École Nationale des Sciences Appliquées de Kénitra

Exploratory Data Analysis

& Modeling Decisions

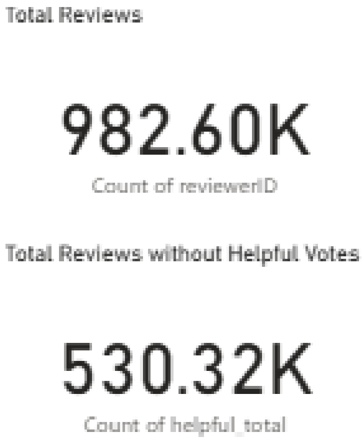
Noha LAKHDIMI
Jihane FETTOUKH

Année universitaire : 2025–2026

Table des matières

1 Univariate Analysis

1.1 Helpfulness Information



Total Reviews

982.60K

Count of reviewerID

Total Reviews without Helpful Votes

530.32K

Count of helpful_total

Insight

Helpfulness information is heterogeneous across the dataset : some reviews receive helpful votes, some are explicitly marked as unhelpful, while a large proportion have no helpfulness feedback at all. This means that a zero helpfulness value may represent either lack of user interaction or negative feedback, and treating these cases identically would introduce noise. Moreover, helpfulness reflects review quality and user agreement, not sentiment polarity itself.

Decision : Using Helpfulness as Loss Weights in ANN Training

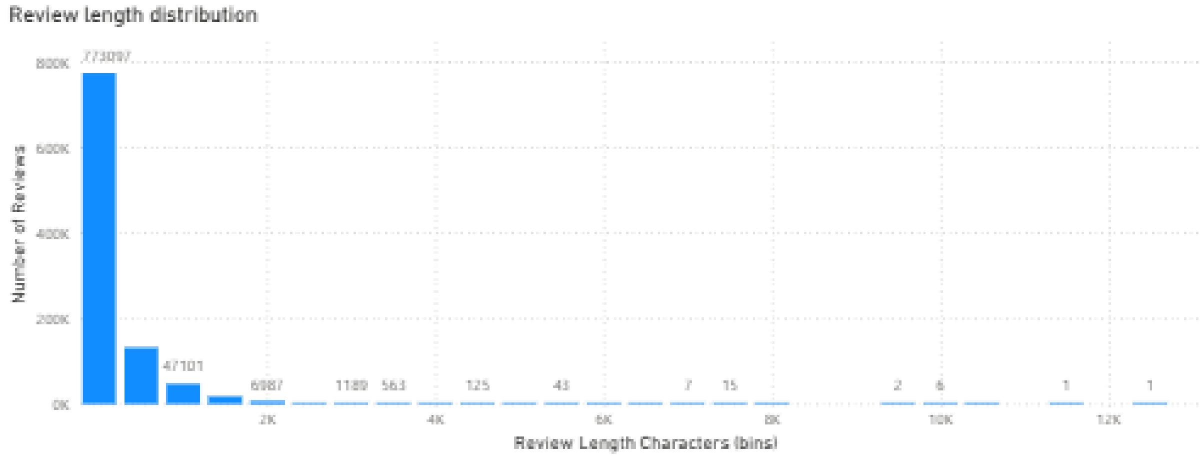
Helpfulness information is incorporated by weighting the loss function during ANN training, rather than being used as an input feature or influencing the predicted sentiment directly. Reviews that have received helpful votes are assigned a higher loss weight, while reviews without helpfulness feedback are assigned a neutral weight. Specifically, reviews with no helpful votes receive a weight of 1.0, whereas reviews with at least one vote are weighted proportionally to their helpfulness ratio (e.g., weight = 1.0 + helpful_ratio). This ensures that misclassifications of more helpful reviews contribute more strongly to the optimization objective.

Why this decision is strongly justified

- Maintains semantic purity of sentiment learning : Helpfulness measures perceived review quality, not sentiment polarity ; applying it to the loss avoids distorting the sentiment representation learned by the ANN.
- Prioritizes high-quality supervision : Reviews validated as helpful by users are likely clearer and more informative, making their errors more costly during training.
- Handles missing helpfulness explicitly : Reviews without votes are treated as neutral rather than unhelpful, preventing incorrect assumptions and noise.

- — Avoids feature leakage : Sentiment prediction relies solely on textual content, while helpfulness only affects training emphasis.
- — Optimizes ANN training effectively : Loss weighting integrates naturally into gradient-based optimization and is a standard, well-accepted practice.

1.2 Review Length Distribution



Insight

More than 770,000 reviews have a length below 400 tokens, indicating that the vast majority of the dataset consists of short to medium-length reviews. Long reviews are relatively rare and represent only a small fraction of the data. This suggests that sentiment in this dataset is generally expressed concisely, and most reviews contain sufficient information within a limited number of tokens.

Decision (For ANN Sentiment Analysis)

The ANN input sequence length is capped at 400 tokens, with shorter reviews padded and longer reviews truncated.

- — A maximum sequence length of 400 tokens covers the vast majority of reviews, ensuring minimal information loss.
- — Truncation affects only a small subset of long reviews, reducing computational cost without significantly harming sentiment representation.
- — Padding shorter reviews enables consistent input dimensions required for ANN batching and optimization.

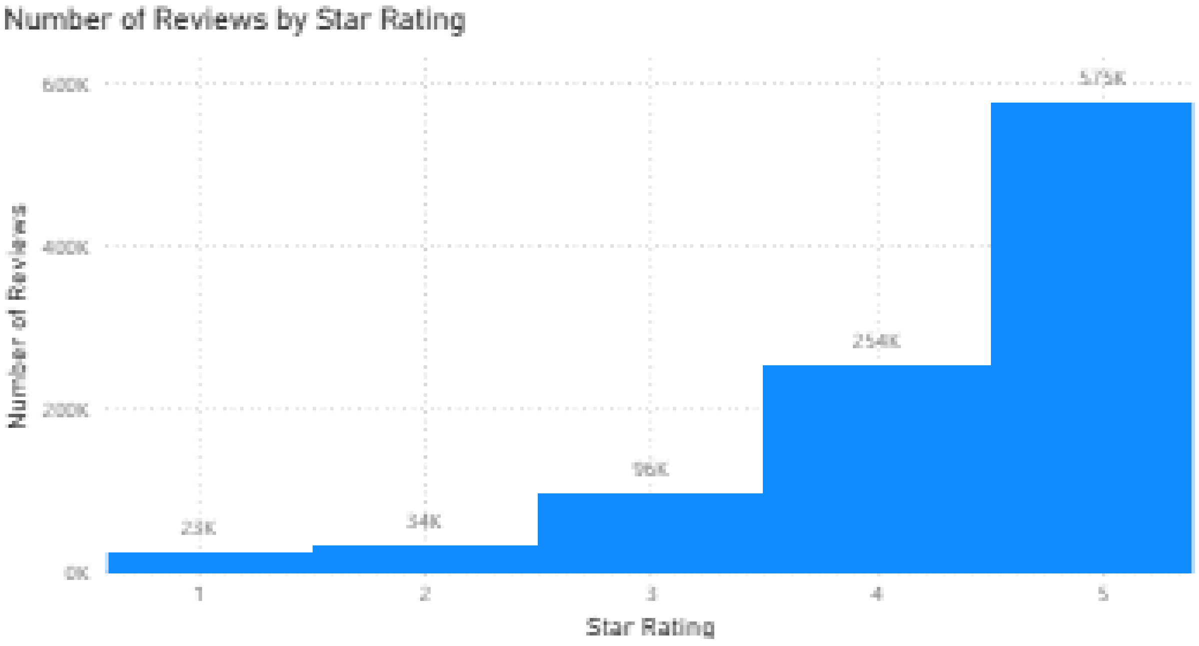
This choice balances model expressiveness, efficiency, and generalization.

Why this decision is strongly justified

- — Data-driven architecture choice : The max length is selected based on empirical review-length distribution, not arbitrarily.
- — Efficient training : Shorter sequences reduce memory usage and training time.

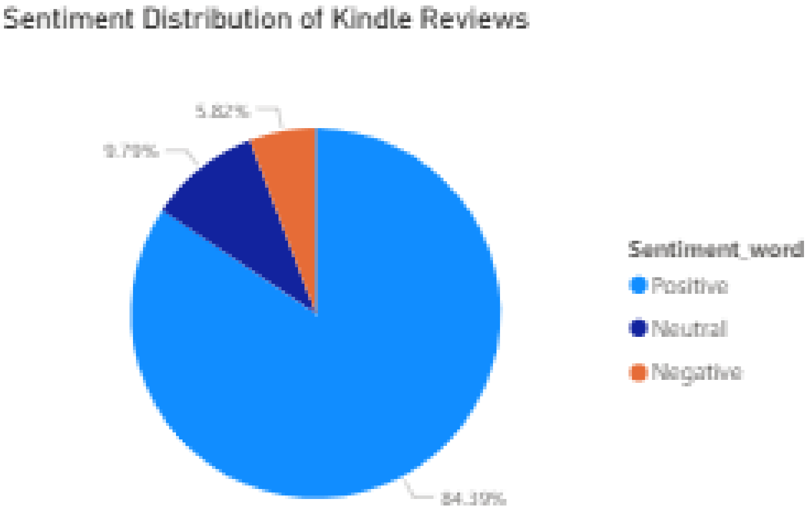
- — Minimal sentiment loss : Sentiment is usually conveyed early in the review.
- — Improved stability : Uniform input size stabilizes gradient updates and convergence.

1.3 Star Rating and Sentiment Distribution



Insight

The distribution of reviews across star ratings is highly imbalanced. High ratings dominate the dataset, with 5-star reviews (∼575k) and 4-star reviews (∼254k) accounting for the majority of samples, while 1-star (∼23k) and 2-star (∼34k) reviews are comparatively rare. This reflects a strong positive bias typical of online review platforms.



Insight

The distribution of sentiment labels derived from star ratings is highly skewed :

- Positive reviews (4–5 stars) : 84.39%
- Neutral reviews (3 stars) : 9.79%
- Negative reviews (1–2 stars) : 5.82%

This confirms that the dataset is strongly dominated by positive reviews, with negative reviews representing only a small fraction. Such imbalance can cause the ANN to overfit the majority class, leading to poor detection of negative sentiment.

Decision (For ANN Sentiment Analysis)

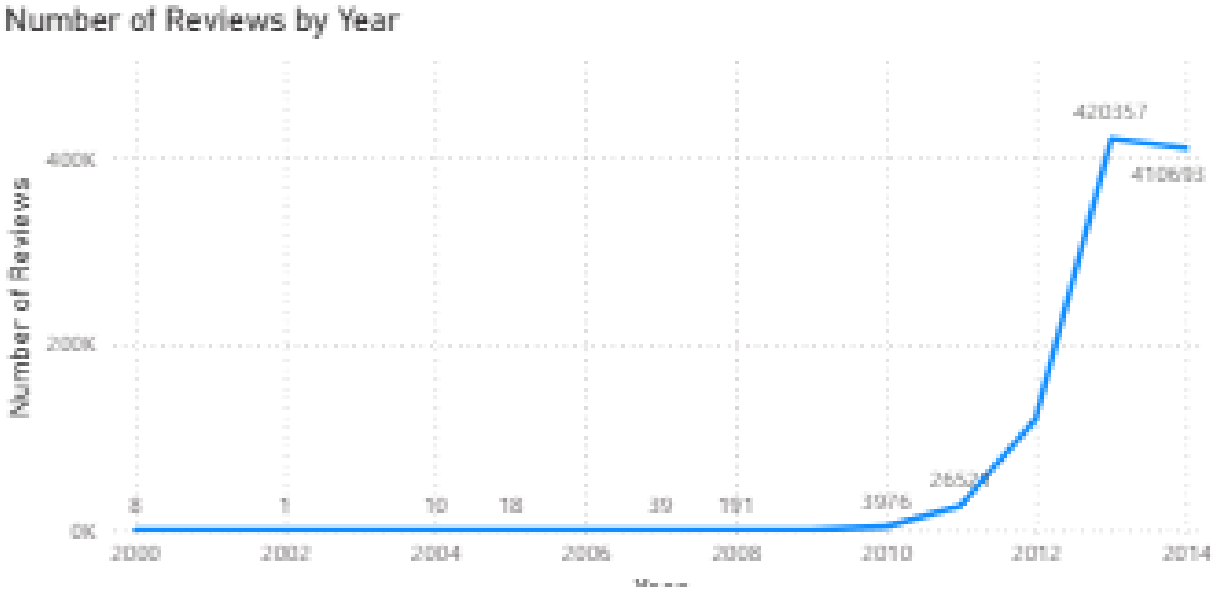
Given the severe class imbalance reflected in star ratings, explicit measures are taken to prevent the ANN from being biased toward positive sentiment.

- Star ratings are not used directly as input features to avoid sentiment leakage.
- When star ratings are mapped to sentiment labels, class imbalance is addressed through :
  - Class weighting in the loss function, or
  - Resampling strategies to balance positive and negative sentiment.
- Model evaluation prioritizes precision, recall, and F1-score rather than accuracy, ensuring fair assessment across sentiment classes.

Why this decision is strongly justified

- Prevents majority-class dominance.
- Improves minority-class detection.
- Avoids label leakage.
- Aligns with deep learning best practices.
- Ensures generalization.

1.4 Number of Reviews by Year



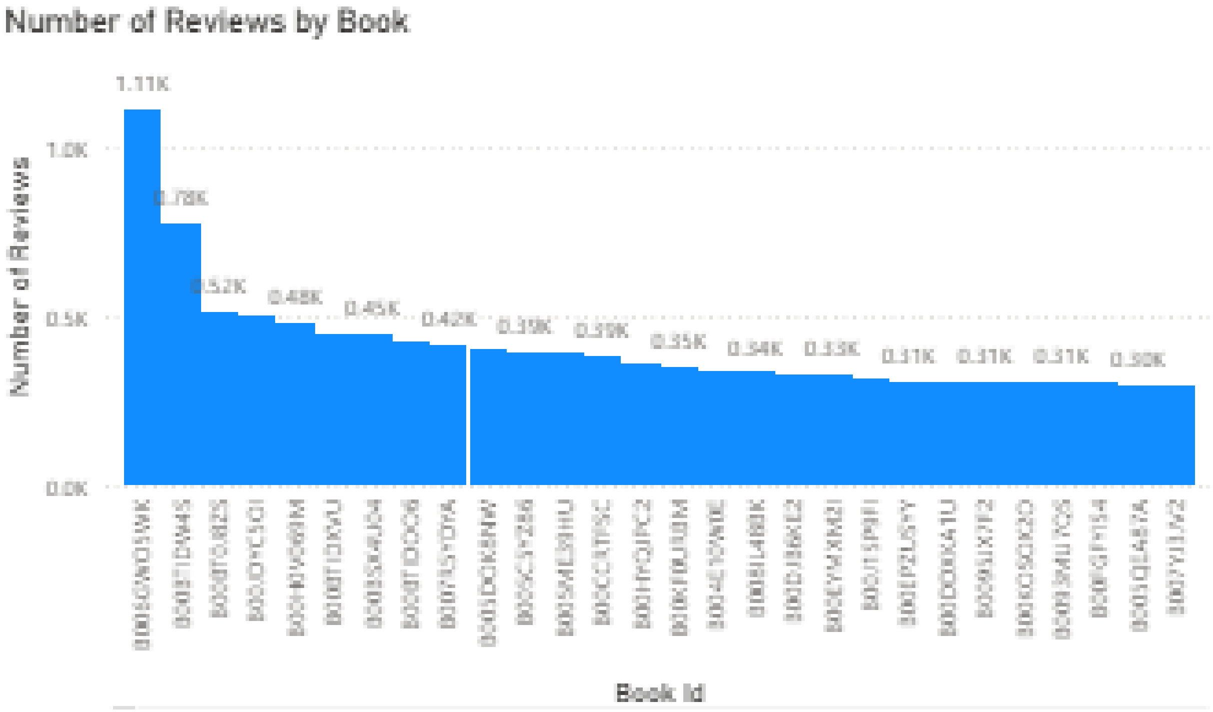
Insight

The dataset shows that review activity peaked around 2013, with fewer reviews in earlier and later years. This indicates that the majority of the dataset comes from a specific time period, and there may be temporal biases in the language, topics, or user behavior reflected in the reviews.

Decision (For ANN Sentiment Analysis)

- Temporal distribution is noted but not used as a primary feature : The ANN should focus on textual sentiment rather than time trends, as the goal is general sentiment prediction.
- Optional preprocessing consideration : we need to be aware that language patterns may have shifted since 2013 ; including a small subset of more recent reviews in validation can help assess generalization.
- No need for heavy weighting by year : Unlike helpfulness or sentiment imbalance, year does not directly impact label reliability, so it can largely be ignored during model training.

1.5 Number of Reviews per Book



Insight

The distribution of reviews per book is highly skewed :

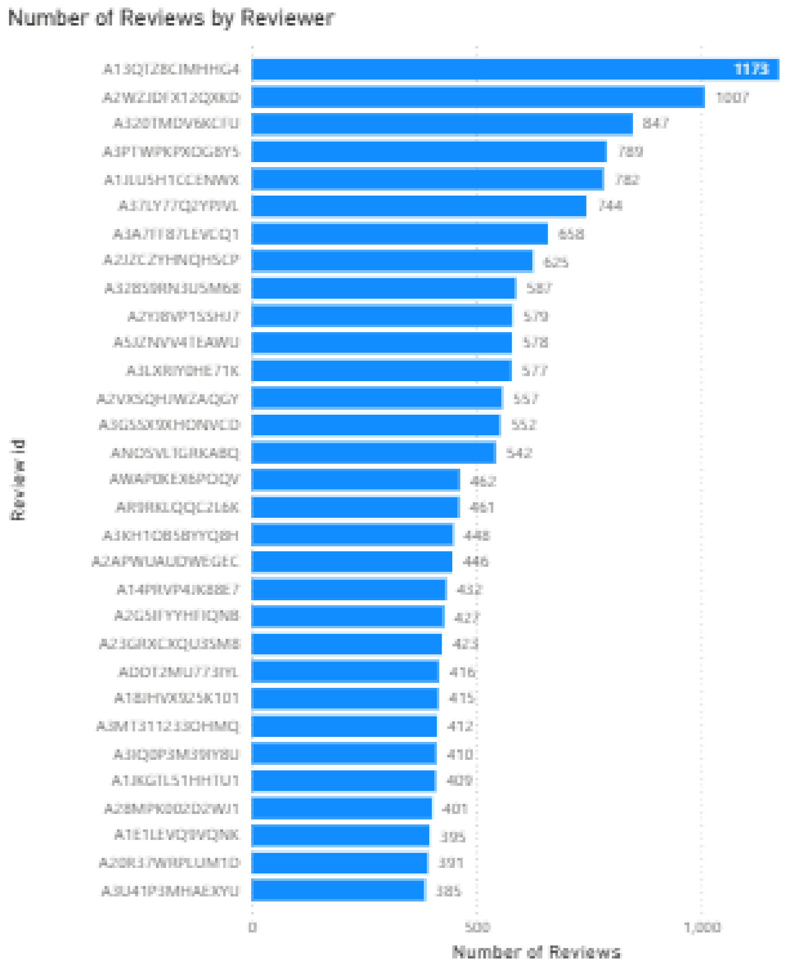
- A few books have very high numbers of reviews (max : 1110, 780),
- Then the count drops quickly to ~ 520, and gradually decreases to as few as 10 reviews per book.

This indicates that most books have relatively few reviews, while a small subset of popular books dominates the dataset. Consequently, the dataset contains a long-tail distribution with many books contributing minimal data.

Decision (For ANN Sentiment Analysis)

- Not oversampling or overweighting popular books : To prevent the ANN from overfitting to the sentiment patterns of a few highly-reviewed books.
- Random shuffling and batching : Ensure that reviews from both popular and less-reviewed books are evenly distributed across training, validation, and test sets.

1.6 Number of Reviews per Reviewer



Insight

The distribution of reviews per reviewer is highly skewed :

- A very small number of reviewers have submitted hundreds to ~ 1000 reviews,
- The majority of reviewers contribute only a few reviews, with the count gradually dropping toward 30 reviews per reviewer.

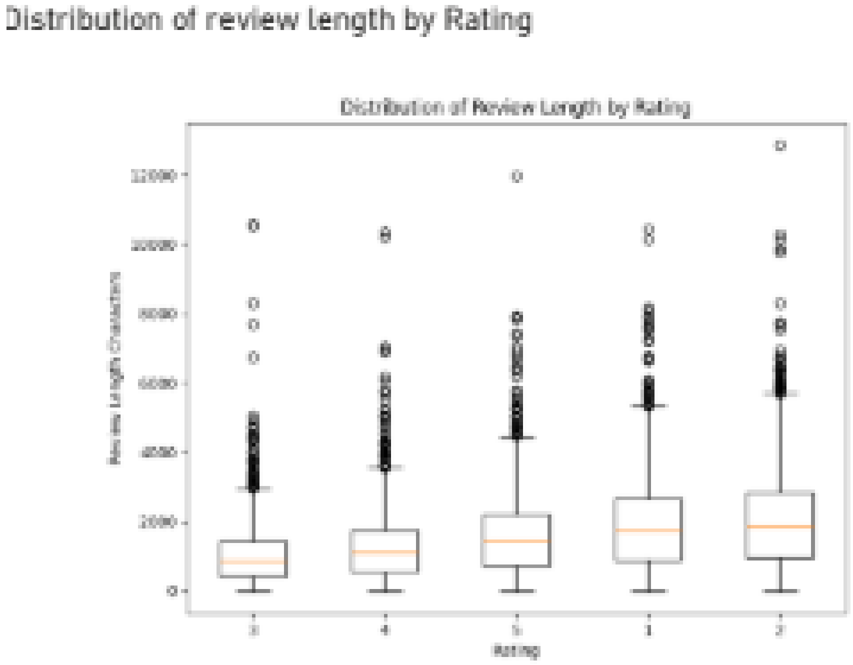
This shows that most reviewers are occasional, while a few highly active reviewers dominate the review activity. This long-tail behavior is typical in online review datasets.

Decision (For ANN Sentiment Analysis)

- Not overweighting prolific reviewers : Avoiding giving excessive influence to a few highly active reviewers, which could bias the ANN toward their writing style or sentiment patterns.
- Random shuffling and batching : Ensure that reviews from both prolific and occasional reviewers are evenly represented across training, validation, and test sets.

2 Bivariate Analysis

2.1 Review Length by Rating



Insight

Overall distribution : Most reviews are relatively short, with medians between 1,000–2,000 characters, and an interquartile range under 3,000 characters. The distribution is positively skewed with long upper whiskers and many extreme outliers (up to $\sim 12,000$ characters).

Differences across ratings :

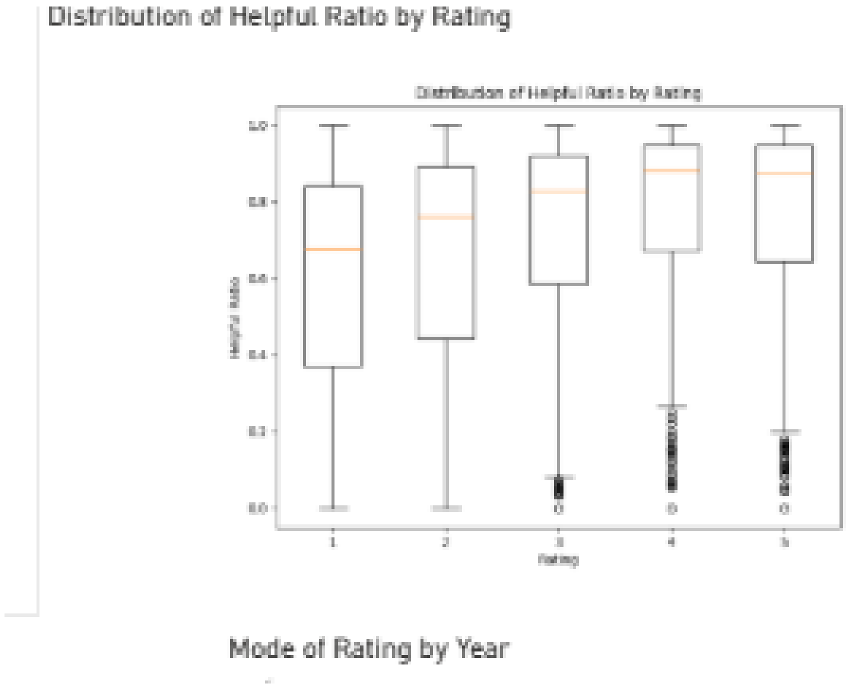
- Ratings 1, 2, and 5 (negative and very positive reviews) have slightly higher medians and more variability.
- Ratings 3 and 4 (neutral or moderately positive) tend to be slightly shorter.
- Extremely long reviews exist across all ratings, indicating that detailed feedback is a small but consistent phenomenon.

No strong linear relationship between review length and rating — median lengths are broadly similar across all ratings.

Decision (For ANN Sentiment Analysis)

Focus on textual content : The primary sentiment signal comes from the review text ; length differences are minor and serve only as a weak secondary cue.

2.2 Helpful Ratio by Rating



Insight

High overall helpfulness : Across all ratings, the median Helpful Ratio is consistently high, generally above ~ 0.75, indicating that most reviews receiving votes are considered helpful by the majority.

Highest medians : Ratings 3, 4, and 5 cluster around ~ 0.82–0.88, suggesting middle-ground and positive reviews are slightly more consistently deemed helpful.

Lower helpfulness for extreme negatives (Rating 1) :

- Median ~ 0.68 with a wide interquartile range (0.38–0.85).
- Some negative reviews are highly helpful, but others are much less useful, leading to a more variable distribution.

Consistent helpfulness for Ratings 2–5 :

- Narrower IQRs, with most reviews having helpful ratios above 0.50, indicating more stable perceived quality.

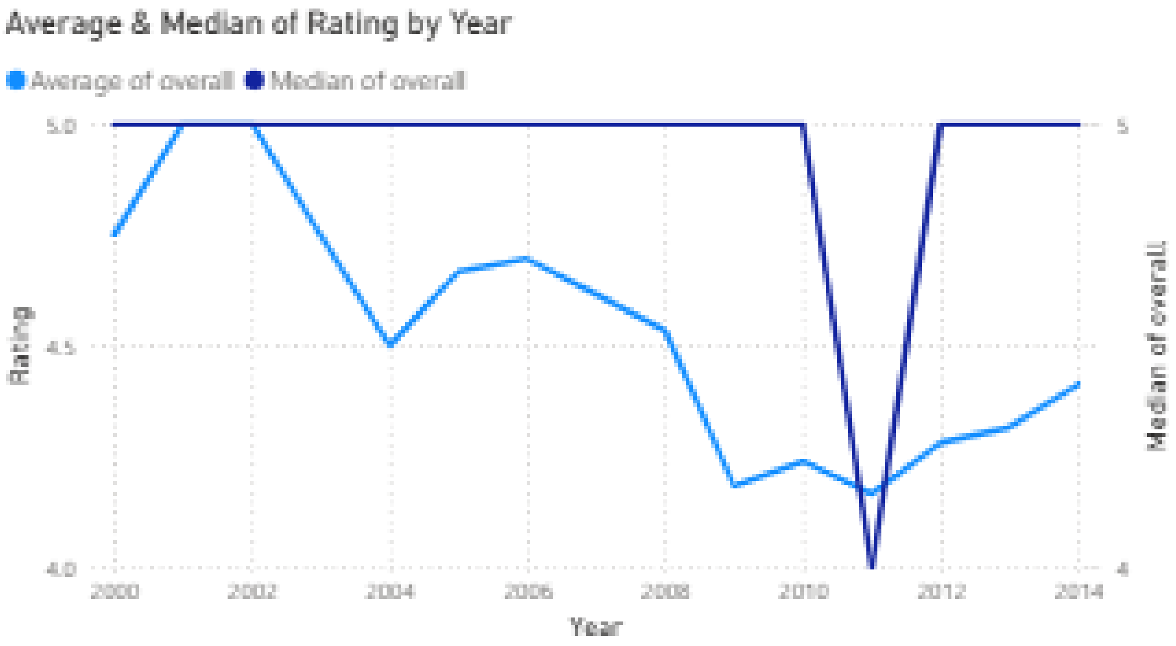
Overall, helpfulness is generally high regardless of rating, but negative reviews are more variable in perceived usefulness.

Decision (For ANN Sentiment Analysis)

- Using helpfulness as a loss weighting factor rather than as an input feature.
- Reviews with higher helpful ratios contribute more to the loss, increasing their influence during training.

- Reviews with no helpful votes or lower ratios contribute less, but are not ignored entirely.
- Account for variability in negative reviews.
- Not using helpful ratio as a direct predictor of sentiment.

2.3 Average and Median Rating by Year



Insight

Long-term downward trend in average rating : The average rating declines from around 4.75 (2000), peaks at 5.0 (2002), then gradually decreases, reaching its lowest point around 2011 (~ 4.1). A mild recovery is observed by 2014 (~ 4.4).

Median rating stability : The median remains consistently at 5.0 for most years (2000–2009 and 2012–2014), indicating that the majority of reviews are highly positive throughout the dataset.

2011 anomaly : Both the average and median sharply drop in 2011 (median = 4.0), signaling a substantial shift in user rating behavior during that year.

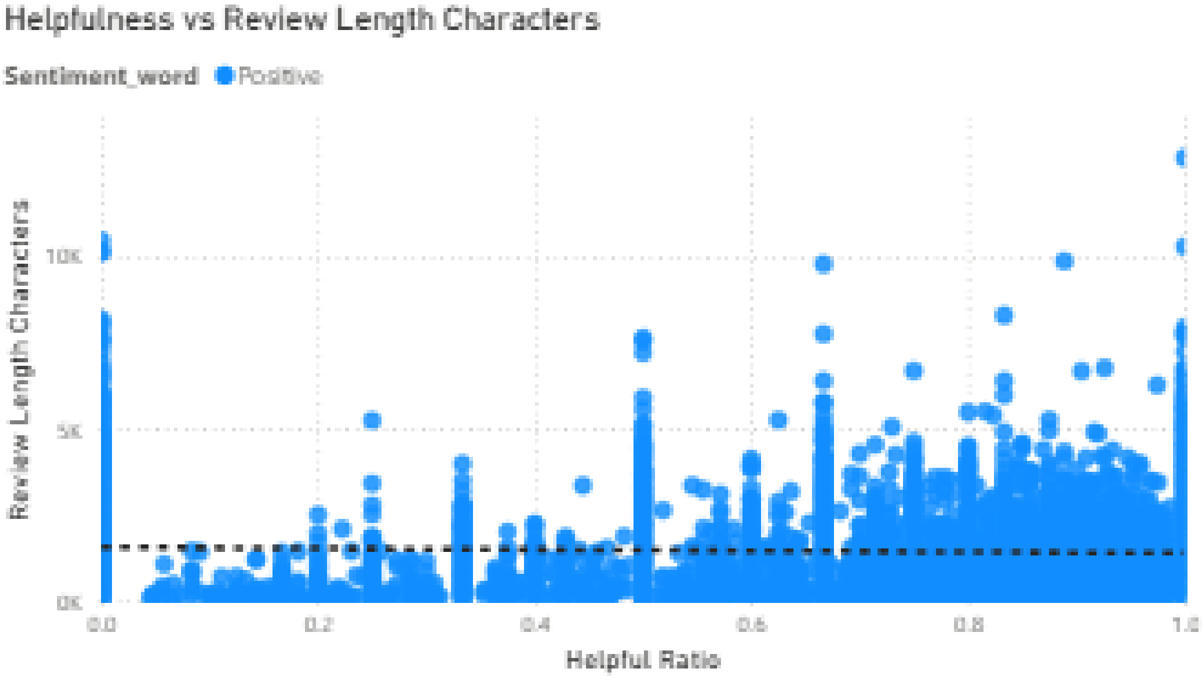
Overall, while most reviews are positive across years, sentiment distribution is not temporally stationary, with 2011 representing a significant deviation.

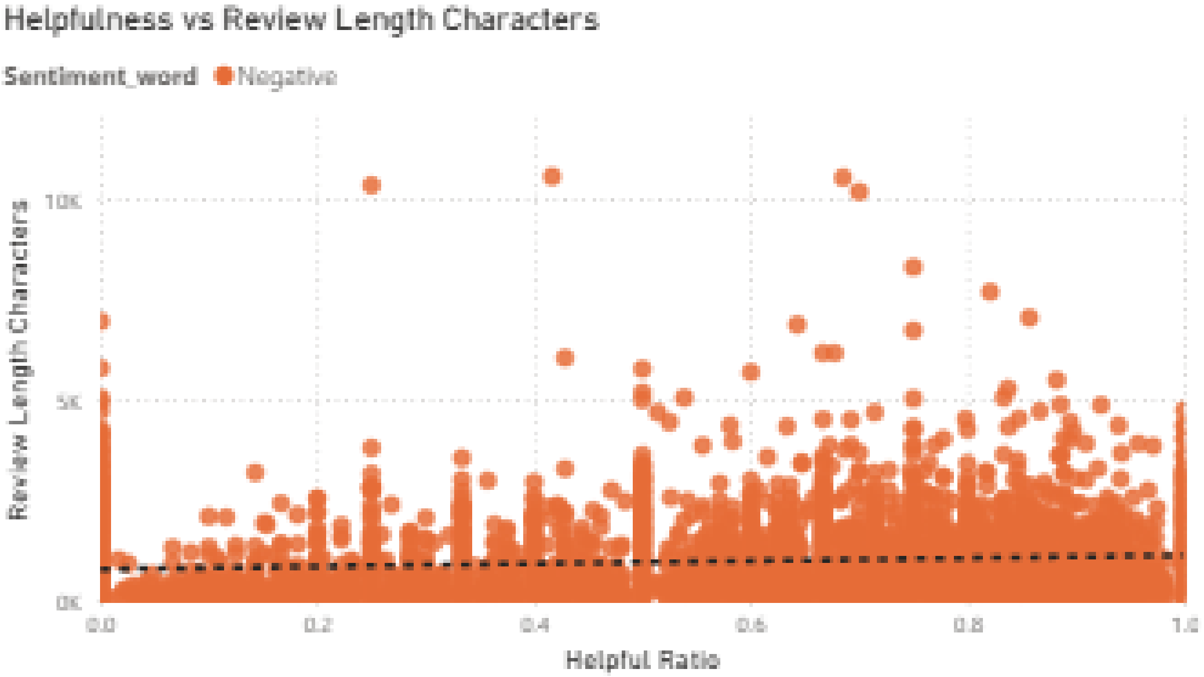
Decision (For ANN Sentiment Analysis)

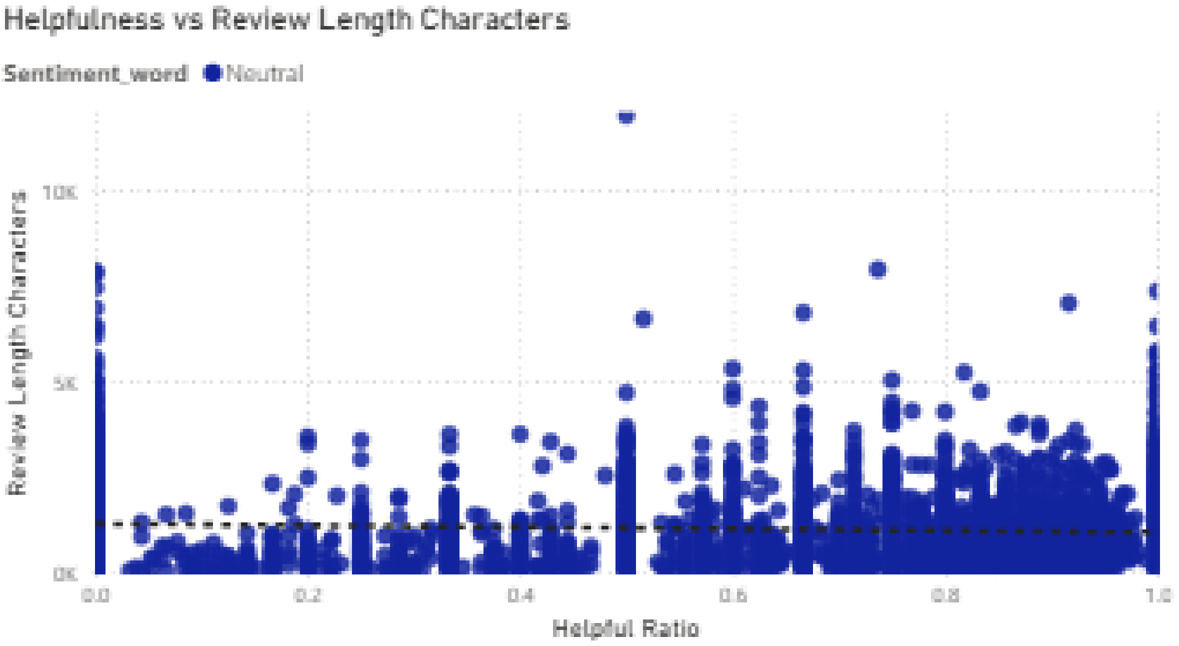
- Including temporal robustness in training.
- Not assuming uniform sentiment distribution across years.
- Avoidinging using year as a direct predictive feature.
- Monitor performance across time slices.
- Treating 2011 as a natural stress test.

3 Multivariate Analysis

3.1 Helpfulness vs Review Length by Sentiment







Insight

Weak overall relationship : Across positive, neutral, and negative reviews, the trend lines are nearly flat, indicating no strong linear correlation between review length and helpful ratio.

Consistency across sentiments : The same weak relationship holds for all three sentiment categories, suggesting that helpfulness depends more on content quality than on verbosity or sentiment polarity.

High-helpfulness cases (Helpful Ratio ≈ 1.0) : Most highly helpful reviews are short, but long reviews exist in both the positive and negative classes.

Neutral reviews : Neutral reviews show fewer long, highly helpful instances, indicating they are less likely to contain detailed, impactful feedback.

Decision (For ANN Sentiment Analysis)

- Not to use review length as a proxy for helpfulness : Since length does not reliably predict helpfulness, it should not be used to weight or bias sentiment predictions directly.
- Separate concerns : semantics vs. quality : Sentiment should be learned from textual semantics, while helpfulness should be handled externally via loss weighting, not through feature engineering.
- Maintain sequence-length constraints : Apply truncation/padding (e.g., 500 tokens) for efficiency, without fear of losing helpfulness-related signal.
- Avoid interaction features : Features combining length × helpfulness × sentiment are unnecessary and risk overfitting, given the weak observed relationships.

4 Conclusion

Based on the exploratory analysis, the ANN sentiment analysis framework is designed with the following decisions :

- **Loss weighting using helpfulness information.**
- **Capped sequence length at 400 tokens.**
- **Class-weighted training to address sentiment imbalance.**
- **Exclusion of star ratings and temporal variables as input features.**
- **Randomized batching to avoid dominance by popular books or prolific reviewers.**
- **Evaluation using class-sensitive metrics.**

Together, these decisions ensure a robust, efficient, and methodologically sound sentiment analysis model that learns semantic sentiment patterns while accounting for data imbalance, quality variability, and temporal shifts.