Project Title:

Advanced Data Analytics – Assignment 2:

Customer Sentiment Analysis & Book Recommendation System

Submitted by Group 3:

Aamena Zakir Shaikh (NF1005115)

Fabio dos Santos Prumucena (NF1002000)

Renato Hiroyuki Oshiro (NF1011996)

Rohit Kumar (NF1003118)

Instructor Name:
Ali El-Sharif

Submission Date:
30th November, 2025

Institution:
University of Niagara Falls, Canada

1. Business Challenge 1 — Yelp Customer Sentiment Analysis

1.1 Objective

Yelp receives millions of textual reviews daily, making manual sentiment evaluation impractical.
Goal: automatically classify reviews as **Positive** or **Negative** to help businesses improve service quality.

1.2 Exploratory Data Analysis (EDA)

- The dataset includes **balanced classes**, slightly more positive reviews.
- Random samples displayed from both classes highlighted:
  - Positive: praise related to food quality, staff friendliness
  - Negative: complaints about service delays, poor taste
- Review length distribution:
  - Most reviews are **short (20–80 words)** → ideal for transformer token limits
  - A wide tail of longer reviews → supports rich sentiment cues

Insight: Preprocessing must retain enough text detail to capture sentiment-bearing phrases.

1.3 Baseline Machine Learning Model

Model: **TF-IDF + Random Forest Classifier**

- Train/Validation split applied
- Initial evaluation metrics:

Metric	**Precision**	**Recall**	**F1-Score**
Positive	~0.89	~0.90	~0.90
Negative	~0.88	~0.87	~0.88

A **confusion matrix** was generated: errors mainly occurred with **neutral-toned** negative reviews misclassified as positive.

1.4 Hyperparameter Optimization — Grid Search

Parameters tuned:

- Number of trees: **100 vs 500**
- Max depth: **None, 20, 50**
- Min samples split: **2, 5**
- Min samples per leaf: **2, 4**
- Max features: **sqrt, log2**

Best Model Outcome

- Improved recall on negatives (fewer undetected unhappy customers)
- Better precision reduces escalation on borderline reviews

Business Value
Supports **service quality improvement**, alerting managers faster to customer dissatisfaction trends.

1.5 Advanced Transformer Models

Two pretrained Hugging Face models implemented:

Model	**Rationale**
BERT-Base-Uncased	Industry standard benchmark
RoBERTa-Base	Enhanced training → higher downstream accuracy

Performance Comparison

Model	**Accuracy**	**Key Benefit**
TF-IDF + RF	~0.90	Fast & scalable
BERT	~0.93	Better semantic understanding
RoBERTa	**~0.95**	Best sentiment capture and fewer false negatives

Business Impact

- **Improved identification of unhappy customers**
- Escalations can be prioritized → cost reductions
- Better insights into **brand reputation** trends

Final Recommendation: Deploy **RoBERTa** model for production.

2. Business Challenge 2 — Book Recommendation System (GoodBooks-10K)

2.1 Objective

Help users discover books aligned with their interests by learning from:

- Personal rating history
- Similar users' preferences

Dataset Summary:

- **6M+ ratings**
- **50K+ users, 10K books**
- Ratings **1–5 scale**

Major sparsity issue:

- Most users rate **few books** → requires matrix factorization approach

2.2 Collaborative Filtering with ALS

ALS = **Alternating Least Squares**
(Implemented using the implicit library)

What ALS Solves

- Finds hidden patterns between readers and books
- Recommends books not yet rated → increases user engagement

How ALS Works

- Builds **two latent factor matrices**:
    - User preferences
    - Item/book attributes
- Optimizes them **alternately** until reconstruction error minimizes

Main Hyperparameters

Hyperparameter	**Role**
Factors	Number of hidden preference dimensions
Regularization	Prevents overfitting
Iterations	Controls model convergence
Alpha (implicit)	Strength of confidence in ratings

2.3 Implementation & Results

Steps performed:
Created **user–item sparse interaction matrix**
Trained ALS only on the **training set**
Extracted recommended books per user from learned latent vectors

Outcome Example
Users receive **personalized recommendations**:

- "Because you loved *Book X*, here are 10 similar titles."

Business Benefit

- Keeps users engaged longer
- Boosts platform revenue through:
  - Subscription retention
  - Higher interaction levels

Conclusion & Actionable Insights

Business Impact Area	**Sentiment Analysis**	**ALS Recommender**
Customer Satisfaction	Detect service failures faster	Suggest relevant content
Competitive Advantage	AI-powered understanding of customer voice	Better user retention

Business Impact Area	**Sentiment Analysis**	**ALS Recommender**
Scalability	Handles millions of reviews	Handles millions of ratings
Cost Reduction	Fewer negative reviews missed	Less churn, more engagement