Project Title

Customer Review Intelligence for Profit-Driven Product Investment Strategy in the Cosmetics Industry

Group Details

Nada Mahmoud - nam298
Mariam Khan - mk2199

I. Project Definition

The motivation behind this project is the need to convert large volumes of unstructured customer feedback into structured intelligence that can support decision-making within research and development (R&D) teams. The development of this Decision-Support System (DSS) would entail the translation of qualitative review text into a quantitative framework based on volume, risk, and cost. In the modern cosmetics and beauty industry, companies collect hundreds of thousands of written product reviews across online retail platforms such as Sephora, Ulta, and Amazon. These reviews contain qualitative insights about customers' experiences, but they are rarely analyzed in a meaningful way. Most companies rely on surface-level metrics such as average star ratings, simple keyword searches, or broad sentiment summaries. While these tools capture general satisfaction trends, they lack the precision required for real product improvement. They do not reveal which aspects of the product fail, how many customers experience the same issue, or how emotional or severe the dissatisfaction is. Even more critically, they do not help companies understand which problems are worth fixing given real-world budget constraints.

This project responds directly to this challenge by proposing an end-to-end pipeline designed to translate unstructured review text into structured data, categorize specific consumer pain-points, and prioritize them using both emotional intensity and operational cost considerations. Instead of treating customer reviews as static information, the system transforms them into a decision-support tool that identifies what issues matter most to customers and which improvements will produce the greatest business impact. The project aims to empower product teams by giving them a rational, quantified, and transparent basis for choosing which problems to fix first.

At its core, the project seeks to address a widespread gap in industry practice, which is the disconnect between customer feedback availability and customer feedback usability. Companies have enormous amounts of data, but without intelligent processing pipelines, this information remains difficult to interpret. Human teams cannot manually read tens of thousands of reviews, nor can they reliably estimate which issues are most important based solely on intuition or anecdotal summaries. This project demonstrates how algorithmic text analysis can bridge this gap and support data-driven, proactive product development rather than reactive problem-solving.

II. Introduction

The importance of this project comes from its integrated, action-focused approach to analyzing customer reviews. While natural language processing (NLP) is widely used in industry, most applications remain descriptive. They summarize sentiment polarity or visualize frequent words, but fail to deliver prescriptive recommendations. Managers may learn that customers are "unhappy," but they do not learn why, or what to fix, or how to allocate budgets strategically.

Our project moves beyond descriptive analytics by integrating multiple techniques: topic modeling, sentiment scoring, cost estimation, and financial impact forecasting into a single continuous workflow. This design is not common in academic coursework or in many corporate dashboards. Most systems stop at clustering or sentiment analysis, leaving the interpretation and prioritization to human judgment. In contrast, our methodology formalizes a prioritization framework that combines three essential signals:

1. Volume - How frequently a complaint appears in reviews.

2. Risk - How emotionally negative customers feel about that issue.

3. Cost - How expensive it is operationally to fix the underlying problem.

The priority score formula we used as a final ranking metric, which mathematically formulates the trade-off between risk, volume, and cost:

$$\textit{Priority Score} = (\textit{Frequency} \times |\textit{Avg Sentiment}|) / \textit{Cost of Fix (M)}$$

Essentially, this formula aims to address issues with the highest combination of negative volume and low cost, bringing them to the top of the action list. This optimizes the investment for maximum risk reduction per dollar spent.

By integrating financial modeling, the project also ensures that recommendations reflect real-world economic constraints. That's normally a dimension that many academic NLP projects overlook. A problem that bothers many customers may be too expensive to fix relative to its expected revenue return, while a smaller, cheaper issue may produce a much higher ROI. This balancing method mirrors the true decision-making environment of product development teams.

Another key dimension of novelty lies in linking text analytics directly to organizational workflows. Instead of ending with a list of topics, the project maps each complaint cluster to the department responsible for fixing it. It can fall under formulation scientists, packaging engineers, logistics teams, or quality assurance. It demonstrates how automated analytics can guide operational processes, trigger alerts, and shape R&D agendas rather than simply inform them.

Overall, the project contributes to broader course themes surrounding unstructured data management, algorithmic classification, and the creation of dashboards that translate complex analytics into accessible information for nontechnical business stakeholders.

III. Methodologies

a. Data Science Component

Data were sourced from a publicly available Kaggle dataset focused on cosmetics and Sephora reviews. These datasets include consumer-written reviews, star ratings, product categories, brand metadata, and SKU identifiers. Reviewing the raw data showed many issues: inconsistent text formatting, slang, abbreviations, typos, and duplicated entries. Due to the written reviews varying drastically in length and style, preprocessing was an important first step.

Data cleaning involved merging multiple datasets containing overlapping fields, meshing encoding formats, and filtering out non-English entries. Duplicate detection techniques were used to prevent over-representation of repeated reviews. Text normalization workflows included lowercasing, removal of punctuation and stop words, tokenization (‘nltk’), and lemmatization (‘WordNetLemmatizer’). These steps ensured the dataset was standardized enough to support reliable downstream NLP analysis.

From the initial master dataset, about 1 million reviews were filtered to include only those rated three stars or below, focusing the topic modeling analysis entirely on customer frustration. This filtered dataset represents the subset of customers actively expressing concerns or frustrations. A sentiment analysis model, specifically VADER (Valence Aware Dictionary and sEntiment Reasoner), then assigned each review a continuous negativity score, capturing emotional intensity beyond binary positive/negative labels. From the initial filtered dataset (ratings <= 3 stars), a 25% sample was utilized for a stable BERTopic cluster, especially because the large size of the dataset dedicated a large portion of time to generating the vector embeddings.

Simulated datasets were created during early development so that dashboard and modeling logic could be fully tested while real-data processing continued. These simulated datasets mimicked real distributions of review lengths, sentiment ranges, and category frequencies, ensuring that the pipeline behaved realistically before full deployment.

b. Database Component

Although the project did not require full relational-database implementation, we treated the cleaned and engineered data as structured tables containing:

- Review IDs
- Product identifiers (SKU, category, brand)

- Sentiment scores
- Topic probabilities
- Topic labels
- Cost estimates - Cost of Fix (M) was simulated based on topic keywords (e.g. ‘packaging’ keywords received a low-cost estimate of $1.5M - $3.5M, while ‘formula’ keywords received a high-cost estimate of $3.0M - $7.0M).
- Priority scores
- Action-mapping fields

Organizing the data in this way allowed seamless movement across phases of the pipeline and enabled clear visualizations within the dashboard. Thoughtful structuring also ensured that future work - including real-time streaming ingestion - can be accommodated easily.

We used this dataset from Kaggle, which consists of 6 different files. The first file contains the product information, while the other five files contain customer reviews (which we consolidated into 1 dataset). Thus, we used 2 separate datasets to complete our analysis.

c. Machine Learning Component

For topic modeling, we used a hybrid approach: Sentence-BERT(all-MiniLM-L6-v2) was used for high-quality semantic embeddings, followed by UMAP for dimensionality reduction ($n_{neighbors}$ = 15, components = 5, min_{dist} = 0.0), and finally K-Means clustering (where k = 50) to enforce a fixed, stable number of complaint topics.

Deep-diving into these concepts: BERTopic is an advanced framework that leverages sentence-BERT embeddings to cluster semantically similar reviews. Unlike traditional clustering methods such as LDA, which rely solely on word co-occurrence, BERTopic captures deeper contextual meaning, particularly valuable given the informal, expressive language common in beauty-product reviews.

BERTopic identified complaint clusters, such as:

- Packaging defects and leakage
- Smudging or poor wear-time
- Skin irritation or allergic reactions
- Unpleasant scent characteristics

- Texture inconsistencies
- Color mismatch
- Pricing frustrations

Each review was assigned topic probabilities, enabling multi-label representation when reviews described more than one issue.

The final step involved financial modeling that estimated potential revenue retention attributable to fixing each issue. Using simulations based on market share, customer lifetime value, and the assumed elasticity of repeat-purchase likelihood, the model projected long-term financial impact and compared it against expected remediation costs. This transformed raw analytics into measurable business predictions.

d. Dashboard Implementation

Using Streamlit, Pandas, NumPy, and Plotly, we built an interactive dashboard enabling stakeholders to explore insights intuitively. It also features AI-generated business summaries and a comprehensive executive narrative, powered by an external API by OpenAI (model version: gpt-3.5-turbo), ensuring that complex analytical rankings are instantly translated into accessible C-suite language. Users can filter the final roadmap using four inputs: Recommended Action Type, Minimum Net Impact (M), Max Average Sentiment, and Minimum Topic Frequency. The dashboard includes:

- Total negative review count
- Distribution of sentiment scores
- Number of high-priority topics
- Tables mapping each complaint topic to the responsible departments
- Filters for product category and timeframe

This design ensures that managers, product designers, and R&D leads can interpret results without technical knowledge.

IV. Results

Across multiple simulation tests and validation cycles, the system generated several clear insights. First, the total financial opportunity quantified across all topics yielded a potential net impact of approximately $99.83 million and a global value-to-cost ratio (ROI) of 1.62:1.Second, dissatisfaction tends to concentrate around a small number of recurring product failure themes. Identifying these high-frequency topics early enables companies to channel resources effectively rather than addressing a scattered set of minor issues.

Incorporating cost-of-fix values significantly reshaped complaint rankings. Some issues that appeared critical - such as major reformulations - dropped in priority due to high production costs. Meanwhile, smaller issues like packaging adjustments rose to the top because they required little investment yet produced noticeable improvements in customer satisfaction.

Formulation-related complaints consistently delivered the highest projected ROI. Specifically, the topic ‘acne pimple scar cystic’ demonstrated the highest investment efficiency, with an ROI exceeding 11:1 and a substantial net impact of over $12 million. Issues such as irritation, longevity failure, and texture problems produced strong emotional responses from customers and directly affected repeat-purchase behavior. Fixing these issues offered substantial long-term financial benefits.

Logistics-related issues, while annoying to customers, typically impacted a smaller share of the user base and therefore generated lower overall return.

Evaluation efforts confirmed the system’s reliability and usability. Running BERTopic with different embeddings showed stable clusters, validating the consistency of topic modeling. Sentiment scores aligned with star-rating trends, demonstrating basic validity. Sensitivity testing showed that small adjustments in cost assumptions produced logical shifts in priority rankings without destabilizing the system.

Finally, usability assessments with peers and the instructor demonstrated that dashboard users could identify top issues and recommended actions within seconds - one of the core goals of the project.

V. Contributions

Mariam

- Focused on code compilation as well as making adjustments to ensure our model runs smoothly.
- Developed the concept of beauty product analysis.
- Assisted in review filtering logic, sentiment assignment, and dataset simulation.
- Assisted in the presentation

Nada

- Researched multiple datasets suitable for the project as well as the criteria within them.
- Assisted with code implementation
- Focused on report concepts such as project introduction, definition, and methodologies used.
- Designed and implemented the presentation to reflect the use case of our model.