# Semester Overview

*Goals: (1) To make review database more robust (2) To add product feature information to the database*

* [Figma board](https://www.figma.com/board/bVOxLt6SvVk4A4G8AJO5ks/Derrick-F24-Plan?node-id=0-1&t=NXdArmj4g1afeK5v-1) with overview figure
* Existing [review database](https://amazon-reviews-2023.github.io/) from UCSD
* ~~Ask for OpenAI playground login~~

| **Week** | **Progress** | **Actual** |
| --- | --- | --- |
| 9/16 | * Make table with current product reviews (fill in below) * Is it possible to make a jupyter notebook with the full pipeline of code? * Randomize selection of reviews (BIFMA) |  |
| 9/23 | * Choose products based off of category * Extract reviews based off stars and other criteria (in Figma) |  |
| 9/30 | * Consider how to cohesively extract product features in a clean way * Run a small trial comparing performance of BERT to GPT 4.0 * Google Collab |  |
| 10/7 | * Try with GPT (sentiment analysis, affordances, and product extraction) * Check in on product features and dataset * Extracting product affordances * Make list of products being chosen   + 1) Identifying domains being analyzed   + 2) Nicole will run human study on extracting features |  |
| 10/14 | * Look for examples of how other people use GPT for sentiment analysis (esp. for customer reviews) and borrow their instructions * Test using API for features/affordances |  |
| 10/21 | * Prompt without examples * Add all features, affordances, and sentiment by end of next week * Comparing GPT vs BERT |  |
| 10/28 | **Goal**: finish database by end of this week |  |
| 11/4 | * Execute topic modeling using LDA, BERTopic, and GPT 4o on reviews * Try NMF |  |
| 11/11 | * Start comparing product features + affordances to topics |  |
| 11/18 | Finish ABSA  NMF and LDA on reviews that mentioned sustainability for product category  Chart of reviews mentioning price vs other sustainability keywords for product category |  |
| 11/25 | * Combine charts across the same product category |  |
| 12/2 |  |  |

# Current Reviews (as of 9/16/24)

| **Certification** | **Certification Description** | **Number of reviews scraped** | **Review selection method (Most recent, random, etc.)** | **Notes** |
| --- | --- | --- | --- | --- |
| RCS 100 (Recycled Claim Standard 100) | The product must contain at least 95-100% recycled material | 8234 | Most Recent 100 | Personal Item: Textile (Clothes and Curtains), Jewelry, Phone Cases |
| FSC (Forest Stewardship Council) | Forests are managed to be environmentally appropriate, socially beneficial, and economically viable. | 6800 | Most Recent 100 | Forest Products: Wood, paper, furniture |
| FFL (Fair for Life) | Certification Programmes for Fair Trade within responsible supply chains and Corporate Social Responsibility. | 6463 | Most Recent 100 | Household Items: Food, personal care items, agricultural products |
| BA (Blue Angel) | The entire life cycle of the product is taken into account and any impacts on the environment and health are fully considered during the development of the criteria ("multi-criteria evaluation"). | 4274 | Most Recent 100 | Office Products: Office equipment, appliances, building materials, paper products |
| BIFMA (Business and Institutional Furniture Manufacturers) | The development and maintenance of the safety, performance, and sustainability standards (packaging, energy optimization, environment impact, bio-based reduction)for furniture. | 3351 | Most Recent 100 | Furniture |

Google Drive:

<https://drive.google.com/drive/u/1/folders/1BLzW0bJO0kHZtt5N8C62ca2G42v3bkoU>

~~Jupyter Notebook:~~ [~~https://jupyter.org/try-jupyter/tree/~~](https://jupyter.org/try-jupyter/tree/)

~~Amazon Product Scraper:~~ [~~https://jupyter.org/try-jupyter/notebooks/index.html?path=Amazon+Product+Scraper.ipynb~~](https://jupyter.org/try-jupyter/notebooks/index.html?path=Amazon+Product+Scraper.ipynb)

~~Amazon Review Scraper:~~ [~~https://jupyter.org/try-jupyter/notebooks/index.html?path=Amazon+Review+Scraper.ipynb~~](https://jupyter.org/try-jupyter/notebooks/index.html?path=Amazon+Review+Scraper.ipynb)

~~ABSA BIFMA:~~ [~~https://jupyter.org/try-jupyter/notebooks/index.html?path=ABSA\_BIFMA.ipynb~~](https://jupyter.org/try-jupyter/notebooks/index.html?path=ABSA_BIFMA.ipynb)

Google Collab: <https://drive.google.com/drive/folders/1vJKRp35y9BR5Q0PZmUKwVOik6-rw6l8w>

Amazon Product Scraper: <https://colab.research.google.com/drive/1MB__cckXirTisgok08Rht40NBnj0MOIy#scrollTo=TgXfJCXMricJ>

Amazon Review Scraper: <https://colab.research.google.com/drive/1TOjgMWDkn8aJpSREimPAdPB6-Goj3t_p>

Amazon Merger 1: <https://colab.research.google.com/drive/1TsnXoO0y6KHvplnCJTL8srU7yrSie-9U>

Amazon Merger 2: <https://colab.research.google.com/drive/1gmnN3xtP38iuLoY5I1ann766cx7jvhv1?usp=sharing>

Amazon Filter: <https://colab.research.google.com/drive/1sBwirz7wTQgC6-CqlKtmJbe14zJAq56X?usp=sharing>

Product Description Analyzer: <https://colab.research.google.com/drive/1TeC4y3NVet23r1mx1E7AZAdmFrFTIrto#scrollTo=CBsavCqx8umx>

Sentiment Analysis with BERT: <https://colab.research.google.com/drive/1p3nf0-KXuiGPaoo6OJ94j60Yvrp8DDhP>

Sentiment Analysis with GPT:<https://colab.research.google.com/drive/1d3TfzFs7JH9-PIM201j3pU_hhWbVAe1O?usp=sharing>

Topic Modelling with GPT:<https://colab.research.google.com/drive/1Smljozt7j6tyFaSmNGVhHx-yfaOZosF0?usp=sharing>

Topic Modelling with LDA: <https://colab.research.google.com/drive/1XjV0Ntt8kywzAH0lpLfByJKIfj9tv3p-?usp=sharing>

Topic Modelling with BERTopic:<https://colab.research.google.com/drive/1xZt4bPvTQDTNNzvS5zl0zCee5X4DJwfA?usp=sharing>

Topic Modelling with NMF: <https://colab.research.google.com/drive/10549WGxJJLJc3ThBq1z5-FpfQtiYXKax?usp=sharing>

ABSA with GPT: <https://colab.research.google.com/drive/1AR0hZMN0WKYsmruomn_tZKeP8P2nJ4Ef?usp=sharing>

Sustainability Info Summary (Per Cert & Product Category): <https://colab.research.google.com/drive/1ClN03fwHMQGgLJuiwXn1RTgTCFU_P6VW?usp=sharing>

Sustainability Info Summary (Per Product Category): <https://colab.research.google.com/drive/1-wX1LbChULBIL9in2bWJV62bPUmQbipo?usp=sharing>

| **Methodology** | **Key Features** | **Applications** | **Strengths** | **Weaknesses** |
| --- | --- | --- | --- | --- |
| **Latent Dirichlet Allocation (LDA)** | - Probabilistic generative model  - Bag-of-words assumption | - Large datasets with longer documents  - Broad topic identification | - Interpretable topics  - Extensive library support | - Requires pre-specifying topics  - Struggles with short texts  - Ignores context |
| **BERTopic** | - Uses transformer embeddings  - Clustering with UMAP & HDBSCAN | - Short texts  - Nuanced customer feedback analysis | - Captures semantic nuances  - Auto-determines topics  - Handles overlapping topics | - Computationally intensive  - High resource requirements |
| **ChatGPT & Large Language Models (LLMs)** | - Deep learning transformer models  - Generates human-like text | - Summarizing customer feedback  - Exploratory data analysis | - Understands context and nuance  - Flexible across tasks  - Coherent summaries | - Less structured topic groupings  - Computationally expensive  - Risk of inaccuracies |
| Non-negative Matrix Factorization (NMF) | - Decomposes term-document matrix  - Additive topic combinations | - Parts-based text representations | - Interpretable components  - Works with sparse data | - Requires parameter tuning  - Less robust for complex topics |
| Probabilistic Latent Semantic Analysis (pLSA) | - Models words as samples from mixture models  - Probabilistic topics | - Clustering  - Dimensionality reduction | - Handles synonymy and polysemy | - Overfitting issues  - Can't assign probabilities to new docs |
| Structural Topic Modeling (STM) | - Incorporates document metadata  - Topics vary with metadata | - Analyzing topic correlation with external variables | - Contextual insights  - Topic dynamics over metadata | - Complex and resource-intensive  - Needs relevant metadata |
| Top2Vec | - Generates embeddings (Doc2Vec)  - Identifies dense topic areas | - Automatic topic discovery  - Large datasets | - Captures semantic relationships  - No pre-set topic number | - Embedding quality dependent  - High computational resources |
| Correlation Explanation (CorEx) | - Maximizes total correlation  - Non-probabilistic model | - Discovering informative, coherent topics | - Effective with short texts  - Incorporates domain knowledge | - Less interpretable foundation  - Scalability issues |