From Amazon Reviews to Business Intelligence:

A Knowledge Discovery Pipeline for Review Helpfulness, Text Mining, and Temporal Analysis

Riccardo Kiefer | Alessio Bufano

Business Intelligence course project

Abstract

Online product reviews combine numerical ratings, temporal and product metadata, community evaluations of review usefulness, and unstructured natural-language content. This project investigated whether these heterogeneous signals could be integrated into a single knowledge-discovery workflow and converted into actionable information for customers, manufacturers, and an e-commerce platform. The dataset contained 4,000 Amazon reviews, equally distributed across four product categories. Following data cleaning and feature engineering, the project analysed the relationship between rating-derived sentiment and helpfulness, tested whether review length carried predictive information, compared supervised classifiers for early usefulness estimation, explored semantic structure through clustering and association-rule mining, analysed reviews belonging to the same product, and modelled review activity over time.

Three classifiers—Decision Tree, k-Nearest Neighbours, and Random Forest—were compared using 20-fold cross-validation in the recovered RapidMiner workflow, with Random Forest retained in the final analysis. Text preprocessing used tokenisation, case normalisation, English stop-word filtering, Snowball stemming, token-length filtering, and TF-IDF vectorisation. K-means was the primary clustering method, supported by Singular Value Decomposition for visual interpretation and compared with K-medoids and DBSCAN. FP-Growth was used to investigate associations between words, ratings, labels, and recurring product-related interests. An ARIMA(1,0,1) model was used to forecast annual review volume.

The project found that positive product evaluation and informational usefulness should not be treated as equivalent concepts: sufficiently evaluated negative reviews could be as useful as positive ones. Review length emerged as an informative feature in the usefulness-prediction workflow. Text mining identified both category-specific language and semantically ambiguous terms spanning multiple categories. These analyses motivated four business applications: early review-usefulness ranking, product-level review summarisation, recommendation hypotheses derived from textual co-occurrence, and review-volume forecasting as a potential proxy for demand. The reconstruction also reveals important limitations: the usefulness target was operationally defined through a dataset-dependent threshold, textual association rules do not prove transactional cross-selling relationships, and review counts cannot be equated with sales without external validation.

Keywords: Amazon reviews; Business Intelligence; Knowledge Discovery; review helpfulness; text mining; Random Forest; K-means; association rules; FP-Growth; ARIMA.

> **Source note.** No external literature has been added to manufacture an academic bibliography. The purpose of this document is to present the original project in publication-style form while remaining faithful to the supplied materials.

Contents

1 Introduction **4**

2 Research Questions and Project Logic **4**

3 Dataset and Preprocessing **5**

3.1 Initial structure and cleaning . . . 5
3.2 Temporal feature engineering . . . 6
3.3 Rating-derived sentiment . . . 6
3.4 Review length . . . 6
3.5 Helpfulness variables and target construction . . . 6
3.6 Subsample construction . . . 7

4 Text Preprocessing **7**

5 Analysis of Rating Sentiment and Review Helpfulness **7**
5.1 Motivation . . . 8
5.2 Observed pattern . . . 8

6 Review Length as an Informative Feature **8**

7 Predicting Review Usefulness **9**
7.1 Operational problem . . . 9
7.2 Compared algorithms . . . 9
7.3 Application to unclassified reviews . . . 9
7.4 Business interpretation . . . 10

8 Multiple Reviews for the Same Product **10**
8.1 Product-level sparsity . . . 10
8.2 Contrasting sentiment within products . . . 10

9 Product-Level Review Summarisation as a Challenge **11**

10 Textual Structure and Clustering **11**
10.1 Initial lexical evidence . . . 11
10.2 K-means configuration . . . 11
10.3 Interpretation of clusters . . . 12
10.4 SVD visualisation . . . 12
10.5 Alternative clustering methods . . . 12

11 Association-Rule Mining **12**
11.1 Words and rating levels . . . 12
11.2 Associations among review terms . . . 12
11.3 Association rules as interpretive support for clustering . . . 13

12 Temporal Analysis and Forecasting **13**
12.1 From dates to annual review counts . . . 13
12.2 ARIMA configuration . . . 13
12.3 Correct interpretation of the forecast . . . 13

13 Integrated Business Applications **14**

14 Critical Discussion **14**
14.1 What the project does well . . . 14
14.2 Where the original interpretation overreaches . . . 15

15 Limitations **15**

16 How the Project Could Be Improved Today **16**
16.1 Model helpfulness without an arbitrary binary threshold . . . 16
16.2 Use a modern evaluation protocol . . . 16
16.3 Modernise text representations . . . 16
16.4 Evaluate clustering independently . . . 16
16.5 Validate recommendation hypotheses with transactional data . . . 17
16.6 Validate review volume as a demand proxy . . . 17

17 Conclusion **17**

A Reconstruction of the Project's Original Development **18**
A.1 Stage 1: text as the original analytical opportunity . . . 18
A.2 Stage 2: conceptual structure before implementation . . . 18
A.3 Stage 3: observing other presentations changes the emphasis . . . 18
A.4 Stage 4: from business applications to analytical motivation . . . 18
A.5 Stage 5: internal correction of the sales-forecast interpretation . . . 18

B Division of Work and Contributions **19**
B.1 Riccardo Kiefer . . . 19
B.2 Alessio Bufano . . . 19

C How to Present the Project Today **19**

D Source Materials Used for This Reconstruction **20**

1 Introduction

Online reviews are unusually rich data objects because they combine explicit evaluation and free-form explanation. A single review may contain a numerical rating, a written description of the user experience, a product identifier, a publication date, and community feedback indicating whether other users found the review useful. These components support different analytical questions and different forms of business value.

The original project was designed around a simple but ambitious idea: user-generated reviews should not be treated merely as text to be described, but as a potential *knowledge layer* connecting customer experience, product requirements, review quality, recommendation opportunities, and temporal market activity. The final presentation explicitly described a funnel consisting of dataset characterisation, attribute selection, preprocessing, data transformation, knowledge extraction, and the conversion of extracted knowledge into business applications.

The analysed dataset contained 4,000 Amazon reviews, with 1,000 observations for each of four product categories:

- Sport & Outdoor;
- Film & TV;
- Cell Phones & Accessories;
- Health & Personal Care.

The project was therefore broader than a single classification or clustering exercise. It combined structured-data analysis and natural-language processing in order to answer several related questions: What makes a review useful? Can usefulness be estimated before the community has voted on a review? Can the language of reviews reveal product categories or customer interests? Can reviews of the same product be summarised? Can temporal review activity support forecasting?

The development conversation is important for understanding this architecture. Early messages focused on extracting information from review text and identifying keywords associated with different levels of satisfaction and quality. Later, after observing other project presentations, the team explicitly reframed the work around business relevance: each analytical technique had to be motivated by a question and connected to a practical application rather than presented as an isolated algorithmic exercise.

2 Research Questions and Project Logic

The reconstructed project can be organised around five research questions.

Table 1: Research questions and corresponding analytical components.

	Research question	Main analytical component
RQ1	Are product rating and review helpfulness measuring the same underlying phenomenon?	Descriptive and comparative analysis of Overall, OverallSentiment, and helpfulness.
RQ2	Can observable review characteristics support an early prediction of usefulness?	Review-length analysis and supervised classification.
RQ3	Does review language contain enough structure to distinguish categories and reveal user interests?	TF-IDF, K-means, SVD, alternative clustering, FP-Growth.
RQ4	Can multiple reviews of the same product be converted into a compact product-level information layer?	Same-product filtering, sentiment comparison, keyword-oriented summarisation concept.
RQ5	Can review activity be modelled over time and used as a signal related to market demand?	Annual aggregation and ARIMA forecasting.

These research questions support a single conceptual chain:

$$\boxed{\text{Business question} \rightarrow \text{Data transformation} \rightarrow \text{Method} \rightarrow \text{Evidence} \rightarrow \text{Application}}$$

This chain is the central idea of the project. The algorithms are instruments, not the organising principle.

3 Dataset and Preprocessing

3.1 Initial structure and cleaning

The starting dataset contained 4,000 reviews and ten initial attributes per record. The project selected a subset of variables considered useful for the intended analyses and removed eight records containing missing values in the selected fields. This yields a cleaned working population of 3,992 observations, although the final presentation continued to refer to the original dataset size when introducing the project.

The dataset exhibited three structural imbalances that influenced the analysis:

1. ratings were strongly skewed toward positive values;
2. many reviews had never received a helpfulness evaluation;
3. the distribution of reviews across individual products was sparse, with relatively few products receiving multiple reviews.

3.2 Temporal feature engineering

The original `reviewTime` field was decomposed into three new attributes:

$$review_year, \qquad review_month, \qquad review_day.$$

This transformation supported both descriptive analysis and time-series modelling.

3.3 Rating-derived sentiment

The numerical rating `Overall` was discretised into a categorical variable, `OverallSentiment`:

$$OverallSentiment = \begin{cases} positive, & Overall \geq 4, \\ neutral, & Overall = 3, \\ negative, & Overall < 3. \end{cases}$$

This variable is best described as *rating-derived sentiment*. It was not generated by a natural-language sentiment classifier.

3.4 Review length

A numerical feature was created to track the length of each review text. In the RapidMiner process this feature appears under the misspelled name `lenght_reviewText`. In this report it is referred to simply as *review length*.

The feature was introduced because length is available immediately when a review is published and could therefore serve as an early signal of informational richness.

3.5 Helpfulness variables and target construction

The Amazon `Helpful` attribute contained two counts. The recovered preprocessing process explicitly renamed them as:

$$U_i = \texttt{useful_reviewEvaluations}, \qquad T_i = \texttt{total_reviewEvaluations}.$$

The helpfulness percentage was then calculated as:

$$P_i = \frac{U_i}{T_i}.$$

This point matters because the oral presentation contained an ambiguous verbal description of the two values. The process file resolves the ambiguity: the implemented ratio was useful evaluations divided by total evaluations.

A categorical target, `Usefulness`, was subsequently created using the threshold 0.682:

$$Usefulness_i = \begin{cases} useful, & P_i \geq 0.682, \\ not\ useful, & P_i < 0.682, \\ not\ classified, & T_i = 0. \end{cases}$$

The threshold corresponded to the mean helpfulness percentage adopted in the original workflow. It should therefore be understood as an *operational definition* created within the project, not as an externally validated threshold for review quality.

3.6 Subsample construction

Because many reviews lacked helpfulness evaluations, two restricted datasets were generated:

$$D_1 = \{i : T_i \geq 1\},$$

containing 1,774 evaluated reviews, and

$$D_4 = \{i : T_i \geq 4\},$$

containing 644 reviews with at least four evaluations.

The threshold of four was motivated by the average number of evaluations among evaluated reviews, reported in the project as slightly above three. The purpose was to compare patterns derived from minimal evidence with patterns supported by a larger number of community evaluations.

4 Text Preprocessing

The recovered RapidMiner workflow shows that both review summaries and full review texts were processed through a standard text-mining pipeline:

Tokenisation → Lower case → English stop-word removal → Snowball stemming → Token-length filter → TF-IDF.

The final process filtered tokens shorter than two characters or longer than 25 characters. TF-IDF vectors were retained for clustering and related text analyses.

This distinction between structured feature engineering and text vectorisation is central to the project. The numerical branch addressed rating, time, helpfulness, and review length, whereas the text branch addressed category-specific vocabulary, semantic overlap, and recurring user interests.

5 Analysis of Rating Sentiment and Review Helpfulness

5.1 Motivation

The first analytical question asked whether a positive review was necessarily an informative review. This distinction is important because platforms often display rating and helpfulness in close proximity, yet the two variables answer different questions:

- **Rating**: How satisfied was the reviewer with the product?
- **Helpfulness**: Did other users consider the review informative or useful?

5.2 Observed pattern

On the complete dataset, average helpfulness appeared closely related to positive ratings: higher-rated reviews showed higher average useful-vote percentages. However, the pattern changed when the analysis was restricted first to evaluated reviews and then to reviews with at least four evaluations.

In the more strongly evaluated subset, both positive and negative reviews tended to display high average helpfulness. The project interpreted this as evidence that the number of evaluations improves the information available about review quality and reduces the misleading association between positivity and usefulness.

The central conceptual result can therefore be stated as:

$$\boxed{\text{Product satisfaction} \neq \text{Review informational utility}}$$

A negative review may be highly useful precisely because it provides detailed evidence about a defect, limitation, or mismatch between expectations and actual performance.

6 Review Length as an Informative Feature

The project then examined whether review length could serve as a simple metric for distinguishing useful from less useful reviews.

On the complete dataset, the relationship between review length and helpfulness percentage was described as relatively weak. In the subset containing reviews with at least four evaluations, the relationship became more pronounced. The original interpretation was that longer reviews contain more detail and are therefore more useful.

A stricter interpretation is preferable. The analysis supports the claim that *review length contained predictive information with respect to the project's helpfulness target*. It does not establish that long reviews are intrinsically better. Length may capture detail, but it may also capture verbosity, repetition, product complexity, or reviewer style.

The project also reported category differences, with the relationship appearing stronger in Health & Personal Care and Cell Phones & Accessories. The proposed explanation was that buyers may seek more detailed information for products with greater personal, health, or economic relevance. This interpretation

is plausible but was not independently tested and should therefore be regarded as a hypothesis generated by the observed pattern.

7 Predicting Review Usefulness

7.1 Operational problem

A newly published review has no community evaluation history. The platform therefore faces a cold-start problem: how should a recent review be ranked before enough users have voted on its helpfulness?

The project addressed this problem by training a classifier on evaluated reviews and applying it to reviews whose `Usefulness` value was `not classified`.

7.2 Compared algorithms

Three supervised models were compared:

- Decision Tree;
- k-Nearest Neighbours (k-NN);
- Random Forest.

The recovered process files show a 20-fold cross-validation procedure. The classification performance operator used accuracy as the main criterion and also calculated weighted mean recall and weighted mean precision.

For k-NN, the final configuration used:

$$k = 20,$$

with weighted voting and a mixed Euclidean distance measure.

The Decision Tree used gain ratio, maximal depth 20, minimal gain 0.01, and pre-pruning/pruning settings.

The final Random Forest configuration used:

$$100 \text{ trees}, \qquad maximal\ depth = 20,$$

with gain ratio and confidence voting.

The final presentation reports Random Forest as the best-performing model among the tested alternatives. The preserved materials do not provide a complete table of the numerical validation results, so this report does not reconstruct values that cannot be verified.

7.3 Application to unclassified reviews

The selected model was applied to the reviews that lacked community-based classification. The presentation reported the following output:

$$1,864 \text{ reviews predicted as useful},$$

354 reviews predicted as not useful.

The total, 2,218, is consistent with the cleaned dataset once the 1,774 already evaluated reviews are removed.

Review length was described in the final presentation as the most selective feature in the Random Forest model.

7.4 Business interpretation

The proposed application was an early review-ranking system. The expected value was different for each stakeholder:

- **Customers**: place potentially informative reviews in a more visible position without waiting for a large number of votes;
- **Manufacturers**: prioritise useful reviews for requirements analysis and targeted text mining;
- **Amazon**: improve navigation and reduce information overload.

The original presentation also referred to discarding non-useful reviews. A more defensible production design would use predicted usefulness for *ranking and prioritisation*, not permanent suppression, because false negatives could hide legitimate minority opinions.

8 Multiple Reviews for the Same Product

8.1 Product-level sparsity

The dataset was strongly imbalanced with respect to the product identifier `asin`. Only 275 products had more than one review. Among these, only four products had five reviews and ten had four reviews.

The project sorted observations by product and date, created a flag identifying repeated product codes, and extracted 617 records associated with multiply reviewed products. The presentation correctly noted that the same result could have been obtained more directly with an SQL `GROUP BY` and `HAVING COUNT > 1` query.

8.2 Contrasting sentiment within products

A further flag identified cases in which reviews associated with the same product had different values of `OverallSentiment`. The reported result was that 95 of the 275 products with multiple reviews exhibited contrasting sentiment.

This result is especially interesting in a dataset dominated by positive ratings. It suggests that product-level averages or aggregate ratings may conceal heterogeneous user experiences.

At the same time, the result must be interpreted cautiously because most products in the subset were represented by only a very small number of reviews.

9 Product-Level Review Summarisation as a Challenge

The same-product analysis motivated one of the project's most forward-looking applications: a product-level review summary.

The proposed system would identify the top N keywords across all reviews of a product and provide links to the reviews containing those terms. Conceptually:

$$\text{Product reviews} \rightarrow \text{Top topics / keywords} \rightarrow \text{Topic-oriented navigation.}$$

The intended advantages were:

- a faster purchasing process for customers;
- easier identification of customer requirements and recurring complaints for manufacturers;
- improved navigation, profiling, and competitive positioning for the platform.

The final presentation explicitly described this application as a *challenge*. The reason was empirical rather than conceptual: the dataset did not contain enough reviews per product to test the approach adequately. This is an important methodological strength of the project narrative because it recognises the difference between a plausible application and a validated one.

10 Textual Structure and Clustering

10.1 Initial lexical evidence

The most frequent words included terms with positive connotations such as *great*, *good*, and *love*, consistent with the rating imbalance toward positive reviews.

The analysis then moved beyond raw word frequency to a more difficult question: could review vocabulary reconstruct the four product categories or reveal category-specific features?

10.2 K-means configuration

The project tested multiple initialisation settings and reported better results for Max Runs equal to 80 or 100. Max Runs = 80 was retained because it reduced computation time.

The elbow method indicated a promising number of clusters close to four. The final presentation focused on $K = 3$, $K = 4$, and $K = 5$, ultimately retaining $K = 4$.

This decision had two justifications: the elbow analysis suggested a similar region, and four clusters preserved the original four-category structure for interpretation. The second justification creates a methodological risk: unsupervised clustering should ideally be selected independently of the known labels and evaluated against them only after the clustering solution has been fixed.

10.3 Interpretation of clusters

The resulting clusters showed partial correspondence with the original categories. One cluster contained observations from more than one category. Inspection of its vocabulary revealed terms such as *blade* and *sharp*, which can describe both technical knives and personal-care razors.

This is not merely a failure of classification by clustering. It is also substantive evidence that consumer vocabulary may organise products according to functional similarity rather than catalogue taxonomy.

10.4 SVD visualisation

Singular Value Decomposition was applied to a sample of 1,000 observations to obtain a lower-dimensional visual representation. The final presentation noted that the cluster separation remained difficult to interpret in the reduced space, illustrating the complexity and sparsity of the textual representation.

10.5 Alternative clustering methods

K-medoids and DBSCAN were explored as alternatives. K-medoids produced acceptable computation times only on relatively small samples, creating concerns about representativeness. DBSCAN classified a large proportion of observations as noise even with comparatively permissive parameter settings.

K-means was therefore retained as the most practical method for the available dataset and tooling.

11 Association-Rule Mining

11.1 Words and rating levels

FP-Growth was first used to investigate associations between summary terms and `Overall`. With a minimum frequency of 50, words such as *amazing*, *love*, and *excellent* were associated with the highest rating level, whereas *ok* was associated with an Overall value of 3. Lowering the minimum-frequency threshold allowed rarer terms associated with lower ratings to appear.

This analysis offered a rule-based complement to the rating-derived sentiment variable by showing which linguistic elements co-occurred with specific rating levels.

11.2 Associations among review terms

A second FP-Growth analysis was performed on full review text to identify terms that tended to occur together within each product category. In Cell Phones & Accessories, for example, the project interpreted the rules as evidence that users were interested in device protection and sound quality.

These patterns motivated product suggestions such as covers, screen protectors, and headphones.

The business idea is reasonable, but the inferential boundary is important:

$$\text{textual co-occurrence} \neq \text{purchase co-occurrence}.$$

The rules show that concepts appear together in reviews. They do not demonstrate that users who buy one

product also buy another. Transaction-level basket data would be needed to validate a true cross-selling model.

11.3 Association rules as interpretive support for clustering

The team also used word-to-label associations to investigate categories that were ambiguous in the initial clustering. This helped identify more specific vocabulary for the labels that were mixed in the K-means solution.

The two techniques therefore played complementary roles:

$$\text{Clustering} \rightarrow \text{discover broad latent structure},$$

$$\text{Association rules} \rightarrow \text{interpret specific recurrent relationships}.$$

12 Temporal Analysis and Forecasting

12.1 From dates to annual review counts

The final analytical branch investigated review activity over time. The monthly distribution was described as comparatively homogeneous, while the annual distribution displayed much stronger variation.

The recovered RapidMiner process aggregated `reviewerID` by year and modelled the resulting annual count series.

12.2 ARIMA configuration

The forecasting workflow used:

$$ARIMA(p, d, q) = ARIMA(1, 0, 1),$$

with the annual count of reviewer IDs as the time-series attribute and year as the index.

The forecast-validation process used a window size of 15 observations and a one-period forecast horizon. The final application generated a two-period forecast.

The original presentation described the resulting pattern as surprisingly accurate and comparable with the expansion of e-commerce.

12.3 Correct interpretation of the forecast

The original oral narrative called this a forecast of sales. The project chat itself shows that this interpretation was questioned during development: the observed variable was correctly recognised as the number of reviews, not sales.

The scientifically defensible statement is therefore:

$$\boxed{\text{The model forecasts review activity.}}$$

A second, explicitly conditional claim can then be made:

> Review activity may act as a proxy for demand if its relationship with sales is validated externally.

Without transaction data, the project cannot establish a direct sales forecast.

13 Integrated Business Applications

The project concluded by mapping the analytical results onto three stakeholders: customers, manufacturers, and Amazon itself.

Table 2: Four proposed business applications and their analytical basis.

Application	**Analytical basis**	**Intended value**
Early review-usefulness ranking	Supervised classification of Usefulness	Prioritise informative reviews before sufficient community votes accumulate.
Product-level review summary	Same-product analysis, keywords, association patterns	Reduce information overload and expose recurring product themes.
Recommendation hypotheses	Textual association rules	Identify complementary needs or product-related interests to be validated with transaction data.
Review-activity forecasting	Annual aggregation and ARIMA(1,0,1)	Monitor future review volume and test its usefulness as a demand proxy.

The unifying theme is not any one algorithm. It is the transformation of reviews into a decision-support resource. The review is treated simultaneously as a product evaluation, an information object, a natural-language document, a signal of customer priorities, and a temporally located market event.

14 Critical Discussion

14.1 What the project does well

The strongest feature of the project is its attempt to connect analysis to a decision problem. During development, the team explicitly reacted against presenting “random” analytical exercises. The final structure instead asks why each analysis is being performed and only then links it to a possible business application.

The usefulness analysis is the clearest example. A descriptive observation—positivity and helpfulness are not the same—leads to a feature investigation, then to supervised modelling, then to an operational ranking concept. This is a coherent knowledge-discovery chain rather than a collection of disconnected algorithms.

The project also demonstrates useful intellectual flexibility. When product-level data were too sparse, the review-summary concept was labelled a challenge rather than presented as fully validated. When

DBSCAN and K-medoids were less suitable, the project retained K-means for practical reasons instead of assuming that more complex algorithms must be superior.

14.2 Where the original interpretation overreaches

Some claims require stricter wording.

First, the helpfulness threshold is endogenous to the dataset. The classifier therefore learns a project-specific label, not an objective universal definition of useful review.

Second, review length is associated with the helpfulness target but is not a causal measure of review quality.

Third, textual association rules reveal semantic co-occurrence rather than actual co-purchasing behaviour.

Fourth, the ARIMA model forecasts review counts rather than sales.

These are not minor semantic corrections. They distinguish what is directly measured from what is inferred for business purposes.

15 Limitations

The project has several limitations that should be stated explicitly in an academic presentation.

1. **Dataset size and representativeness.** The dataset contains 4,000 initial reviews and only four product categories. It is balanced by category but strongly imbalanced in ratings, helpfulness availability, and reviews per product.

2. **Operational target definition.** `Usefulness` is generated by thresholding a helpful-vote percentage at 0.682. The model predicts this constructed target rather than an externally validated concept of review quality.

3. **Vote-count uncertainty.** A helpfulness percentage based on one vote is much less reliable than the same percentage based on many votes. The four-evaluation subset partly addresses this issue but reduces sample size.

4. **Incomplete preserved performance reporting.** The process files confirm 20-fold cross-validation and performance operators, but the supplied final materials do not preserve a complete table of model metrics. Quantitative superiority among classifiers therefore cannot be independently reconstructed here.

5. **Clustering confirmation risk.** Choosing $K = 4$ partly because four known categories exist risks allowing labels to influence an unsupervised procedure.

6. **Sparse product-level data.** Very few products have enough reviews to validate a robust product-level summarisation system.

7. **Recommendation evidence.** Textual co-occurrence does not demonstrate co-purchasing and should be interpreted as hypothesis generation.

8. **Forecast target.** Review volume is not equivalent to sales volume. A direct commercial forecast would require transaction data.

16 How the Project Could Be Improved Today

A contemporary redesign could preserve the original logic while strengthening its statistical and machine-learning foundations.

16.1 Model helpfulness without an arbitrary binary threshold

Instead of reducing the helpful-vote ratio to a binary class at the global mean, the model could predict either the probability that a future vote will be positive or a posterior distribution over the latent helpfulness rate. A binomial or beta-binomial formulation would naturally account for both the number of useful votes and the total number of votes.

This would distinguish, for example, 1/1 useful votes from 100/100 useful votes even though both produce a raw ratio of 1.0.

16.2 Use a modern evaluation protocol

The supervised task should report at minimum:

- accuracy;
- precision;
- recall;
- F1 score;
- ROC-AUC or PR-AUC where appropriate;
- confusion matrix;
- confidence intervals or repeated validation results.

A held-out test set would provide an additional safeguard against tuning decisions being influenced by cross-validation performance.

16.3 Modernise text representations

TF-IDF remains interpretable and useful, but a modern replication could compare it with sentence embeddings or transformer-based representations. This would allow semantic similarity to be captured even when reviews use different words to describe the same concept.

16.4 Evaluate clustering independently

The number of clusters should be selected using internal criteria such as silhouette score, stability analysis, or gap statistics. Known product labels should then be used only for external evaluation, for example with

adjusted mutual information or adjusted Rand index.

16.5 Validate recommendation hypotheses with transactional data

Text mining can identify what customers discuss together. To establish actual complementarity, the analysis should be joined with baskets, sequential purchases, or clickstream data.

16.6 Validate review volume as a demand proxy

A direct comparison between sales and review counts would allow estimation of whether review activity is a leading, coincident, or lagging indicator. Without such validation, forecasting should remain explicitly limited to review activity.

17 Conclusion

This project investigated how structured and unstructured components of Amazon reviews could be transformed into business-relevant knowledge through an integrated data-mining pipeline.

The workflow combined data cleaning, feature engineering, helpfulness analysis, supervised classification, text clustering, association-rule mining, same-product analysis, and time-series forecasting. Its most important conceptual result was the separation between product satisfaction and review informational utility. A negative review can be useful, and usefulness can potentially be estimated before community votes accumulate.

The text-mining branch showed that product categories possess partially distinctive vocabularies but also share functional language that crosses catalogue boundaries. Association-rule mining generated interpretable hypotheses about user interests, while the product-level analysis motivated a review-summary concept. The temporal branch demonstrated that review activity can be modelled, while also revealing the importance of not conflating review volume with sales.

The broader methodological lesson is that Business Intelligence does not become valuable merely by applying multiple algorithms. Value emerges when the analysis maintains a defensible connection between question, data, method, evidence, and decision:

$$\boxed{\text{Question} \rightarrow \text{Data} \rightarrow \text{Method} \rightarrow \text{Evidence} \rightarrow \text{Decision}}$$

The original work is therefore best understood not as a collection of disconnected exercises, but as an early end-to-end knowledge-discovery project whose strongest contribution is the attempt to convert review data into a coherent decision-support framework.

A Reconstruction of the Project's Original Development

The final presentation makes the project appear as a pre-planned analytical pipeline, but the project chat reveals a more informative development process.

A.1 Stage 1: text as the original analytical opportunity

One of the earliest explicit ideas was to extract information from review text and search for keywords associated with different levels of satisfaction and quality. This establishes text mining as one of the project's original conceptual cores rather than a technique added only at the end.

A.2 Stage 2: conceptual structure before implementation

The project was deliberately structured conceptually before the algorithms were fully implemented. The stated intention was to define the project logic first and then implement the necessary methods. This explains why the final work contains several distinct techniques without being reducible to any single one.

A.3 Stage 3: observing other presentations changes the emphasis

A decisive change occurred after observing other course presentations. Particular attention was paid to whether teams clearly explained their business objectives and whether they could connect those objectives to clustering, classification, or association rules. A question reportedly asked during another presentation—how the proposed applications would be presented to Amazon's CEO—made the business translation requirement explicit.

The team subsequently emphasised the need to "clarify the business idea" and to avoid presenting analyses without explaining why they had been performed.

A.4 Stage 4: from business applications to analytical motivation

The most mature framing appears later in the conversation: the objective slide should not merely list business applications, but explain the reason that motivated each analysis and then identify the business application it could support.

This distinction is subtle but important. It yields the final architecture:

$$\text{Analytical question} \rightarrow \text{Method} \rightarrow \text{Result} \rightarrow \text{Business use}.$$

A.5 Stage 5: internal correction of the sales-forecast interpretation

The chat also contains an early methodological correction. When the temporal model was described as "sales forecast", the response was that it was actually a forecast of reviews. The final presentation nevertheless returned to stronger sales language. The reconstructed report restores the more accurate interpretation already recognised during development.

B Division of Work and Contributions

The following division reflects the final presentation script and the project-development conversation. It should not be interpreted as a strict separation of ownership: parameter choices, interpretation, slide construction, and troubleshooting were discussed collaboratively.

B.1 Riccardo Kiefer

The material supports a primary contribution in the following areas:

- initial framing of the dataset-analysis funnel;
- preprocessing and feature-engineering presentation;
- text mining and the conceptual focus on extracting information from reviews;
- K-means analysis and parameter exploration;
- SVD visualisation and interpretation of lexical ambiguity;
- comparison with alternative clustering approaches;
- association-rule analysis of textual data;
- temporal analysis and forecasting;
- explicit structuring of the presentation around the reason for each analysis before its business application.

B.2 Alessio Bufano

The material supports a primary contribution in the following areas:

- analysis of helpfulness, sentiment, and number of evaluations;
- review-length analysis;
- usefulness-prediction workflow and comparison of classifiers;
- same-product analysis and contrasting sentiment;
- development of the product-level review-summary challenge;
- final synthesis of stakeholder needs and business applications.

C How to Present the Project Today

A strong contemporary oral summary would be:

> "In a Business Intelligence project, we analysed approximately 4,000 Amazon reviews across four product categories. The work was designed as an end-to-end knowledge-discovery pipeline rather than a

> single modelling task: we performed data cleaning and feature engineering, analysed review helpfulness, compared supervised classifiers, applied TF-IDF text mining, K-means clustering, association rules, and time-series forecasting.
>
> One central question was whether we could estimate how useful a review would be before it accumulated community votes. We operationalised helpfulness from Amazon useful-vote data, compared Decision Tree, k-NN and Random Forest with 20-fold cross-validation, and retained Random Forest in the final workflow. Review length emerged as an informative feature.
>
> On the text side, we investigated whether vocabulary could recover product categories and reveal recurring customer interests. The project was always tied back to applications: review ranking, customer-requirement analysis, product-level summarisation, recommendation hypotheses, and temporal monitoring.
>
> If I presented the work today, I would make some interpretations more conservative. The temporal model forecast review counts rather than sales, textual association rules generated cross-selling hypotheses rather than transactional evidence, and the usefulness label was an operational project definition rather than a universal ground truth. That reassessment is part of what I learned from the project."

This version communicates technical breadth without overstating the evidence. It also demonstrates the ability to reassess an earlier project critically—an important signal of analytical maturity.