2025

CS210 Final Presentation

Beauty Products:
Your Profit-Driven Investment Strategist

By Nada Mahmoud and Mariam Khan

Table of Contents

Table of Contents

Project Overview

Problem:
Companies collect thousands of customer reviews but usually rely only on star ratings or simple sentiment scores. These methods fail to explain *why* customers are dissatisfied or what issues deserve immediate investment

Goal:
Build an analytics system that transforms written customer reviews into **clear, prioritized improvement actions** for R&D teams

Solution Overview:
Using NLP and financial modeling, we:

- Extract major complaint themes from review text
- Measure dissatisfaction intensity through sentiment scoring
- Rank issues using a cost-adjusted priority algorithm
- Present recommendations through an interactive business dashboard

Outcome:
A decision-support tool that guides R&D investment using real customer data rather than intuition

Problem

Limitations of Current Approaches:

- Star ratings summarize opinion but offer no problem explanation
- Manual review reading is time-consuming and inconsistent
- Keyword tracking misses nuanced complaints
- Basic sentiment analysis ignores root causes
- No built-in prioritization method for selecting improvement targets

Business Impact:

- Companies fix the wrong problems first
- Investment decisions are often emotion-driven instead of data-driven
- High-impact issues remain hidden in large data volumes

Research Question:
How can review text be transformed into actionable R&D guidance that considers both customer urgency and business feasibility?

Our Solution

1_data_processor.py

↓

2a_embedding_generator.py

↓

2b_topic_clustering.py

↓

3_priority_ranking.py

↓

4_financial_modeling.py

↓

app.py

Step 1: Review cleaning and sentiment scoring

- **Business value:** Cleaned, tagged dataset quantifying the **emotional intensity** of dissatisfaction

Step 2: Topic modeling to identify major complaint themes

- **Business value:** Identified 50 major, recurring complaint themes (e.g., 'acne pimple', 'sunscreen white cast')

Step 3: Priority scoring combining:

- Complaint frequency
- Emotional negativity
- Cost of product or operational fixes

Step 4: Financial modeling to justify investments

- **Business value:** Converts topics into quantifiable metrics: **Projected Gain ($M)** and **Net Impact ($M)**, justifying investment return

Step 5: Dashboard visualization

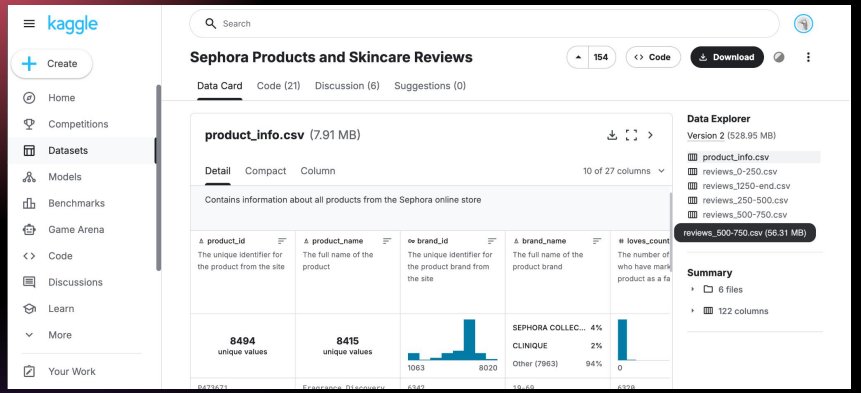
- **Business Value:** actionable roadmap for stakeholders, driving quick, data-informed investment decisions

Result: a structured pipeline that converts unstructured customer language into ranked, measurable R&D investments

Data Sources

Primary Datasets:

- Sephora product & skincare reviews (Kaggle) - 1 .csv file
- Customer reviews (Kaggle) - 5 .csv files

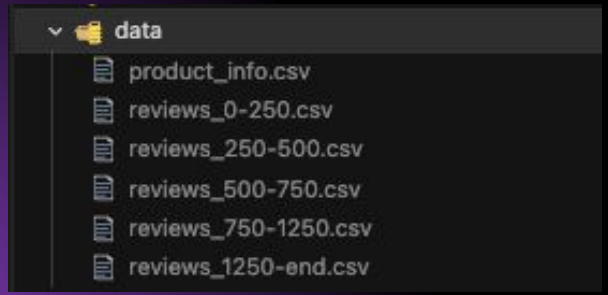


Data Types:

- Review text (unstructured) - used for topic modeling
- Star ratings (numeric) - used for risk segmentation
- Product Metadata: Product SKUs, brand IDs, category labels - provided context for financial modeling

Key Challenges:

- Inconsistent language - typos, slang, inaccurate grammar, etc.
- Data redundancy - duplicates can cause over-weighting
- Data heterogeneity - irregular lengths, mixed formatting across files



[Link to dataset!]

Phase 1: NLP Cleaning & Sentiment Analysis

Data Preprocessing & Normalization::

- **Cleaning & Standardization:** Removed punctuation, corrected spacing, and normalized capitalization to ensure consistency.
- **Tokenization & Lemmatization:** Broke down sentences into individual words and reduced them to their root form to aggregate related terms accurately.

Sentiment Analysis:

- Each review assigned a continuous negativity score.
- Scores quantify dissatisfaction intensity instead of binary labels.
- Allows comparison of emotional impact across topics.

Core Data Filtering:

- **Risk Segmentation:** Filtered to include only reviews with ratings <= 3 stars. This helps us focus only on high-risk and dissatisfied customer feedback.

Output:

- Cleaned, sentiment-tagged review records ready for clustering.

Phase 2: Embedding Generation & Topic Modeling

Method: BERTopic

- → Uses transformer models (BERT) to create clusters
- → Converts text reviews into dense semantic embeddings
- → Groups reviews into complaint categories based on meaning

Advantage: CAPTURES SEMANTIC MEANING!

Example Topics

- → Acne pimple scar cystic (Targeting R&D)
- → Eye cream circle dark (Targeting New Product Development)
- → Pump packing bottle product (Targeting Operations & Supplier Review)

Benefits

- → Eliminates noise
- → Quantifiable risk
- → Actionable alignment

Outputs

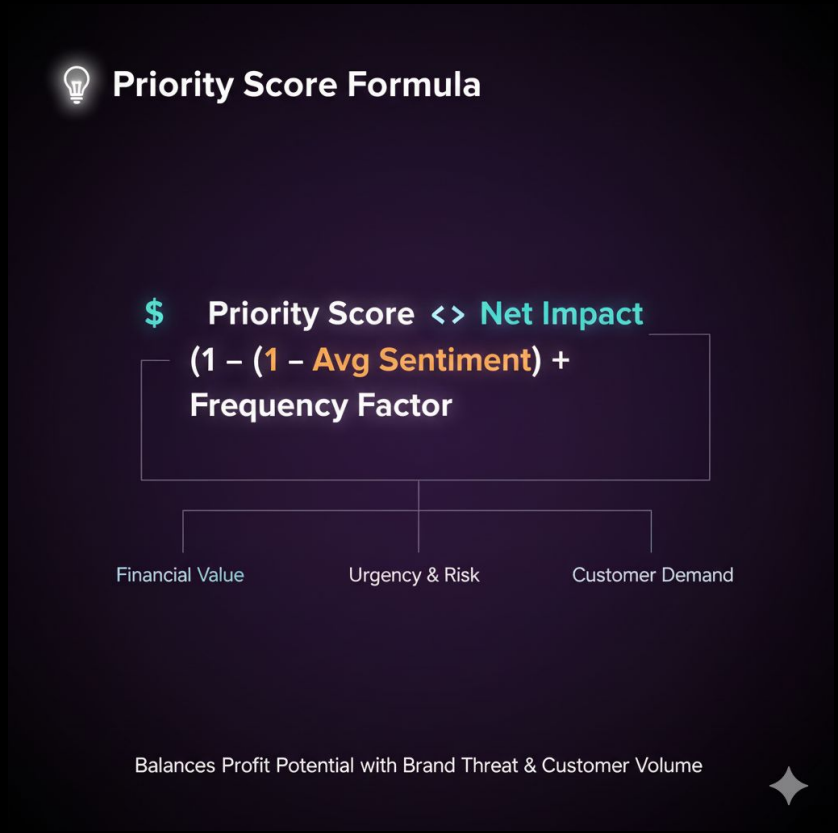
- → 50 verified, semantically distinct topic drivers - each quantified by its list of representative words (c-TF-IDF score) and assigned to over 48,000 sampled negative reviews
- → A segmented dataset - every <= 3 star rated review is linked to a primary business problem

Phase 3: Prioritization Algorithm

Why?

To define the weighted logic that converts customer dissatisfaction into a single, comprehensive score for strategic ranking. This formula ensures that not all complaints are treated equally. Next, issues with higher impact, frequency, or severity automatically rise to the top of the priority list. By quantifying qualitative feedback, the team can allocate resources objectively rather than relying on intuition or anecdotal evidence.

This step calculates the priority scores and ranks them, displaying the top 5 topics.



Phase 4: Financial Modeling & Action Mapping

1. Financial Quantification

- **Measure Value:** Calculated **Net Impact (M)** to justify large budgets
- **Measure Efficiency:** Calculated **Value-to-Cost Ratio (ROI)** to prioritize "quick win" projects
- **Final Data:** Every customer issue now has a quantified profit and efficiency score

2. Action Mapping

- **R&D Mandate:** Topics with High Net Impact are mapped to **Product R&D / New Product** action
- **Operations Mandate:** Topics containing keywords like 'pump' or 'packaging' are mapped to **Packaging & Supplier Review**
- **Strategic Mandate:** High ROI topics are flagged for **Immediate Implementation** regardless of department

Strategic Outcome:

- The strategy is fully decentralized: any manager can filter the final dashboard by **Recommended Action** to find their mandate
- The conversation shifts from "Do we fix this?" to **"How soon can we fund this?"**

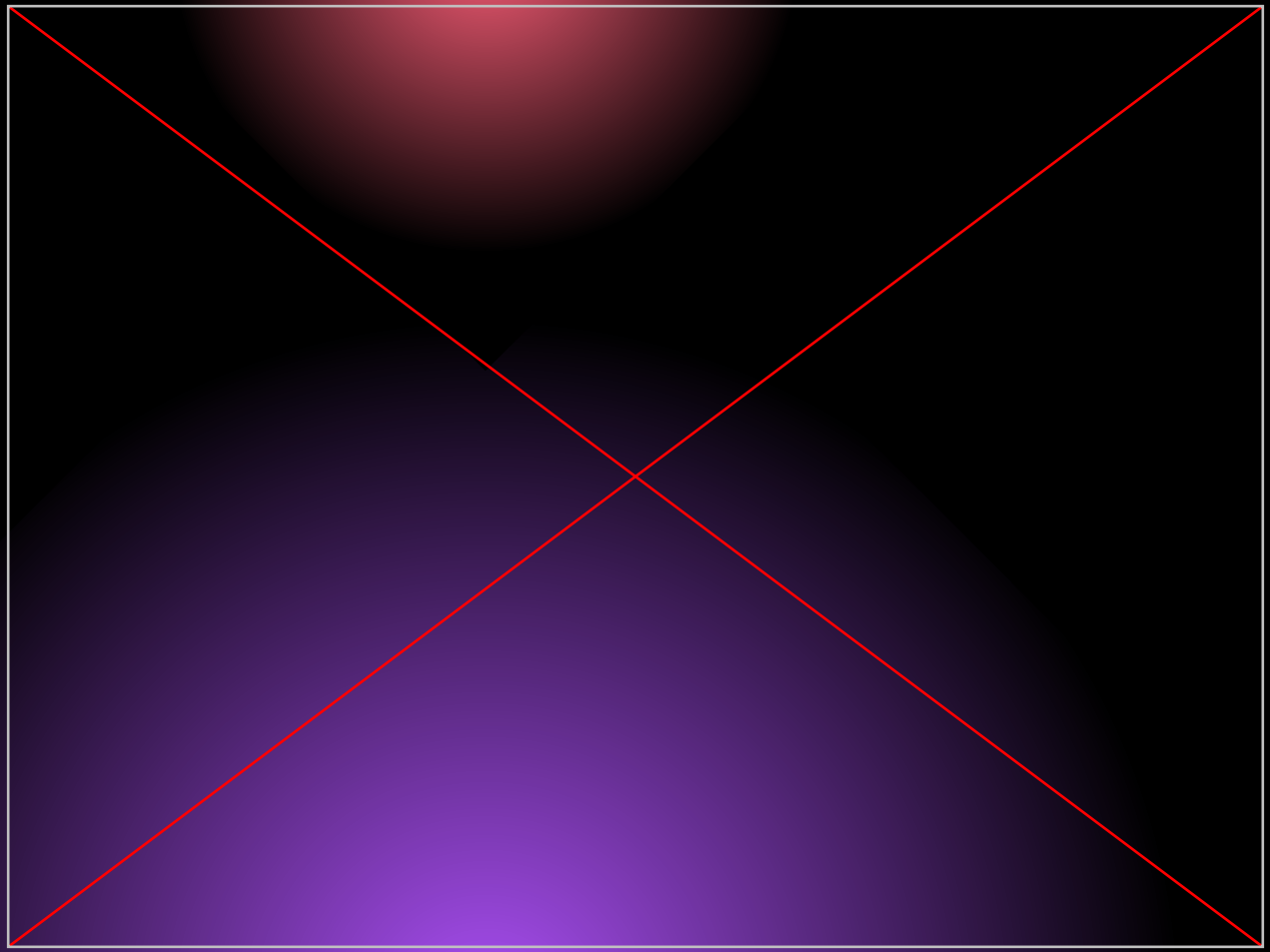
Implementing OpenAI API

Used advanced external models (via openai_api.py) to automatically translate complex data analysis into clear, actionable executive narratives

Key Features:

- Generated the concise, plain-language Business_Summary for all 50 financial topics
  - complex findings instantly digestible and convenient for any
  - at a glance
- **Action Mapping Validation:** Reviewed calculated financial metrics and keywords to assign the definitive Recommended_Action_Type (e.g., "Product R&D"), ensuring every dollar figure had a concrete next step
- **Executive Narrative:** Produced the overall high-level summary (e.g., Global ROI, Top Threats) displayed on the dashboard, providing the C-suite with an immediate strategic conclusion

Demo



Thank you!