Applied Text Mining Research Project

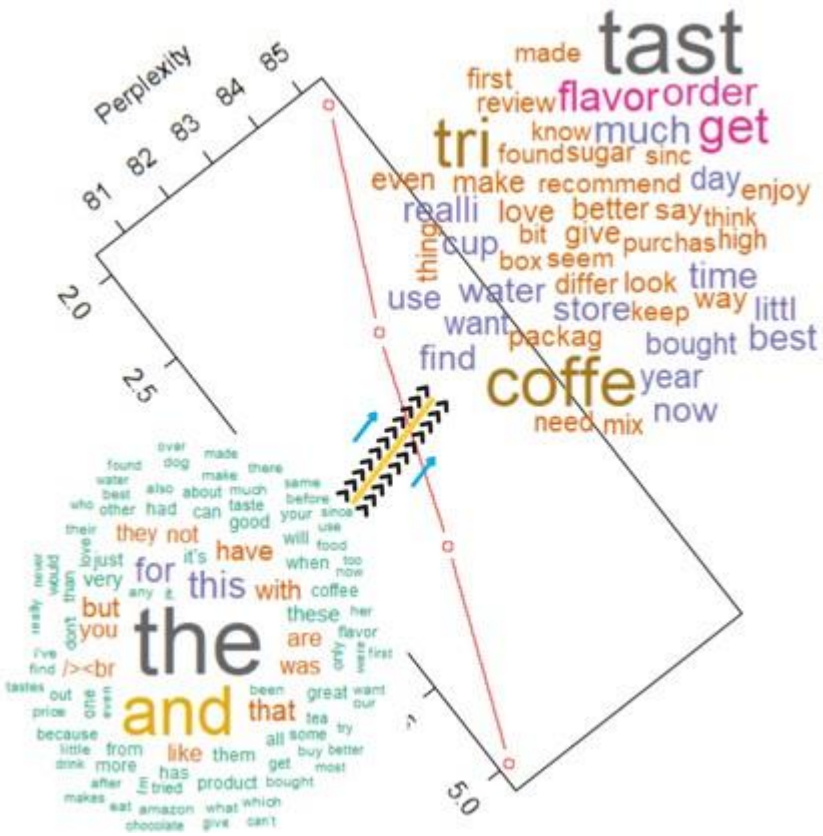


Table of Contents

1. Introduction

Understanding customer opinions and market trends is critical for businesses to make informed decisions making. Online shopping platforms like Amazon get thousands of reviews every day. These reviews contain important information about product quality, features, and customer experiences. Manually reading all reviews is difficult because there is too much text. This is where text mining come in for business analytics. Text mining helps us automatically analyse large amounts of unstructured text data. It can find patterns, trends, and important topics in customer feedback.

The main goal of this project is **to analyse Amazon product reviews to** find out what customers talk about most. We also want to see if customers are happy or unhappy with the products. For doing this, we use **Latent Dirichlet Allocation (LDA)** to discover the main topics discussed in reviews, and **sentiment analysis** to understand customer opinions as positive, negative, or neutral.

This study is an *application of text mining on market intelligence*, which is a broader business context where companies use data from product reviews, social media, or market reports to make strategic decisions. The Market intelligence also helps businesses improve products, services, and customer satisfaction by transforming unstructured text into actionable insights.

The dataset used for this project is the Amazon Product Reviews dataset available on Kaggle[1]. It has many product reviews, which are suitable for topic modeling and sentiment analysis in R.

The main objectives of this project are:

- To identify the main topics discussed by customers.
- To analyze sentiment trends to understand customer satisfaction.
- To provide useful insights for improving products and business decisions, and marketing strategies.

We chose this approach because Amazon reviews are easy to access and provide detailed information about customer experiences. Text mining in R is also simple and effective for working with large datasets. It allows us to automatically extract useful insights from a large amount of text data. This process supports market intelligence, which is very important for business planning and decision-making. In this project, all the analysis is done using R, including data cleaning, preprocessing, topic modeling, sentiment analysis, and visualization.

[1] *Amazon Product Reviews dataset, Arham-Rumi, 2023.*

2. Literature Review

Text mining is now very common in marketing and business research. It helps companies understand what customers think and what trends are happening in the market. By using text mining, businesses can turn a lot of customer comments into useful information. In this assignment, we also use text mining to study customer reviews. To understand how others used it, we looked at two research papers. Both papers used text mining to study product reviews and customer opinions. Their work is very similar to this assignment, which uses topic modeling and sentiment analysis on Amazon product reviews.

Study 1: Understanding Consumer Sentiments in E-Commerce Reviews

Reference:

Walaa Medhat, Ahmed Hassan, Hoda Korashy. (2014). *Sentiment analysis algorithms and applications: A survey.* Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Background:

This study reviewed how sentiment analysis, a branch of text mining, is used to analyze customer reviews and opinions across e-commerce platforms like Amazon and Yelp. The goal was to help companies understand what makes customers happy or unhappy.

Dataset and Method:

The study used many Amazon product reviews. It tested two ways to study the text:

1) *lexicon-based* and

2) *machine learning* methods. The text was cleaned by removing stopwords, splitting sentences into words (tokenization), and using stemming to make words simpler.

Findings and Conclusions:

The study found that machine learning models like Support Vector Machines (SVM) and Naïve Bayes gave higher accuracy in detecting customer sentiment compared to simple lexicon-based methods. These models understood customer emotions more accurately. The study showed that analyzing online reviews provides valuable information about product quality, brand reputation, and consumer needs.

Implications for This Project:

This study helps our project because it shows that using sentiment analysis is useful to understand customer opinions. It also reminds us to clean and prepare the text carefully before analysis, which we did in R.

.

Study 2: Analyzing Online Reviews to Understand Product Returns

Reference:

Andrea Mor, Carlotta Orsenigo, Mauricio Soto Gomez, Carlo Vercellis (2024). *Shaping the causes of product returns: Topic modeling on online customer reviews.* Electronic Commerce Research. https://doi.org/10.1007/s10660-024-09901-x

Background:

This study focused on understanding why customers return products by analyzing online reviews. he authors wanted to know what problems make people unhappy and returns. They used text mining and topic modeling to find common patterns in customer feedback. The study helps businesses improve products and reduce returns.

Dataset and Method:

The researchers collected thousands of customer reviews from multiple e-commerce platforms across different product categories. They used LDA topic modeling to find main topics, like product quality, delivery problems, and wrong descriptions. They also analyzed which topics were common in returned products.

Findings and Conclusions:

The study found that product quality, wrong sizes, and delivery issues were the main causes of returns. Topic modeling helped summarize many reviews automatically from large text datasets. The authors said text mining can help businesses understand problems and fix them.

Implications for This Project:

This study is useful for our assignment because we also use LDA topic modeling on Amazon reviews. It shows that topic modeling can reveal hidden patterns and main issues in customer feedback. The approach confirms that analyzing large text data helps businesses gain. The study reminds us to clean text carefully and show topics clearly for better understanding.

Both studies show that text mining is very useful. It helps to study many customer reviews at a time. They used methods like topic modeling and sentiment analysis. These methods can find patterns in text. This helps businesses make better decisions. These findings helped our assignment. We used LDA in R to find main topics from Amazon reviews and interpret customer opinions. By comparing with previous studies, this project contributes a practical

example of using open-source **R tools** for text mining. This makes it easier for businesses to perform low-cost market intelligence analysis using publicly available data.

3. Methodology

3.1. Business Understanding

Understanding customer feedback is very important for product quality, improving customer satisfaction, and helping smart business decisions. By analyzing product reviews, companies can:

- Detect common problems or concerns that customers mention.
- Identify key product strengths highlighted by customers
- Support marketing strategies based on what customers prefer.
- Make recommendations to improve products quality and services.

The main audience for this report includes marketing teams, product managers, and data analysts who can use these insights to make better decisions.

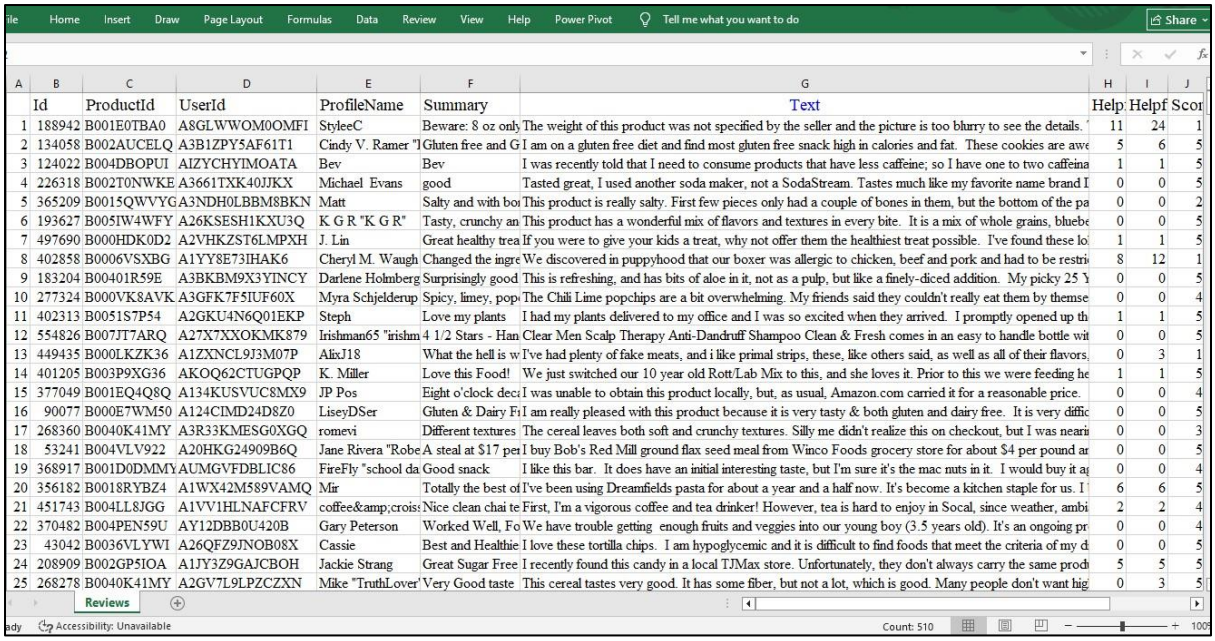
3.2. Data Understanding

The dataset used for this assignment is the **Amazon Product Reviews dataset**, available on Kaggle. It contains many product reviews, which are suitable for **topic modeling** and **sentiment analysis** in R.

For this assignment, **510 reviews were randomly sampled** from the full dataset to make the analysis manageable while retaining variety in content. Each review contains a **Text column**, representing the customer's written feedback.

Key observations:

- Some reviews have missing text, which were replaced with empty strings.
- Reviews vary in **length, style, and detail**.
- The text needs **pre-processing** to clean and standardize it for analysis.



	Id	ProductId	UserId	ProfileName	Summary	Text	Help	Helpf	Scor
1	188942	B001E0TBA0	A8GLWWOM0OMFI	StyleeC	Beware: 8 oz only	The weight of this product was not specified by the seller and the picture is too blurry to see the details.	11	24	1
2	134058	B002AUCELQ	A3B1ZPY5AF61T1	Cindy V. Ramer "J	Gluten free and G	I am on a gluten free diet and find most gluten free snack high in calories and fat. These cookies are awe	5	6	5
3	124022	B004DBOPUI	AIZYCHYIMOATA	Bev	Bev	I was recently told that I need to consume products that have less caffeine; so I have one to two caffeina	1	1	5
4	226318	B002T0NWKE	A3661TXK40JJKX	Michael Evans	good	Tasted great, I used another soda maker, not a SodaStream. Tastes much like my favorite name brand I	0	0	5
5	365209	B0015QWVYG	A3NDH0LBBM8BKN	Matt	Salty and with bo	This product is really salty. First few pieces only had a couple of bones in them, but the bottom of the pa	0	0	2
6	193627	B005IW4WFY	A26KSESH1KXU3Q	K G R "K G R"	Tasty, crunchy an	This product has a wonderful mix of flavors and textures in every bite. It is a mix of whole grains, bluebe	0	0	5
7	497690	B000HDK0D2	A2VHKZST6LMPXH	J. Lin	Great healthy trea	If you were to give your kids a treat, why not offer them the healthiest treat possible. I've found these lo	1	1	5
8	402858	B0006VSXBG	A1YY8E73IHAK6	Cheryl M. Waugh	Changed the ingre	We discovered in puppyhood that our boxer was allergic to chicken, beef and pork and had to be restri	8	12	1
9	183204	B00401R59E	A3BKBM9X3YINCY	Darlene Holmberg	Surprisingly good	This is refreshing, and has bits of aloe in it, not as a pulp, but like a finely-diced addition. My picky 25 Y	0	0	5
10	277324	B000VK8AVK	A3GFK7F5IUF60X	Myra Schjelderup	Spicy, limey, pop	The Chili Lime popchips are a bit overwhelming. My friends said they couldn't really eat them by themse	0	0	4
11	402313	B0051S7P54	A2GKU4N6Q01EKP	Steph	Love my plants	I had my plants delivered to my office and I was so excited when they arrived. I promptly opened up th	1	1	5
12	554826	B007JT7ARQ	A27X7XXOKMK879	Irishman65 "irishm	4 1/2 Stars - Han	Clear Men Scalp Therapy Anti-Dandruff Shampoo Clean & Fresh comes in an easy to handle bottle wit	0	0	5
13	449435	B000LKZK36	A1ZXNCL9J3M07P	AlixJ18	What the hell is w	I've had plenty of fake meats, and i like primal strips, these, like others said, as well as all of their flavors,	0	3	1
14	401205	B003P9XG36	AKOQ62CTUGPQP	K. Miller	Love this Food!	We just switched our 10 year old Rott/Lab Mix to this, and she loves it. Prior to this we were feeding he	1	1	5
15	377049	B001EQ4Q8Q	A134KUSVUC8MX9	JP Pos	Eight o'clock deci	I was unable to obtain this product locally, but, as usual, Amazon.com carried it for a reasonable price.	0	0	4
16	90077	B000E7WM50	A124CIMD24D8Z0	LiseyDSer	Gluten & Dairy Fi	I am really pleased with this product because it is very tasty & both gluten and dairy free. It is very diffic	0	0	5
17	268360	B0040K41MY	A3R33KMESG0XGQ	romevi	Different textures	The cereal leaves both soft and crunchy textures. Silly me didn't realize this on checkout, but I was neari	0	0	3
18	53241	B004VLV922	A20HKG24909B6Q	Jane Rivera "Robe	A steal at $17 per	I buy Bob's Red Mill ground flax seed meal from Winco Foods grocery store for about $4 per pound ar	0	0	5
19	368917	B001D0DMMY	AUMGVFDBLIC86	FireFly "school da	Good snack	I like this bar. It does have an initial interesting taste, but I'm sure it's the mac nuts in it. I would buy it ag	0	0	4
20	356182	B0018RYBZ4	A1WX42M589VAMQ	Mir	Totally the best of	I've been using Dreamfields pasta for about a year and a half now. It's become a kitchen staple for us. I t	6	6	5
21	451743	B004LL8JGG	A1VV1HLNAFCFRV	coffee&crois	Nice clean chai te	First, I'm a vigorous coffee and tea drinker! However, tea is hard to enjoy in Socal, since weather, ambi	2	2	4
22	370482	B004PEN59U	AY12DBB0U420B	Gary Peterson	Worked Well, Fo	We have trouble getting enough fruits and veggies into our young boy (3.5 years old). It's an ongoing pr	0	0	4
23	43042	B0036VLYWI	A26QFZ9JNOB08X	Cassie	Best and Healthie	I love these tortilla chips. I am hypoglycemic and it is difficult to find foods that meet the criteria of my di	0	0	5
24	208909	B002GP5IOA	A1JY3Z9GAJCBOH	Jackie Strang	Great Sugar Free	I recently found this candy in a local TJMax store. Unfortunately, they don't always carry the same prodi	5	5	5
25	268278	B0040K41MY	A2GV7L9LPZCZXN	Mike "TruthLover"	Very Good taste	This cereal tastes very good. It has some fiber, but not a lot, which is good. Many people don't want hig	0	3	5

Figure 1: Sample Records from Amazon Product Reviews Dataset (Kaggle, 2023)

3.3. Data Preparation

Before performing text mining, the review texts were cleaned and processed to ensure quality analysis. The main steps include:

1. **Text Cleaning:**
    - Convert all text to **lowercase**.
    - Remove **punctuation, numbers, and extra whitespace**.
    - Remove common **stopwords** (e.g., “the”, “and”) and custom words that are not meaningful for analysis (e.g., “product”, “amazon”).
    - Apply **stemming** to reduce words to their root form (e.g., “running” → “run”).

Text Cleaning	Example Review Text
Before Cleaning	[1] "arrived on time, and is a great tasting item, will order it again and again. convinient too!!!!!!" [2] "This product is fantastic. It tastes like a sweet snack but is 100% apples."
After Cleaning	"arriv time tast item order convini" "fantast tast sweet snack appl"

Table 1: Before and After Cleaning table

2. **Corpus Creation:**
    - The reviews were converted into a **text corpus** using the tm package in **R**
    - Convert the cleaned text into a **corpus**, which is a collection of documents used for text analysis.

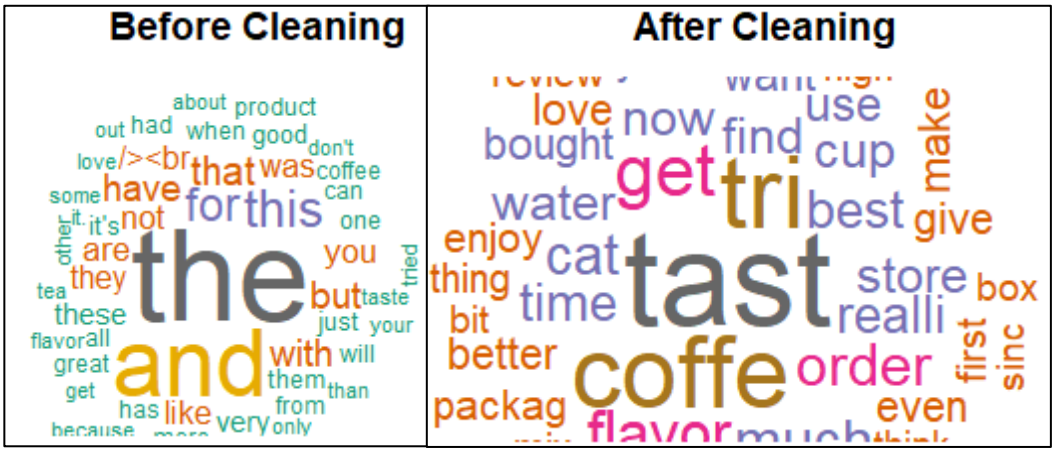


Figure 2: Word Cloud of Most Frequent Words Before and After Cleaning

3. **Document-Term Matrix (DTM):**

A Document-Term Matrix (DTM) was created where each row represents a review and each column represents a word. Sparse terms that appear in very few reviews were removed using a 0.95 threshold to reduce noise, and any empty documents were deleted to ensure all rows contain text. These steps make the DTM clean and standardized, ready for topic modeling (LDA), sentiment analysis, and visualizations such as word clouds.

```
inspect(dtm_sparse[1:5, 1:10])
```

	Terms									
Docs	actual	add	alway	back	bad	best	better	bit	bought	box
1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0

Figure 3: Sample of Document-Term Matrix (First 5 Documents × 10 Terms)

4. Exploratory Analysis

This section explores the key patterns in the cleaned Amazon review data using text mining techniques. The purpose is which words and phrases appear most often, how they are linked,

and what feelings customers show in their reviews. These analyses help identify the main topics and emotions before applying topic modeling.

4.1. Term Frequency & Wordclouds

Term frequency analysis shows which words appear the most in customer reviews. Words like “taste,” “flavor,” and “coffee” come up many times. This means customers mostly talk about the product’s quality and flavor. A wordcloud was created to visualize these frequent terms, where larger words indicate higher frequency (see Figure 4 and Figure 5).

```
top_10_terms
tast tri coffe get flavor order much time best realli
 229 173   172 114     93    86   83   75   74     73
```

Figure 4: Top Terms by Frequency

```
set.seed(123)
wordcloud(names(term_freq), term_freq, max.words = 50, colors = brewer.pal(8, "Dark2"))
```

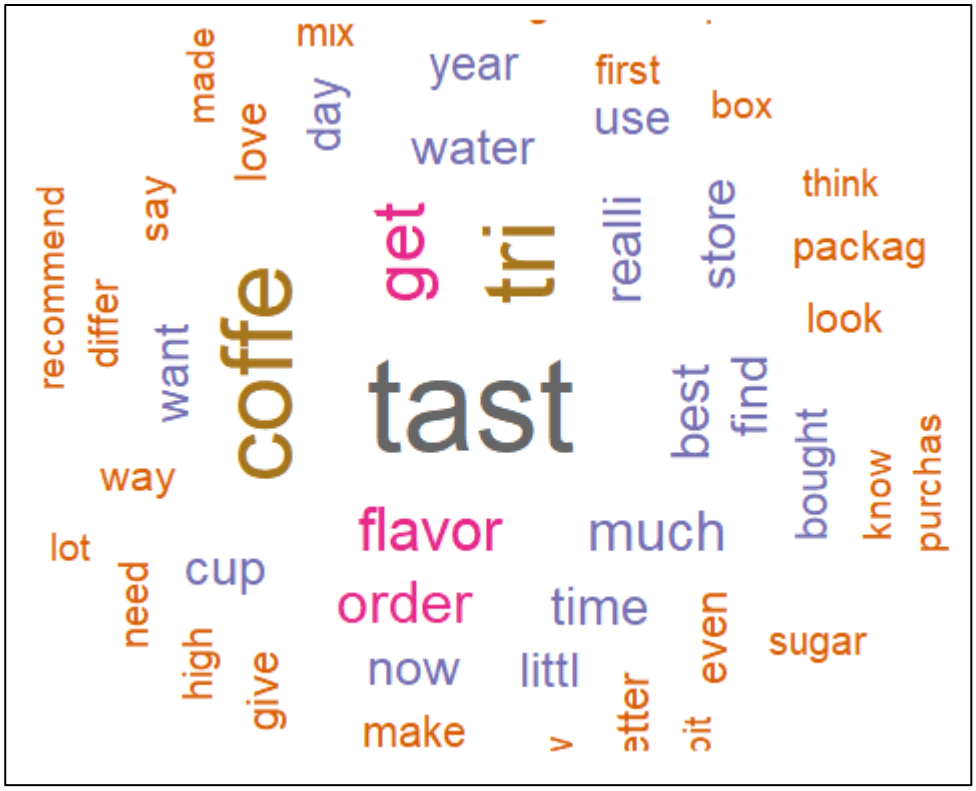


Figure 5: Wordcloud showing the most frequent words in the cleaned Amazon reviews

4.2. Bigrams

Bigrams are pairs of words that often appear together in reviews, Its help us understand how words are connected. For example, common pairs such as “read review”, “high recommend”, “much better”, “littl bit”, “tast better” or “cup coffe” can provide more context than single words. This analysis helps identify key product issues or customer preferences mentioned in the reviews.

Top Bigrams
groceri store
high recommend
cup coffe
gluten free
peanut butter
year ago
give tri
green mountain
hard find
subscrib save
groceri store
high recommend
give tri
green mountain
hard find
subscrib save
dark roast
go back
br br
im sure
much better
read review

Table 2: Top 20 most frequent bigrams

4.3. Sentiment Analysis

Sentiment analysis was done to find out if the customer reviews were positive, negative, or neutral. The sentiment scores were calculated using the “bing” method. The results show that most reviews are positive, with only a few negative ones. The histogram shows how sentiment scores are distributed across reviews. This analysis helps understand overall customer satisfaction and emotional tone in the dataset.

```
> summary(sentiment_scores)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.0000  0.0000  1.0000  0.9431  2.0000 10.0000
```

Figure 6: Summary of Sentiment analysis

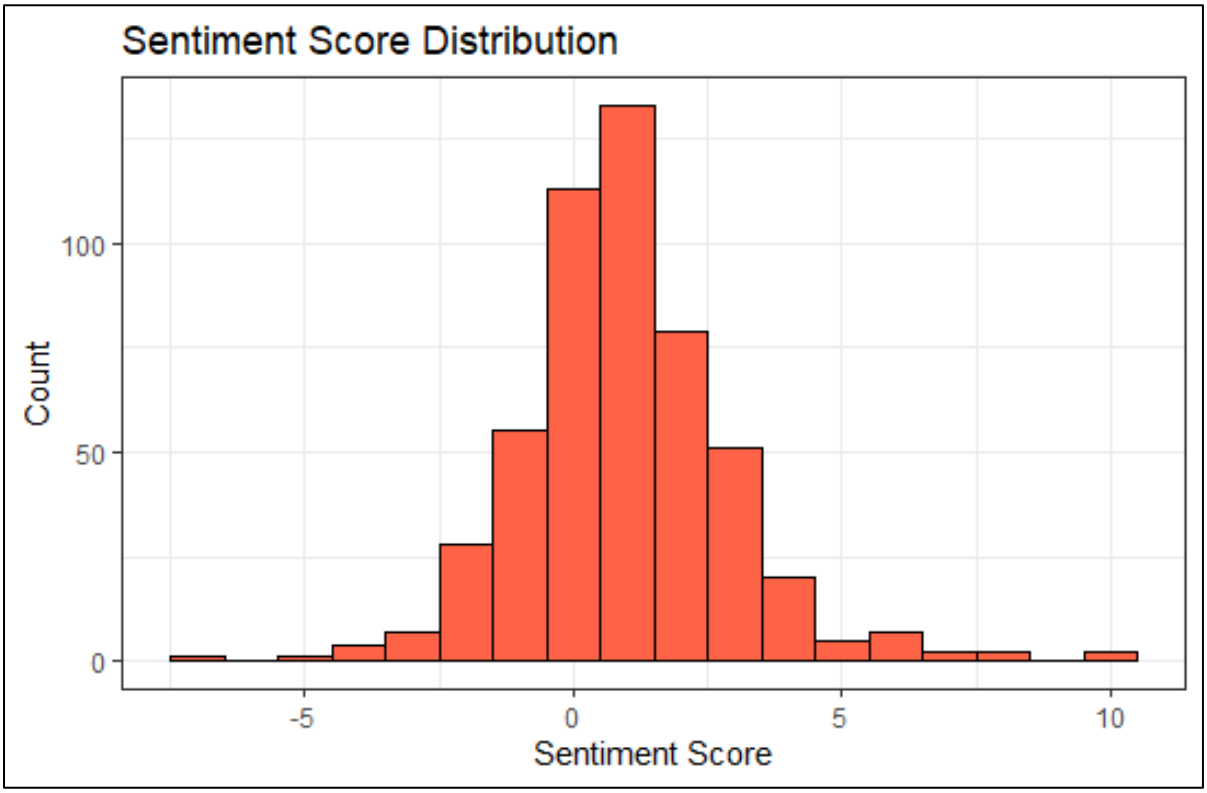


Figure 7: Distribution of Sentiment Scores in Customer Reviews

4.4. Heatmap of Top Terms

A heatmap of the top 20 terms was created to visualize how frequently these words co-occur across reviews. This helps identify clusters of related terms and patterns in customer discussions.

For example, words related to taste, smell, and packaging often appear together, showing that these product features are important to customers.

Figure 3: Heatmap of top 20 terms

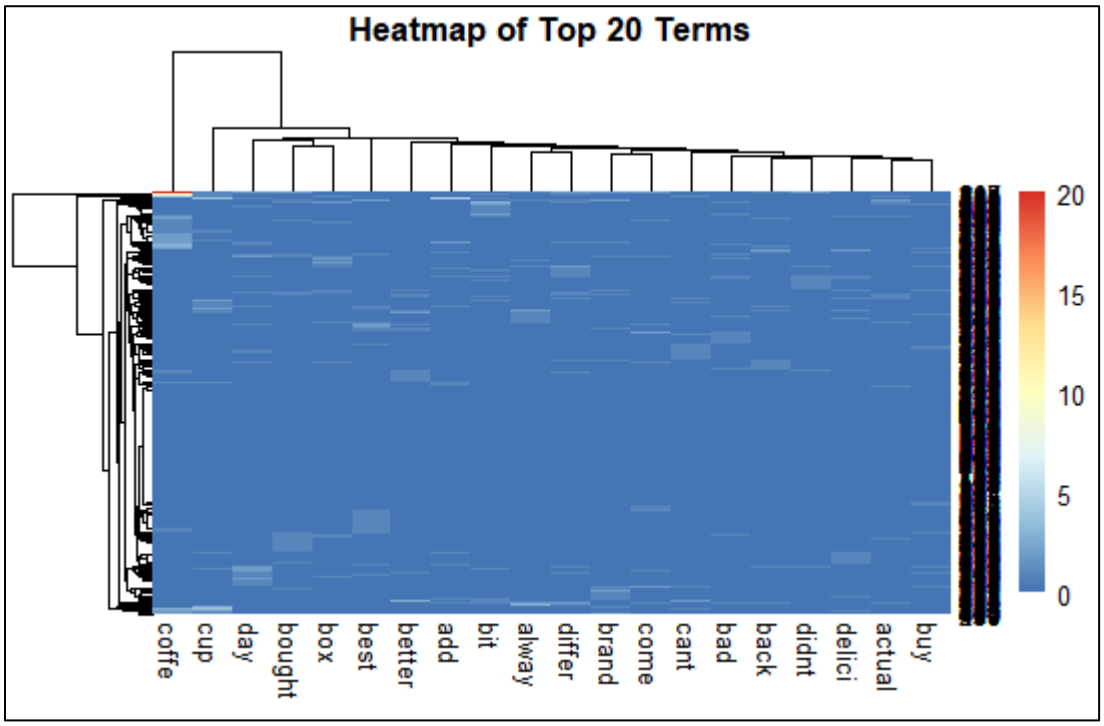


Figure 8: Heatmap of top 20 terms

5. Topic Modeling

5.1. Model Tuning

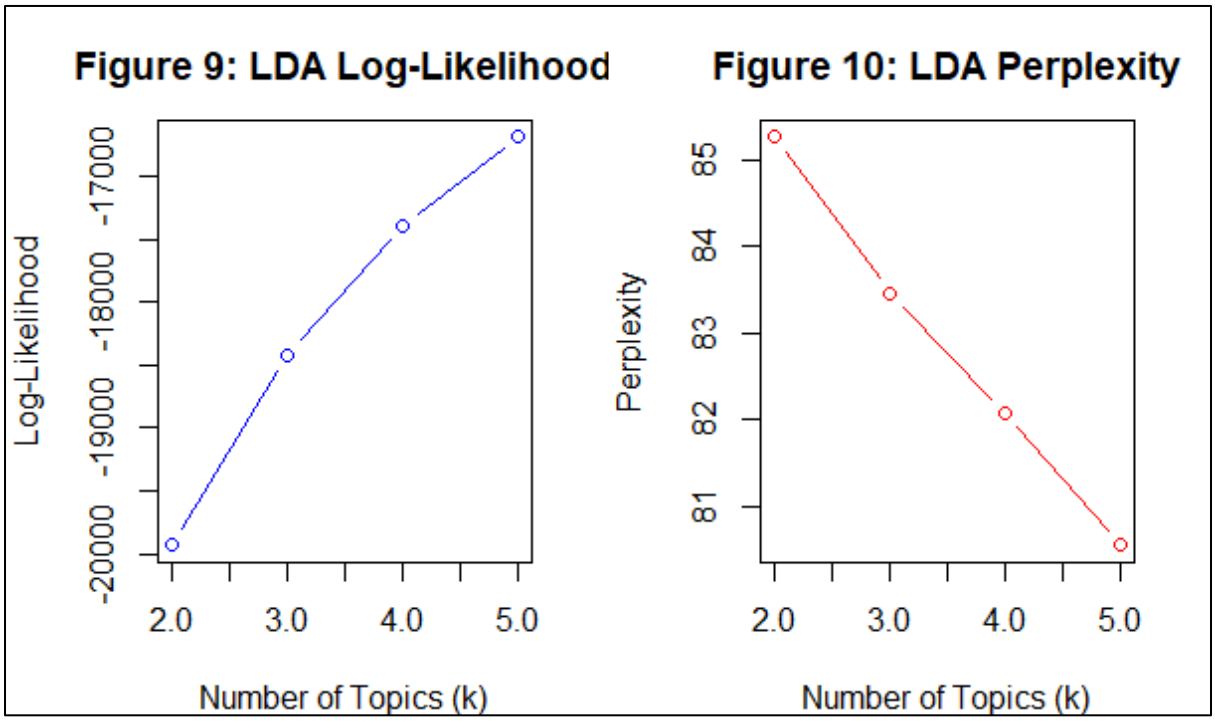
We used Latent Dirichlet Allocation (LDA) on the cleaned Amazon review dataset to find hidden topics. The model assumes that each review can have a mix of topics. Each topic is a set of related words. The purpose of model tuning is to select an appropriate number of topics (k). This helps balance how easy the topics are to understand and how well the model works. Multiple LDA models were built with k = 2, 3, 4, 5. For each model, we calculated log-**likelihood and perplexity**. Log-likelihood shows how well the model explains the data. Perplexity shows how well the model predicts new data. Higher log-likelihood and lower perplexity mean a better model.

k	Log-Likelihood	Perplexity
2	-19933.02	85.26
3	-18415.81	83.46
4	-17387.78	82.09

5	-16688.75	80.55

Table 3: Model tuning results show log-likelihood and perplexity for each k

Figure 8 visualizes log-likelihood by topic count. The figure shows that log-likelihood gets better as k increases. Perplexity goes down at the same time (Figure 9). Considering both how easy it is to understand and these metrics, we chose k = 3 for the final LDA model.



Top terms for each topic were also examined to understand the content and context of topics. This helps management see the main product features, strengths, and areas to improve. By summarizing customer opinions into a few main topics, this analysis gives useful insights for making data-driven decisions in product development, marketing, and quality control.

5.2. Wordcloud per Topic

Wordclouds were created for each topic to visualize the most important terms. They show the most important words. Bigger words appear more often. Wordclouds help visualize the most important words in each topic. Larger words shows higher frequency within that topic. This gives us a quick visual understanding by the topic.

Topic 1 – Trials & Purchase Time

Top words: *tri, now, littl, day, give, year, thing, need, differ, made.*

This topic is about trying the product and when people buy it. Customers talk about trying samples, buying small packs, and if the product met their needs over time.

Topic 2 – Coffee & Packaging

Top words: *coffe, get, realli, cup, make, packag, high, say, cant, two.*

This topic is about the coffee and its packaging. People talk about how the coffee tastes, how it is packed, how they make it, and the quality or quantity.

Topic 3 – Taste & Use

Top words: *tast, water, purchas, mix, sweet, smell, use, bad, drink, wonder.*

This topic is about taste and how people use the product. Reviews mention flavour, sweetness, smell, and how it mixes with water. Some say it tastes good, others say it's bad.

Topic 4 – Order & Store Experience

Top words: *order, find, store, time, want, love, first, bought, bit, box.*

This topic talks about ordering and shopping. Customers mention how they ordered, found it in stores, got delivery, and their first reaction when they opened or bought it.

Topic 5 – Product Satisfaction & Recommendation

Top words: *flavor, much, best, better, even, look, enjoy, review, think, recommend.*

This topic shows how happy customers are with the product. They talk about good flavour, nice look, and often recommend it to others.

Topic	**Top 10 Terms**
Topic 1 – Trials & Purchase Time	tri, now, littl, day, give, year, thing, need, differ, made
Topic 2 – Coffee & Packaging	coffe, get, realli, cup, make, packag, high, say, cant, two
Topic 3 – Taste & Use	tast, water, purchas, mix, sweet, smell, use, bad, drink, wonder
Topic 4 – Order & Store Experience	order, find, store, time, want, love, first, bought, bit, box

Topic 5 – Product Satisfaction & Recommendation	flavor, much, best, better, even, look, enjoy, review, think, recommend

Table 4: T op 10 Terms per Topic (from LDA Final Model, k = 5)

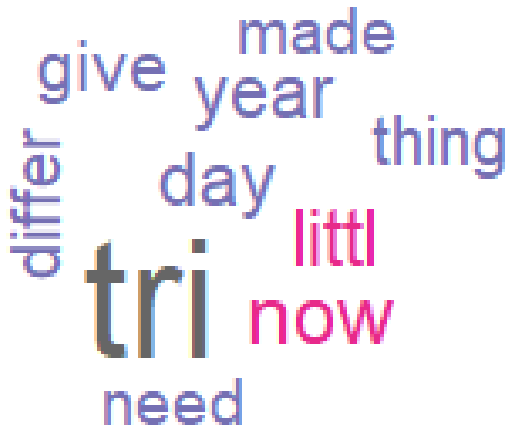
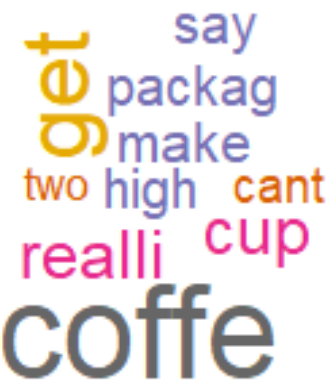
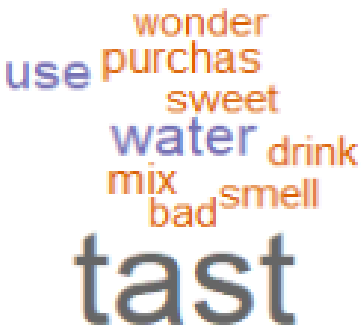
Topic 1 – Trials & Purchase Time	**Topic 2 – Coffee & Packaging**
	
Topic 3 – Taste & Use	**Topic 4 – Order & Store Experience**
	
Topic 5 – Product Satisfaction & Recommendation	

Figure 11: Wordclouds for the Top 10 Terms per Topic

Figure 11 shows wordclouds for the top 10 words in each topic. Each wordcloud highlights the most common and important words found by the LDA model. Bigger words mean those words appear more often in that topic.

5.3. Visualization of Top Terms per Topic with Probabilities

The purpose of this section is to identify and visualize the most representative terms for each topic found by the LDA model. Examining these top terms and their probability weights (beta values), and we can easily understand what each topic means.

After fitting the LDA model, the top 10 terms with the highest beta values were taken for each topic. These terms represent the words that contribute most strongly to defining that topic. A higher beta value means the word is more strongly linked to that topic.

A bar plot was created using ggplot2, where each facet represents one topic. The bars show the probability weights of the top 10 terms. This visualization helps in interpreting the topics more naturally by showing which words dominate within each one.

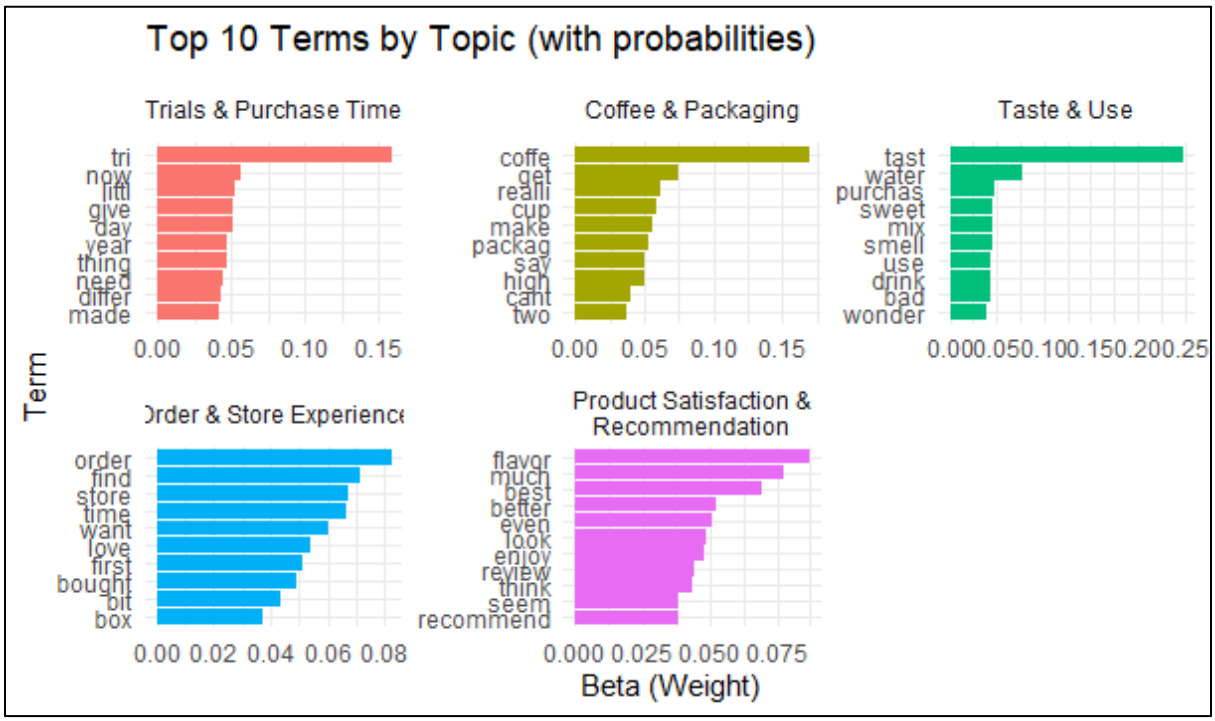


Figure 12: Top 10 Terms in Each Topic Showing Their Probabilities (Beta Values)

Topic	Term	Beta (Weight)
Topic 1 – Trials & Purchase Time	tri	0.158255623
	now	0.056774547
Topic 2 – Coffee & Packaging	coffe	0.169757349
	get	0.074077727
Topic 3 – Taste & Use	tast	0.249346974
	water	0.076295168
Topic 4 – Order & Store Experience	order	0.082363563
	find	0.071328250
Topic 5 – Product Satisfaction & Recommendation	best	0.069330089
	recommend	0.038454341

Table 5: Sample of Top Terms per Topic with Probabilities

5.4. Topic Coherence

Topic Coherence shows how related the words in a topic are. It checks if the top words in a topic make sense together. A higher score means the topic is clear and meaningful. A lower score means the topic may be confusing or have unrelated words.

Coherence helps us see if the LDA model is good. It also gives confidence that the topics reflect real patterns in customer reviews. This can help managers make decisions about product quality, customer preferences, and improvements.

In this analysis, coherence scores were calculated using the **textmineR** package. The top words of each topic were compared to the rest of the dataset. **Table 7** shows the coherence scores for the five topics. The average topic coherence is **0.0573**, which gives an overall idea of how good the model is.

Topic	Coherence Scores
Topic 1 – Trials & Purchase Time	0.0730
Topic 2 – Coffee & Packaging	0.0747
Topic 3 – Taste & Use	0.0364
Topic 4 – Order & Store Experience	0.0614
Topic 5 – Product Satisfaction & Recommendation	0.0409
Average	**0.0573**

Table 6: Topic Coherence Scores

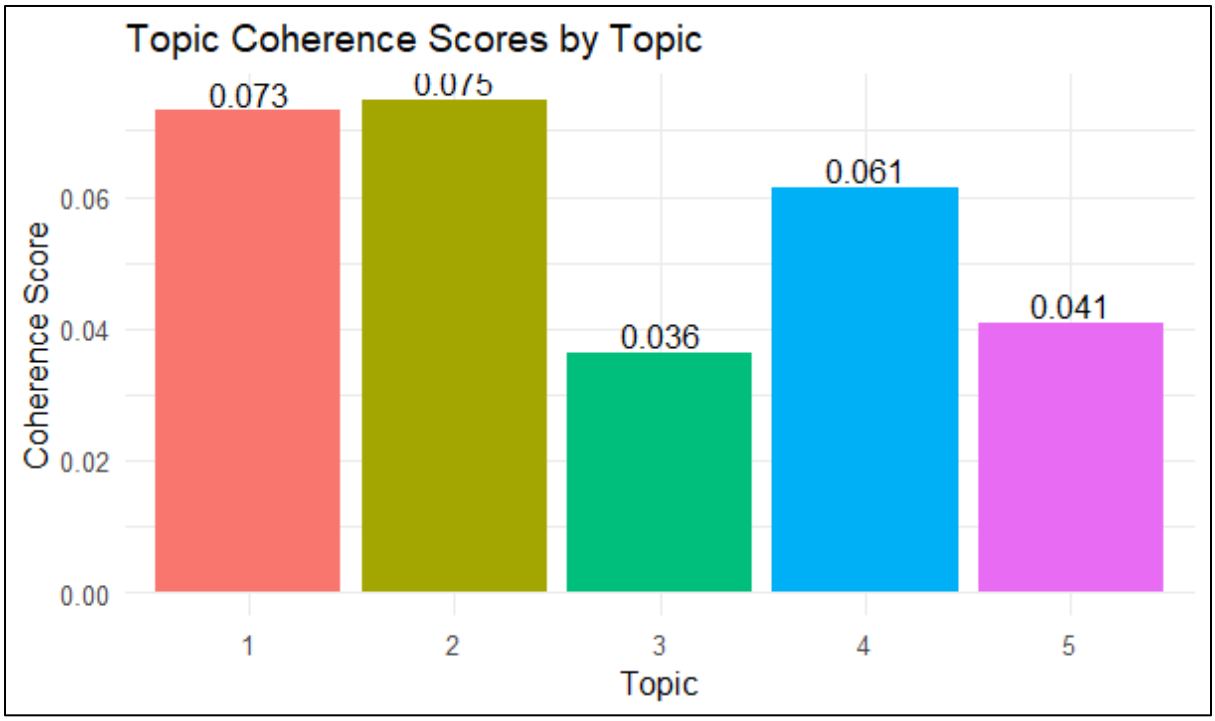


Figure 13: Topic Coherence Scores by Topic

Table 6 shows, Topics 1 and 2 have the highest coherence. This means their words are closely related and the themes are clear, like trials, purchases, coffee, and packaging. Topics 3 and 5 have lower coherence. Their words are less connected, showing more varied discussions about usage, recommendations, and product enjoyment.

Overall, the average coherence shows the LDA model is fairly good. The topics can help find patterns in customer feedback for product improvement and decisions.

5.5. Topic Distribution per Document

The purpose of this analysis is to understand how the identified topics are distributed across all customer reviews. Each review can have more than one topic, each with different probabilities. By examining these probabilities, we can see which topics are more common in each review and how topics appear together. This information is useful for management to identify key areas of customer focus, guide product improvements, and plan marketing strategies.

The topic probabilities for each document were extracted from the LDA model (gamma). Each row represents a review, and each column shows the probability of that review belonging to a particular topic. The first few rows of the Topic Distribution per Document are shown in Table 7.

Id	Topic 1 – Trials & Purchase Time	Topic 2 – Coffee & Packaging	Topic 3 – Taste & Use	Topic 4 – Order & Store Experience	Topic 5 – Product Satisfaction & Recommendation
1	0.196	0.214	0.196	0.179	0.214
2	0.228	0.193	0.175	0.193	0.211
3	0.281	0.193	0.175	0.175	0.176
4	0.196	0.179	0.25	0.179	0.196
5	0.2	0.182	0.182	0.236	0.2

Table 7: Sample of topic probabilities per review

To visualize the distribution, a stacked bar chart was created (Figure 14) where each bar represents a review, and colors indicate topic probabilities. This allows us to see which topics are most prevalent across reviews.

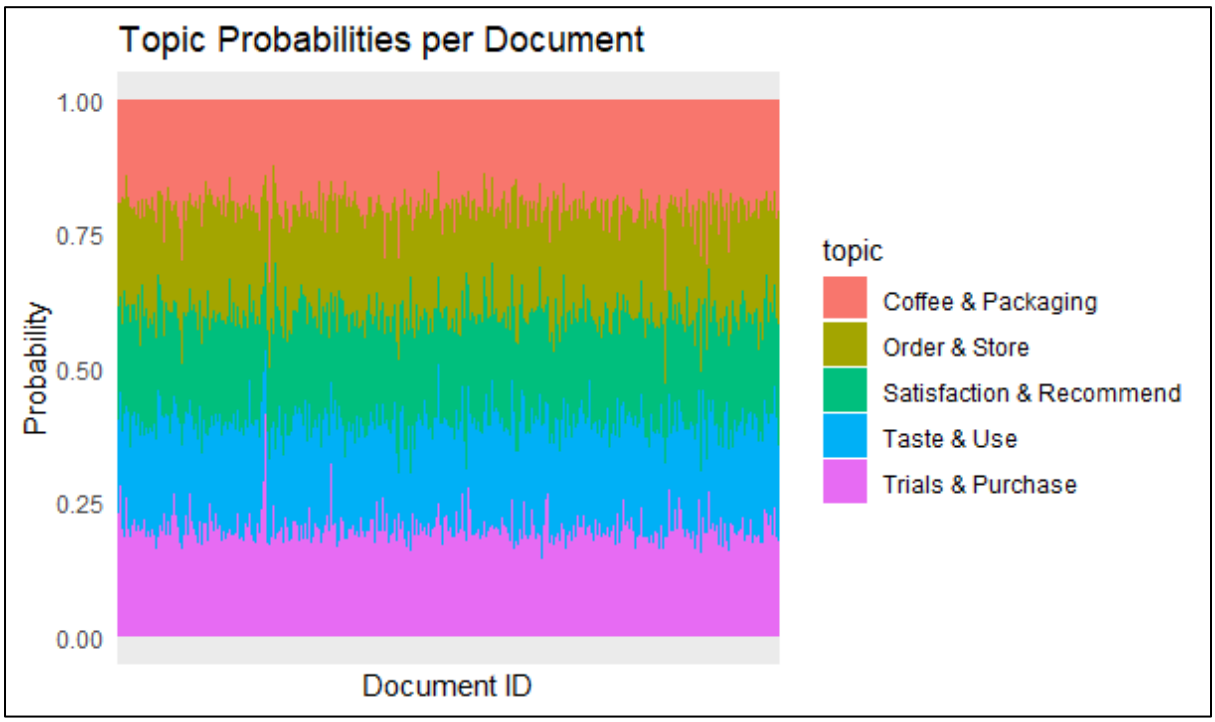


Figure 14: Topic Probabilities per Document (Stacked Bar Chart)

Figure 15 a heatmap shows the same topic probabilities in a matrix form, highlighting patterns and clusters of topics across multiple reviews.

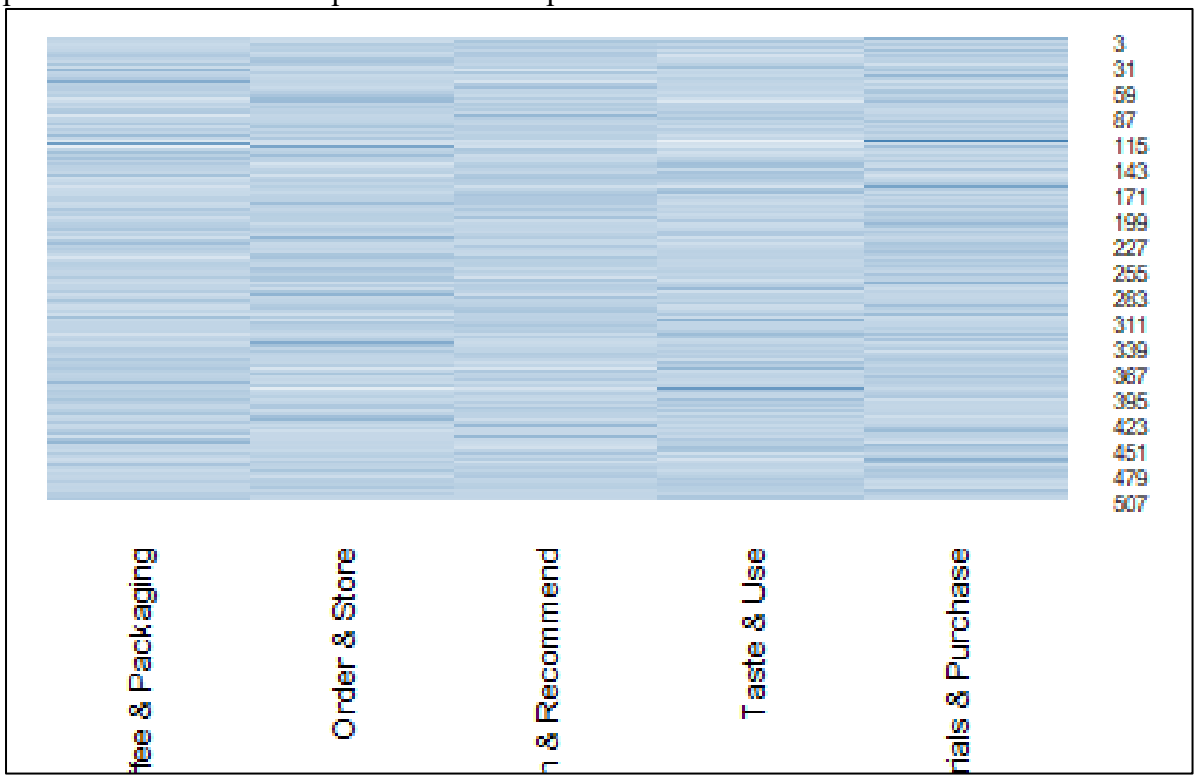


Figure 15: Heatmap of Topic Probabilities across Documents

5.6. Top 5 Documents per Topic

The main objective of this section is to show example reviews that best represent each topic discovered by the LDA model. Earlier sections showed topics using words and coherence scores, this section presents actual text excerpts from the dataset to help connect the topics to real customer experiences

By reading the top 5 reviews most strongly associated with each topic, we can validate whether the topics make sense in real life. This helps us see if the topics really reflect what customers talk about, such as taste, packaging, and satisfaction. Mention that the top 5 reviews per topic were extracted using the highest topic probabilities (gamma) and exported as separate files.

```
# ------------------------------
# Top 5 Texts per Topic
# ------------------------------
top_docs_per_topic <- lapply(1:k, function(t) {
  top_docs <- gamma_df[order(-gamma_df[[as.character(t)]]), ][1:5, "doc_id"]
  data$Text[top_docs]
})
names(top_docs_per_topic) <- paste0("Topic_", 1:k)
```

Topic	Example Review (Excerpt)
Topic 1 – Order & Store Experience	I ordered 3 boxes of Bob's candy canes because I could not find them in stores near me and I wanted the real thing, Bob's,not Spanglers, to help in the celebration of my father's last Christmas. We wanted an old fashioned, basic heart of America Christmas with him. We were willing to pay almost any price to get original Bob's candy canes for this purpose, hence ordering them at such an inflated price on-line from Amazon. More than half came smashed to smithereens. I sent a note to customer service alerting them to this and heard nothing back.
Topic 2 – Product Use & Pet Experience	These chews are great. My little Shi-Poo puppy Oliver loves them. Instead of chewing on my hand (which he kept wanting to do), I can hold this in my hand and he chews on this. He also does a good job of holding it himself because the spirals make it easier for him to do so. I will always make sure I have some on hand. Thanks for offering them 5 in a package.
Topic 3 – Product Convenience & Cost	After years of dealing with increasing price of soda, finding place to store it and returning empty bottles decided to give this a try. We bought ours at Costco which came with 1-130L Co2 Carbonator, 2 1-Liter Stainless Steel Bottles and trial flavors pack for $99. First let me say I'm very

	pleased, the flavors are not bad at all and i can use my own filtered water from either the fridge or Pure faucet attachment and no longer need to pay or bring back bottles for the deposit. They also have much less sodium than store bought soda with the exception of Diet Rite. I have tried many diet flavors and the ones i like the best are Orange, Caffeine free cola, dr pete, Fountain Mist, Cream Soda and Root Beer. As for the cost it has changed everything, no longer have to pay deposit, return empties or search for sale price on soda which means lot less trips to the store saving me from impulse buying and gasoline usage. I can now make fresh bottles of soda whenever I need it (unlike store bought soda which can lose carbonation) and any flavor I feel like at the time. Already used up 1- 130L Co2 carbonator bottle so ordered 1 full bottle + 1 exchange for empty bottle from SodaStream web site this way I'll always have one full bottle when one is done. You should always search the web for coupons for any online store before ordering and SodaStream is no exception, saved $15 which is the price of the license on a new bottle. So it cost me the same as if ordering to exchange 130L Co2 Carbonators, pretty sweet. If you order 10 bottles of syrup at a time you get one free and with on line coupons and cashing in points your earn from them with every purchase can save you more than just the shipping cost, the cost of a trip to the store. For me the SodaStream is well worth the cost.
Topic 4 – Coffee Quality & Taste	I was going to re-order the Newman's Own K-Cups I normally order, but the price jumped for whatever reason, so I went searching for a replacement. I found these, tried them out, and I'm happy. Good bold taste with no bitterness. These will be in my regular morning coffee rotation.
Topic 5 – Premium Coffee Experience	Illy medium roast is one the best ground coffees around. I brew it drip-style with a Melitta 102 filter cone and get a perfect brew every time. This is one of the rare coffees that can be brewed very strong and still taste smooth. I pay about 6 euro a can locally but is well worth it. Illy medium roast is a great coffee that I highly recommend.

Table 8: Top 5 Reviews per Topic (Excerpts)

Top 5 Reviews per Topic help us understand what each topic is about:
Topic 1 is about ordering and shopping experiences, especially delivery and packaging.
Topic 2 talks about how people use the product, mainly sharing happy pet experiences.
Topic 3 focuses on price, convenience, and good value for money.
Topic 4 is about the taste and quality of the coffee.
Topic 5 shows premium products and very positive customer feelings.
This helps us confirm that the topics accurately reflect what customers talk about, such as taste, packaging, and satisfaction.

6. Summary

The text mining analysis of customer reviews provided valuable insights into the main themes discussed by consumers. Using Latent Dirichlet Allocation (LDA), we identified **five coherent topics**, each showing different parts of customer experiences:

1. **Trials & Purchase Time** – Customers discuss trying products, small purchases, and their experiences over time. Words like *tri*, *now*, *day*, and *made* show initial interactions and trial behaviors.
2. **Coffee & Packaging** – This topic is about packaging, product condition, and coffee preparation. Words like *coffe, cup, packag, make* show attention to how the product is presented and used.
3. **Taste & Use** – Consumers comment on taste, usage, and sensory qualities of the products. Words like *tast*, *sweet*, *smell*, and *mix* indicate how products are consumed and perceived.
4. **Order & Store Experience** – Reviews talk about ordering and shopping. Words like *order, store, find, time* highlight delivery, store availability, and first impressions.
5. **Product Satisfaction & Recommendation** – This topic shows satisfaction and willingness to recommend. Words like *flavor, enjoy, recommend, best* reflect positive experiences and customer loyalty.

Key findings from the analysis:

- **Model Evaluation:** Perplexity and log-likelihood suggested that 3–5 topics provide a reasonable representation of the data. Coherence scores were moderate, confirming that topics are interpretable while allowing for minor refinement.

- **Topic Interpretability:** Top terms per topic and wordcloud visualizations provided a clear understanding of the themes. The top 5 reviews per topic confirm that the topics match real customer experiences.

Practical Insights:

- Consumers are highly focused on **taste and flavor**, indicating that product quality is a critical factor in satisfaction.
- **Packaging and delivery** are also important, affecting customer convenience and perception of value.
- Positive experiences lead to **recommendations** and repeat purchases, helping with customer retention and brand loyalty

Overall, the text mining analysis successfully captured the main themes in customer reviews. The results give actionable insights to improve products, packaging, and customer satisfaction. The combination of quantitative evaluation (log-likelihood, perplexity, coherence) and qualitative validation (top reviews per topic) ensures that the findings are both **statistically sound and practically meaningful**.

7. Conclusion

The text mining analysis successfully identified the main themes in customer reviews. These topics include taste and flavor, packaging and delivery, usage and recommendations, order and store experience, and overall product satisfaction. Evaluation of the LDA model using log-likelihood, perplexity, and coherence scores showed that the selected model (**k = 5**) is reasonable and easy to interpret. Examining the top reviews for each topic confirmed that the topics reflect real customer experiences.

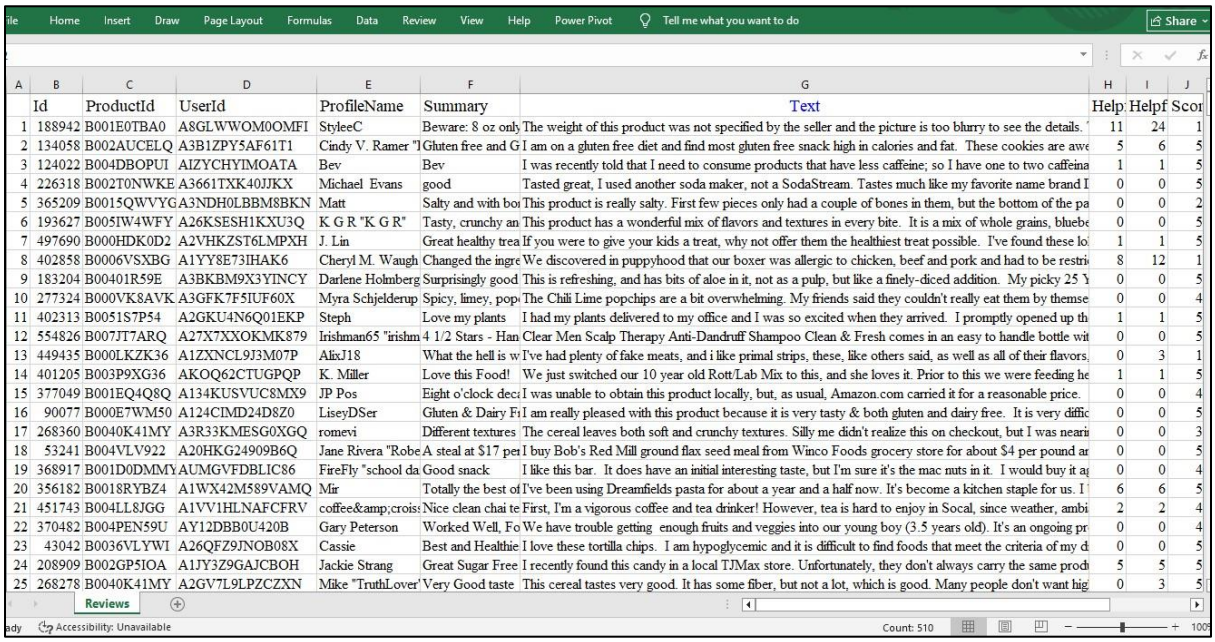
Evaluation of the LDA model using log-likelihood, perplexity, and coherence scores confirmed that the selected model (k = 5) is reasonable and interpretable. Qualitative assessment of the top documents per topic further validated the semantic relevance of each topic.

Overall, this analysis provides useful and actionable insights into customer preferences and behaviour. These insights can help guide product improvements, marketing strategies, and ways to enhance customer engagement.

8. Appendix

8.1. Figures

Figure 1: Sample Records from Amazon Product Reviews Dataset (Kaggle, 2023)



	Id	ProductId	UserId	ProfileName	Summary	Text	Help	Helpf	Scor
1	188942	B001E0TBA0	A8GLWWOM0OMFI	StyleeC	Beware: 8 oz only	The weight of this product was not specified by the seller and the picture is too blurry to see the details.	11	24	1
2	134058	B002AUCELQ	A3B1ZPY5AF61T1	Cindy V. Ramer "I	Gluten free and G	I am on a gluten free diet and find most gluten free snack high in calories and fat. These cookies are awe	5	6	5
3	124022	B004DBOPUI	AIZYCHYIMOATA	Bev	Bev	I was recently told that I need to consume products that have less caffeine; so I have one to two caffeina	1	1	5
4	226318	B002T0NWKE	A3661TXK40JJKX	Michael Evans	good	Tasted great, I used another soda maker, not a SodaStream. Tastes much like my favorite name brand I	0	0	5
5	365209	B0015QWVYG	A3NDH0LBBM8BKN	Matt	Salty and with bo	This product is really salty. First few pieces only had a couple of bones in them, but the bottom of the pa	0	0	2
6	193627	B005IW4WFY	A26KSESH1KXU3Q	K G R "K G R"	Tasty, crunchy an	This product has a wonderful mix of flavors and textures in every bite. It is a mix of whole grains, bluebe	0	0	5
7	497690	B000HDK0D2	A2VHKZST6LMPXH	J. Lin	Great healthy trea	If you were to give your kids a treat, why not offer them the healthiest treat possible. I've found these lo	1	1	5
8	402858	B0006VSXBG	A1YY8E73IHAK6	Cheryl M. Waugh	Changed the ingre	We discovered in puppyhood that our boxer was allergic to chicken, beef and pork and had to be restri	8	12	1
9	183204	B00401R59E	A3BKBM9X3YINCY	Darlene Holmberg	Surprisingly good	This is refreshing, and has bits of aloe in it, not as a pulp, but like a finely-diced addition. My picky 25 Y	0	0	5
10	277324	B000VK8AVK	A3GFK7F5IUF60X	Myra Schjelderup	Spicy, limey, pop	The Chili Lime popchips are a bit overwhelming. My friends said they couldn't really eat them by themse	0	0	4
11	402313	B0051S7P54	A2GKU4N6Q01EKP	Steph	Love my plants	I had my plants delivered to my office and I was so excited when they arrived. I promptly opened up th	1	1	5
12	554826	B007JT7ARQ	A27X7XXOKMK879	Irishman65 "irishm	4 1/2 Stars - Han	Clear Men Scalp Therapy Anti-Dandruff Shampoo Clean & Fresh comes in an easy to handle bottle wit	0	0	5
13	449435	B000LKZK36	A1ZXNCL9J3M07P	AlixJ18	What the hell is w	I've had plenty of fake meats, and i like primal strips, these, like others said, as well as all of their flavors,	0	3	1
14	401205	B003P9XG36	AKOQ62CTUGPQP	K. Miller	Love this Food!	We just switched our 10 year old Rott/Lab Mix to this, and she loves it. Prior to this we were feeding he	1	1	5
15	377049	B001EQ4Q8Q	A134KUSVUC8MX9	JP Pos	Eight o'clock deca	I was unable to obtain this product locally, but, as usual, Amazon.com carried it for a reasonable price.	0	0	4
16	90077	B000E7WM50	A124CIMD24D8Z0	LiseyDSer	Gluten & Dairy Fi	I am really pleased with this product because it is very tasty & both gluten and dairy free. It is very diffic	0	0	5
17	268360	B0040K41MY	A3R33KMESG0XGQ	romevi	Different textures	The cereal leaves both soft and crunchy textures. Silly me didn't realize this on checkout, but I was neari	0	0	3
18	53241	B004VLV922	A20HKG24909B6Q	Jane Rivera "Robe	A steal at $17 per	I buy Bob's Red Mill ground flax seed meal from Winco Foods grocery store for about $4 per pound ar	0	0	5
19	368917	B001D0DMMY	AUMGVFDBLIC86	FireFly "school da	Good snack	I like this bar. It does have an initial interesting taste, but I'm sure it's the mac nuts in it. I would buy it ag	0	0	4
20	356182	B0018RYBZ4	A1WX42M589VAMQ	Mir	Totally the best of	I've been using Dreamfields pasta for about a year and a half now. It's become a kitchen staple for us. I	6	6	5
21	451743	B004LL8JGG	A1VV1HLNAFCFRV	coffee&croiss	Nice clean chai te	First, I'm a vigorous coffee and tea drinker! However, tea is hard to enjoy in Socal, since weather, ambi	2	2	4
22	370482	B004PEN59U	AY12DBB0U420B	Gary Peterson	Worked Well, Fo	We have trouble getting enough fruits and veggies into our young boy (3.5 years old). It's an ongoing pr	0	0	4
23	43042	B0036VLYWI	A26QFZ9JNOB08X	Cassie	Best and Healthie	I love these tortilla chips. I am hypoglycemic and it is difficult to find foods that meet the criteria of my di	0	0	5
24	208909	B002GP5IOA	A1JY3Z9GAJCBOH	Jackie Strang	Great Sugar Free	I recently found this candy in a local TJMax store. Unfortunately, they don't always carry the same prodi	5	5	5
25	268278	B0040K41MY	A2GV7L9LPZCZXN	Mike "TruthLover"	Very Good taste	This cereal tastes very good. It has some fiber, but not a lot, which is good. Many people don't want hig	0	3	5

Figure 2: Word Cloud of Most Frequent Words Before and After Cleaning

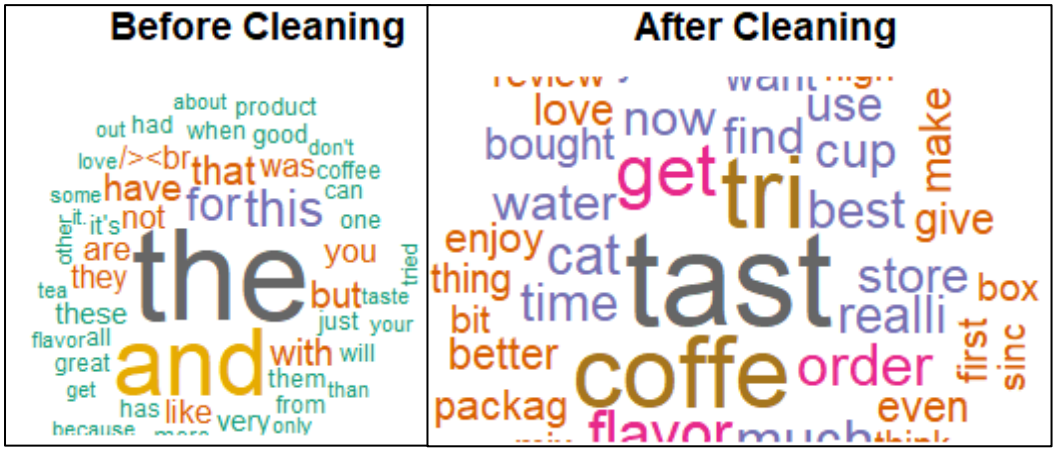


Figure 3: Sample of Document-Term Matrix (First 5 Documents × 10 Terms)

```
inspect(dtm_sparse[1:5, 1:10])
```

```
    Terms
Docs actual add alway back bad best better bit bought box
   1      1   0     0    0   0    0      0   0      0   0
   2      0   0     0    0   0    0      0   0      0   0
   3      0   0     0    0   0    0      0   0      0   0
   4      0   0     0    0   0    0      0   0      0   0
   5      0   0     0    0   0    0      0   0      0   0
```

Figure 4: Top Terms by Frequency

```
top_10_terms
tast tri coffe get flavor order much time best realli
 229 173   172 114     93    86   83   75   74     73
```

Figure 5: Wordcloud showing the most frequent words in the cleaned Amazon reviews

```
set.seed(123)
wordcloud(names(term_freq), term_freq, max.words = 50, colors = brewer.pal(8, "Dark2"))
```

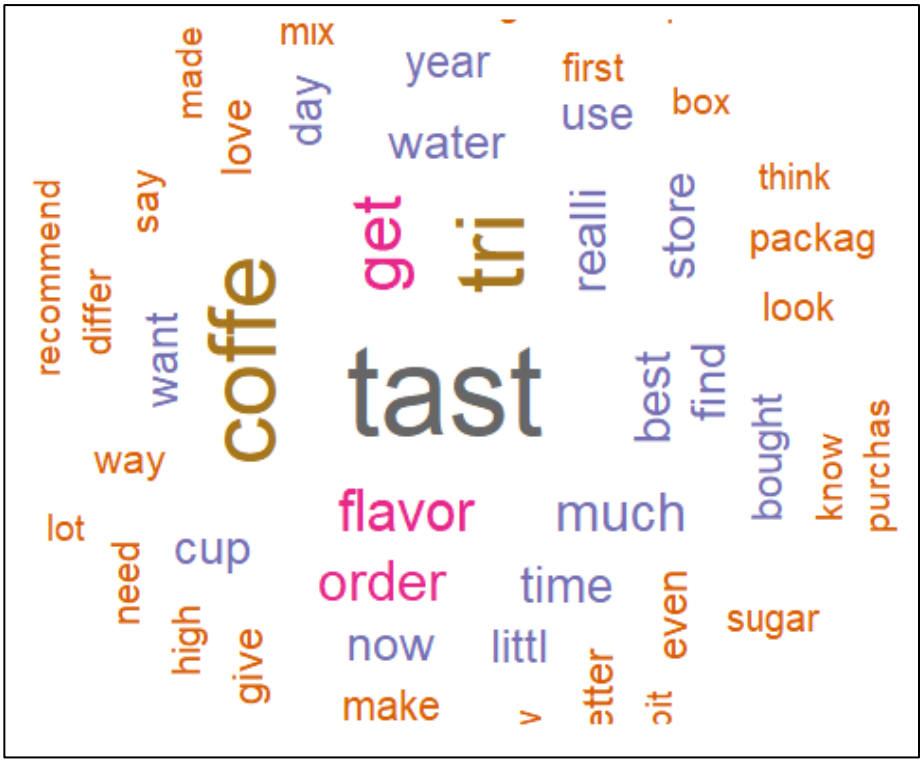


Figure 6: Summary of Sentiment analysis

```
> summary(sentiment_scores)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.0000  0.0000  1.0000  0.9431  2.0000 10.0000
```

Figure 7: Distribution of Sentiment Scores in Customer Reviews

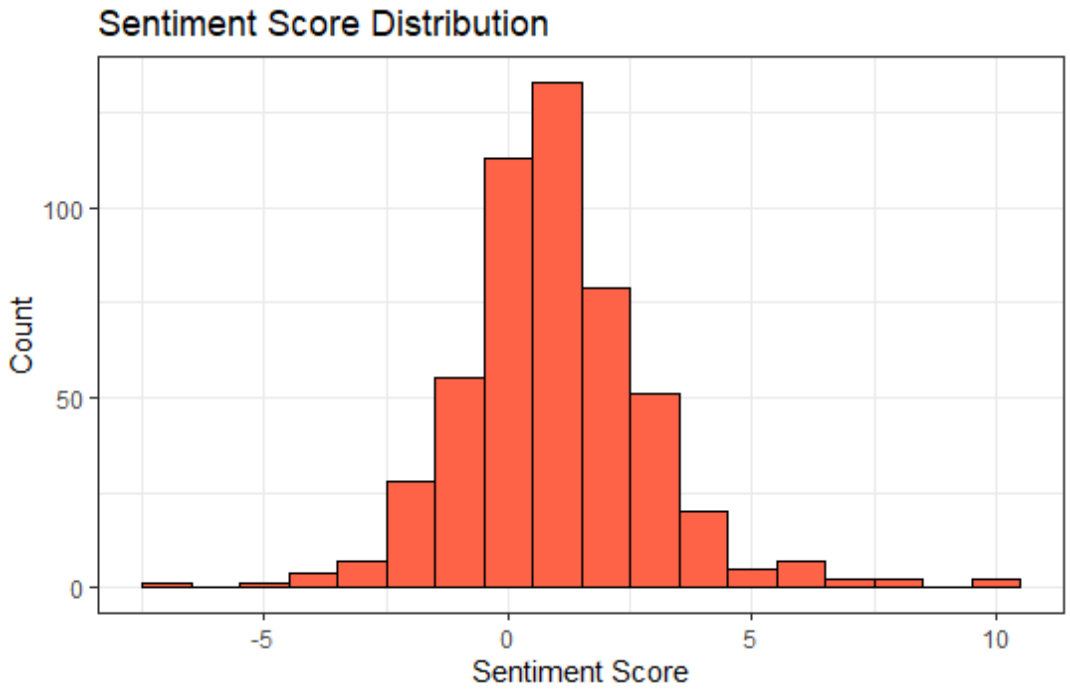


Figure 8: Heatmap of top 20 terms

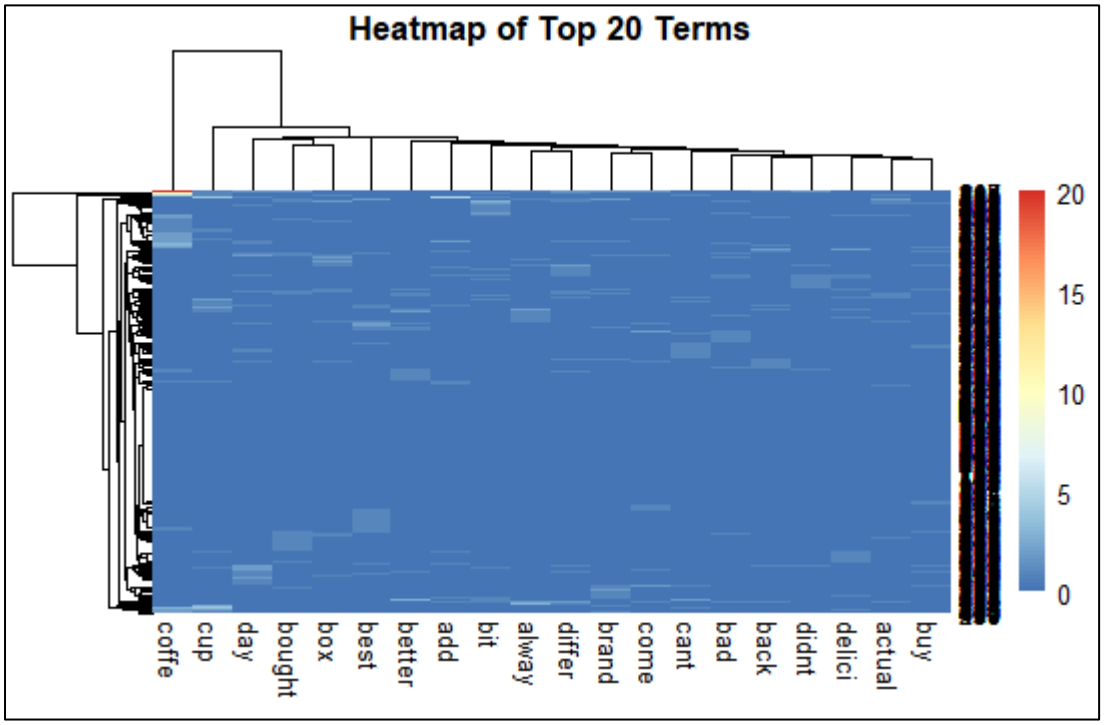


Figure 9 & 10: log-likelihood & LDA Perplexity

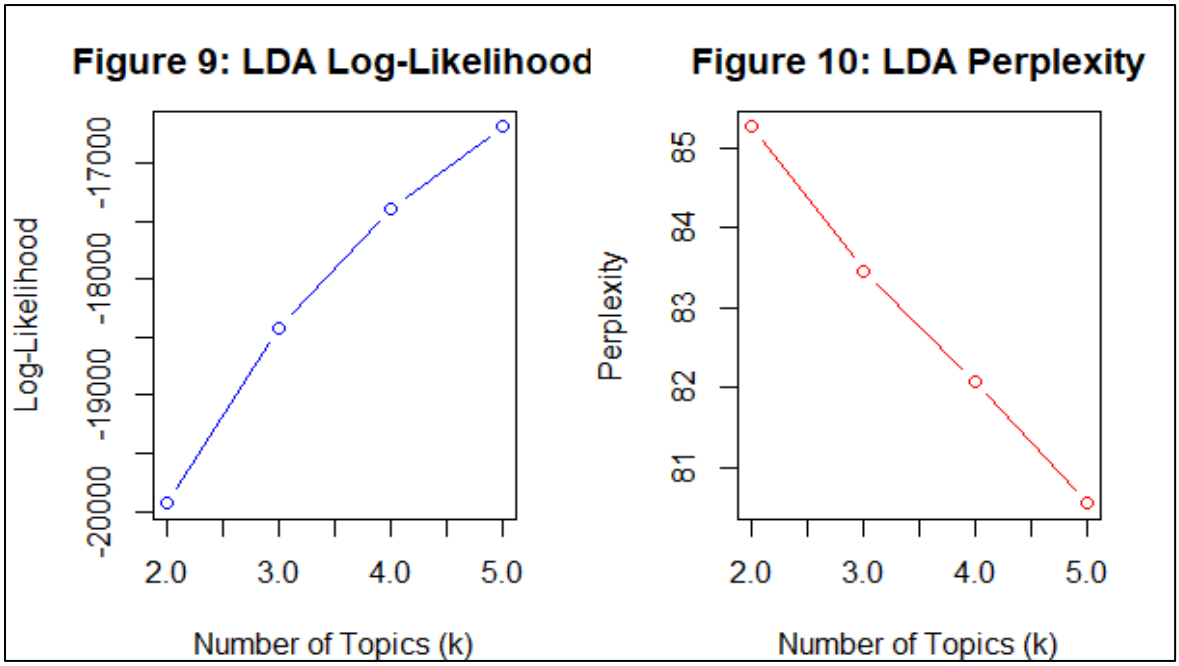
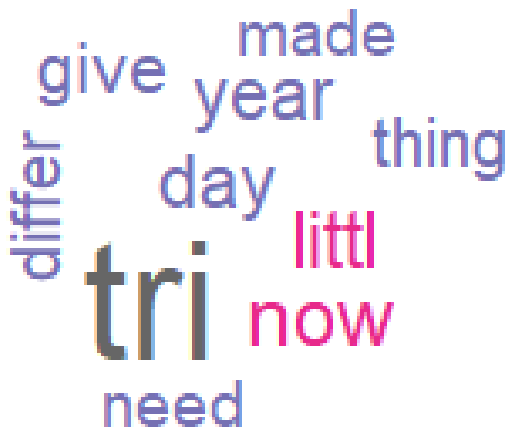
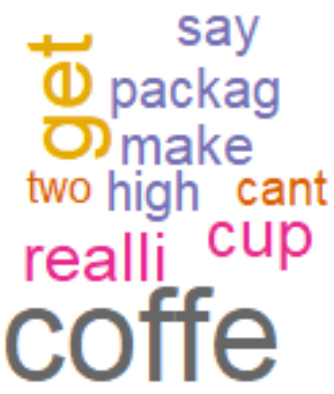
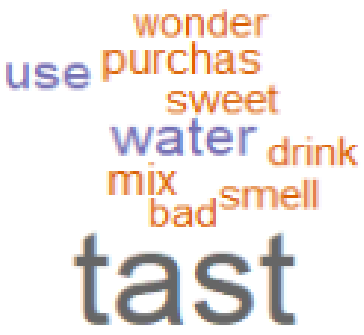
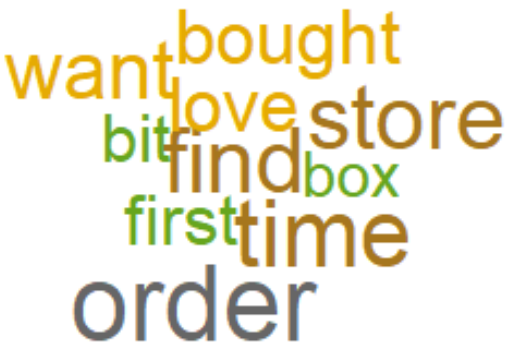


Figure 11: Wordclouds for the Top 10 Terms per Topic

Topic 1 – Trials & Purchase Time	**Topic 2 – Coffee & Packaging**
	
Topic 3 – Taste & Use	**Topic 4 – Order & Store Experience**
	

Topic 5 – Product Satisfaction & Recommendation	

Figure 12: Top 10 Terms in Each Topic Showing Their Probabilities (Beta Values)

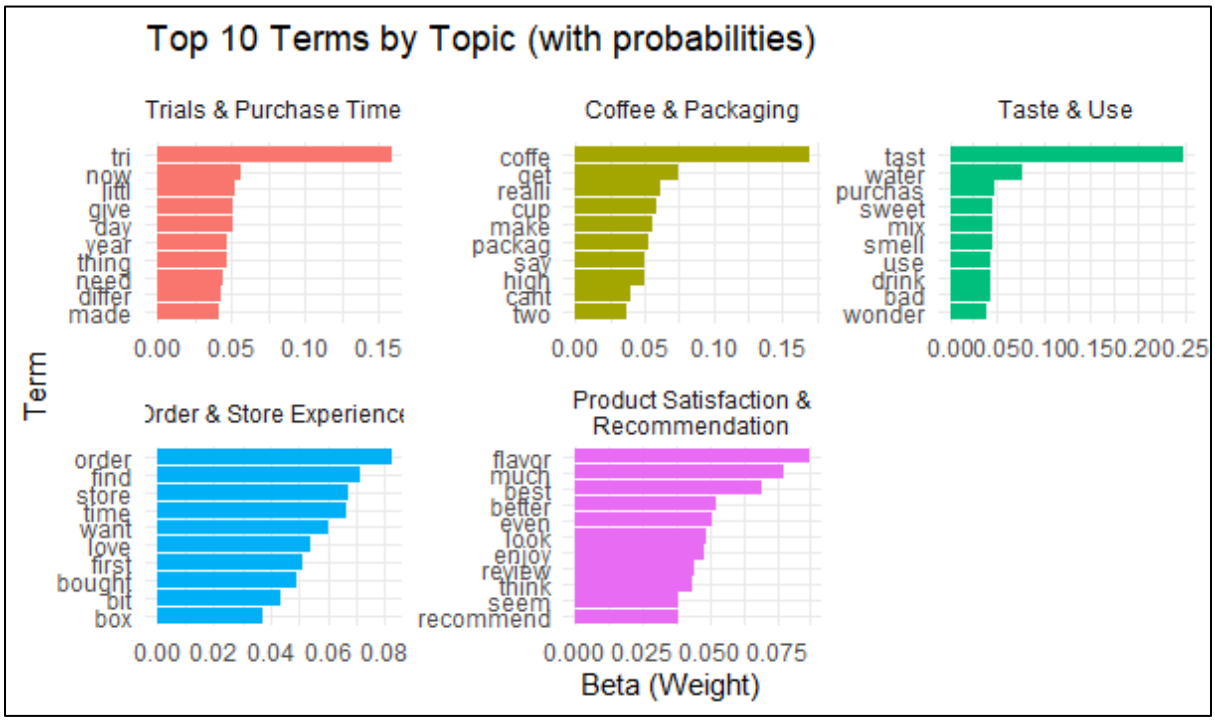


Figure 13: Topic Coherence Scores by Topic

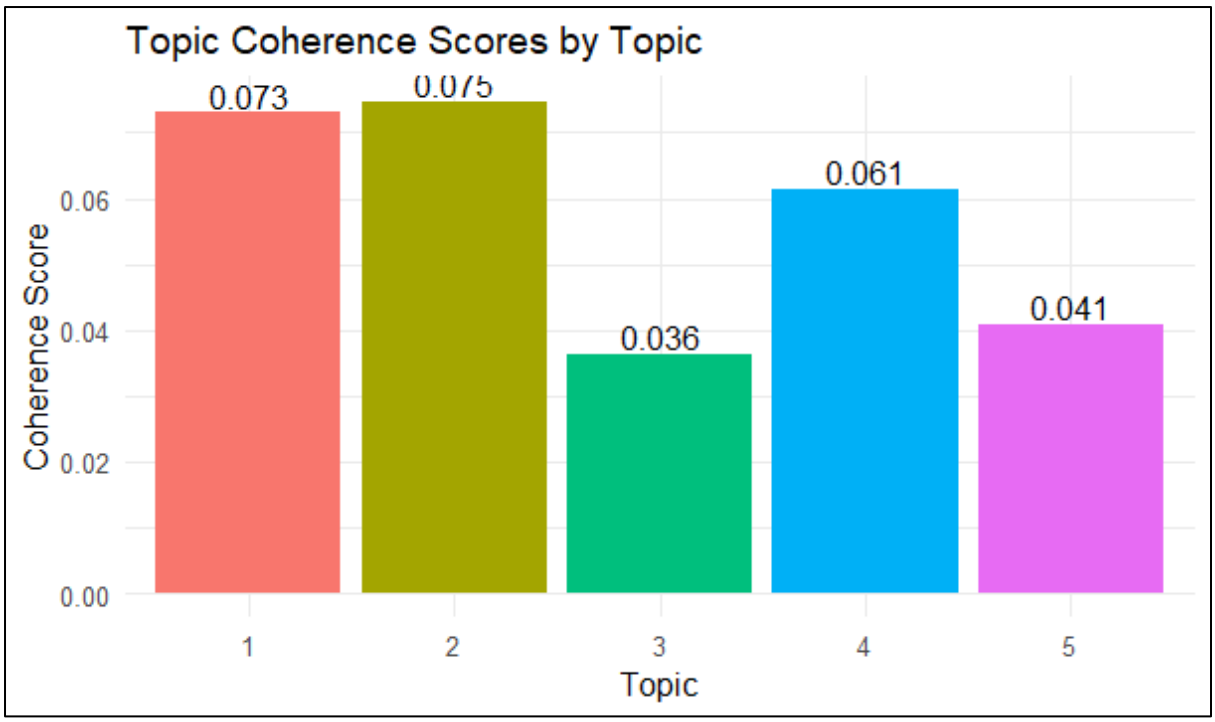


Figure 14: Topic Probabilities per Document (Stacked Bar Chart)

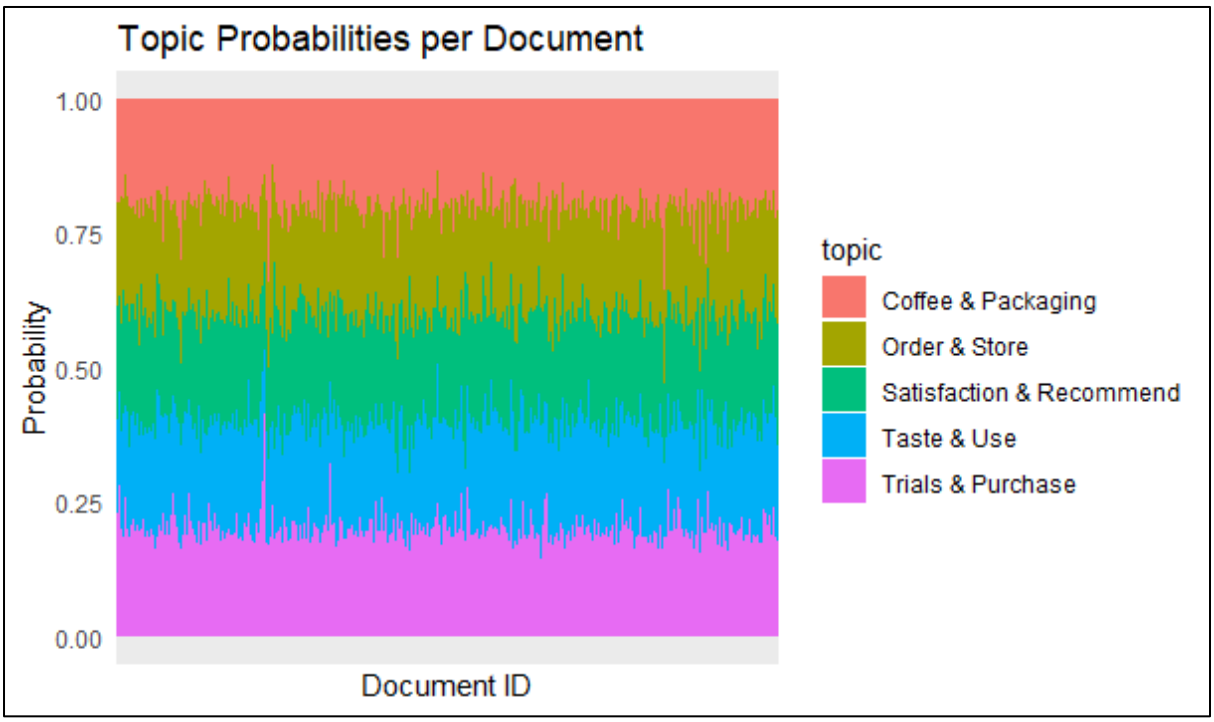
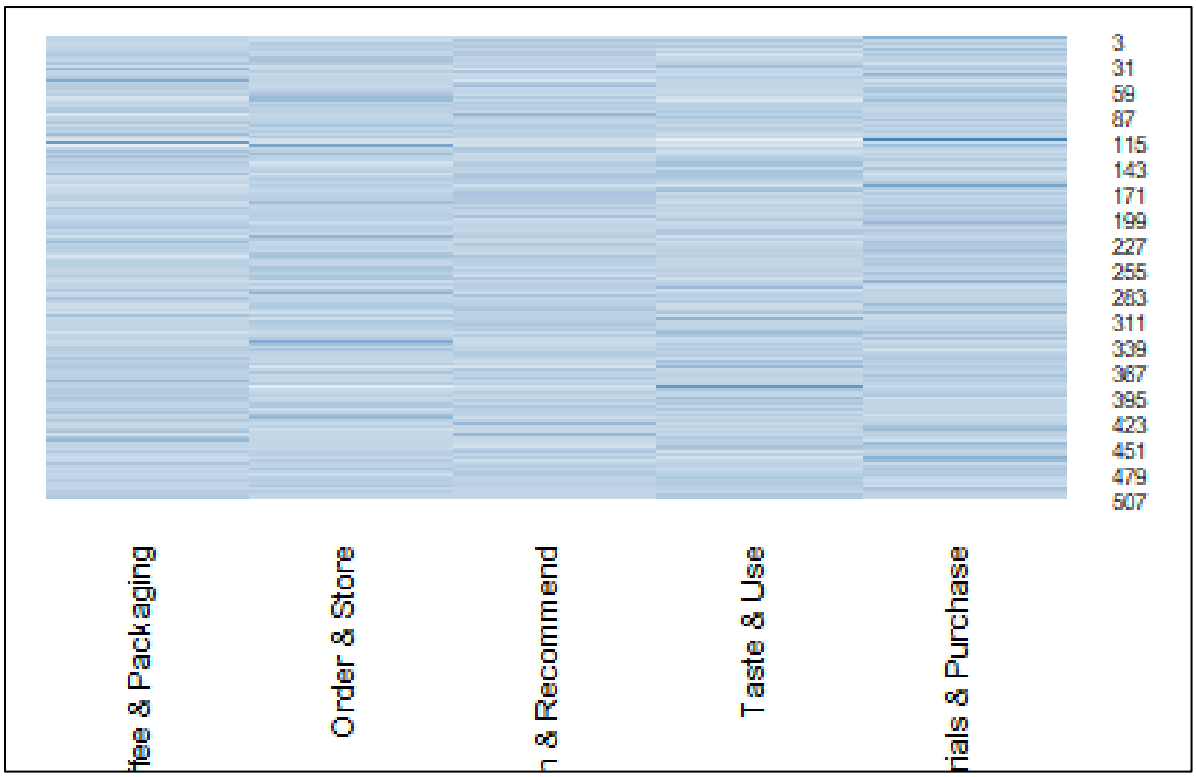


Figure 15: Heatmap of Topic Probabilities across Documents



8.2. Tables

Table 1: Before and After Cleaning table

Text Cleaning	Example Review Text
Before Cleaning	[1] "arrived on time, and is a great tasting item, will order it again and again. convinient too!!!!!!" [2] "This product is fantastic. It tastes like a sweet snack but is 100% apples."
After Cleaning	"arriv time tast item order convini" "fantast tast sweet snack appl"

Table 2: Top 20 most frequent bigrams

Top Bigrams
groceri store
high recommend
cup coffe
gluten free
peanut butter
year ago
give tri
green mountain

hard find
subscrib save
groceri store
high recommend
give tri
green mountain
hard find
subscrib save
dark roast
go back
br br
im sure
much better
read review

Table 3: Model tuning results show log-likelihood and perplexity for each k

k	Log-Likelihood	Perplexity
2	-19933.02	85.26
3	-18415.81	83.46
4	-17387.78	82.09
5	-16688.75	80.55

Table 4: T op 10 Terms per Topic (from LDA Final Model, k = 5)

Topic	**Top 10 Terms**
Topic 1 – Trials & Purchase Time	tri, now, littl, day, give, year, thing, need, differ, made
Topic 2 – Coffee & Packaging	coffe, get, realli, cup, make, packag, high, say, cant, two
Topic 3 – Taste & Use	tast, water, purchas, mix, sweet, smell, use, bad, drink, wonder
Topic 4 – Order & Store Experience	order, find, store, time, want, love, first, bought, bit, box
Topic 5 – Product Satisfaction & Recommendation	flavor, much, best, better, even, look, enjoy, review, think, recommend

Table 5: Sample of Top Terms per Topic with Probabilities

Topic	Term	Beta (Weight)
Topic 1 – Trials & Purchase Time	tri	0.158255623
	now	0.056774547
Topic 2 – Coffee & Packaging	coffe	0.169757349
	get	0.074077727
Topic 3 – Taste & Use	tast	0.249346974
	water	0.076295168
Topic 4 – Order & Store Experience	order	0.082363563
	find	0.071328250
Topic 5 – Product Satisfaction & Recommendation	best	0.069330089
	recommend	0.038454341

Table 6: Topic Coherence Scores

Topic	Coherence Scores
Topic 1 – Trials & Purchase Time	0.0730
Topic 2 – Coffee & Packaging	0.0747
Topic 3 – Taste & Use	0.0364
Topic 4 – Order & Store Experience	0.0614
Topic 5 – Product Satisfaction & Recommendation	0.0409
Average	**0.0573**

Table 7: Sample of topic probabilities per review

Id	Topic 1 – Trials & Purchase Time	Topic 2 – Coffee & Packaging	Topic 3 – Taste & Use	Topic 4 – Order & Store Experience	Topic 5 – Product Satisfaction & Recommendation
1	0.196	0.214	0.196	0.179	0.214
2	0.228	0.193	0.175	0.193	0.211
3	0.281	0.193	0.175	0.175	0.176
4	0.196	0.179	0.25	0.179	0.196
5	0.2	0.182	0.182	0.236	0.2

Table 8: Top 5 Reviews per Topic (Excerpts)

Topic	Example Review (Excerpt)
Topic 1 – Order & Store Experience	I ordered 3 boxes of Bob's candy canes because I could not find them in stores near me and I wanted the real thing, Bob's,not Spanglers, to help in the celebration of my father's last Christmas. We wanted an old fashioned, basic heart of America Christmas with him. We were willing

	to pay almost any price to get original Bob's candy canes for this purpose, hence ordering them at such an inflated price on-line from Amazon. More than half came smashed to smithereens. I sent a note to customer service alerting them to this and heard nothing back.
Topic 2 – Product Use & Pet Experience	These chews are great. My little Shi-Poo puppy Oliver loves them. Instead of chewing on my hand (which he kept wanting to do), I can hold this in my hand and he chews on this. He also does a good job of holding it himself because the spirals make it easier for him to do so. I will always make sure I have some on hand. Thanks for offering them 5 in a package.
Topic 3 – Product Convenience & Cost	After years of dealing with increasing price of soda, finding place to store it and returning empty bottles decided to give this a try. We bought ours at Costco which came with 1-130L Co2 Carbonator, 2 1-Liter Stainless Steel Bottles and trial flavors pack for $99. First let me say I'm very pleased, the flavors are not bad at all and i can use my own filtered water from either the fridge or Pure faucet attachment and no longer need to pay or bring back bottles for the deposit. They also have much less sodium than store bought soda with the exception of Diet Rite. I have tried many diet flavors and the ones i like the best are Orange, Caffeine free cola, dr pete, Fountain Mist, Cream Soda and Root Beer. As for the cost it has changed everything, no longer have to pay deposit, return empties or search for sale price on soda which means lot less trips to the store saving me from impulse buying and gasoline usage. I can now make fresh bottles of soda whenever I need it (unlike store bought soda which can lose carbonation) and any flavor I feel like at the time. Already used up 1- 130L Co2 carbonator bottle so ordered 1 full bottle + 1 exchange for empty bottle from SodaStream web site this way I'll always have one full bottle when one is done. You should always search the web for coupons for any online store before ordering and SodaStream is no exception, saved $15 which is the price of the license on a new bottle. So it cost me the same as if ordering to exchange 130L Co2 Carbonators, pretty sweet. If you order 10 bottles of syrup at a time you get one free and with on line coupons and cashing in points your earn from them with

	every purchase can save you more than just the shipping cost, the cost of a trip to the store. For me the SodaStream is well worth the cost.
Topic 4 – Coffee Quality & Taste	I was going to re-order the Newman's Own K-Cups I normally order, but the price jumped for whatever reason, so I went searching for a replacement. I found these, tried them out, and I'm happy. Good bold taste with no bitterness. These will be in my regular morning coffee rotation.
Topic 5 – Premium Coffee Experience	Illy medium roast is one the best ground coffees around. I brew it drip-style with a Melitta 102 filter cone and get a perfect brew every time. This is one of the rare coffees that can be brewed very strong and still taste smooth. I pay about 6 euro a can locally but is well worth it. Illy medium roast is a great coffee that I highly recommend.

9. References

Walaa Medhat, Ahmed Hassan, Hoda Korashy. (2014). *Sentiment analysis algorithms and applications: A survey.* Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Andrea Mor, Carlotta Orsenigo, Mauricio Soto Gomez, Carlo Vercellis (2024). *Shaping the causes of product returns: Topic modeling on online customer reviews.* Electronic Commerce Research. https://doi.org/10.1007/s10660-024-09901-x

Arham Rumi. (n.d.). *Amazon Product Reviews* [Dataset]. Kaggle. https://www.kaggle.com/datasets/arhamrumi/amazon-product-reviews

Grün, B., & Hornik, K. (2011). *topicmodels: An R package for fitting topic models*. Journal of Statistical Software, 40(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Mullen, K. (2019). *textmineR: Text Mining and Topic Modeling in R*. CRAN R Project. https://cran.r-project.org/package=textmineR

Wickham, H., & Girlich, M. (2023). *tidyverse: Easily install and load the tidyverse*. R package version 2.0.0. https://CRAN.R-project.org/package=tidyverse

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

Hornik, K., Grün, B., & Others. (2011). *Latent Dirichlet allocation in R: Implementation and applications*. *Journal of Statistical Software*, 40(13), 1–30.

Fellows, I. (2020). *Wordcloud: Word clouds in R*. CRAN R Project. https://cran.r-project.org/package=wordcloud