**Capstone Project**

**Turning Book Reviews into Business Opportunities**

**Background Context**  
 In the global book market, Amazon is the undisputed leader—selling approximately 300 million print books annually and generating $28 billion in book sales worldwide each year. It also controls over 50 % of the U.S. e-book market

With such scale, Amazon receives millions of customer reviews, which have a direct impact on purchase decisions: studies show that over 90 % of shoppers read reviews before buying.

However, manually sifting through large volumes of unstructured review text is time-consuming, error-prone, and often misses critical patterns—leading to slower issue resolution, higher return rates, and lost upsell opportunities. Business teams (Customer Support, Marketing, Product, and Publisher Relations) need a faster, more reliable way to flag “urgent negatives,” identify key themes in complaints, and surface top-rated feedback for promotional campaigns

# Problem statement

Manual triage of customer reviews is slow, inconsistent, and miss critical feedback, leading to higher support costs, slower issue resolution, and elevated return rates

* What is the problem or the opportunity that the project is investigating?
* Why is this problem valuable to address?
* What is the current state (e.g. unsatisfied customers, lost revenue)?
* What is the desired state?
* Has this problem been addressed by other research projects? What were the outcomes?

# Industry/ domain

* **Industry:** E-Commerce
* **Challenges:** Increasing competition, reputation management, customer retention

# Stakeholders

* Customer Support & Experience: Quickly surface truly negative or unhelpful reviews to cut resolution time.
* Marketing & Promotions: Identify enthusiastic, positive reviewers for targeted campaigns.
* Product Management & Merchandising: Spot recurring complaint themes (e.g. “missing pages,” “plot issues”) to drive quality improvements.
* Authors & Publishers: Understand which writing styles and themes resonate (or backfire) with readers.

# Business Question

How can we automatically detect and categorize negative Amazon book reviews so that Customer Support can triage issues faster, Product & Editorial teams can fix root causes, and Marketing can surface top‐positive feedback for upsells?

**Business Value**

**1. Support Efficiency**

* **Automate review triage**  
  Reduce manual tagging, freeing support agents to focus on complex issues and increase savings

**2. Rapid Negative-Review Response**

* **Flag 90 % of urgent negatives within 24 hours**  
  Exceed customer expectations (53 % expect reply within a week) and convert 18 % of detractors into repeat buyers—while intercepting return-triggering issues early to shave 1–2 % off your $28 B book-sales return rate.

**3. Positive-Sentiment Amplification**

* **Surface your happiest readers**  
  Identify and engage the top 5 % of “Recommendation & Enjoyment” reviewers for VIP campaigns, driving a 5–10 % lift in CTR and up to $2.5 M incremental revenue.

**4. Proactive Issue Resolution**

* **Pinpoint and prioritize complaints**  
  Negative-sentiment alerts let Product & Editorial teams zero in on real-time pain points (e.g. “format” or “plot” issues), enabling targeted fixes that boost quality and cut returns.

**5. Actionable Topic Insights**

* **Translate topics into actions**  
  Map topic labels to concrete initiatives—e.g.
  + “Format & Missing-Pages” → switch print partner
  + “Plot & Engagement” → refine marketing synopses
  + “Gender/Thematic” → add sensitivity reviews  
    This ensures every theme drives a measurable business improvement.

**Required accuracy & trade-offs:**

**Target:** F1-score ≥ 0.70 on the negative class, ROC-AUC ≥ 0.80.

**False positives:** Flagging a positive review as negative wastes support time and may annoy happy customers.

**False negatives:** Missing real negatives delays resolution, risking customer churn and higher return costs.

# Data question

Which features (textual and metadata) best predict a review’s sentiment (positive vs. negative) and, for negatives, which topic/theme it belongs to? What is the data required to answer the question?

# Data Answer

**Review text:** raw and preprocessed (tokenized, lemmatized) content.

**Explicit rating:** 1– 5 star value as a proxy for sentiment.

**Book metadata:**

* Category/genre (e.g. Fiction, Self-Help)
* **Publication date**
* Price (where available)
* Review metadata:

# Data

# Data Source

* **Kaggle “Amazon Books Reviews” dataset** by Mohamed Bakhet
* Collected from public Amazon pages (scraped) and uploaded to Kaggle

**Data Attributes**

* **Review Text** – Book Purchaser feedback (raw text)
* **Rating** – Numerical score given by the guest
* **Date of Review** – Timestamp of when the review was posted
* **Sentiment Score** – Computed sentiment classification (Positive/Negative)

## Data analysis

**Data Pipeline**

1. **Data Cleaning** – Removing special characters and formatting issues.
2. **Manual Labeling** – Assigning sentiment labels (Positive/Negative).
3. **Sentiment Analysis** – Applying ML models to classify reviews.
4. **Categorization** – Grouping reviews by Positive & Negative topic/theme.
5. **Summarization** – Extracting key insights from categorized reviews.
6. **Data Visualization** – Create charts and plots to show key insights

**Exploratory Data Analysis (EDA)**

**A blue circle with orange and white text

AI-generated content may be incorrect.**

**A graph with blue and orange bars

AI-generated content may be incorrect.**

**Positive sentiment has higher number of reviews than Negative Sentiment across all categories.**

**A graph of a number of numbers

AI-generated content may be incorrect.**

**Positive reviews have a higher word count than Negative reviews**

**A close-up of words

AI-generated content may be incorrect.  
  
Negative‐Sentiment Cloud**

**Terms:** whale, page, Moby Dick, Hemingway, classic

**Pain‐point theme:** Pacing & genre fatigue—some readers find *Moby Dick* overly long and “too classic” in style.

**A close-up of words

AI-generated content may be incorrect.**

**Positive‐Sentiment Cloud**

**Terms:** love, Moby Dick, Melville, enjoy, Maugham

**Strength Theme:** Readers love and enjoy Moby Dick & Melville and Maugham books

A pie chart with numbers and text

AI-generated content may be incorrect.

A pie chart with numbers and text

AI-generated content may be incorrect.

## Modeling

Processed Review is used as input features

Sentiment is the label we are predicting.

**Train/test split** partition your labelled dataset (X,y) into two disjoint subsets:

1. **Training set**  
   – Typically 70–80 % of your data.  
   – Used to fit your model’s parameters (e.g. learn the weights in a logistic regression or the splits in a random forest).
2. **Test set**  
   – The remaining 20–30 %.  
   – Held out during training and only used once at the end to get an unbiased estimate of **how well your model generalizes** to new, unseen reviews.

TF-IDF used **to evaluate the importance of a word in a document relative to a collection of documents (corpus).**

MaxAbsScaler is used to rescales each feature by dividing by its largest absolute value so that after scaling, every feature’s values range between –1 and 1. It doesn’t shift the data (no centering), so any zeros in your sparse inputs stay zeros.

Truncated Singular Value Decomposition (SVD), is a dimensionality reduction technique used on TF-IDF matrices to extract the underlying semantic structure of a text corpus. It transforms the high-dimensional TF-IDF matrix into a lower-dimensional representation that captures the relationships between terms and documents

The features were then trained using the Random Forest, XGBoost, Logistic Regression.

But the three model are not satisfactory, because the recall for the negative class is 49% to 50%.  
  
I decided to Combine Random Forest, XGBoost, SVM with Logistic Regression as meta Learning.  
Only then , the result is improve to 73% recall for negative class, without not much of a loss to positive class performance metrics.

Random Forest Classifer  
A screenshot of a graph

AI-generated content may be incorrect.

XGBoost

A screenshot of a graph

AI-generated content may be incorrect.

Logistic Regression  
A screenshot of a graph

AI-generated content may be incorrect.

Stacking Classifier (Most Suitable Model)

A diagram of a graph

AI-generated content may be incorrect.

# Business Recommendation to stakeholders

|  |  |  |
| --- | --- | --- |
| **Negative Issues** | **Recommendation** | **Stakeholders** |
| **Copy Quality & Pricing Issues** | Streamline replacement workflows  Implement clearer price-match guarantees to address  complaints about poor copies and  perceived overpricing. | **Customer Support & Experience**: |
| **Political / Historical Discomfort** | Add content advisories  Refine category tags for politically or historically heavy titles—  so readers know what to expect before purchase. | **Marketing & Promotions**: |
| **Length & Writing Style Critique** | Offer “Quick-Read” abridged editions and  tighten editorial guidelines to improve pacing and prose clarity. | **Authors & Publishers**: |
| **Character & Plot Engagement Issues** | Enrich book descriptions with detailed plot/character  summaries and test synopses for emotional hooks to  boost reader engagement. | **Product Management & Merchandising**: |

# References

* Where are the data and code used in the project?

<https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews/data>  
  
<https://github.com/shahriar-rahman/EDA-Amazon-Books-Reviews/blob/main/notebooks/feature_exploration.ipynb>  
  
<https://github.com/rollyjohn/Topic-Modelling/blob/main/topic_model_V3.ipynb?fbclid=IwZXh0bgNhZW0CMTAAAR2Iw-KoPY1BjGQgBPmvv7x324m-6juHJgFVO7RTtz7C60gGf7oAZ5wJhlI_aem_vGREf6PLWVVqWcd72EyB6Q>

* What are the resources used in the project? (libraries, algorithms, etc)

Sklearning