topic are more closely related, reflecting a coherent and distinctive theme.
Coherence values are used to optimize the number of topics in a model, as higher coherence values typically correspond to more meaningful and easily interpretable topics.
A function named ‘compute_coherence_values’ was defined to train LDA topic models with different numbers of topics and compute coherence values for each model. This function uses the Gensim library for topic modeling.
The function takes as input the following parameters:

- dictionary: Gensim dictionary - A mapping of words to their integer ids.
- corpus: Gensim corpus - The document-term matrix.
- texts: List of input texts - The list of input documents.

- limit: Max num of topics - The maximum number of topics to consider.
- start: Min num of topics - The minimum number of topics to consider.
- step: Step size for the number of topics - The step size between the number of topics.

And Returns:

- model_list: List of LDA topic models - A list containing the trained LDA models.
- coherence_values: Coherence values corresponding to the LDA model with respective number of topics - A list containing coherence values for each model.

4.6.2 Temu Coherence Values

Based on the values below in Table 5 and Figure 8, the highest coherence is reached for 3 topics, suggesting that this number of topics may offer the most interpretable and semantically meaningful representation of the data. However, it's essential to consider the overall context and explore neighboring values, for example the second highest score that is reached at 4 topics, for a more comprehensive understanding of the optimal number of topics.

Number of Topics	Coherence Values
2	0.47130670209153347
3	0.4947196438858761
4	0.4765197444406982
5	0.44617688439206366
6	0.4286070514719385
7	0.4639901160658694
8	0.4133910490630932
9	0.4401114674621025

Table 5: Coherence values for each number of topic

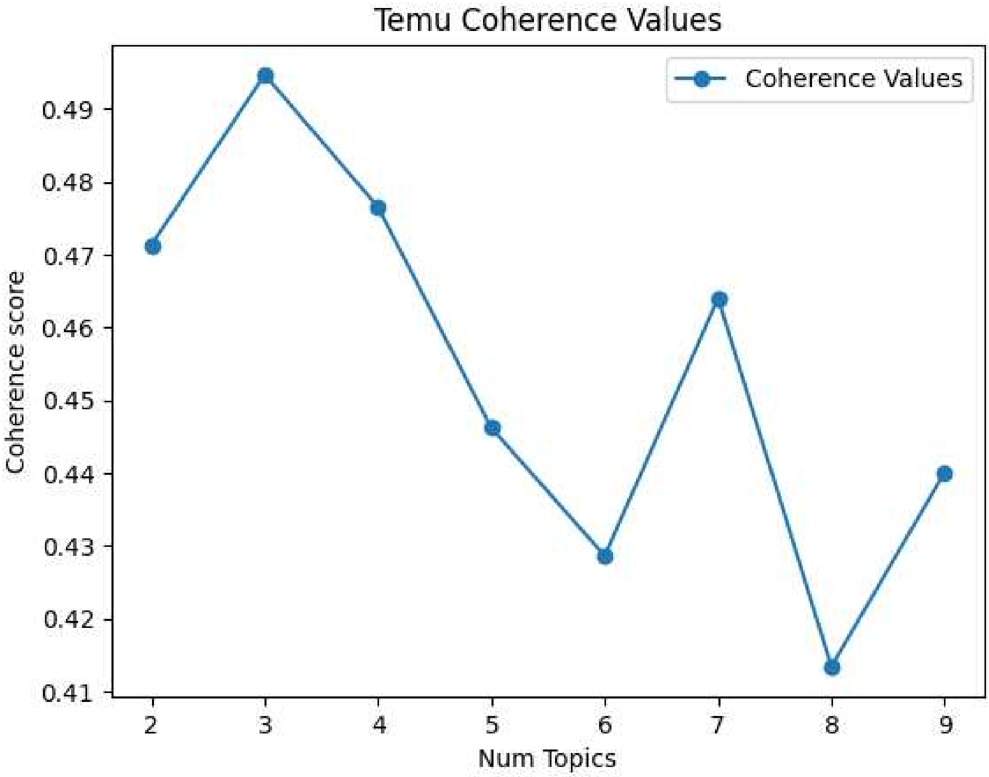


Figure 8: Coherence value scores

4.6.3 Topic Visualization

Visualize the topics is useful for interpretability and for choosing the most meaningful number of topics. To do so, a visualization package, pyLDAvis, is used to help interactively with:

- Better understanding and interpreting individual topics, and
- Better understanding the relationships between the topics.

For the first point, each topic can be manually selected to view its top most frequent and/or "relevant" terms, using different values of the λ parameter. This can help in trying to assign a human interpretable name or "meaning" to each topic.

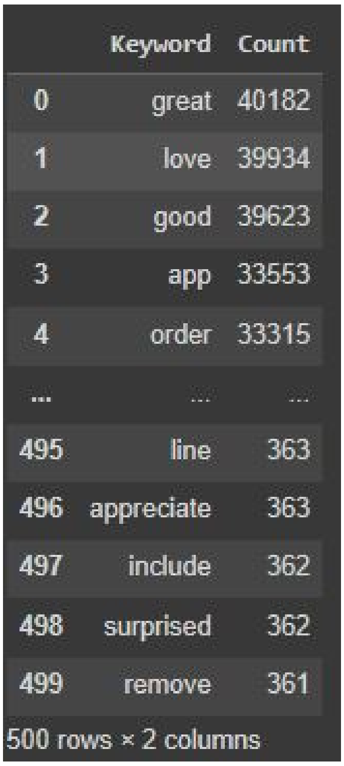
For the second point, exploring the Intertopic Distance Plot can help in learning about how topics relate to each other, including potential higher-level structure between groups of topics.

Within the code, interactive visualizations for LDA models with both 3 and 4 topics can be found. After a careful examination of the terms comprising each topic, the model with 4 topics offers a more robust interpretability of the underlying data.

All observed words were evaluated and interpreted based on the terms contained in each group and were given a topic label.
The obtained topics with relative keywords will be described in the ‘Results’ section.

4.7 Association Rules Mining

Next, the top 500 frequently occurring words were extracted from the reviews and labelled according to its closely resembled e-commerce feature.
Frequent words that were too general and therefore do not resemble any e-commerce feature (for example words as ‘great’, ‘love’, etc...) were not considered in the analysis.
Below, in Figure 9, some keywords and their frequency.

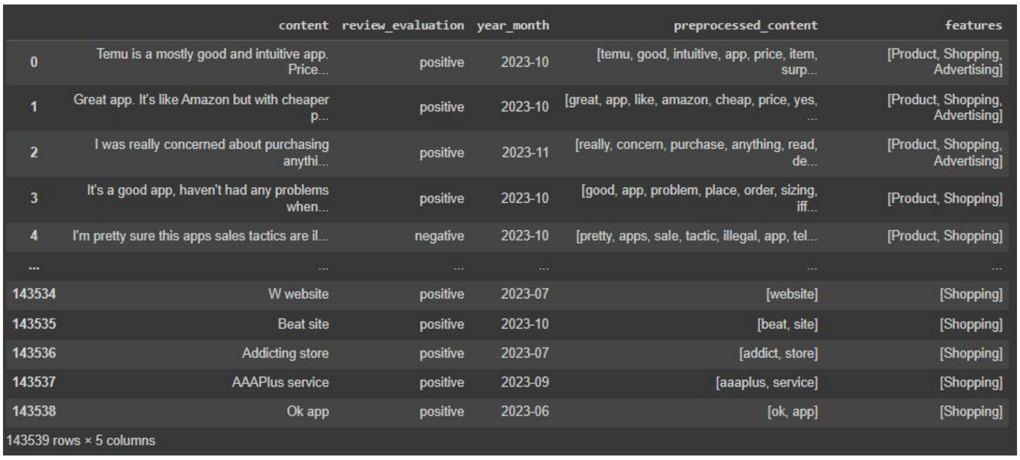


	Keyword	Count
0	great	40182
1	love	39934
2	good	39623
3	app	33553
4	order	33315
...	...	...
495	line	363
496	appreciate	363
497	include	362
498	surprised	362
499	remove	361

500 rows × 2 columns

Figure 9: Keywords

Each keyword was then matched and compared with the words in the ‘preprocessed_content’ column. Only those reviews containing the above mentioned keywords were retained and labelled with the corresponding e-commerce feature. Reviews without keywords in the text were removed. Only 143539 reviews remained out of 174554.
Below, in Figure 10, the obtained DataFrame with the new column ‘features’:



	content	review_evaluation	year_month	preprocessed_content	features
0	Temu is a mostly good and intuitive app. Price...	positive	2023-10	[temu, good, intuitive, app, price, item, surp...	[Product, Shopping, Advertising]
1	Great app. It's like Amazon but with cheaper p...	positive	2023-10	[great, app, like, amazon, cheap, price, yes, ...	[Product, Shopping, Advertising]
2	I was really concerned about purchasing anythi...	positive	2023-11	[really, concern, purchase, anything, read, de...	[Product, Shopping, Advertising]
3	It's a good app, haven't had any problems when...	positive	2023-10	[good, app, problem, place, order, sizing, iff...	[Product, Shopping]
4	I'm pretty sure this apps sales tactics are il...	negative	2023-10	[pretty, apps, sale, tactic, illegal, app, tel...	[Product, Shopping]
...	...	...	...	...	...
143534	W website	positive	2023-07	[website]	[Shopping]
143535	Beat site	positive	2023-10	[beat, site]	[Shopping]
143536	Addicting store	positive	2023-07	[addict, store]	[Shopping]
143537	AAAPlus service	positive	2023-09	[aaaplus, service]	[Shopping]
143538	Ok app	positive	2023-06	[ok, app]	[Shopping]

143539 rows × 5 columns

Figure 10: DataFrame

To acquire association rules the DataFrame is transformed into a list of tuples. Each tuple corresponds to a row in the original DataFrame. The subsequent steps involve extracting the lists of items and labels from each tuple, forming two lists containing the features (for example the features [Product, Shopping, Advertising]) and the review evaluation (for example [Positive]) for each review. These two lists are then combined to be compatible with the A-Priori algorithm. Finally, the A-Priori algorithm is applied to the transactions using the 'apriori' function of the 'efficient apriori' package, with a specified minimum support threshold of 0.001.

This process resulted in the generation of itemsets and rules, providing insights into associations and patterns within the transactions, which will be described and discussed in the next section 'Results'.

5 Results

The chosen LDA model was 4 topics with the second highest coherence score of 0.4765. Table 6 below shows the obtained topics.

Topic Number	Security concerns	Keywords
1	Security concerns	spyware, malware, intrusive, data, steal, fraud...
2	Product Quality	quality, merchandise, item, fit, size...
3	Shopping Experience	app, deal, service, fast, delivery...
4	Advertising	advertising, pyramid, real, false, scammer...

Table 6: Topics

Topic 1 contains words such as ”spyware,” ”malware,” and ”intrusive” highlight the potential threats posed to online transactions and user privacy. The presence of terms like ”steal” and ”fraud” underscores the apprehensions related to unauthorized access and financial deceit. The term ”data” emphasizes the worry about the compromise of personal and sensitive information during online interactions. In summary, this topic revolves around the multifaceted security challenges within the e-commerce landscape, including both technological threats and apprehensions related to data integrity.

Topic 2 shows terms such as ”quality,” ”merchandise,” and ”item” emphasizes the central theme of assessing and ensuring the overall quality and variety of products offered online. The terms ”fit” and ”size” highlight the importance of accurate product descriptions and sizing information, addressing concerns related to customer satisfaction and expectations. This topic suggests a focus on the tangible attributes of goods, emphasizing the need for reliability and precision in product representation to foster consumer trust.

Topic 3 with words as ”app,” ”site,” and ”store” suggests a focus on the platforms through which users engage in online shopping. ”Deal” and ”price ” underscore the significance of competitive pricing and discounts. The terms ”fast”, delivery” and ”shipping” emphasize the importance of prompt and reliable order fulfillment, contributing to a positive customer experience. The inclusion of ”addictive” suggests a consideration of features or strategies that make the online shopping experience engaging and compelling for users. In summary, this topic encompasses a broad spectrum of factors, ranging from platform usability and pricing strategies to delivery efficiency, all of which collectively shape the quality and appeal of the e-commerce shopping experience.

Topic 4 includes terms like ”pyramid,” ”scheme,” and ”scammer” that hint at potential concerns related to fraudulent or deceptive advertising practices, suggesting a need for vigilance in discerning legitimate offerings from dubious ones. The contrasting terms ”real” and ”fake” highlight the ongoing challenge of distinguishing authentic advertisements from misleading ones. The inclusion of ”legit” suggests a consideration for genuine and trustworthy promotional efforts. In essence, this topic encompasses the spectrum of authenticity, trustworthiness, and transparency in the realm of online advertising.

For a further exploration into the aspects commonly perceived as positive or negative, association rule mining was performed to analyze the relationships between the features and corresponding review scores.
All relevant metrics including support, confidence, and lift were calculated. These metrics play a crucial role in determining the strength and significance of the relationships between different features.
In the context of confidence, high values (close to 1) suggest a strong association between the antecedent and the consequent. However, confidence alone doesn’t consider how common the consequent is in general. It’s possible to have a high confidence value even if the overall support of the consequent is high.
Therefore, while confidence provides information about the strength of a rule, it is often used in conjunction with lift. A high-confidence rule with a lift greater than 1 indicates a strong association that is not just due to the general occurrence of the consequent. It suggests that the antecedent has a significant positive impact on the occurrence of the consequent beyond what would be expected by chance.
For the sake of focusing on meaningful and actionable insights, only the association rules with a lift greater than 1 were considered. By filtering rules with lift greater than 1, associations that are genuinely influential and not merely occurring by chance were prioritized. Table 7 below shows the obtained association rules.

LHS	RHS	Support	Confidence	Lift
Security, Advertising	Negative	0.002	0.883	4.599
Security, Shopping, Advertising	Negative	0.013	0.839	4.370
Security	Negative	0.010	0.834	4.347
Security, Shopping	Negative	0.019	0.726	3.784
Product, Security, Shopping, Advertising	Negative	0.023	0.686	3.576
Product, Security	Negative	0.001	0.618	3.220
Product, Shopping	Positive	0.321	0.918	1.136
Product	Positive	0.059	0.900	1.114
Shopping	Positive	0.277	0.886	1.096

Table 7: Association rules sorted by Lift

- $('Security', 'Advertising') \to negative$:

  Customer reviews that discuss both 'Security' and 'Advertising' features are likely to convey a 'negative' sentiment. The strong lift value of 4.599 indicates that the combination of security and advertising concerns significantly influences negative sentiments in customer reviews. This implies that customers may associate security issues with negative perceptions of advertising on the e-commerce platform.

- $('Security', 'Shopping', 'Advertising') \to negative$:

  Reviews mentioning 'Security,' 'Shopping,' and 'Advertising' together are likely to express a 'negative' sentiment. The strong association with a lift value of 4.370 suggests that customers discussing security, shopping, and advertising features concurrently are more inclined to have negative opinions in their reviews. This indicates a potential link between these features and overall dissatisfaction.

- $('Security') \rightarrow negative$:

  Reviews specifically discussing the ’Security’ feature tend to convey a ’negative’ sentiment. The high lift value of 4.347 indicates that customers expressing concerns or experiences related to security features in their reviews are strongly associated with negative sentiments. Security issues seem to play a critical role in shaping negative opinions.

- $('Security','Shopping') \rightarrow negative$:

  Reviews discussing both ’Security’ and ’Shopping’ features are likely to express a ’negative’ sentiment. The lift value of 3.784 suggests a significant association between discussions about security and shopping features with negative sentiments. This implies that security concerns impact the overall shopping experience negatively.

- $('Product','Security','Shopping','Advertising') \rightarrow negative$:

  Reviews discussing ’Product,’ ’Security,’ ’Shopping,’ and ’Advertising’ together are associated with a ’negative’ sentiment. The high lift value of 3.576 indicates a strong correlation between these features and negative sentiments in customer reviews. This suggests that when customers discuss specific products along with security and advertising, they are more likely to express dissatisfaction.

- $('Product','Security') \rightarrow negative$:

  Reviews mentioning both ’Product’ and ’Security’ features are likely to convey a ’negative’ sentiment. The lift value of 3.220 suggests that products associated with security concerns tend to lead to negative opinions in customer reviews. Security issues significantly impact the perceived quality or satisfaction with certain products.

- $('Product','Shopping') \rightarrow positive$:

  Reviews mentioning both ’Product’ and ’Shopping’ features are associated with a ’positive’ sentiment. The strong lift value of 1.136 indicates a positive association between discussions about specific products and the shopping experience. Customers generally have positive experiences when discussing products in the context of their overall shopping experience.

- $('Product') \rightarrow positive$:

  Description: Reviews specifically discussing ’Product’ tend to convey a ’positive’ sentiment. The high lift value of 1.114 suggests a strong positive association between discussions about products and positive sentiments in customer reviews. Customers express satisfaction or positive experiences related to the products they have purchased.

- $('Shopping') \rightarrow positive$:

  Reviews specifically discussing the 'Shopping' feature alone are associated with a 'positive' sentiment. The strong lift value of 1.096 suggests that customers generally express positive sentiments when discussing their overall shopping experiences on the e-commerce platform.

6 Conclusions

Throughout this study, the goal has been to uncover valuable insights hidden within users' reviews of the Temu e-commerce app, with particular attention to addressing the two research questions:

- What are the main topics discussed in the user reviews?
- Which features do users view positively and which negatively?

This chapter addresses these questions, providing a concise overview of the findings and their implications for the e-commerce landscape.
Leveraging Latent Dirichlet Allocation, four key themes have emerged: Security Concerns, Product Quality, Shopping Experience and Advertising. Security Concerns stand out as a top priority, reflecting users' anxieties regarding online safety and privacy. This highlights the importance of implementing robust security measures to safeguard user trust in the digital realm. Product Quality has garnered significant attention, with users stressing the need for accurate and reliable product information, emphasizing the importance of transparency and authenticity in product representation. The Shopping Experience encompasses various elements, including platform usability, pricing strategies, and delivery efficiency, all crucial factors in shaping user satisfaction and loyalty. Lastly, our examination of Advertising reveals concerns surrounding authenticity and transparency, emphasizing the need for ethical advertising practices to maintain consumer trust. Overall, these themes underscore the multifaceted nature of the e-commerce landscape and highlight areas where businesses can focus to enhance user experience and build trust.
Through association rule mining, valuable insights can be gained on how specific features influence the sentiment of user reviews on the e-commerce platform, thereby addressing the second research question. Interestingly, discussions revolving around product quality and the shopping experience consistently garner positive feedback from users. This underscores the paramount importance of ensuring top-notch product offerings and seamless shopping experiences to keep customers happy and engaged. Conversely, security-related topics tend to evoke negative sentiments among users, particularly when discussed alongside advertising or shopping features. This highlights the pressing need for e-commerce platforms to prioritize cybersecurity measures to reassure users

and maintain their trust. Moreover, it underscores the importance of transparent and ethical advertising practices to mitigate negative perceptions and uphold positive review scores.

Looking ahead, there are several opportunities for further development and expansion of this work. One potential direction is the exploration of alternative techniques beyond LDA and association rule mining to gain deeper insights into user sentiments and preferences within e-commerce platforms. Additionally, future research could incorporate comparative analyses, either between different markets for the same e-commerce platform or between different e-commerce platforms altogether. By examining user reviews across various markets or platforms, researchers could identify patterns and trends, highlighting the strengths and weaknesses of each platform. This comparative approach could offer valuable insights for e-commerce companies seeking to refine their strategies, enhance user experiences, and maintain a competitive edge in the dynamic digital landscape.

References

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. “Mining association rules between sets of items in large databases”. In: *SIGMOD Rec.* 22.2 (June 1993), pp. 207–216. ISSN: 0163-5808. DOI: 10.1145/170036.170072. URL: https://doi.org/10.1145/170036.170072.

[2] Rakesh Agrawal and Ramakrishnan Srikant. “Fast Algorithms for Mining Association Rules”. In: *Proc. 20th Int. Conf. Very Large Data Bases VLDB* 1215 (Aug. 2000).

[3] Rubayyi Alghamdi and Khalid Alfalqi. “A Survey of Topic Modeling in Text Mining”. In: *International Journal of Advanced Computer Science and Applications* 6 (Jan. 2015). DOI: 10.14569/IJACSA.2015.060121.

[4] Assem Alhawari, Najadat Hassan, and Raed Shatnawi. “Classification of application reviews into software maintenance tasks using data mining techniques”. In: *Software Quality Journal* 29 (Sept. 2021). DOI: 10.1007/s11219-020-09529-8.

[5] R. Arun et al. “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations”. In: June 2010, pp. 391–402. ISBN: 978-3-642-13656-6. DOI: 10.1007/978-3-642-13657-3_43.

[6] Junghwa Bahng and Chang Heon Lee. “Topic Modeling for Analyzing Patients’ Perceptions and Concerns of Hearing Loss on Social Qamp;A Sites: Incorporating Patients’ Perspective”. In: *International Journal of Environmental Research and Public Health* 17.17 (2020). ISSN: 1660-4601. DOI: 10.3390/ijerph17176209. URL: https://www.mdpi.com/1660-4601/17/17/6209.

[7] André Bergholz et al. “Improved Phishing Detection using Model-Based Features.” In: Jan. 2008.

[8] Fernando Berzal and Nicolfás Matín. “Data mining: concepts and techniques by Jiawei Han and Micheline Kamber”. In: *ACM SIGMOD Record* 31 (June 2002), pp. 66–68. DOI: 10.1145/565117.565130.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent dirichlet allocation”. In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[10] Luca Cagliero and Alessandro Fiori. “Analyzing Twitter User Behaviors and Topic Trends by Exploiting Dynamic Rules”. In: *Behavior Computing: Modeling, Analysis, Mining and Decision*. Ed. by Longbing Cao and Philip S. Yu. London: Springer London, 2012, pp. 267–287. ISBN: 978-1-4471-2969-1. DOI: `10.1007/978-1-4471-2969-1_17`. URL: `https://doi.org/10.1007/978-1-4471-2969-1_17`.

[11] Juan Cao et al. “A density-based method for adaptive LDA model selection”. In: *Neurocomputing* 72.7 (2009). Advances in Machine Learning and Computational Intelligence, pp. 1775–1781. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2008.06.011`. URL: `https://www.sciencedirect.com/science/article/pii/S092523120800372X`.

[12] Chaffey. *Digital Marketing*. Pearson, 2019.

[13] Chow. “The Truth About Temu, the Most Downloaded New App in America”. In: *Time* (2022). URL: `https://time.com/6243738/temu-app-complaints/`.

[14] Conrad. “How Retail App Temu Lures US Shoppers With Mind-Bending Prices: The new ecommerce platform can beat Amazon on price by shipping direct from China. It’s already racing up the charts”. In: *Wired* (2022).

[15] Romain Deveaud, Eric Sanjuan, and Patrice Bellot. “Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval”. In: *Document numérique* 17 (June 2014). DOI: `10.3166/dn.17.1.61-84`.

[16] Thomas L. Griffiths and Mark Steyvers. “Finding scientific topics”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004), pp. 5228–5235. URL: `https://api.semanticscholar.org/CorpusID:15671300`.

[17] Justin Grimmer and Brandon M. Stewart. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. In: *Political Analysis* 21.3 (2013), pp. 267–297. DOI: `10.1093/pan/mps028`.

[18] Pritam Gundecha and Huan Liu. “Mining Social Media: A Brief Introduction”. In: Oct. 2012, pp. 1–17. ISBN: 9780984337835. DOI: `10.1287/educ.1120.0105`.

[19] Yue Guo, Stuart J. Barnes, and Qiong Jia. “Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation”. In: *Tourism Management* 59 (2017), pp. 467–483. ISSN: 0261-5177. DOI: `https://doi.org/10.1016/j.tourman.2016.09.009`. URL: `https://www.sciencedirect.com/science/article/pii/S0261517716301698`.

[20] Jiawei Han and Jian Pei. “Mining frequent patterns by pattern-growth: methodology and implications”. In: *SIGKDD Explor. Newsl.* 2.2 (Dec. 2000), pp. 14–20. ISSN: 1931-0145. DOI: 10.1145/380995.381002. URL: https://doi.org/10.1145/380995.381002.

[21] Hamed Jelodar et al. *Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey*. 2018. arXiv: 1711.04305 [cs.IR].

[22] J. Kacprzyk and S. Zadrozny. “Linguistic summarization of data sets using association rules”. In: *The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ ’03*. Vol. 1. 2003, 702–707 vol.1. DOI: 10.1109/FUZZ.2003.1209449.

[23] V. Kakulapati and S. Mahender Reddy. “Mining Social Networks: Tollywood Reviews for Analyzing UPC by Using Big Data Framework”. In: *Smart Innovations in Communication and Computational Sciences*. Ed. by Shailesh Tiwari et al. Singapore: Springer Singapore, 2019, pp. 323–334. ISBN: 978-981-13-2414-7.

[24] Andreas M. Kaplan and Michael Haenlein. “Users of the world, unite! The challenges and opportunities of Social Media”. In: *Business Horizons* 53.1 (2010), pp. 59–68. ISSN: 0007-6813. DOI: https://doi.org/10.1016/j.bushor.2009.09.003. URL: https://www.sciencedirect.com/science/article/pii/S0007681309001232.

[25] Kiattipoom Kiatkawsin, Ian Sutherland, and Jin-Young Kim. “A Comparative Automated Text Analysis of Airbnb Reviews in Hong Kong and Singapore Using Latent Dirichlet Allocation”. In: *Sustainability* 12.16 (2020). ISSN: 2071-1050. DOI: 10.3390/su12166673. URL: https://www.mdpi.com/2071-1050/12/16/6673.

[26] En-Gir Kim and Se-Hak Chun. “Analyzing Online Car Reviews Using Text Mining”. In: *Sustainability* 11.6 (2019). ISSN: 2071-1050. DOI: 10.3390/su11061611. URL: https://www.mdpi.com/2071-1050/11/6/1611.

[27] Sotiris B. Kotsiantis and Dimitris N. Kanellopoulos. “Association Rules Mining: A Recent Overview”. In: 2006. URL: https://api.semanticscholar.org/CorpusID:671244.

[28] Kenneth C. Laudon and Carol Guercio Traver. *E-commerce 2021-2022: Business, Technology, Society*. Pearson Education Limited, 2022.

[29] D. N. Lawley and A. E. Maxwell. “Factor Analysis as a Statistical Method”. In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 12.3 (1962), pp. 209–229. ISSN: 00390526, 14679884. URL: http://www.jstor.org/stable/2986915 (visited on 02/24/2024).

[30] KGIII March. *Google suspends Chinese E-commerce app pinduoduo over malware*. Mar. 2023. URL: https://krebsonsecurity.com/2023/03/google-suspends-chinese-e-commerce-app-pinduoduo-over-malware/.

[31] M. J. Martín-Bautista et al. "Text Mining using Fuzzy Association Rules". In: *Fuzzy Logic and the Internet*. Ed. by Vincenzo Loia, Masoud Nikravesh, and Lotfi A. Zadeh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 173–189. ISBN: 978-3-540-39988-9. DOI: 10.1007/978-3-540-39988-9_9. URL: https://doi.org/10.1007/978-3-540-39988-9_9.

[32] Andrew Mccallum, Andrés Corrada-Emmanuel, and Xuerui Wang. "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email". In: *Andrew McCallum* 30(1) (Jan. 2005).

[33] Kyoung-Bok Min, Sung-Hee Song, and Jin-Young Min. "Topic Modeling of Social Networking Service Data on Occupational Accidents in Korea: Latent Dirichlet Allocation Analysis". In: *J Med Internet Res* 22.8 (Aug. 2020), e19222. ISSN: 1438-8871. DOI: 10.2196/19222. URL: http://www.ncbi.nlm.nih.gov/pubmed/32663156.

[34] T.K. Moon. "The expectation-maximization algorithm". In: *IEEE Signal Processing Magazine* 13.6 (1996), pp. 47–60. DOI: 10.1109/79.543975.

[35] Murray. "What To Know About Temu: New Chinese-Owned Fast Fashion App Draws Comparisons (Good And Bad) To Shein". In: *Forbes* (2023). URL: https://www.forbes.com/sites/conormurray/2023/02/17/what-to-know-about-temu-new-chinese-owned-fast-fashion-app-draws-comparisons-good-and-bad-to-shein/?sh=351587c23f14.

[36] Dennis Pagano and Walid Maalej. "User feedback in the appstore: An empirical study". In: *2013 21st IEEE International Requirements Engineering Conference (RE)*. 2013, pp. 125–134. DOI: 10.1109/RE.2013.6636712.

[37] Renjith Paulose, B. Gopal Samy, and K. Jegatheesan. "Text mining and Natural Language Processing on Social Media Data giving Insights for Pharmacovigilance: A Case Study with Fentanyl". In: *Indian Journal of Pharmaceutical Sciences* 80 (2018), pp. 762–766. URL: https://api.semanticscholar.org/CorpusID:56433109.

[38] Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. “A Tweet Summarization Method Based on Maximal Association Rules”. In: *Computational Collective Intelligence*. Ed. by Ngoc Thanh Nguyen et al. Cham: Springer International Publishing, 2018, pp. 373–382. ISBN: 978-3-319-98443-8.

[39] Rajaraman and Ullman. *Mining of Massive Data Sets*. University of Standford, 2013.

[40] Michael Röder, Andreas Both, and Alexander Hinneburg. “Exploring the Space of Topic Coherence Measures”. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM ’15. Shanghai, China: Association for Computing Machinery, 2015, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

[41] Michal Rosen-Zvi et al. *The Author-Topic Model for Authors and Documents*. 2012. arXiv: 1207.4169 [cs.IR].

[42] Yang Song, Jessica Fernandez, and Tong Wang. “Understanding Perceived Site Qualities and Experiences of Urban Public Spaces: A Case Study of Social Media Reviews in Bryant Park, New York City”. In: *Sustainability* 12.19 (2020). ISSN: 2071-1050. DOI: 10.3390/su12198036. URL: https://www.mdpi.com/2071-1050/12/19/8036.

[43] Ian Sutherland et al. “Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation”. In: *Sustainability* 12 (2020), p. 1821. URL: https://api.semanticscholar.org/CorpusID:216266816.

[44] Pang-Ning Tan et al. *Introduction to Data Mining (2nd Edition)*. 2nd. Pearson, 2018. ISBN: 0133128903.

[45] Hannu Toivonen. “Sampling Large Databases for Association Rules”. In: *Proceedings of the 22th International Conference on Very Large Data Bases*. VLDB ’96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996, pp. 134–145. ISBN: 1558603824.

[46] Yuren Wang, Xin Lu, and Yuejin Tan. “Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines”. In: *Electronic Commerce Research and Applications* 29 (2018), pp. 1–11. ISSN: 1567-4223. DOI: https://doi.org/10.1016/j.elerap.2018.03.003. URL: https://www.sciencedirect.com/science/article/pii/S1567422318300279.

[47] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. USA: Cambridge University Press, 2014. ISBN: 1107018854.

[48] Zuraini Zainol et al. “Association Analysis of Cyberbullying on Social Media using Apriori Algorithm”. In: *International Journal of Engineering and Technology* 7 (Nov. 2018), pp. 72–75. DOI: 10.14419/ijet.v7i4.29.21847.