Final Report

Matthew Gauden

MSDS 453 - Section 56

8/30/2025

## Introduction & Problem Statement

The goal of this final project was to build a recommender system in Python using NLP methods and practices. The dataset that will be used is Amazon product reviews data. The dataset contains 11,399 unique products that have been reviewed by actual Amazon customers.

This project aims to fulfill a business need that is more prevalent in today's age than ever before. As the economy becomes increasingly focused on e-commerce, consumer satisfaction and personalization has been the primary goal of a lot of online retailers. Authors Jiayin Lin, Geng Sun, Ghassan Beydoun, and Li Li emphasize this message in their article Applying Machine Translation and Language Modelling Strategies for the Recommendation Task of Micro Learning Service. The authors write: "Hence, filtering out irrelevant information and picking the one that matches the learner's learning requirement is the key to such a personalized online learning service. Especially for the online service/application that deploys in the context of big data, a sophisticated recommender system is a key factor to guarantee efficiency and personalization" (2022).

Recommender systems, especially those based in NLP methods, are able to extract and analyze user feedback and tailor suggestions for purchases based on text similarity. The crux of the project is content-based filtering. In their same paper listed above, the authors identify three key types of recommender strategies: collaborative filtering, content-based filtering, and hybrid recommending strategies (2022). This project aims to apply content-based filtering by using text embeddings and similarity search to retrieve products that are semantically similar based on content features.

## Research Design and Modeling Methods

Data Processing

The first step before modeling could take place was data wrangling. This included inspecting the data itself and understanding the scale, and the nuances. There were initially two datasets. The first was Amazon metadata. This data contained information on the products themselves, and I deemed it unnecessary for this report. The second dataset contained the actual Amazon reviews from real users who bought and used the products. This was the dataset of focus for the experiments.

To build the dataframe that would be used in the following models, filtering of the data was required. Columns such as "ReviewerName", "unixReviewTime", etc. were deemed irrelevant. These columns were dropped, and the focus of the dataframe was "asin" (product id), "reviewText" (user reviews), and "title" (product name). The title of the product was contained in the metadataset. A merge of the two dataframes was required in order to display the title of the product alongside the review.

FAISS

FAISS stands for Facebook AI Similarity Search. It is an open-source library that is often combined with machine-learning models. Its primary objective is to search for similar items in a large collection of vectors.

BERT

A BERT-based sentence transformer was used in order to convert the product reviews into vector embeddings. BERT, as seen in previous assignments, is able to capture contextual meaning in bodies of text. Each review in the dataset is analyzed by BERT to produce an embedding which is able to represent the semantic meaning of the reviews. This has large advantages in user recommendation systems because it focuses on user meaning rather than wording. I think it is vital to account for human-error, and it is plausible that a user could make a grammatical mistake, or use the "wrong" word for what they are searching for. However, BERT is able to avoid this being an issue because it is trained on meaning rather than the specific words themselves.

t-SNE

The final step in the pipeline is to simply plot a visual representation of semantic search results. T-SNE is chosen due to its ability to plot higher-dimension embeddings – such as BERT – into a reduced 2D space. It is also easy to discern clusters of similarity along with outliers.

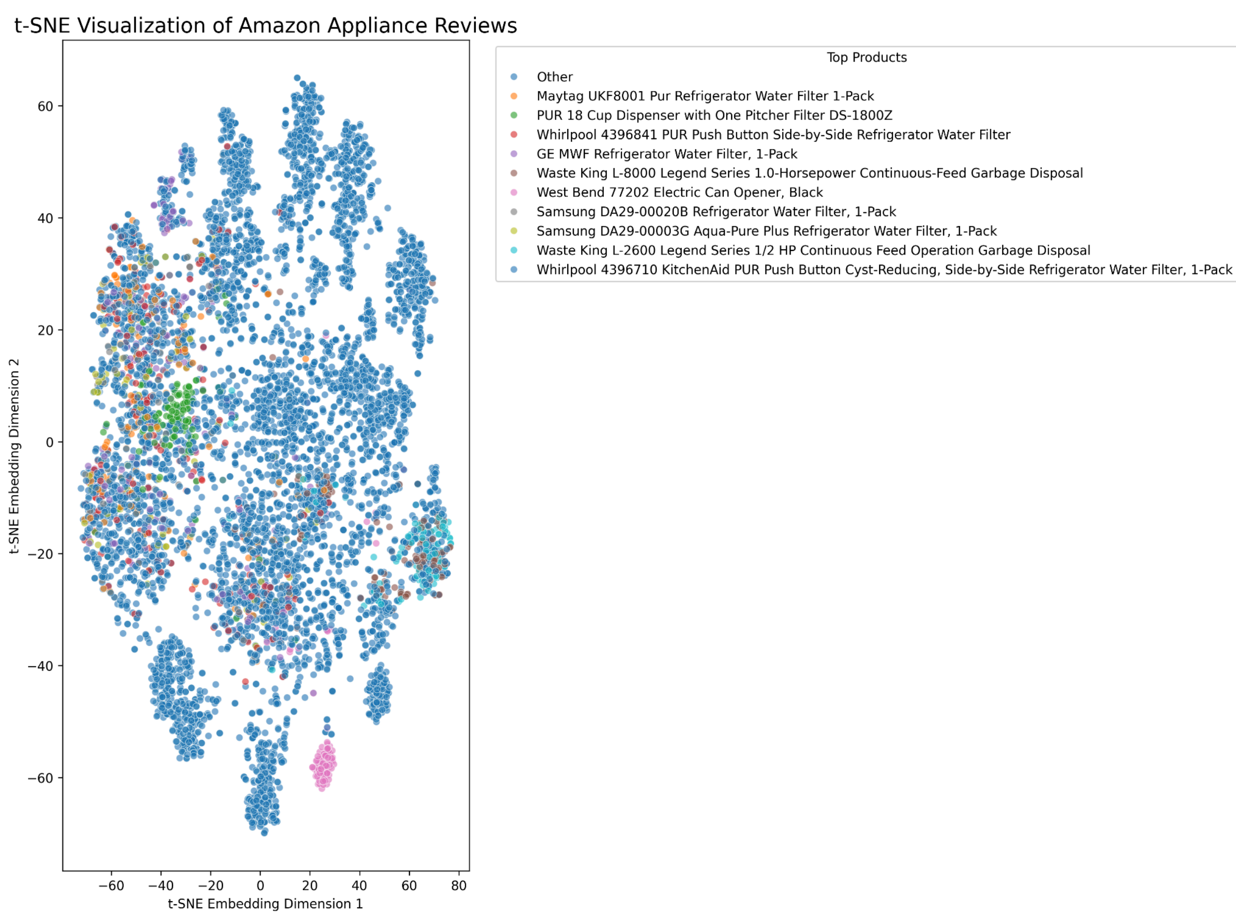
## **Results**

The FAISS library specializes in fast nearest-neighbor search. The library builds its own index that is able to perform quick vector similarity searches. The results of using this library tool is seen in the test query below. Instead of comparing the query against every single review, it just searches its aforementioned index to find the top 5 similar results.

```
# test with a query
query = "Best heavy-duty washing machine"
query_vec = model.encode([query], normalize_embeddings=True)
query_vec_pca = pca.transform(query_vec)

# Search top 5 reviews
D, I = index.search(query_vec_pca, k=5)
results = df.iloc[I[0]][['title', 'reviewText']]
for idx, row in results.iterrows():
  print(f"Title: {row['title']}")
  print(f"Review: {row['reviewText']}")
  print()
```

```
Title: Whirlpool WTW5500XW 27 3.6 cu. Ft. Top-Load Washer - White
Review: We purchased this machine in July 2010 and I've been struggling to stick with it ever since.  Washing small loads of laundry have been okay.  But I have

Title: LG WT5070 4.7 Cu. Ft. Ultra-Large Capacity High Efficiency Top Load Washer with WaveForce, White
Review: We've only had this machine for a short while but I've done many loads. Some of these loads were all pants (ie: long pieces that could twist) and some we

Title: Speed Queen AWN542S 26'' Top-Load Washer 3.3 cu. ft. Capacity
Review: After trying to give Whirlpool a chance and using their new low-water cabrio system for a while, I found that just wasn't cutting it.Sure, if you only ha

Title: Maytag MVWB750WQ 28 5 cu. Ft. Top-Loader Washer - White
Review: If I could marry this washing machine I would!  Not a week goes by that I do not say, "I love my new washing machine".  I did extensive research before b

Title: Whirlpool Washing Machine - Complete Tub Assembly
Review: We were quite disgusted when our Whirlpool Duet washing machine gave out after six years.  The bearings were not the only thing to go, after we disassemb
```

The next results come from a t-SNE visualization:




In the visualization, each dot represents a single review from the data. Due to there being over 11,000 unique products, the top 10 most-reviewed products are given colors to differentiate them. If a product is not in the top 10 most-reviewed, then it is labeled with a blue "Other" dot. Since t-SNE is a two-dimensional space, reviews that are deemed semantically similar by the model are clustered together.

## Analysis

### BERT

One of the issues that could arise from this experiment is that of scaling issues. It was mentioned in the Introduction, that the dataset containing Amazon product reviews was quite large. There were 11,399 unique products reviewed by users, and the dataframe consisted of around 150,000 reviews. In experimentation with BERT, various RAM issues were present. A downside of BERT, and using transformers in general, is that it is computationally expensive, so batch processing was used to encode the reviews. Multiple batch size parameters were used and experimented with. Initially, a batch size of 128 was used. This proved to be computationally inefficient for the system on which the experiment was being run. After 30 minutes of execution time, the batch processing was only around 11% complete.

After this experiment, a batch size of 64 was used and was still running inefficiently. This led to the conclusion that the data itself was too large. A subset of the data was then taken from the initial dataset. Instead of 143,630 reviews, a subset of 7,142 reviews were used in the development of this project. Then, a batch size of 64 ran much more smoothly and a lot quicker than previously. ChatGPT was used to troubleshoot this problem, and it was extremely beneficial to see how changing the batch size, reducing the dimensions with Principal Component Analysis (PCA), and then fitting those reduced dimensions to our embeddings could help with RAM and efficiency obstacles.

### t-SNE

The t-SNE visualization tells us about similar themes as it relates to the products themselves. The pink cluster near the bottom right is the tightest cluster out of the top 10 products shown in the legend. The pink cluster is for the West Bend electric can opener. What this tells us is that nearly all of the reviews for that specific can opener are written using highly similar language from users who reviewed it. That cluster itself also stands at a distance from the clusters around it. This leads me to believe that the language used by reviewers for that product is much different than the other products in the plot, since there is such distance between the can opener and other types of products.

There are also concentrated clusters for the Waste King Garbage Disposal, and PUR 18 Cup Dispenser. Like the pink cluster for the West Bend can opener, this shows that those specific products have great similarity among their documents.

A fascinating insight to add to this analysis is the evolution of technology when it comes to NLP practices. In the book Product Recommendations in E-Commerce Retailing Applications by Nicholas Knotzer, it is written “Suggestions of items may be supplemented by text comments. Because text comments are not completely machine--understandable, many e-vendors require the user to give an additional numerical rating to indicate the direction of the comment (i.e. pro or against the item)” (2018). This book was published in 2018, a relatively short time ago. We can see that at that time text comments were seen as supplemental, and not sufficient for recommender systems. Thanks to evolving work being done in NLP and machine-learning, we can now use tools such as BERT to extract semantic meaning from text comments.

## Conclusions

In the Introduction, it was pointed out how the FAISS library, in conjunction with BERT, can be a novel way to build a recommender system. A user is able to query what they’re looking for, and the tools can recommend the top 5 – or more if so desired – products available to purchase, ranked by most-reviewed. This project tackles the current challenge of online retailers to ever-increase the personalization of their products using foundational NLP methods such as converting text into vector embeddings (BERT), and using t-SNE to visualize higher-dimensional data and semantic similarity.

## Future Work

There are many possible avenues that one could take to improve, or expand on the work performed in this project. One of those possibilities is to incorporate user ratings data for the recommender system to take into account. In the original Amazon product reviews metadata, there is a column that indeed contains user rating data for the products. The users were asked to give an overall score of the product, from 1 to 5. This can be used by the recommender system to make recommendations based on similar scores, or products of the same score.

In the article, Recommendation Systems with Purchase Data, it is noted that this method can produce unwanted complexities, but that there may be a work around. The article states, “In many contexts, this can be unattractive because getting good ratings data entails a proper survey mechanism that can collect such data accurately and efficiently” (Bodapati, 2008). The article highlights this proposes logistical problems such as a resistance from customers to participate in said surveys.

However, the article mentions that it could be useful to use binary data instead of a score on a given scale. An example of binary user data would be a simple “thumbs up, thumbs down” feature. Bodapti then points out, “A commonly advanced argument is that thumbs down does not require survey mechanisms but can be inferred from just purchase behavior” (2008).

Given this background, future work could see BERT’s binary classification capabilities be used to transform user sentiment in the review text into a binary positive or negative score. This is similar to what we saw from BERT in Assignment 2 in this course.

References

Bodapati, A. V. (2008). Recommendation Systems with Purchase Data. *Journal of Marketing Research*, *45*(1), 77–93. https://doi.org/10.1509/jmkr.45.1.77

Knotzer, N. (2018). *Product Recommendations in E-Commerce Retailing Applications*.

Lin, J., Sun, G., Beydoun, G., & Li, L. (2022). Applying Machine Translation and Language Modelling Strategies for the Recommendation Task of Micro Learning Service. *Educational Technology & Society*, *25*(1), 205–212. JSTOR. https://doi.org/10.2307/48647041