UCONN
UNIVERSITY OF CONNECTICUT

OPIM 5671 **Data Mining and Business Intelligence**

EVALUATING AND ENHANCING FLIPKART CUSTOMER EXPERIENCE: STRATEGIC INSIGHTS FROM FEEDBACK FOR BOOSTING SATISFACTION AND REPURCHASE RATES

Abstract

The growth of online shopping platforms has generated vast amounts of data that offer insights into consumer behavior. Analyzing this data, especially customer reviews and ratings, is crucial for businesses to understand customer satisfaction, predict repurchase intentions, and improve product offerings. This study leverages a dataset from Flipkart, one of the leading e-commerce platforms in India, to analyze customer reviews and ratings. These evaluations aid in identifying both positive and negative attributes of a product, enabling users to compare the standout features of a top-selling item with similar offerings in the market. Understanding customer perspectives is necessary for businesses, and feedback garnered from e-commerce platforms allows companies to gauge customer engagement and devise strategies to enhance business value for future products. Through text analysis and sentiment analysis, this project aims to uncover patterns in customer feedback that could predict their repurchase behavior and overall sentiment towards the products.

CONTENTS

1. Executive Summary

In the competitive arena of e-commerce, the customer experience stands at the forefront of determining a business's success and longevity. This project leverages advanced analytical techniques to dive deep into the customer feedback ecosystem of Flipkart, one of India's premier online shopping platforms. This project aims to distill actionable insights from customer reviews and ratings, focusing on enhancing the overall customer experience, satisfaction, and encouraging repurchase behavior.

Our analysis is rooted in a robust dataset obtained from Kaggle, featuring an extensive collection of customer feedback on products purchased through Flipkart. This dataset not only serves as a mirror reflecting the multifaceted experiences of consumers but also as a foundation for our analytical endeavors. By meticulously examining this dataset, we aim to uncover underlying patterns, sentiments, preferences, and dissatisfactions that customers express, providing a granular view of the factors that drive customer experience on the platform.

The methodology of our project employs a two-pronged approach: utilizing Python for preliminary exploratory analysis and SAS for advanced modeling and text analytics. The objective is to furnish a comprehensive understanding of our findings, their implications for Flipkart's business strategy, and recommendations for enhancing the customer experience. Through this analysis, we endeavor to illuminate the path toward not just meeting but exceeding customer expectations, thereby fostering a loyal customer base and driving repurchase rates in the competitive e-commerce landscape.

2. Business Case

The primary objective of this study is to analyze the Flipkart dataset to predict customer repurchase intentions and extract sentiment from reviews and ratings. The analysis seeks to answer the following questions:

1. Can customer reviews and ratings predict their likelihood of repurchasing a product?
2. What is the overall sentiment of customers towards the products they purchased?
3. How do these insights inform business strategies for Flipkart?

3. Dataset Details

Name: Flipkart Product reviews with sentiment Dataset

Data source: Kaggle

Data link: [Flipkart Product reviews with sentiment Dataset (kaggle.com)]()

About data: This dataset compiles customer reviews for products sold on Flipkart. It is ideal for sentiment analysis, opinion mining, and natural language processing tasks. The dataset includes a variety of features related to customer feedback and product information.

Data is formatted in CSV, suitable for text processing and analytics tasks. Records contain detailed review texts and associated metrics such as ratings, review titles and product information. There are 104 products in the dataset and has 205053 rows and 6 columns.

The primary variables include:

Product_name	Name of the product
Product_price	Price of the product
Rate	Customer's rating on product (Between 1 to 5)
Review	Customer's review on each product
Summary	This column includes descriptive information of customer's thoughts on each product
Sentiment	This column contains 3 labels such as Positive, Negative and Neutral (Which was given based on summary)

Table 1. Details of the dataset

The KPI's (Key Point Indicators) that have been considered for the analysis are Rate (Ratings of the product), Review (Reviews of the product), Summary (Summary of the reviews) and Sentiment. During supervised modelling, 'Sentiment' has been considered as a target variable. These KPI's helped us dissect the Flipkart's customer feedback at a granular level, thus allowing us to draw correlations between customer satisfaction and repurchase behaviour.

4. Exploratory Analysis

4.1 Visualization using WordCloud

Word clouds are an intriguing and perceptive way to visually represent textual data in exploratory analysis. Word clouds provide a quick overview of the terms that appear most frequently in a particular text corpus, revealing important themes, subjects, or patterns. The steps involved in creating a WordCloud are Data Preparation, Frequency Calculation, WordCloud Generation and Visualization Interpretation.

Exploratory analysis using word clouds helps in gaining quick insights into textual data, enabling identification of patterns and trends in an intuitive and visually appealing manner.

Fig 1. WordCloud based on word frequencies

4.2 Visualization using Pie Chart

A pie chart is a useful visualization tool employed to represent categorical data and display the proportional distribution of different categories within a dataset. In the context of representing ratings ranging from 1 to 5, a pie chart offers a clear and concise depiction of the distribution of ratings across the entire range.

The largest portion of the chart is designated for the rating of 5, which constitutes 57.4% of the responses, indicating a majority of the customers are highly satisfied. Following this, the second-largest segment is for the rating of 4, taking up 20.4% of the pie, suggesting a sizable fraction of customers are quite content with their experience.

Moving to the lower end of the scale, 10.6% of the ratings are at 3, reflecting a moderate level of satisfaction. Ratings of 1, which typically indicate dissatisfaction, make up 10.6% of the chart, hinting that a notable minority of users were not pleased with their experience. Finally, the smallest slice represents the rating of 2, which is 3.3%, suggesting a small number of customers found the service or product below average.

Overall, the chart suggests that the website is generally well-regarded, with a significant number of users rating the experience as either good or excellent. However, there is a clear indication that there is room for improvement, as indicated by the ratings of 3 and below.

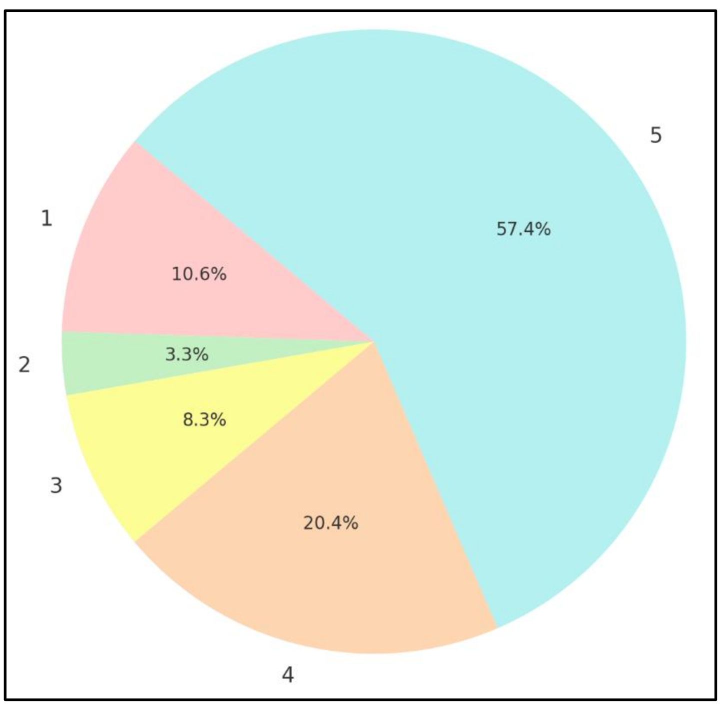


Fig 2. Insights into the distribution of ratings

4.3 Comparison of Corpus

The bar charts compare the most common words found in customer feedback for ratings 1-3 and 4-5. Lower ratings feature a mix of positive ("good," "nice") and negative words ("bad," "waste," "poor"), indicating mixed feedback. In contrast, higher ratings predominantly contain positive words like "good," "nice," "awesome," and "excellent," signifying overall customer satisfaction. This word frequency analysis highlights the contrast in customer sentiments between the two rating groups.

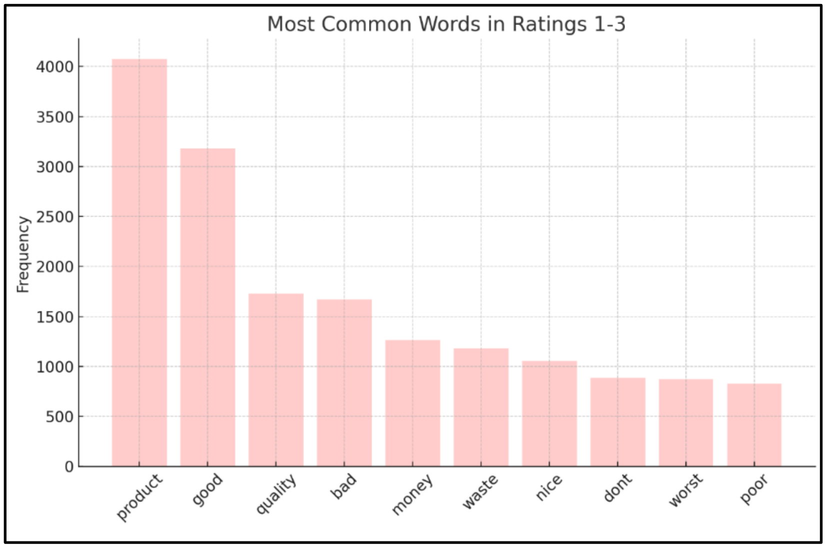


Fig 3. Frequently used words by customers who rate the products 1-3

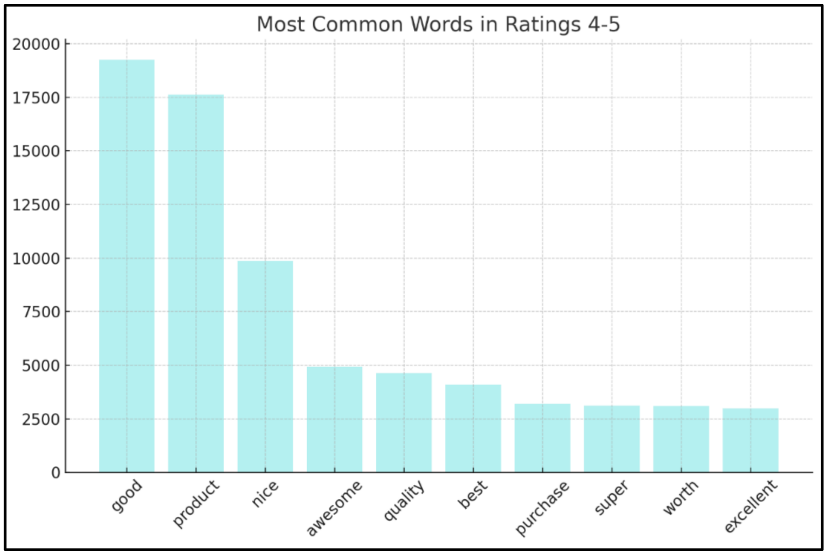


Fig 4. Frequently used words by customers who rate the products 4-5

A polarized tag plot, as shown in the image below, is a visual representation used to compare the frequency of specific terms between two distinct sets of data. In this context, the plot compares word frequencies in product reviews with negative (ratings 1-3) and positive (ratings 4-5) sentiments. It creates a mirrored effect that easily highlights differences and similarities in word usage between the two polarities of reviews.

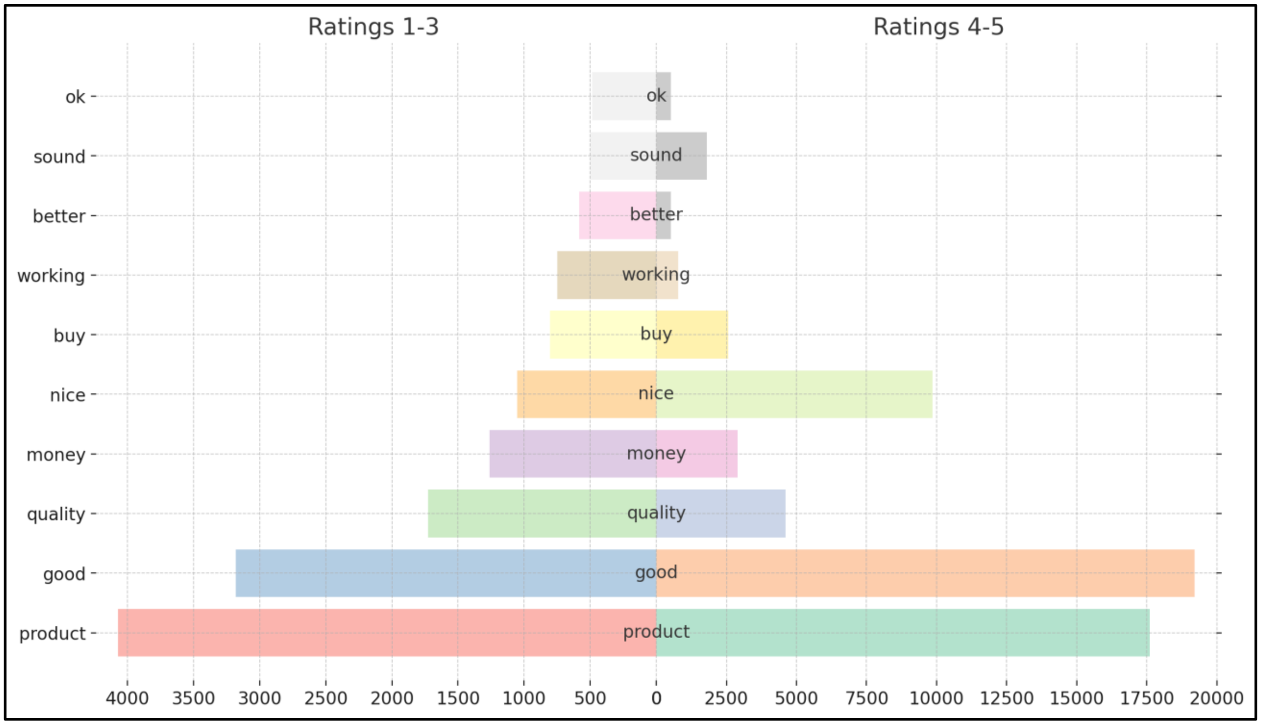


Fig 5. Polarized tag plot

This polarized tag plot compares the frequency of specific words in product reviews with lower (1-3) and higher (4-5) ratings. Words more frequent in positive reviews (4-5) like "good" and "nice" suggest satisfaction, while the prevalence of "money" and "quality" in negative reviews (1-3) could indicate concerns about product value and standards. "Product" is the most mentioned term in both, underscoring its centrality to customer feedback.

5. Modelling

For the analysis, two distinct products from five different categories have been selected and modelled. Each product has been selected such that it is from a different vendor and has different price. Reviews for each of these products have been sourced from the Flipkart website and will be utilized for further analysis using SAS® Enterprise Miner.

The selected products for the analysis are:

S.no.	Category	Product Name
1.	Home Theatre	PHILIPS MMS8085B94 Convertible 80 W Bluetooth Home Theatre
2.	Footwear	Men's Black Sandals
3.	Cooler	MAHARAJA WHITELINE 65 L Desert Air Cooler
4.	Hard Disk	Seagate Portable 2 TB External Hard Disk Drive (HDD)(Light Blue)
5.	Perfume	Bella vita organic Luxury Unisex Perfume Gift Set

Table 2. Details of products considered for the analysis

5.1 Unsupervised Modelling

The process involves importing the SAS dataset for each product into SAS Enterprise Miner 15.1. The data had ratings ranging from 1 to 5. (1 Being the lowest and 5 being the highest.) Subsequently, the dataset is partitioned into three distinct sets using a filter node. The ratings were divided into three sets. Ratings 1 & 2, Rating 3 and Ratings 4 & 5. This was done to analyze which topics, clusters, and words separate these ratings. As the problem statement is to gain business value after analyzing the review text, breaking down into these 3 categories would give insights into areas to improve upon.

The figure below illustrates the methodology employed for the analysis. The diagram is a workflow for processing and analyzing customer reviews. The process begins with **'File Import'** wherein the dataset is being imported from a file. Following the file import, the workflow splits into three paths, each one corresponding to different ranges of review scores: 1-2, 3, and 4-5. This is done to segment the reviews by their score before further processing.

Each path then follows a similar sequence of steps:

1. Text Parsing: This step involves breaking down the text into more manageable pieces, such as sentences or words, and includes removing punctuation, capitalization, and possibly stopwords (common words that are usually irrelevant to the analysis).
2. Text Filter: This is a step where specific criteria are applied to include only relevant pieces of text, or to exclude non-informative content.
3. Text Topic: This node identifies and categorizes main topics or themes present in a body of text within a body of text, focusing on high-level topics.
4. Text Cluster: This phase involves grouping similar pieces of text together, to identify common themes or topics within each review score range.

5.1.1 Category 1 (Home Theatre):

PHILIPS MMS8085B/94 Convertible 80 W Bluetooth Home Theatre

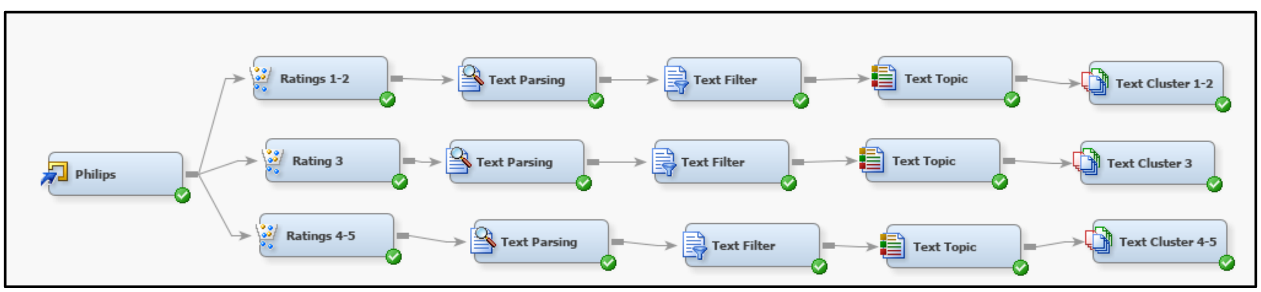


Fig 6. Flow diagram with unsupervised model

The text parsing node generated a term-by-document matrix, which assists in identifying the most frequently occurring words and the frequency of occurrence across comments. For the Philips Home Theatre system receiving ratings of 1, 2, and 3, this matrix will reveal patterns and commonalities in language that reflect customer dissatisfaction, such as recurring issues or shortcomings mentioned across multiple reviews. This analytical approach enables businesses to pinpoint problem areas and prioritize them for improvement, as the frequency and distribution of terms can highlight the most critical aspects that affect customer experience.

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
not	... Adv	Alpha	72	60	N		675	1
+ product	... Noun	Alpha	77	59	Y	+	193	2
very	... Adv	Alpha	60	52	N		666	3
+ be	... Verb	Alpha	63	50	N	+	692	4
quality	... Noun	Alpha	38	37	Y		101	5
bad	... Adj	Alpha	37	34	Y		82	6
sound	... Adj	Alpha	25	25	Y		262	7
+ sound	... Noun	Alpha	21	21	Y	+	335	8
i	... Noun	Alpha	26	20	N		715	9
poor	... Adj	Alpha	20	20	Y		418	9
good	... Adj	Alpha	20	19	Y		30	11
dont	... Noun	Alpha	18	18	Y		334	12
money	... Noun	Alpha	16	16	Y		309	13
no	... Adv	Alpha	15	14	N		706	14
+ buy	... Verb	Alpha	15	13	Y	+	17	15
working	... Adj	Alpha	14	13	Y		72	15
sound qu...	Noun Gr...	Alpha	12	12	Y		345	17

Fig 7. term-by-document matrix for Philips Home Theatre with ratings of 1 and 2

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ be	... Verb	Alpha	53	30	N	+	411	1
not	... Adv	Alpha	30	26	N		401	2
good	... Adj	Alpha	19	18	Y		27	3
good	... Noun	Alpha	14	14	Y		207	4
+ product	... Noun	Alpha	20	14	Y	+	104	4
+ sound	... Noun	Alpha	13	11	Y	+	191	6
quality	... Noun	Alpha	11	10	Y		63	7
sound	... Adj	Alpha	9	9	Y		142	8
good pro...	Noun Gr...	Alpha	8	8	Y		181	9
very	... Adv	Alpha	10	8	N		393	9
+ volume	... Noun	Alpha	8	8	Y	+	106	9
bass	... Noun	Alpha	7	7	Y		194	12
nice	... Adj	Alpha	7	7	Y		82	12
+ speake	... Noun	Alpha	9	7	Y	+	122	12
+ have	... Verb	Alpha	7	6	N	+	398	15

Fig 8. Term-by-document matrix for Philips Home Theatre with ratings of 3

Results - Node: Text Parsing Diagram: Flipkart reviews

File Edit View Window

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ product	... Noun	Alpha	379	346	Y	+	622	1
+ good	... Adj	Alpha	363	322	Y	+	132	2
nice	... Adj	Alpha	219	212	Y		470	3
very	... Adv	Alpha	230	193	N		2051	4
+ be	... Verb	Alpha	279	177	N	+	2096	5
sound	... Adj	Alpha	168	161	Y		842	6
quality	... Noun	Alpha	165	152	Y		351	7
good	... Noun	Alpha	146	136	Y		1205	8
+ sound	... Noun	Alpha	127	117	Y	+	1116	9
bass	... Noun	Alpha	112	106	Y		1131	10
+ good product	... Noun Group	Alpha	95	94	Y	+	1045	11
awesome	... Adj	Alpha	96	91	Y		73	12
i	... Noun	Alpha	102	83	N		2142	13
bass	... Adj	Alpha	68	66	Y		478	14
+ nice product	... Noun Group	Alpha	62	62	Y	+	1162	15
flipkart	... Noun	Alpha	51	50	Y		786	16
+ price	... Noun	Alpha	51	49	Y	+	1773	17
money	... Noun	Alpha	47	47	Y		1023	18
sound quality	... Noun Group	Alpha	45	44	Y		1152	19
value	... Noun	Alpha	39	39	Y		566	20
not	... Adv	Alpha	46	38	N		2065	21
excellent	... Adj	Alpha	38	37	Y		934	22
super	... Adj	Alpha	37	36	Y		1609	23
super	... Noun	Alpha	37	36	Y		1392	23
philips	... Noun	Alpha	36	35	Y		612	25
also	... Adv	Alpha	37	34	N		2026	26
quality	... Adj	Alpha	34	34	Y		874	26
superb	... Adj	Alpha	33	33	Y		1057	28
+ speaker	... Noun	Alpha	35	30	Y	+	728	29
bluetooth	... Noun	Alpha	27	27	Y		704	30
really	... Adv	Alpha	27	25	N		2040	31
best	... Noun	Alpha	26	24	Y		1078	32
happy	... Adj	Alpha	25	24	Y		1367	32
+ have	... Verb	Alpha	27	23	N	+	2062	34
no	... Adv	Alpha	25	23	N		2124	34
+ system	... Noun	Alpha	26	23	Y	+	1306	34
+ go	... Verb	Alpha	21	21	N	+	2044	37
+ love	... Verb	Alpha	23	21	Y	+	247	37
+ buy	... Verb	Alpha	21	20	Y	+	74	39

Fig 9. Term-by-document matrix for Philips Home Theatre with ratings of 4 and 5

Subsequently, the parsing node is connected to a text filter node, as illustrated in the figure below. This node eliminates words that occur less frequently, as specified in the properties panel. The following parameters have been modified from their default settings in the properties panel:

- The minimum number of documents is adjusted to 5.
- The "Check spelling" option is enabled ('yes'), allowing SAS to replace misspelled words with correctly spelled alternatives.

Terms

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
product	77	59	✓	0.213	Noun	Alpha
products	2	2			Noun	Alpha
product	75	58			Noun	Alpha
quality	38	37	✓	0.281	Noun	Alpha
bad	37	34	✓	0.302	Adj	Alpha
sound	25	25	✓	0.358	Adj	Alpha
sound	21	21	✓	0.392	Noun	Alpha
sound	20	20			Noun	Alpha
sounds	1	1			Noun	Alpha
poor	20	20	✓	0.402	Adj	Alpha
good	20	19	✓	0.416	Adj	Alpha
dont	18	18	✓	0.423	Noun	Alpha
money	16	16	✓	0.447	Noun	Alpha
buy	15	13	✓	0.496	Verb	Alpha
working	14	13	✓	0.493	Adj	Alpha
sound quality	12	12	✓	0.504	Noun Group	Alpha

Fig 10. Text filter terms in the Philips Home Theatre reviews, for ratings of 1 and 2

Terms

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
good	19	18	✓	0.326	Adj	Alpha
product	20	14	✓	0.401	Noun	Alpha
product	19	14			Noun	Alpha
products	1	1			Noun	Alpha
good	14	14	✓	0.381	Noun	Alpha
sound	13	11	✓	0.448	Noun	Alpha
sound	11	10			Noun	Alpha
sounds	2	2			Noun	Alpha
quality	11	10	✓	0.467	Noun	Alpha
sound	9	9	✓	0.485	Adj	Alpha
volume	8	8	✓	0.512	Noun	Alpha
volume	6	6			Noun	Alpha
volumes	2	2			Noun	Alpha
good product	8	8	✓	0.512	Noun Group	Alpha
nice	7	7	✓	0.544	Adj	Alpha
speaker	9	7	✓	0.557	Noun	Alpha

Fig 11. Text filter terms in the Philips Home Theatre reviews, for ratings of 3

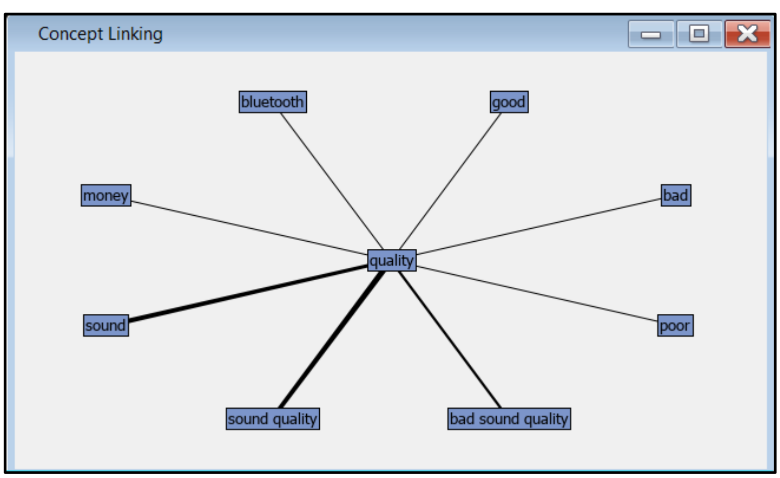
Terms

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
product	379	346	✓	0.153	Noun	Alpha
products	23	23			Noun	Alpha
product	356	327			Noun	Alpha
good	363	322	✓	0.165	Adj	Alpha
good	307	278			Adj	Alpha
better	13	12			Adj	Alpha
best	43	42			Adj	Alpha
nice	219	212	✓	0.221	Adj	Alpha
sound	168	161	✓	0.261	Adj	Alpha
quality	165	152	✓	0.272	Noun	Alpha
good	146	136	✓	0.288	Noun	Alpha
sound	127	117	✓	0.311	Noun	Alpha
sound	119	110			Noun	Alpha
sounds	8	8			Noun	Alpha
bass	112	106	✓	0.323	Noun	Alpha
good product	95	94	✓	0.338	Noun Group	Alpha
awesome	96	91	✓	0.345	Adj	Alpha
bass	68	66	✓	0.391	Adj	Alpha
nice product	62	62	✓	0.398	Noun Group	Alpha
flipkart	51	50	✓	0.431	Noun	Alpha

Fig 12. Text filter terms in the Philips Home Theatre reviews, for ratings of 4 and 5

<u>Concept Linking:</u>

The concept links, accessible through the interactive filter viewer in the properties panel of the text filter node, represent associations between terms within the comments. These links, generated through association analysis, can be created for all terms present in the comments, although it is more meaningful to focus on key terms. The concept link diagram displays the word under analysis at the center, with associated words connected to it via links. The thickness of these links indicates the strength of association between the two words in reviewer comments. Below are the concept links for some of the most frequent terms:



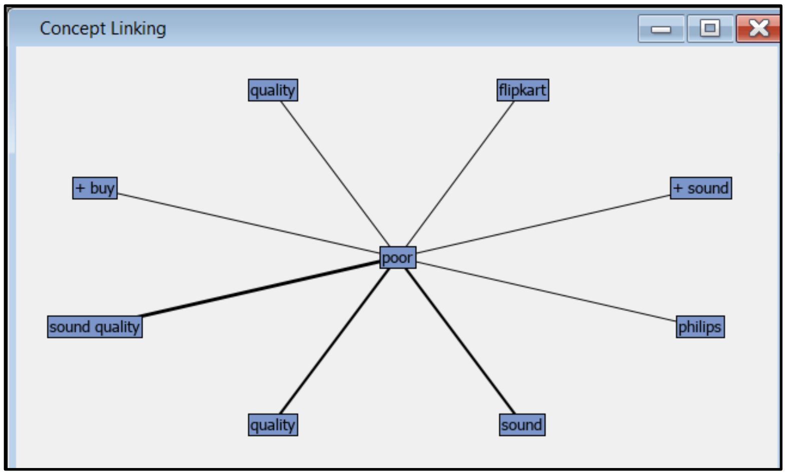


Fig 13. Conceptual associations among terms in the Philips Home Theatre dataset, categorized by ratings of 1 and 2

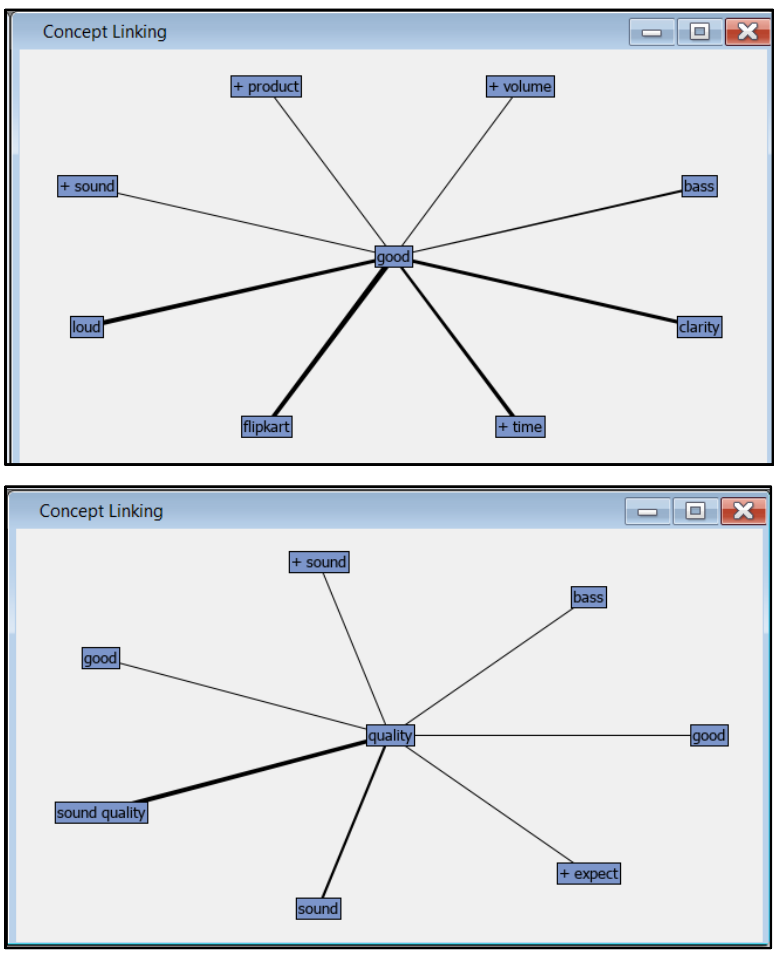


Fig 14. Conceptual associations among terms in the Philips Home Theatre dataset, categorized by ratings of 3

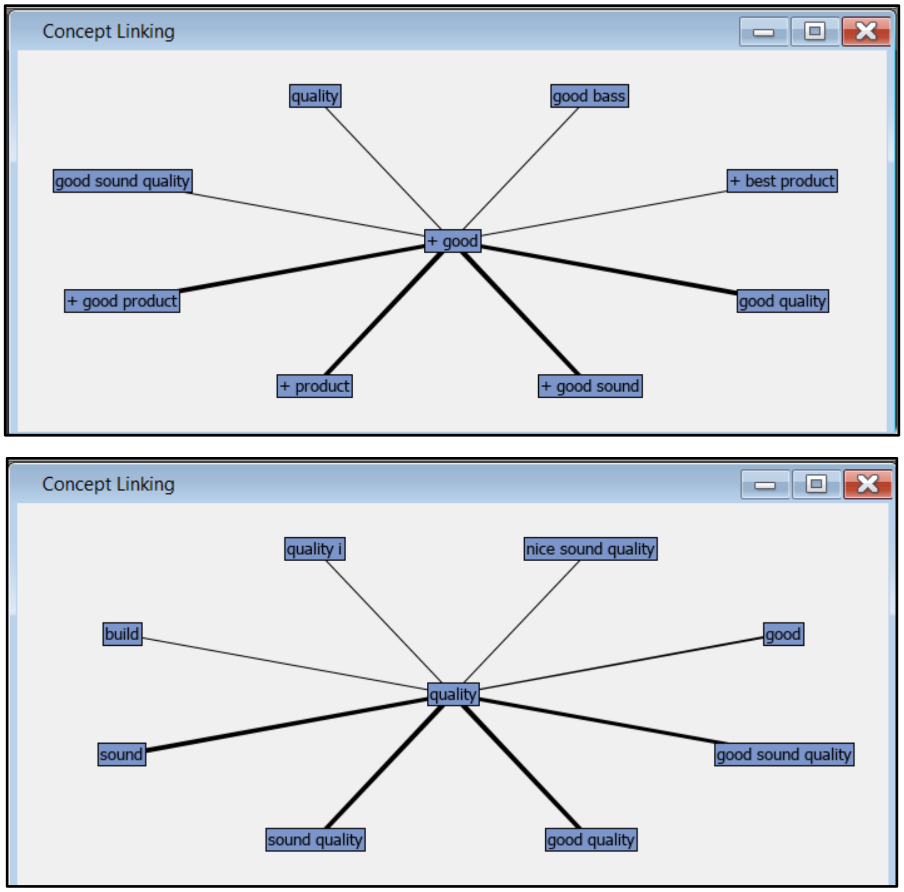


Fig 15. Conceptual associations among terms in the Philips Home Theatre dataset, categorized by ratings of 4 and 5

Following the connection of the Text Filter node to the Text Parsing node in SAS Enterprise Miner, the Text Topic node is linked to the Text Filter node. This allows SAS to aggregate terms into topics, facilitating the extraction of valuable insights from the data. To comprehend the data and identify the features that reviewers are particularly interested in commenting on, the number of Multi-Term Topics has been configured to 10 in the properties panel.

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.381	0.259	quality,sound,sound quality,poor,bad sound quality	4	23
Multiple	2	0.265	0.265	+product,working,+buy,flipkart,worst	3	29
Multiple	3	0.312	0.248	good,+sound,quality,+product,working	2	18
Multiple	4	0.338	0.257	bad,bad product,+product,quality,bad sound quality	3	25
Multiple	5	0.282	0.262	waste,money,dont,buy,quality	5	22

Fig 16. Text topics for the Philips Home Theatre dataset, categorized by ratings of 1 and 2

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.337	0.168	+product,good product,good,bass,price	3	9
Multiple	2	0.265	0.173	quality,sound quality,sound,good,+expect	5	7
Multiple	3	0.273	0.162	good,small,loud,flipkart,bad	2	10
Multiple	4	0.241	0.162	nice,+high,+sound,bass,good sound	1	7
Multiple	5	0.214	0.175	price,high,system,+high,dont	6	6
Multiple	6	0.216	0.180	bass,+volume,audio,poor,+high	8	11
Multiple	7	0.225	0.180	+sound,+speaker,good,+volume,clarity	7	10
Multiple	8	0.202	0.170	money,value,bad,low,noise	5	5
Multiple	9	0.179	0.180	+month,working,connectivity,remote,cable	9	8
Multiple	10	0.202	0.179	music,good,poor,small,normal	8	8

Fig 17. Text topics for the Philips Home Theatre dataset, categorized by ratings of 3

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.323	0.105	+good product,+good,+product,+price,good quality	3	97
Multiple	2	0.301	0.106	nice,+nice product,+product,+nice sound,+sound	3	122
Multiple	3	0.251	0.104	good,quality,bass,+product,delivery	1	125
Multiple	4	0.249	0.116	quality,sound,sound quality,good quality,+good	9	138
Multiple	5	0.219	0.112	+sound,+good sound,+nice sound,superb,nice	7	102
Multiple	6	0.204	0.110	awesome,+awesome product,+product,superb,+love	4	86
Multiple	7	0.173	0.103	super,bass,philips,sound,range	2	36
Multiple	8	0.164	0.113	money,value,best value,worth,+product	2	50
Multiple	9	0.180	0.117	excellent,excellent product,superb,+product,super	10	62
Multiple	10	0.179	0.121	bass,sound,flipkart,quality,super	9	113

Fig 18. Text topics for the Philips Home Theatre dataset, categorized by ratings of 4 and 5

Following the generation of topics by the Text Topic node, the Text Cluster node is connected to further analyze the best and worst features of products based on reviewer ratings. SAS Enterprise Miner facilitates the grouping of closely related terms into distinct clusters of related terms. After experimentation, the properties settings for the Text Cluster node are configured to produce clearly separated clusters in the cluster space. Settings are changed in the properties panel, resulting in a solution with a maximum of 10 clusters and 15 descriptive terms per cluster, utilizing the Expectation-Maximization Cluster Algorithm.

Clusters

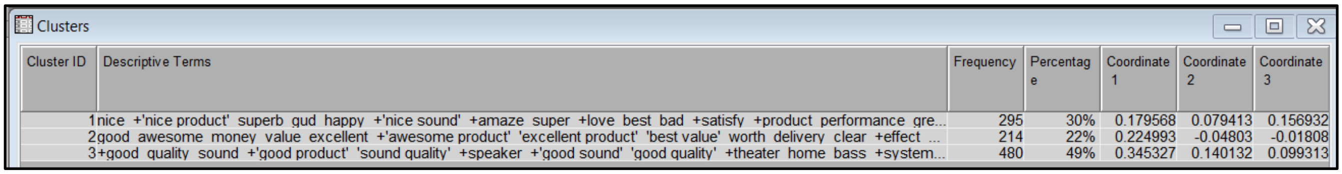
Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordin 4
1	+low +sound good bass poor +satisfy quality +product bad sound ...	48	29%	0.179068	0.148996	-0.00795	-0.03
2	+buy working worst flipkart +speaker 'bad product' 'worst product' +day philips +product money waste +satisfy ...	85	52%	0.459854	-0.00352	0.122936	0.068
3	'sound quality' sound quality 'bad sound quality' buy poor dont bad bluetooth service good waste bass money ...	32	19%	0.380612	0.552247	-0.13154	-0.02

Fig 19. Text clusters for the Philips Home Theatre dataset, categorized by ratings of 1 and 2

Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Co 5
1	sound 'sound quality' +high cable +expect +month +option poor remote +connection best high increase ...	38	49%	0.223765	0.130203	0.067104	-0.02525	0.
2	'good sound' awesome 'good product' nice music good +product +speaker +feature +satellite average n...	21	27%	0.257164	0.230792	0.190972	-0.18104	-
3	+compare bad loud memory value money small good +lack +mention +satellite +setting bluetooth buy ...	18	23%	0.329452	0.042264	-0.14727	-0.16278	0.

Fig 20. Text clusters for the Philips Home Theatre dataset, categorized by ratings of 3



Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3
1	nice +'nice product' superb gud happy +'nice sound' +amaze super +love best bad +satisfy +product performance gre...	295	30%	0.179568	0.079413	0.156932
2	good awesome money value excellent +'awesome product' 'excellent product' 'best value' worth delivery clear +effect ...	214	22%	0.224993	-0.04803	-0.01808
3	+good quality sound +'good product' 'sound quality' +speaker +'good sound' 'good quality' +theater home bass +system...	480	49%	0.345327	0.140132	0.099313

Fig 21. Text clusters for the Philips Home Theatre dataset, categorized by ratings of 4 and 5

5.1.2 Category 2 (Footwear):

<u>Men's Black Sandals</u>

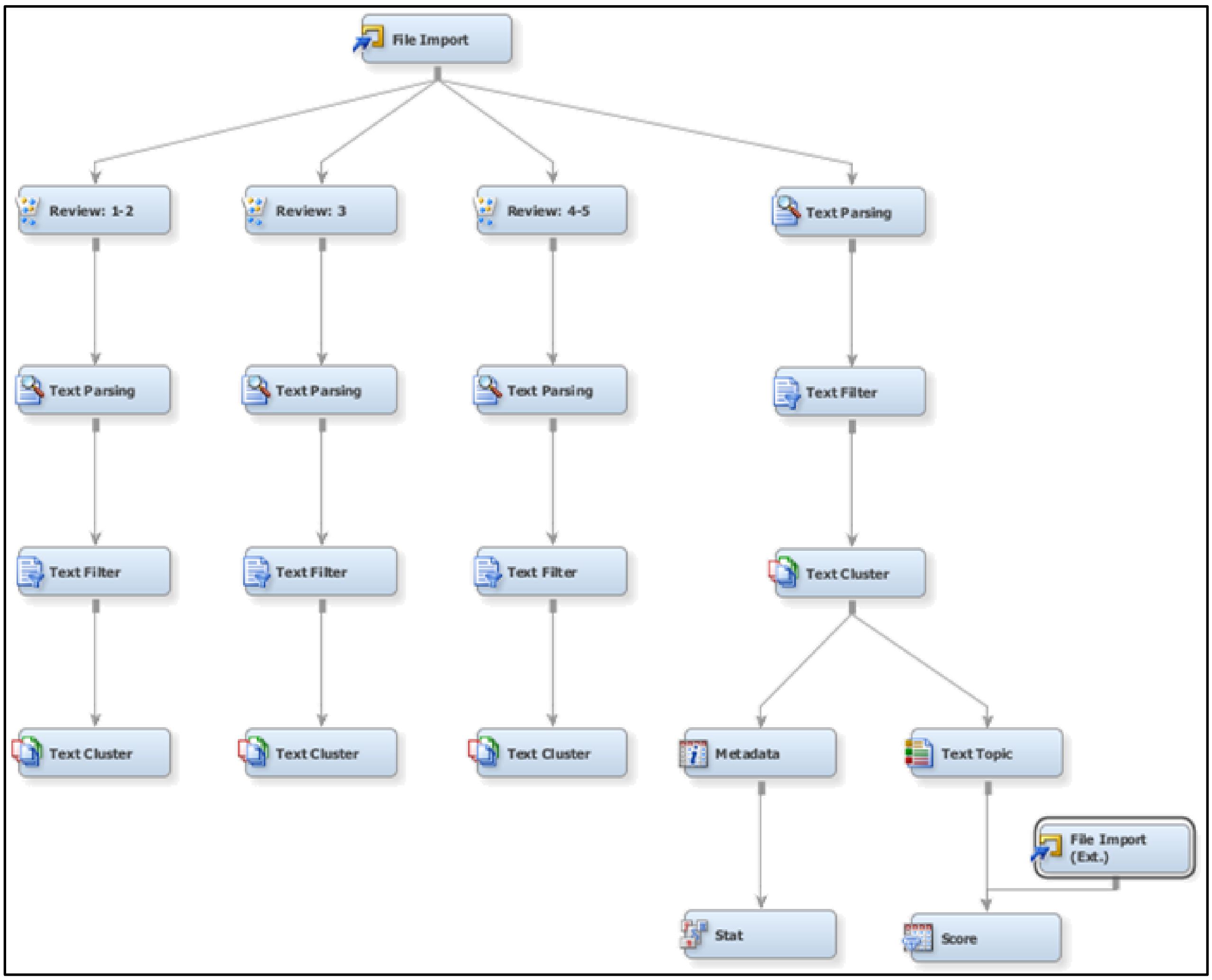


Fig 22. Segment of flow diagram with unsupervised model

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	RMS Std.	Descriptive Term Sample	X	Y
1	+quality +dont buy 'bad quality' item 'low quality' +'product dont' +expect 'dont buy' +break +side purchase 'poor quality' +...	237	59%	0.816401	-0.28546	0.124528	0.222127	+quality...	-0.72309	0.812959
2	good cheap +'cheap quality' +worth +look 'at all' material +sandal comfortable +minute +wear +month bad +worst +qualit...	17	4%	0.241308	-0.51836	-0.78907	0.151951	good c...	1.781316	-0.09595
3	+product +good +'bad product' +'worst product' +bed 'good product' size 'waste product' useless bad +worst +show waste ...	150	37%	0.853172	0.300893	-0.15178	0.208641	+produc...	-1.05822	-0.71701

Fig 23. Text clusters for the Men's Black Sandals dataset, categorized by ratings of 1 and 2

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	RMS Std.	Descriptive Term Sample	X	Y
1	'good product' 'heavy weight' heavy weight size good +nice comfortable +product +average...	61	65%	0.127437	-0.73987	0.02307	0.264069	'good prod...	0.705674	1.005826
2	+quality +look low +sandal week +average +product comfortable +bad +nice good ...	23	24%	0.881233	-0.21453	0.109352	0.268309	+quality +l...	0.951452	-0.91106
3	+bad ...	10	11%	0.193467	-0.00619	-0.98021	0.007968	+bad	-1.65713	-0.09477

Fig 24. Text clusters for the Men's Black Sandals dataset, categorized by ratings of 3

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	RMS Std.	Descriptive Term Sample	X	Y
1	+nice +'nice product' 'best product' amazing gud supar wow +awesome +price +product +best excellent +sandal worth bea...	172	46%	0.023938	-0.4797	-0.40185	0.295878	+nice ...	-0.85204	0.629802
2	+bad +price +good +product ...	8	2%	0.98714	0.036522	-0.0021	0.072195	+bad +...	1.868298	-0.03585
3	+good good +quality +super +'good product' comfortable +delivery 'good i' boy design stylish thanks 'good quality' nice ite...	198	52%	0.198478	-0.32778	-0.85972	0.191189	+good ...	-1.01626	-0.59395

Fig 25. Text clusters for the Men's Black Sandals dataset, categorized by ratings of 4 and 5

5.1.3 Category 3 (Cooler):

MAHARAJA WHITELINE 65 L Desert Air Cooler

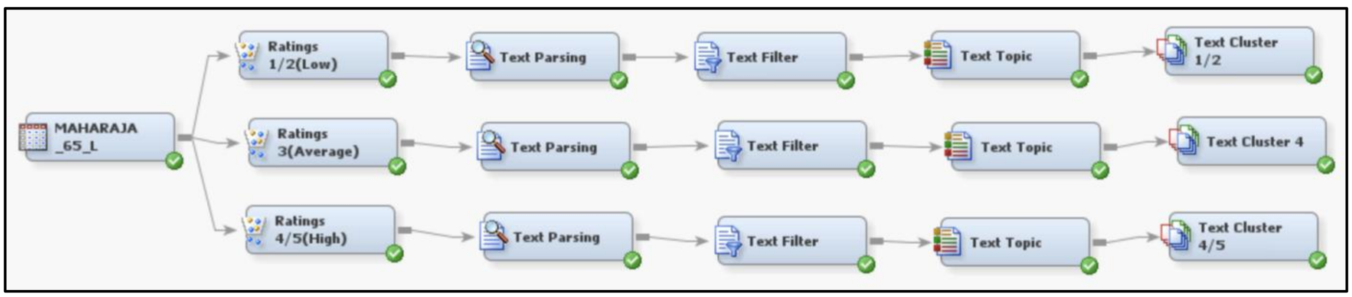


Fig 26. Flow diagram with unsupervised model

Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6	Coord
1	+replace +issue buy dont automatically problem swing cant +day +product +receive +return worst care defective ...	41	26%	0.39726	0.07121	0.057582	-0.1016	-0.01127	0.381129	
2	+cooler +review room cool +cool air +pad honey fan bad +lot +recommend +experience +disappoint +month ...	53	34%	0.611138	-0.21681	-0.16107	0.139902	0.087435	-0.06754	-0
3	speed delivery +service quality money +work waste good flow pump poor back low crompton properly ...	62	40%	0.293634	-0.0901	-0.00929	-0.14259	0.158312	0.280419	0.

Fig 27. Text clusters for the Maharaja Air Cooler dataset, categorized by ratings of 1 and 2

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6	Coordinate
1	body front +write problem summer good quality average better hot motor +month cool fan noise ...	54	47%	0.144332	-0.23985	0.297031	-0.46402	0.278816	0.140278	-0.02
2	+clean sound upto properly difficult high +big fine +wheel +work speed +finish buy design dont ...	25	22%	0.176485	-0.2628	-0.0522	0.021638	0.024749	0.166519	0.0755
3	+look bad +expect +receive +satisfy color time water flipkart value +place +room flow money air ...	35	31%	0.125674	-0.08089	0.256514	-0.31474	0.248631	0.043213	0.2778

Fig 28. Text clusters for the Maharaja Air Cooler dataset, categorized by ratings of 3

Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6
1	good +product nice +cool happy perfect +swing fantastic superb +sound awesome super plastic great capacity ...	306	34%	0.563827	0.214341	0.16078	-0.33945	0.096642	0.047057
2	quality +little flow desert +wheel noisy +big bit tank crompton water overall better superb +day ...	257	28%	0.303008	0.066329	0.064358	0.013294	0.041536	0.087358
3	+cooler delivery buy flipkart +deliver money best +satisfy excellent cool +work +throw awesome ventilation +value ...	350	38%	0.379749	0.024769	0.077447	-0.21562	0.025228	-.000369

Fig 29. Text clusters for the Maharaja Air Cooler dataset, categorized by ratings of 4 and 5

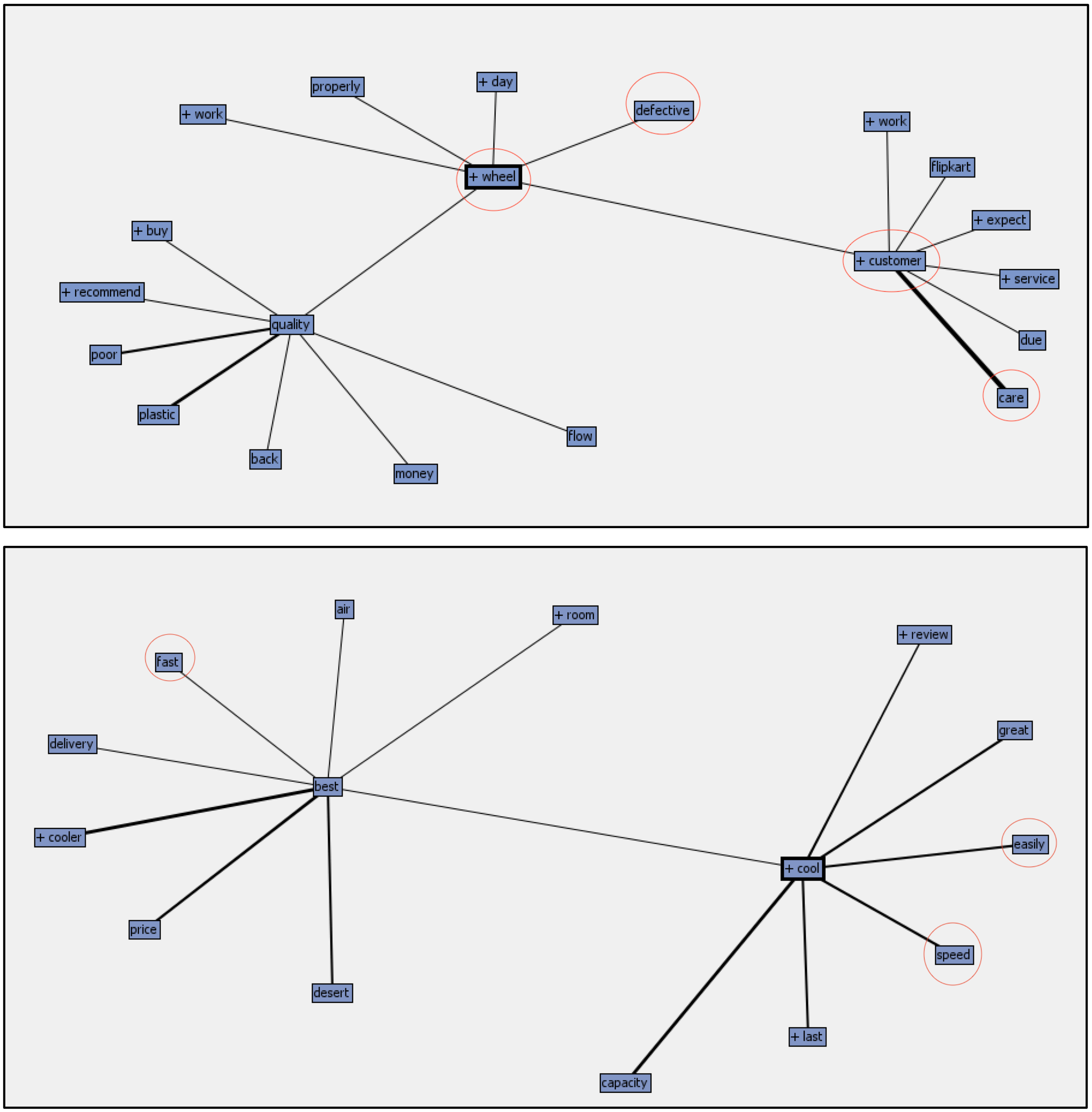


Fig 30. Conceptual associations among terms in Maharaja Air Cooler dataset

5.1.4 Category 4 (Hard Disk):

<u>Seagate Portable 2 TB External Hard Disk Drive (HDD)(Light Blue)</u>

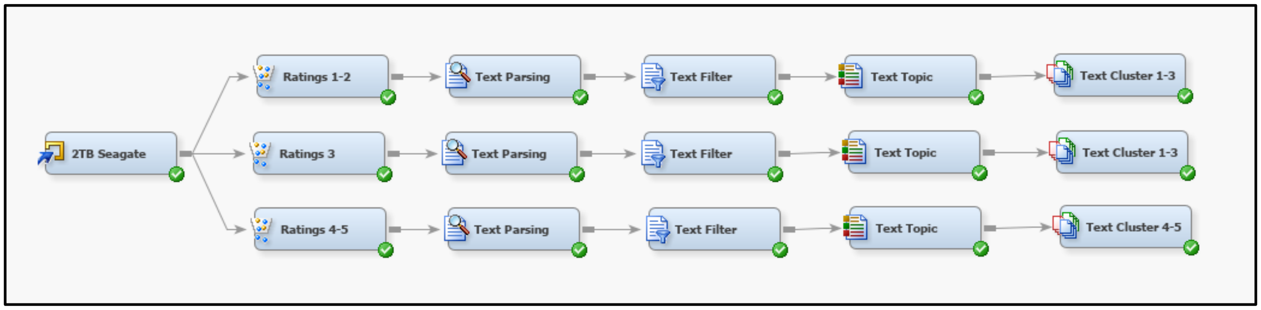
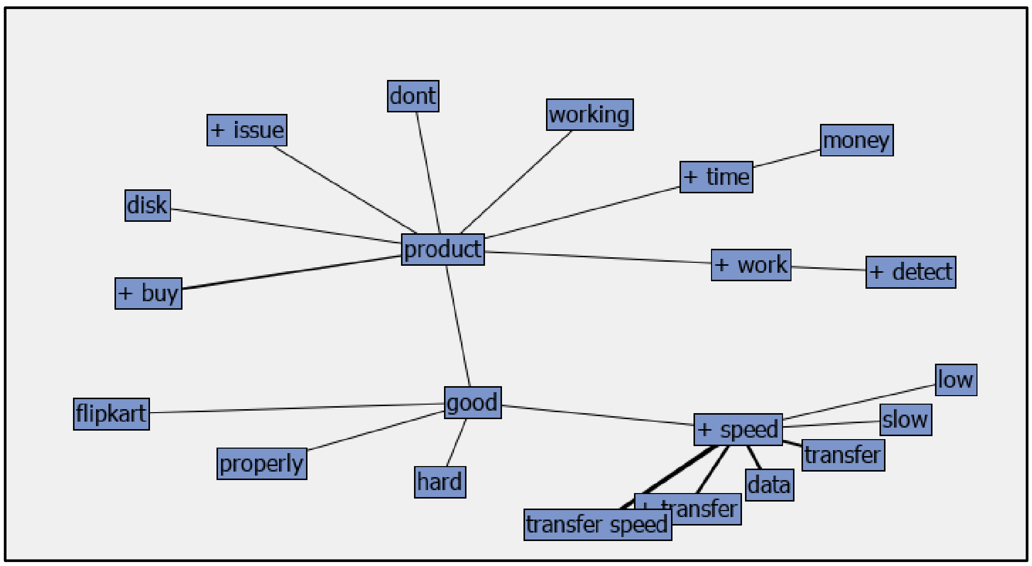
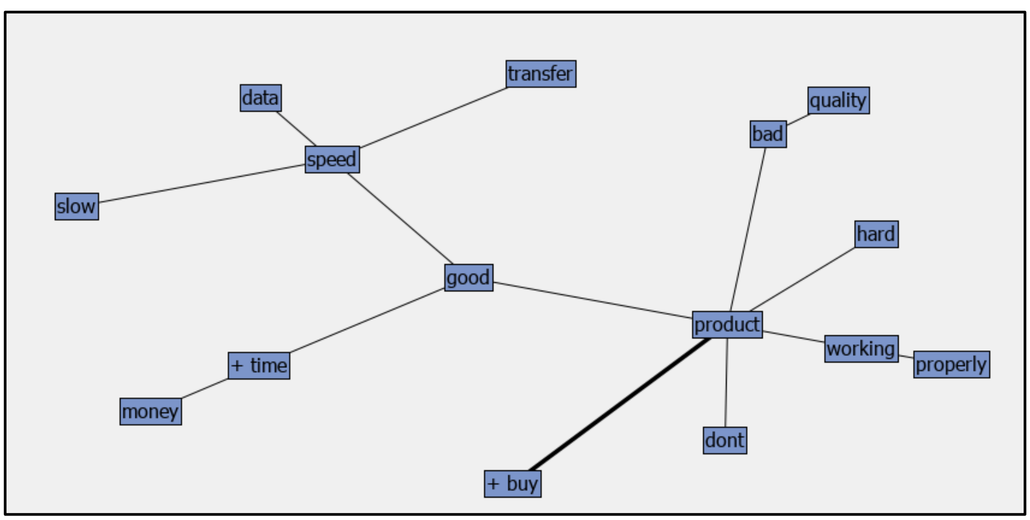


Fig 31. Flow diagram with unsupervised model





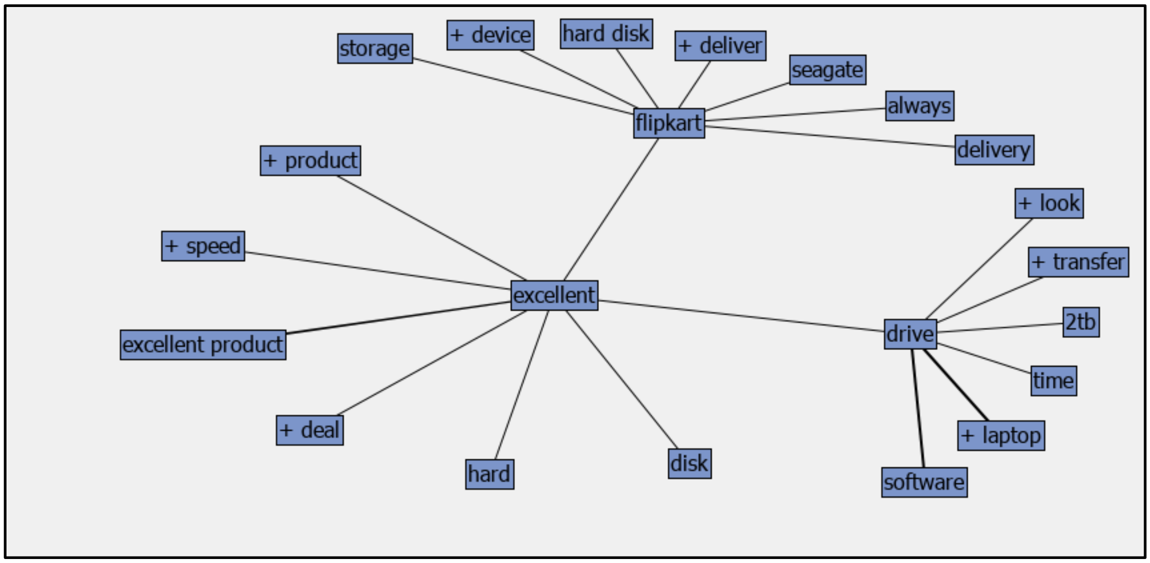


Fig 32. Conceptual associations among terms in Seagate 2TB Hard Disk dataset

1	...	17
2	flipkart nice money product +time good +buy quality working ...	25
3	+speed 'transfer speed' transfer data slow +transfer low +time poor dont +work +detect hdd bad good ...	25
4	disk hard properly working +work +time +detect slow +buy data bad good product +speed ...	9
5	working good ...	15
6	+issue +work product properly +buy dont hard working +speed good ...	7
7	+file hdd +detect +transfer transfer low bad working good ...	11
8	quality bad +buy dont poor low +issue money properly product hdd working +speed ...	20

Fig 33. Text clusters for the Seagate 2TB Hard Disk dataset, categorized by ratings of 3

Clusters

Cluster ID	Descriptive Terms	Frequency
1	quality money poor +time good properly bad product working ...	24
2	transfer slow speed data +time good hdd poor ...	12
3	working hard properly good product +time data slow speed +buy ...	18
4	+detect bad hdd good data hard speed product working ...	14
5	dont +buy product money speed working +time quality bad ...	12

Fig 34. Text clusters for the Seagate 2TB Hard Disk dataset, categorized by ratings of 3

Clusters

Cluster ID	Descriptive Terms	Frequency
1	good ...	168
2	+speed quality +buy excellent great best data seagate +look transfer +transfer delivery money flipkart hard ...	446
3	+product nice +good +'good product' +'nice product' awesome 'awesome product' 'best product' +deliver ...	191

Fig 35. Text clusters for the Seagate 2TB Hard Disk dataset, categorized by ratings of 3

5.1.5 Category 5 (Perfume):

Bella vita organic Luxury Unisex Perfume Gift Set

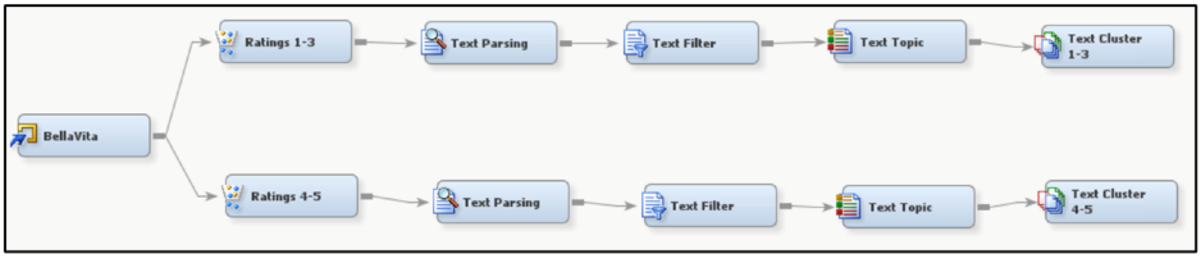


Fig 36. Flow diagram with unsupervised model

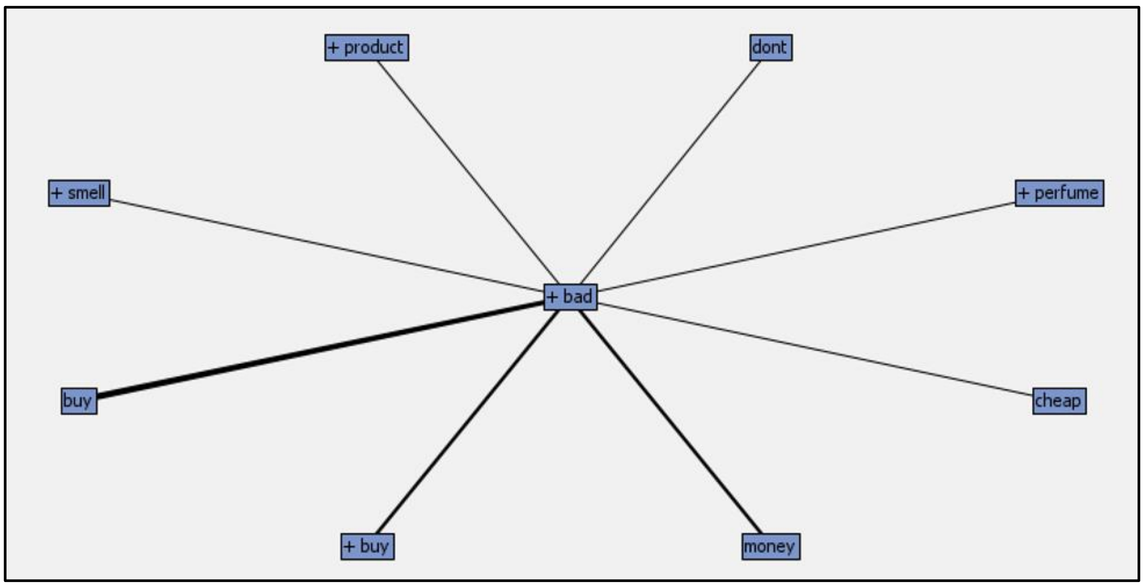


Fig 37. Conceptual associations among terms in Bella Vita Perfume dataset

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	RMS Std.	Descriptive Term Sample	X	Y
1	long +last +perfume +product good +smell ...	40	30%	0.100244	-0.29297	-0.0245	0.014056	-0.05196	0.215412	long +last ...	0.23357	0.841035
2	waste money dont buy +smell +fragrance long +bad +last +product good ...	10	7%	0.622891	0.189047	-0.37456	0.474847	-0.01709	0.21779	waste mon...	1.619619	0.329759
3	good +smell cheap +product ...	18	13%	0.41229	0.05466	0.24572	0.206767	0.832801	0.077283	good +sm...	-1.06504	-1.09466
4	+bad small +product +buy +perfume buy dont cheap money +smell +fragranc...	47	35%	0.572337	0.015842	0.542511	0.071584	-0.08689	0.272967	+bad small...	0.557442	-1.05034
5	+fragrance good +smell +buy +perfume dont waste money +last +product ...	20	15%	0.536657	-0.06023	-0.16709	-0.64874	0.374814	0.158629	+fragrance ...	-1.34559	0.974196

Fig 38. Text clusters for the Bella Vita Perfume dataset, categorized by ratings of 1,2 and 3

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4	Coordinate 5	Coordinate 6	C
1	excellent +'nice fragrance' money happy superb worth +blow fragnance mind osm quality super strong +fragrance +amaze ...	118	21%	0.174472	-0.08637	0.065368	0.077109	-0.06857	0.109847	
2	nice +good +'good product' +'nice perfume' 'nice smell' +'good smell' super 'best product' best +product 'best perfume' 'product i' flipkart ...	152	27%	0.163048	-.000437	0.107306	0.123784	0.03173	0.269208	
3	long +buy budget lasting +last +package +look +hour different +perfume perfect +long friendly +gift bella ...	135	24%	0.361649	-0.0687	0.027193	0.017174	-0.04319	-0.01944	
4	good +'nice product' +product awesome nice packing +'awesome fragrance' awsm range wow price super +gift +love +fragrance ...	92	17%	0.108461	-0.06505	0.132613	0.155239	-0.03261	0.241252	
5	'amazing product' +love amazing +'amazing fragrance' fresh +smell cute 'product i' +order pack +'different fragrance' mild vita awsm bel...	58	10%	0.345199	-0.10955	0.079261	0.023053	-0.1337	0.041077	

Fig 39. Text clusters for the Bella Vita Perfume dataset, categorized by ratings of 4 and 5

5.2 Supervised Modelling

Supervised modeling is an approach where the data is labelled to train the algorithms. This means that each piece of text the training set is associated with a label or outcome, such as a category, class, or numerical value. The process begins with a dataset where each text document is tagged with a label. In the context of sentiment analysis, this label is "positive," "negative," or "neutral." From this text data, features are extracted that the machine learning model can use. Features include the frequency of specific words or phrases, the presence or absence of certain terms, and statistical measures. Common models in text mining include logistic regression, decision trees, support vector machines (SVM), and neural networks. The model learns to associate the extracted features with the labels from the training set.

The model's performance is evaluated using metrics like accuracy, precision, recall, and F1 score. They are used to measure how well the model has learned to predict the labels based on the text's features. Supervised models require a considerable amount of labeled data to learn effectively and generalize to new data. The quality and quantity of the training data significantly influence the performance of the model.

5.2.1 Category 1 (Home Theatre):

PHILIPS MMS8085B/94 Convertible 80 W Bluetooth Home Theatre

The flow diagram provided outlines the flow process for supervised sentiment analysis using text data. Here's a breakdown of the stages and what they represent:

1. **Data Partition:** The process begins with a dataset from a source labeled 'Philips'. This data is partitioned into subsets, for training and testing purposes.
2. **Text Parsing:** The partitioned data is then parsed. Text parsing involves preprocessing steps such as tokenization, stemming, removing stop words, etc., to prepare the text for analysis.
3. **Text Filtering:** After parsing, there are four different filtering techniques applied:

i) **Entropy:** This is filtering based on information gain or entropy, which is used in decision trees to determine the splits that maximize information gain.

ii) **IDF (Inverse Document Frequency):** A weighting measure used in text mining that reduces the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

iii) **No weight:** This suggests that one of the pathways involves using the parsed text without any additional weighting or filtering.

iv) **Mutual Information:** A measure of the mutual dependence between two variables. In this context, it's used to filter out the text based on the relevance of terms to the target sentiment classification.

4. **Text Clustering:** Following the filtering, text clustering is performed, to group similar texts together. This is a part of feature engineering to create new variables that represent these clusters.

5. **Text Topic:** After clustering, a topic modeling step is conducted. Topic modeling is used to find the abstract "topics" that occur in a collection of documents.

6. **Modeling Techniques:** For each of the filtered datasets, several predictive models are built:

   i) **Neural Network:** A deep learning model designed to find patterns in the data.

   ii) **Decision Tree:** A model that uses a tree-like model of decisions and their possible consequences.

   iii) **Regression:** Suitable for classification (like logistic regression), predicting a numeric sentiment score.

7. **Model Comparison:** All the different models generated from different filtered datasets are then compared in the final stage. The comparison is based on various metrics such as accuracy, precision, recall, F1 score, or area under the ROC curve, to determine the best-performing model.

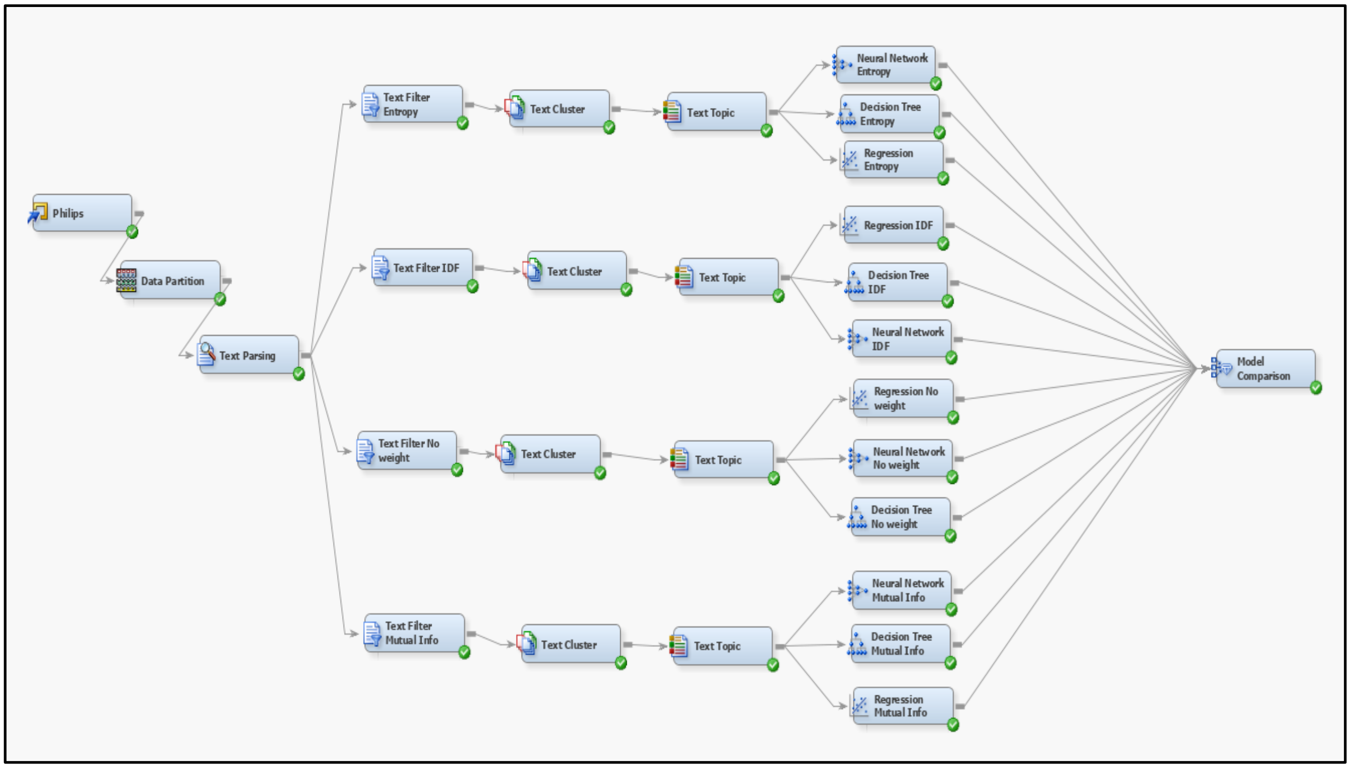


Fig 40. Flow diagram with supervised model for the Philips Home Theatre dataset

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Roc Index	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error
Y	Neural4	Neural4	Neural Network Mutual Info	Sentiment	Sentiment	0.974	864.9382	0.029982	0.117858	1174	302	1476	2214	260.9382	0.045408	0.995917
	Neural3	Neural3	Neural Network No weight	Sentiment	Sentiment	0.973	842.2901	0.026575	0.107629	1174	302	1476	2214	238.2901	0.040247	0.997056
	Reg2	Reg2	Regression IDF	Sentiment	Sentiment	0.968	351.0239	0.037275	0.149514	1466	10	1476	2214	331.0239	0.037784	0.996317
	Reg	Reg	Regression Mutual Info	Sentiment	Sentiment	0.967	350.9561	0.037016	0.15129	1468	8	1476	2214	334.9561	0.03742	0.997541
	Reg3	Reg3	Regression Entropy	Sentiment	Sentiment	0.965	329.3394	0.034086	0.136106	1462	14	1476	2214	301.3394	0.034738	0.993112
	Neural2	Neural2	Neural Network IDF	Sentiment	Sentiment	0.953	867.7724	0.019838	0.070358	1120	356	1476	2214	155.7724	0.032449	0.905817
	Reg4	Reg4	Regression No weight	Sentiment	Sentiment	0.951	354.4512	0.035927	0.149255	1464	12	1476	2214	330.4512	0.036516	0.998791
	Neural	Neural	Neural Network Entropy	Sentiment	Sentiment	0.95	973.7671	0.028278	0.112813	1114	362	1476	2214	249.7671	0.046655	0.994979
	Tree4	Tree4	Decision Tree Mutual Info	Sentiment	Sentiment	0.901	.	0.038413	.	.	.	1476	2214	.	.	0.983165
	Tree2	Tree2	Decision Tree IDF	Sentiment	Sentiment	0.901	.	0.03624	.	.	.	1476	2214	.	.	0.983165
	Tree	Tree	Decision Tree Entropy	Sentiment	Sentiment	0.847	.	0.047668	.	.	.	1476	2214	.	.	0.969388
	Tree3	Tree3	Decision Tree No weight	Sentiment	Sentiment	0.847	.	0.047668	.	.	.	1476	2214	.	.	0.969388

Fig 41. Model Comparison for the Philips Home Theatre dataset

		Miscalculation Rate			ROC		
Model	**Model Description**	**Train**	**Valid**	**Test**	**Train**	**Valid**	**Test**
Neural4	Neural Network Mutual Info	5.83%	6.94%	7.26%	98.90%	96.10%	97.40%
Neural3	Neural Network No weight	4.88%	6.53%	6.85%	99.00%	96.00%	97.30%
Reg2	Regression IDF	7.45%	6.94%	8.47%	97.30%	95.70%	96.80%
Reg	Regression Mutual Info	6.78%	6.94%	7.66%	97.00%	93.20%	96.70%
Reg3	Regression Entropy	6.64%	6.94%	8.47%	97.40%	95.30%	96.50%
Neural2	Neural Network IDF	3.79%	7.35%	10.48%	99.90%	96.00%	95.30%
Reg4	Regression No weight	7.18%	7.35%	7.66%	96.70%	96.60%	95.10%
Neural	Neural Network Entropy	5.56%	8.57%	8.47%	98.80%	96.70%	95.00%
Tree4	Decision Tree Mutual Info	6.64%	6.53%	7.66%	91.20%	89.60%	90.10%
Tree2	Decision Tree IDF	6.10%	6.12%	7.66%	91.20%	89.60%	90.10%
Tree	Decision Tree Entropy	7.59%	6.94%	7.26%	83.20%	83.80%	84.70%
Tree3	Decision Tree No weight	7.59%	6.94%	7.26%	83.20%	83.80%	84.70%

Table 3. Model Comparison Based on Miscalculation Rate and ROC

The table compares different machine learning models based on their miscalculation rates (errors) and ROC (Receiver Operating Characteristic) values, both for training, validation, and testing phases. Models include various neural network, regression, and decision tree configurations, with different feature selection methods like Mutual Information, IDF, and Entropy.

Key observations:

1. Neural Network with No weight has the lowest miscalculation rate in the test set.
2. Neural Network with Mutual Info has the highest ROC value in the test set.
3. Models generally perform better on training data than on validation or test data, indicating potential overfitting.
4. There's a trade-off between miscalculation rate and ROC score across models.

5.2.2 Category 2 (Footwear):

<u>Men's Black Sandals</u>

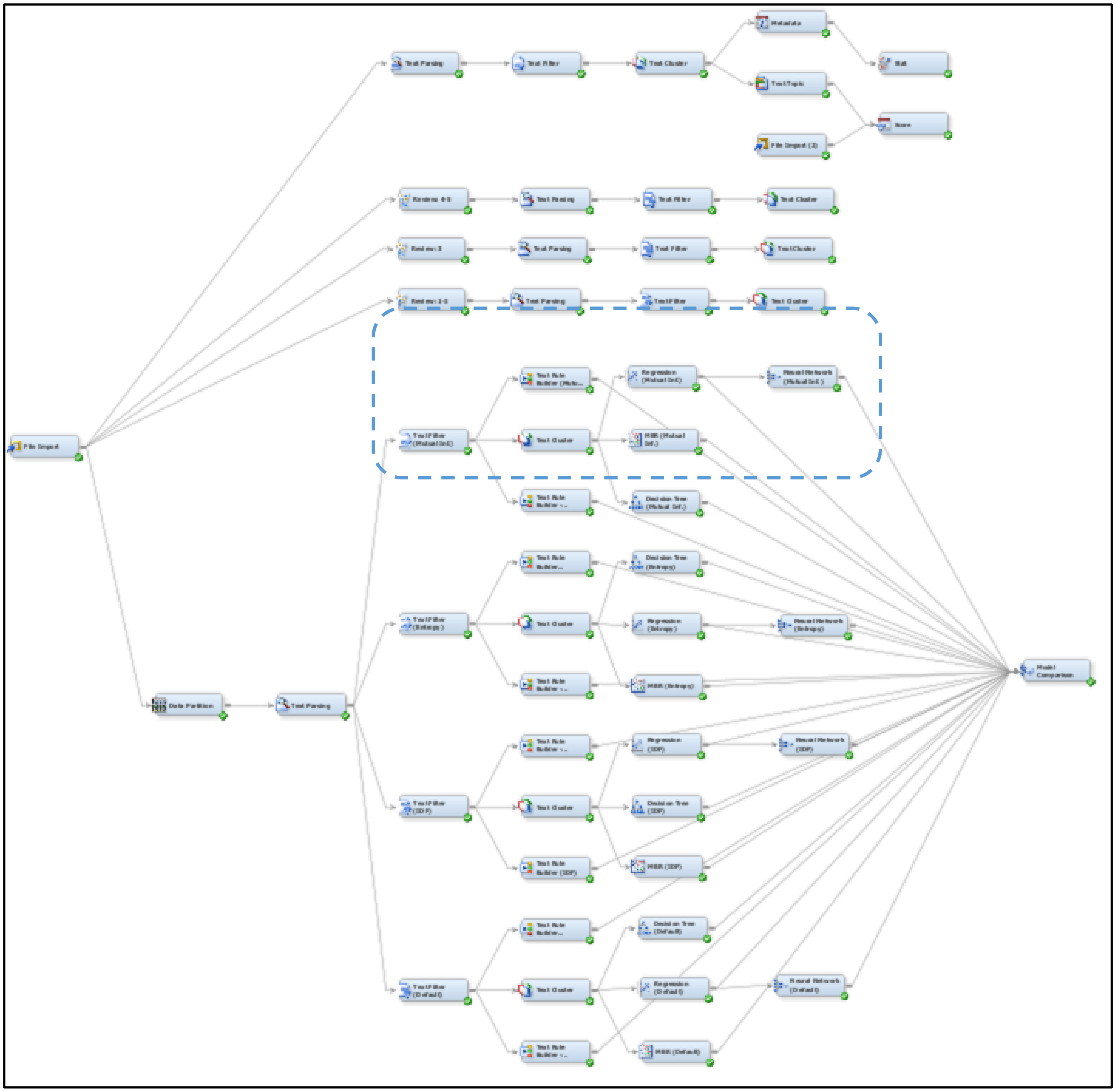


Fig 42. Complete flow diagram with supervised and unsupervised models

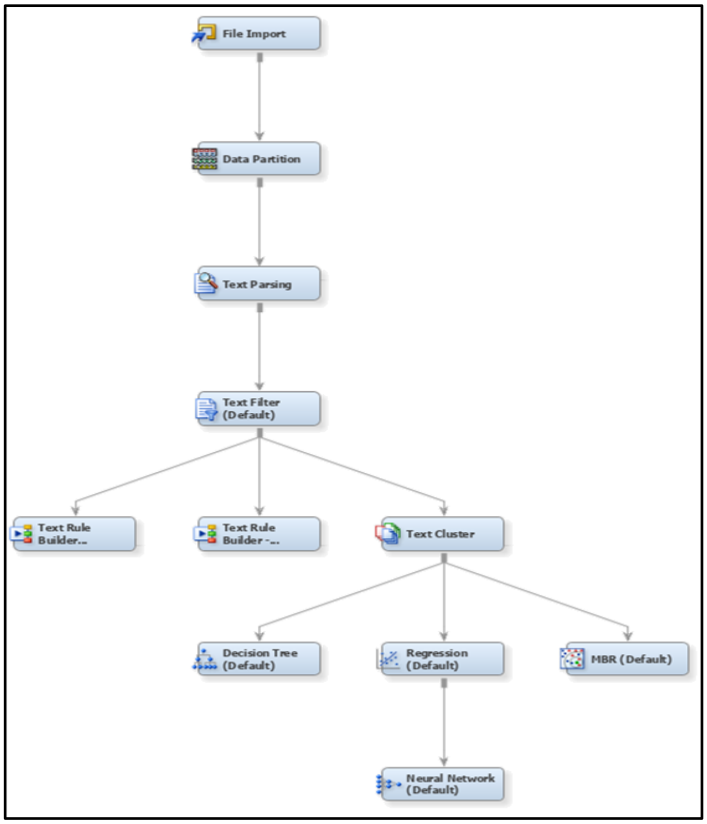


Fig 43. Segment of Flow Diagram with Supervised Model

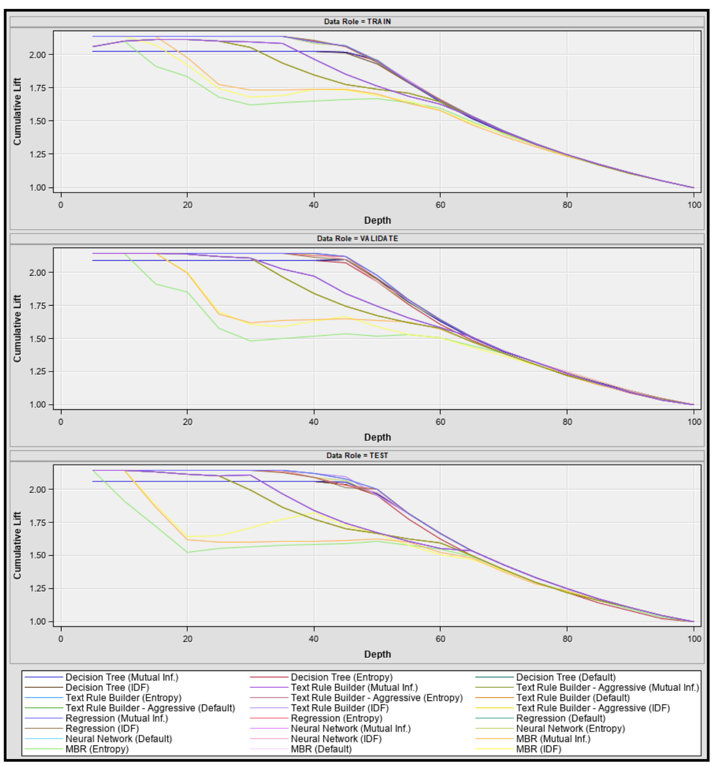


Fig 44. Lift Curve

		Miscalculation Rate			ROC		
Model	**Model Description**	**Train**	**Valid**	**Test**	**Train**	**Valid**	**Test**
Neural2	Neural Network (Default)	9.4%	6.3%	7.9%	99.1%	98.7%	99.5%
Neural4	Neural Network (Mutual Inf.)	9.4%	6.3%	7.9%	99.1%	98.7%	99.5%
Reg2	Regression (Default)	8.8%	5.7%	7.9%	99.0%	98.7%	99.5%
Reg4	Regression (Mutual Inf.)	8.8%	5.7%	7.9%	99.0%	98.7%	99.5%
Neural3	Neural Network (Entropy)	8.2%	7.4%	8.4%	99.1%	98.3%	99.3%
Reg3	Regression (Entropy)	8.4%	8.0%	8.4%	99.1%	98.4%	99.2%
Reg	Regression (IDF)	8.4%	8.5%	9.0%	99.0%	98.4%	99.1%
Neural	Neural Network (IDF)	7.7%	8.0%	9.6%	99.2%	98.8%	99.0%
Tree2	Decision Tree (Default)	8.2%	7.4%	7.9%	96.8%	97.5%	98.1%
Tree4	Decision Tree (Mutual Inf.)	8.2%	7.4%	7.9%	96.8%	97.5%	98.1%
Tree	Decision Tree (IDF)	9.4%	6.8%	9.6%	96.6%	97.5%	98.0%
Tree3	Decision Tree (Entropy)	7.5%	7.4%	9.6%	96.8%	96.4%	95.8%
TextRule2	Text Rule Builder (IDF)	18.6%	21.0%	21.3%	95.3%	94.1%	92.9%
TextRule4	Text Rule Builder (Default)	18.6%	21.0%	21.3%	95.3%	94.1%	92.9%
TextRule6	Text Rule Builder (Entropy)	18.6%	21.0%	21.3%	95.3%	94.1%	92.9%
TextRule8	Text Rule Builder (Mutual Inf.)	18.6%	21.0%	21.3%	95.3%	94.1%	92.9%
TextRule	Text Rule Builder - Aggressive (IDF)	16.1%	20.5%	19.7%	94.1%	91.9%	91.3%
TextRule3	Text Rule Builder - Aggressive (Default)	16.1%	20.5%	19.7%	94.1%	91.9%	91.3%
TextRule5	Text Rule Builder - Aggressive (Entropy)	16.1%	20.5%	19.7%	94.1%	91.9%	91.3%
TextRule7	Text Rule Builder - Aggressive (Mutual Inf.)	16.1%	20.5%	19.7%	94.1%	91.9%	91.3%
MBR	MBR (IDF)	20.1%	23.9%	21.9%	89.1%	85.6%	87.3%
MBR2	MBR (Default)	20.9%	21.6%	21.9%	89.4%	86.8%	84.9%
MBR4	MBR (Mutual Inf.)	20.9%	21.6%	21.9%	89.4%	86.8%	84.9%
MBR3	MBR (Entropy)	20.3%	22.7%	21.3%	87.5%	82.5%	83.7%

Table 4. Model Comparison Based on Miscalculation Rate and ROC

This table compares various models based on miscalculation rate and ROC values across training, validation, and testing datasets. Models include neural networks, regression, decision trees, text rule builders, and MBR (Model-Based Reasoning). Text rule builder models and MBR models have significantly higher miscalculation rates, especially in testing. Neural networks and regression models with default and Mutual Information (Mut. Inf.) settings perform best, with lower miscalculation rates and high ROC scores, suggesting better predictive power and generalization. The consistent test miscalculation rates across several models indicate stable model performance from training to unseen data.

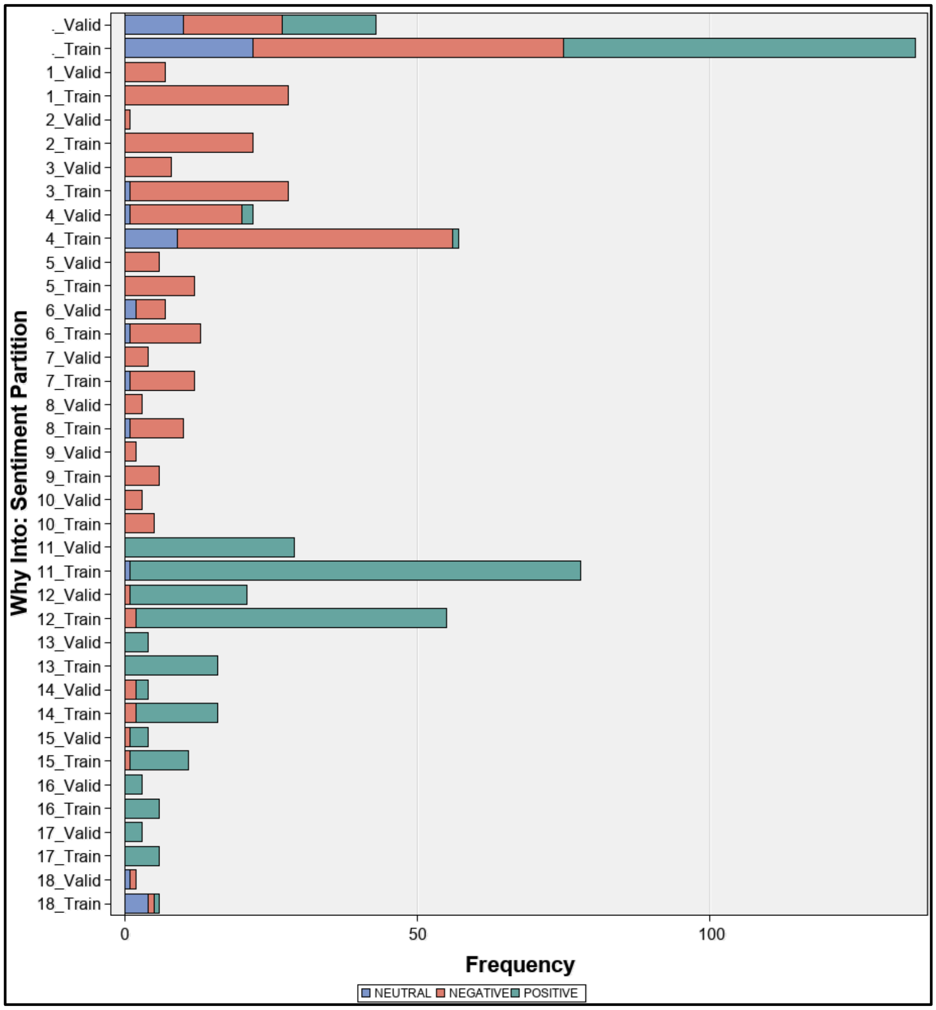


Fig 45. Rule Success Graph

The graph displays the frequency of sentiment predictions made by a text rule builder model, broken down by sentiment category (neutral, negative, positive) and dataset partition (training, validation). Each rule, numbered from 1 to 18, corresponds to a specific set of text-based criteria used to assign a sentiment. The length of each bar represents the frequency of occurrences that a particular rule predicted a certain sentiment. By the frequency distribution, the majority of rules predict the 'neutral' sentiment more often than 'negative' or 'positive', especially in the validation set. This indicates that the rules are either more conservative in their classification or that the dataset contains more neutral instances.

Rule no. 11, 12, 13 and 14 are the best rules classifying reviews as positive. It suggests that these rules likely have criteria or patterns that effectively capture language or expressions commonly found in positive sentiment. This might be due to the presence of positive keywords, phrases, or perhaps even the structure of the sentences that these rules are identifying.

It's also possible that these rules are not overfit to the training data, as they seem to perform consistently on the validation set as well. Their effectiveness would be better judged by looking at precision, recall, or F1 score specifically for the positive class, but from the frequency perspective, they are the most utilized for positive classifications in the model.

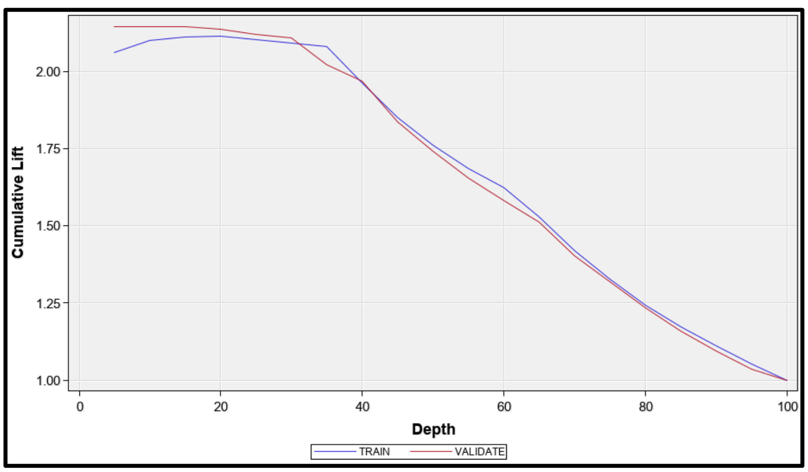


Fig 46. Cumulative Lift

Target Value	Rule #	Rule	Precision	Recall	F1 score
NEGATIVE	1	worst	100.0%	11.76%	21.05%
NEGATIVE	2	bad product	100.0%	21.01%	34.72%
NEGATIVE	3	poor	98.72%	32.35%	48.73%
NEGATIVE	4	bad	91.85%	52.10%	66.49%
NEGATIVE	5	low	92.52%	57.14%	70.65%
NEGATIVE	6	bad	92.50%	62.18%	74.37%
NEGATIVE	7	waste	92.44%	66.81%	77.56%
NEGATIVE	8	dont	92.31%	70.59%	80.00%
NEGATIVE	9	damage	92.55%	73.11%	81.69%
NEGATIVE	10	bed	92.75%	75.21%	83.06%
POSITIVE	11	nice	98.72%	31.56%	47.83%
POSITIVE	12	good	97.74%	53.28%	68.97%
POSITIVE	13	super	97.99%	59.84%	74.30%
POSITIVE	14	good product	96.97%	65.57%	78.24%
POSITIVE	15	good quality	96.59%	69.67%	80.95%
POSITIVE	16	best	96.70%	72.13%	82.63%
POSITIVE	17	awesome	96.81%	74.59%	84.26%
NEUTRAL	18	average	66.67%	10.00%	17.39%

Fig 47. Rules Obtained

5.2.3 Category 3 (Cooler):

MAHARAJA WHITELINE 65 L Desert Air Cooler

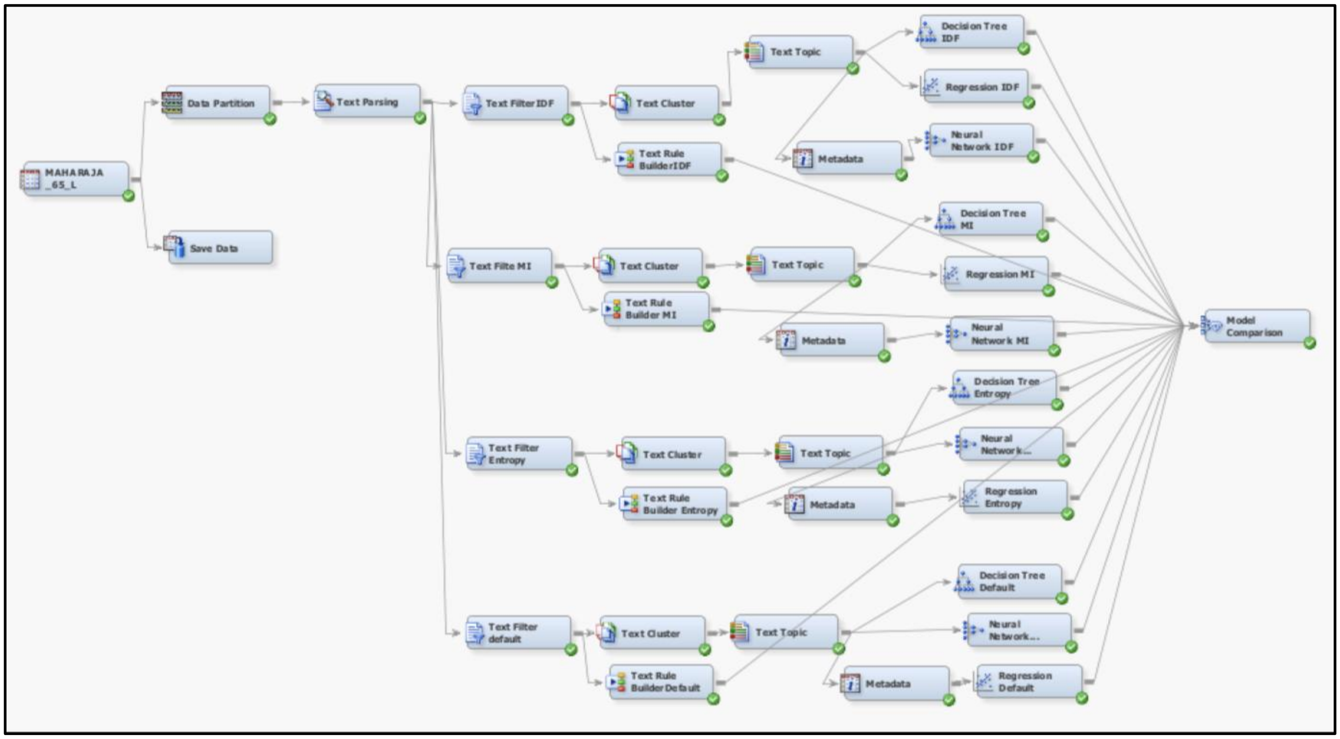


Fig 48. Flow diagram with supervised models

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Roc Index	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Train: Frequency of Classified Cases	Train: Misclassification Rate	Train: Number of Wrong Classifications	Train: Error Function	Train: Average Error Function
Y	Reg5	Reg5	Regression...	Sentiment		0.807	0.061432	1770	0.989273	590	0.247854	108.734	590	0.133898	79	376.3986	0.212655
	Reg6	Reg6	Regression...	Sentiment		0.807	0.061432	1770	0.989273	590	0.247854	108.734	590	0.133898	79	376.3986	0.212655
	Neural2	Neural2	Neural Net...	Sentiment		0.803	0.05148	1770	0.986449	590	0.226891	91.1188	590	0.105085	62	335.199	0.189378
	Reg	Reg	Regression...	Sentiment		0.795	0.065309	1770	0.996467	590	0.255556	115.5963	590	0.132203	78	413.6417	0.233696
	Reg4	Reg4	Regression...	Sentiment		0.784	0.075448	1770	0.987036	590	0.274678	133.5425	590	0.154237	91	477.0478	0.269519
	Neural	Neural	Neural Net...	Sentiment		0.784	0.048303	1770	0.997458	590	0.219779	85.496	590	0.09322	55	315.4184	0.178203
	Neural3	Neural3	Neural Net...	Sentiment		0.781	0.041839	1770	0.990818	590	0.204546	74.05498	590	0.098305	58	262.2002	0.148136
	Neural4	Neural4	Neural Net...	Sentiment		0.781	0.041839	1770	0.990818	590	0.204546	74.05498	590	0.098305	58	262.2002	0.148136
	TextRule2	TextRule2	Text Rule B...	Sentiment		0.778	0.020908	1770	0.82002	590	0.144594	37.00632	590	0.10339	61	353.377	0.199648
	TextRule3	TextRule3	Text Rule B...	Sentiment		0.778	0.020908	1770	0.82002	590	0.144594	37.00632	590	0.10339	61	353.377	0.199648
	TextRule4	TextRule4	Text Rule B...	Sentiment		0.778	0.020908	1770	0.82002	590	0.144594	37.00632	590	0.10339	61	353.377	0.199648
	TextRule5	TextRule5	Text Rule B...	Sentiment		0.778	0.020908	1770	0.82002	590	0.144594	37.00632	590	0.10339	61	353.377	0.199648
	Tree4	Tree4	Decision Tr...	Sentiment		0.69	0.073526	1770	0.96875	590	0.271157	130.1409	590	0.133898	79	469.7332	0.265386
	Tree5	Tree5	Decision Tr...	Sentiment		0.607	0.090875	1770	0.968284	590	0.301455	160.8486	590	0.159322	94	603.4469	0.34093
	Tree6	Tree6	Decision Tr...	Sentiment		0.607	0.090875	1770	0.968284	590	0.301455	160.8486	590	0.159322	94	603.4469	0.34093
	Tree	Tree	Decision Tr...	Sentiment		0.595	0.0929	1770	0.978947	590	0.304795	164.433	590	0.172881	102	600.29	0.339147

Fig 49. Results of the Model Comparison

		Miscalculation Rate			ROC		
Model	**Model Description**	**Train**	**Valid**	**Test**	**Train**	**Valid**	**Test**
Reg5	Regression Entropy	15.48%	9.32%	15.54%	80.70%	91.30%	85.30%
Reg6	Regression Default	16.74%	13.39%	13.56%	80.70%	91.30%	85.30%
Neural2	Neural Network MI	16.74%	13.39%	13.56%	80.30%	94.30%	84.50%
Reg	Regression IDF	17.15%	15.42%	15.82%	79.50%	88.30%	84.80%
Reg4	Regression MI	17.57%	10.51%	16.38%	78.40%	84.90%	79.80%
Neural	Neural Network IDF	18.83%	13.22%	16.95%	78.40%	96.10%	83.40%
Neural3	Neural Network Entropy	19.25%	17.29%	18.08%	78.10%	97.80%	85.40%
Neural4	Neural Network Default	19.25%	15.93%	18.36%	78.10%	97.80%	85.40%
TextRule2	Text Rule BuilderDefault	19.25%	15.93%	18.36%	77.80%	93.10%	84.10%
TextRule3	Text Rule Builder Entropy	19.67%	9.83%	14.41%	77.80%	93.10%	84.10%
TextRule4	Text Rule Builder MI	19.67%	9.83%	14.41%	77.80%	93.10%	84.10%
TextRule5	Text Rule BuilderIDF	20.08%	10.34%	15.54%	77.80%	93.10%	84.10%
Tree4	Decision Tree MI	20.08%	10.34%	15.54%	69.00%	80.00%	72.10%
Tree5	Decision Tree Entropy	20.08%	10.34%	15.54%	60.70%	63.40%	54.70%
Tree6	Decision Tree Default	20.08%	10.34%	15.54%	60.70%	63.40%	54.70%
Tree	Decision Tree IDF	20.08%	13.39%	16.95%	59.50%	66.90%	62.30%

Table 5. Model Comparison Based on Miscalculation Rate and ROC

The table above shows a comparison of various models' performance based on miscalculation rate and ROC (Receiver Operating Characteristic) scores in training, validation, and testing phases. Models with the suffix 'MI' ('Mutual Information'), 'IDF' (Inverse Document Frequency), 'Entropy', and 'Default' indicate different feature selection or processing techniques used. The 'Text Rule Builder' models exhibit higher miscalculation rates in testing than other models, while 'Regression Entropy' and 'Regression Default' lead with the lowest miscalculation rates in the test phase. ROC scores are consistently highest for 'Regression Entropy' and 'Regression Default' in the test phase, indicating a better balance of true positive rate to false positive rate.

5.2.4 Category 4:

Seagate Portable 2 TB External Hard Disk Drive (HDD)(Light Blue)

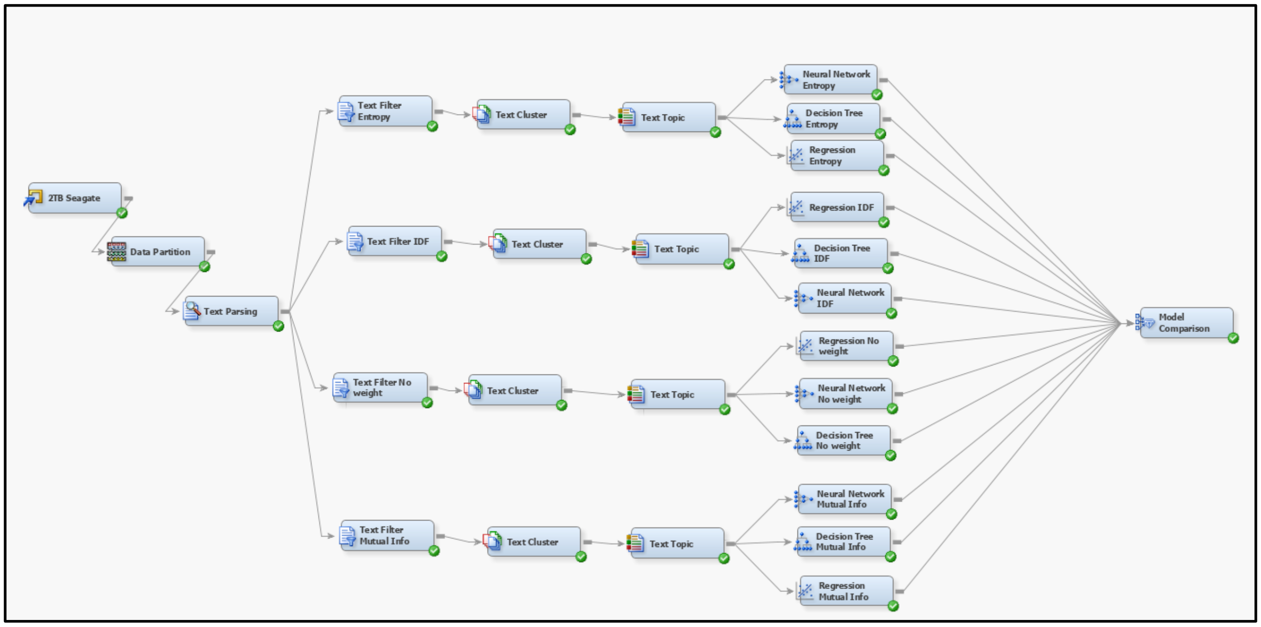


Fig 50. Flow diagram with supervised models

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Test: Roc Index	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error
Y	Reg7	Reg7	Regres...	Sentim...	Sentim...	0.928	225.759	0.0297...	0.1274...	1112	6	1118	1677	213.759	0.0300...	0.9993...	0.0299...	559	6	0.1724...	0.1734...	0.1729...
	Neural8	Neural8	Neural ...	Sentim...	Sentim...	0.926	641.50...	0.0158...	0.06172	849	269	1118	1677	103.50...	0.0259...	0.9972...	0.0209...	559	269	0.1260...	0.1611...	0.1446...
	Neural7	Neural7	Neural ...	Sentim...	Sentim...	0.912	693.65...	0.0145...	0.0713...	831	287	1118	1677	119.65...	0.0245...	0.9999...	0.01951	559	287	0.1204...	0.1565...	0.1396...
	Reg6	Reg6	Regres...	Sentim...	Sentim...	0.911	225.94...	0.0288	0.1275...	1112	6	1118	1677	213.94...	0.0291...	0.9994...	0.0289...	559	6	0.1697...	0.1706...	0.1701...
	Reg5	Reg5	Regres...	Sentim...	Sentim...	0.904	236.38...	0.0316...	0.1361...	1114	4	1118	1677	228.38...	0.0318...	0.99917	0.0317...	559	4	0.1778...	0.17846	0.1781...
	Reg8	Reg8	Regres...	Sentim...	Sentim...	0.904	236.38...	0.0316...	0.1361...	1114	4	1118	1677	228.38...	0.0318...	0.99917	0.0317...	559	4	0.1778...	0.17846	0.1781...
	Neural6	Neural6	Neural ...	Sentim...	Sentim...	0.876	772.13...	0.0213...	0.0823...	801	317	1118	1677	138.13...	0.0382...	0.9490...	0.0298...	559	317	0.1462...	0.1956...	0.1727...
	Neural5	Neural5	Neural ...	Sentim...	Sentim...	0.871	716.11...	0.0110...	0.0418...	795	323	1118	1677	70.112...	0.0200...	0.9562...	0.0155...	559	323	0.1051...	0.1415...	0.1246...
	Tree6	Tree6	Decisio...	Sentim...	Sentim...	0.848	.	0.0263...	.	.	.	1118	1677	.	.	0.9979...	.	559	.	0.1623...	.	.
	Tree5	Tree5	Decisio...	Sentim...	Sentim...	0.805	.	0.0363...	.	.	.	1118	1677	.	.	0.9844...	.	559	.	0.1906...	.	.
	Tree7	Tree7	Decisio...	Sentim...	Sentim...	0.805	.	0.0363...	.	.	.	1118	1677	.	.	0.9844...	.	559	.	0.1906...	.	.
	Tree8	Tree8	Decisio...	Sentim...	Sentim...	0.805	.	0.0363...	.	.	.	1118	1677	.	.	0.9844...	.	559	.	0.1906...	.	.

Fig 51. Results of the Model Comparison

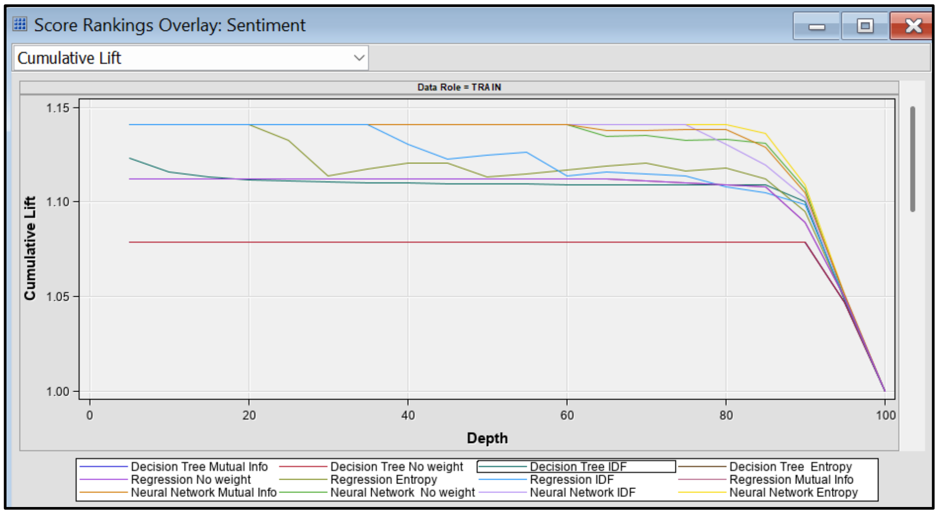


Fig 52. Lift Curve

5.2.5 Category 5:

<u>Bella vita organic Luxury Unisex Perfume Gift Set</u>

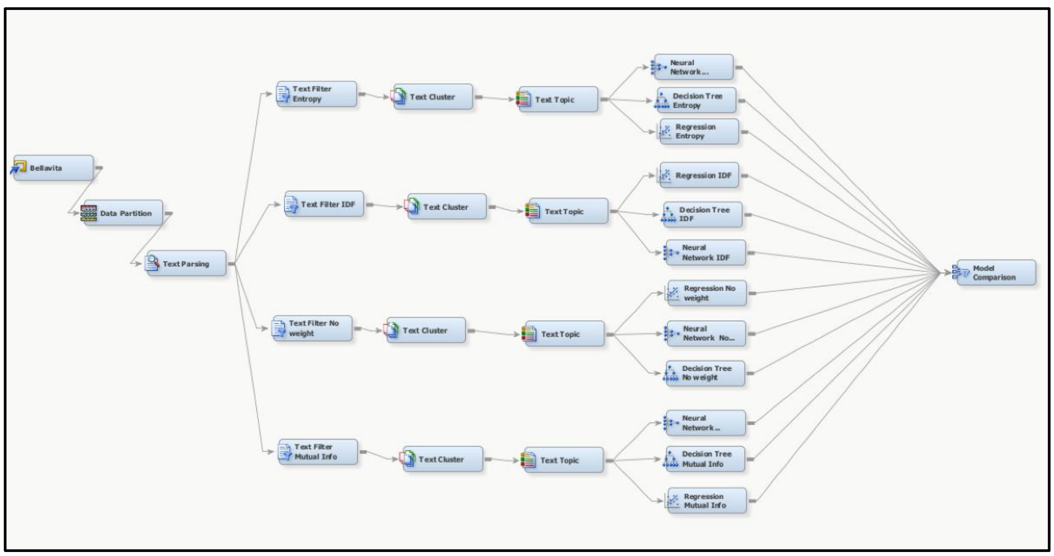


Fig 53. Flow diagram with supervised models

Fit Statistics

Node	Model Description	Selection Criterion: Test: Roc Index	Test: Misclassification Rate	Target Variable	Target Label	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion
2	Neural Network IDF	0.966	0.120567	Sentiment	Sentiment	650.4804	0.020754	0.071413	545	281	826	1239	88.48035	0.042156	0.917103	0.031455	413	281	0.144063	0.205319	0.177356	1975.843
4	Neural Network Mutual Info	0.958	0.141844	Sentiment	Sentiment	548.9724	0.017374	0.066967	593	233	826	1239	82.97241	0.031028	0.998668	0.024201	413	233	0.131811	0.176146	0.155566	1647.939
	Neural Network Entropy	0.953	0.12766	Sentiment	Sentiment	627.7972	0.013216	0.048262	542	284	826	1239	59.79723	0.027067	0.988231	0.020141	413	284	0.114962	0.164519	0.141921	1967.31
3	Neural Network No weight	0.927	0.141844	Sentiment	Sentiment	594.8086	0.019234	0.070063	572	254	826	1239	86.80859	0.036315	0.969541	0.027774	413	254	0.138685	0.190566	0.166657	1792.824
	Regression Mutual Info	0.915	0.099291	Sentiment	Sentiment	202.5867	0.039334	0.157051	822	4	826	1239	194.5867	0.039717	0.998	0.039525	413	4	0.198327	0.19929	0.198809	221.4531
	Regression IDF	0.915	0.099291	Sentiment	Sentiment	202.5867	0.039334	0.157051	822	4	826	1239	194.5867	0.039717	0.998	0.039525	413	4	0.198327	0.19929	0.198809	221.4531
	Regression Entropy	0.915	0.099291	Sentiment	Sentiment	202.5867	0.039334	0.157051	822	4	826	1239	194.5867	0.039717	0.998	0.039525	413	4	0.198327	0.19929	0.198809	221.4531
	Regression No weight	0.915	0.099291	Sentiment	Sentiment	202.5867	0.039334	0.157051	822	4	826	1239	194.5867	0.039717	0.998	0.039525	413	4	0.198327	0.19929	0.198809	221.4531
	Decision Tree No weight	0.905	0.085106	Sentiment	Sentiment	.	0.035802	.	.	.	826	1239	.	.	0.996951	.	413	.	0.189214	.	.	
	Decision Tree Entropy	0.801	0.099291	Sentiment	Sentiment	.	0.043334	.	.	.	826	1239	.	.	0.983471	.	413	.	0.208169	.	.	
	Decision Tree IDF	0.801	0.099291	Sentiment	Sentiment	.	0.043334	.	.	.	826	1239	.	.	0.983471	.	413	.	0.208169	.	.	
	Decision Tree Mutual Info	0.801	0.099291	Sentiment	Sentiment	.	0.043334	.	.	.	826	1239	.	.	0.983471	.	413	.	0.208169	.	.	

Fig 54. Results of the Model Comparison

6. Insights and Recommendations

Category 1: (Home Theatre)

1. **Bluetooth Connectivity Issue:**
   i. Many users report intermittent Bluetooth connectivity, suggesting the wireless feature does not always work as smoothly as expected.
   ii. Complaints often highlight a limited Bluetooth range shorter than advertised, impacting the user's ability to connect devices from a reasonable distance.
   iii. Some reviews point to difficulties in pairing with specific devices (like smartphones or laptops), indicating potential compatibility issues that need addressing.
2. **Delivery Issues:**
   i. Negative delivery experiences, such as longer than promised delivery times, have been mentioned, leading to dissatisfaction even before the product is used.
   ii. Reports of receiving damaged units upon delivery suggest improvements are needed in packaging and handling during transit.
   iii. There are instances where customers received incomplete sets (missing cables, remote controls, etc.), indicating a lapse in the order fulfillment process.
3. **Sound Quality:**
   i. While many users might be satisfied with the sound output, there are notable complaints regarding the clarity and depth of the sound, especially at higher volumes.
   ii. Some reviews express disappointment with the bass performance, feeling that it does not meet their expectations for a home theatre system.
   iii. Feedback on sound distortion at certain levels or imbalance in speaker output suggests the need for better audio engineering and quality control.

Recommendations:

1. Addressing Bluetooth connectivity by enhancing the range and compatibility with a broader array of devices could improve user satisfaction and functionality of the home theatre system.
2. Enhancing the delivery process through better packaging, accurate fulfillment, and timely

shipments could significantly boost initial customer satisfaction.

3. Focusing on improving sound quality, including clarity, bass performance, and eliminating distortion, would likely enhance the overall user experience, meeting or exceeding customer expectations for a home theatre set.

Category 2: (Footwear)

1. **Poor Product Quality:**
   i. Reviews often mention that the materials used in some sandals are not durable, leading to early wear and tear.
   ii. Complaints about the sole coming off or the straps breaking within a few months of purchase indicate a need for better quality control.
   iii. Negative feedback on the finish and aesthetics of the sandals suggests that more attention is needed in the manufacturing process to ensure a premium look and feel.
2. **Heavy Weight:**
   i. A significant number of users find the sandals heavier than expected, which could affect comfort, especially for prolonged use.
   ii. The heavy weight of sandals is frequently attributed to the sole material, indicating a need for lighter alternatives without compromising durability.
   iii. Some reviews suggest that the sandals are not suitable for activities requiring a lot of walking due to their weight, highlighting a potential market segment for lightweight designs.
3. **Fitting Issue:**
   i. Fitting issues, including sandals being too tight, too loose, or not true to size, are common complaints that suggest a discrepancy in sizing standards.
   ii. Reviews indicate problems with the strap adjustments not accommodating different foot widths, leading to discomfort.
   iii. Feedback points to the need for clearer size guidance and possibly the introduction of half sizes or wider/narrower options to better meet customer needs.

Recommendations:

1. The data suggests a critical need for the manufacturer to revisit the design and materials used in their men's black sandals to address these key concerns.
2. Implementing a more rigorous quality control process could significantly improve product durability and customer satisfaction.
3. Considering customer feedback in the product development process could lead to innovations in design, such as lighter materials and more adjustable fittings, which would enhance the overall product appeal and market competitiveness.

Category 3: (Cooler)

1. **Bad Quality:**
   i. Complaints frequently mention the use of inferior materials in the cooler's construction, leading to cracks, leaks, or parts breaking shortly after purchase.
   ii. Negative reviews highlight an overall lack of durability and reliability of the cooler, with some users experiencing issues almost immediately or within a few months of use.
   iii. Users could be dissatisfied with the cooling efficiency, indicating that the product does not perform as advertised under normal conditions.
2. **Motor Issues:**
   i. A significant number of reviews point to motor failures, where the motor burns out or stops working within a short period, suggesting a design flaw/ the use of substandard motor parts.
   ii. There are mentions of the motor overheating, which could also affect the cooler's performance and safety.
   iii. Feedback also includes frustration over the motor's noise level increasing dramatically before it malfunctions or fails, indicating potential longevity and quality issues.
3. **Poor Customer Service:**
   i. Customers often express dissatisfaction with the difficulty in reaching customer service or receiving timely responses to their queries and complaints.
   ii. There are a notable number of reviews stating that the warranty claims or service requests are not handled efficiently or are outright ignored, leading to a trust deficit between

customers and the brand.

iii. Some reviews highlight a lack of available service centers or technicians who can repair the cooler, especially in certain areas, complicating the resolution process.

4. **Bad Customer Experience:**

   i. The cumulative effect of issues such as delivery delays, receiving damaged or defective products, and unhelpful customer service representatives could contribute to a poor overall customer experience.

   ii. Reviews mention incorrect product descriptions or misleading advertising that sets unrealistic expectations, leading to disappointment upon product use.

   iii. There are be instances of customers feeling that their feedback or complaints are not valued, leading to a sense of frustration and helplessness.

5. **Fan Noise:**

   i. Several users report that the cooler's fan is excessively loud, to the point of being disruptive in a home environment, which could indicate a need for design improvements.

   ii. Complaints about the fan noise could also relate to vibration issues, where the cooler is not stable enough, causing additional noise and reducing user satisfaction.

   iii. The noise level could be a significant deterrent for users looking for a quiet cooling solution, impacting the product's appeal to potential buyers.

Recommendations:

1. Addressing these areas of concern requires a multifaceted approach, including enhancing product quality, improving motor design and reliability, overhauling customer service practices, and ensuring accurate product descriptions to avoid mismatched expectations.
2. Focusing on reducing fan noise through better design and stability could enhance the user experience significantly.
3. Implementing rigorous quality control measures and actively engaging with customer feedback could help in identifying and rectifying issues more efficiently, potentially improving brand loyalty and customer satisfaction.

Category 4: (Hard Disk)

1. **Data Corruption:**
   i. Reviews frequently mention instances of data loss or corruption after a short period of use, raising serious concerns about the reliability of the storage device.
   ii. Some customers report that the hard disk fails to retain data, with files becoming corrupted or unreadable, suggesting potential issues with the disk's firmware or manufacturing defects.
   iii. Negative feedback often emphasizes the devastating impact of losing important or irreplaceable data, underscoring the need for improved quality control and reliability testing.
2. **Connectivity Issues:**
   i. Users commonly experience difficulty in connecting the hard disk to their computers, citing inconsistent detection by the operating system, which could point to compatibility or hardware interface problems.
   ii. Complaints about faulty USB cables or ports are notable, indicating that accessories or connectivity components might not meet quality standards.
   iii. The frequency of connectivity complaints suggests that enhancing the plug-and-play experience and ensuring compatibility with a wide range of devices should be a priority.
3. **Transfer Speed:**
   i. Despite advertised high transfer speeds, several reviews describe performance that falls short of expectations, with slow file transfer rates becoming a significant point of frustration.
   ii. Feedback indicates that the actual transfer speed varies greatly from the advertised speed, especially when dealing with large files or volumes of data, hinting at possible optimization issues.
   iii. Some users compare the product unfavorably to competitors in terms of speed, suggesting that improvements in data transfer efficiency could enhance market competitiveness.
4. **Expensive:**
   i. A considerable number of customers feel that the hard disk is overpriced, especially when

considering the issues with data corruption, connectivity, and transfer speeds.

ii. The cost becomes a particular point of contention in reviews when users compare the product to other brands offering similar or better functionality at a lower price point.

iii. Negative perceptions about value for money could be mitigated by addressing the aforementioned technical issues or by offering additional features or services that justify the higher price.

Recommendations:

1. Addressing the root causes of data corruption and ensuring the reliability of the hard disk are crucial steps to regain consumer trust and satisfaction.
2. Improving connectivity options and ensuring that the hard disk is compatible with a broad spectrum of devices can enhance the user experience.
3. Enhancing the efficiency of data transfer rates to meet or exceed advertised speeds could significantly improve customer perceptions of the product.
4. Reevaluating the pricing strategy or providing exceptional features and customer service may be necessary to justify the product's cost and improve its value proposition in a competitive market.

Category 5: (Perfume)

1. Bad Quality:

i. Reviews frequently cite the perfume's packaging as appearing cheap or damaged upon delivery, which raises initial concerns about the overall quality.

ii. Customers express dissatisfaction with the spray mechanism malfunctioning or breaking shortly after use, indicating poor manufacturing standards.

iii. Negative feedback on the alcohol content being too high compared to the fragrance oils, suggesting a formulation that prioritizes cost savings over user experience.

2. Not Long-Lasting:

i. A significant number of users report that the scent fades quickly after application, failing to last through the day, which contradicts claims of longevity.

ii. Complaints highlight a discrepancy between the advertised duration of the fragrance and the actual performance, leading to disappointment.
iii. Feedback suggests that to achieve lasting scent, users have to apply the perfume multiple times throughout the day, which is not economical or convenient.

3. Bad Fragrance:

i. Reviews often describe the fragrance as too strong, synthetic, or not as described, indicating a mismatch between customer expectations and the actual scent profile.
ii. Some customers compare the fragrance unfavorably to similar products in the same price range, questioning the value proposition of the perfume.
iii. Negative reactions to the scent also include physical discomfort, such as headaches or allergic reactions, which could be attributed to the quality of ingredients used.

Recommendations:

1. Addressing quality control issues, both in terms of packaging and the perfume mechanism, could significantly improve the initial customer impression and product satisfaction.
2. Reformulating the perfume to ensure the scent's longevity without requiring frequent reapplication could address one of the primary concerns raised in reviews.
3. Revisiting the fragrance composition to better align with customer expectations and advertised descriptions, potentially through consumer testing or feedback, might help in correcting negative perceptions regarding the scent.