



# **Amazon Electronics Troubleshooting Retrieval System**

## **<Documentation>**

CS 410 – Text Information Systems

Team Member: Fok Chun Chung (ccfok2)

Project Coordinator: Fok Chun Chung (ccfok2)

*Repository: [jasonccfok/amazon-electronics-troubleshooting-retrieval](https://github.com/jasonccfok/amazon-electronics-troubleshooting-retrieval)*

*Project Keywords: #electronics, #retrieval-system, #troubleshooting, #amazon-reviews, #semantic-search*

## Table of Contents

1.	Project Overview .....	1
2.	Motivation .....	1
3.	Technical Approach.....	1
4.	Evaluation .....	2
5.	System Overview .....	2
6.	Project Structure .....	3
7.	Setup .....	4
8.	Usage Guide .....	5
9.	Technologies .....	9

# 1. Project Overview

The Amazon Electronics Troubleshooting Retrieval System is designed to assist users in finding community-sourced solutions hidden within the vast collection of Amazon product reviews for electronic devices. Instead of wading through hundreds of unrelated reviews or generic troubleshooting pages, users can describe their specific issue in natural language — such as:

- *“Bluetooth keeps disconnecting”*
- *“battery drains too quickly”*
- *“screen flickering after update”*

The system intelligently analyzes the input and retrieves reviews where similar problems have been discussed by other customers — often alongside helpful fixes, workarounds, or insights.

## 2. Motivation

When electronics malfunction, users often seek solutions online, but official documentation rarely addresses real-world issues. Amazon reviews are rich with troubleshooting knowledge, where users share problems and fixes in their own words. This project seeks to harness that valuable resource, making it easily searchable to help users quickly determine if others have faced similar issues and how they resolved them.

## 3. Technical Approach

The retrieval pipeline integrates two complementary approaches:

### A. Keyword-based retrieval (BM25 / TF-IDF)

Implements traditional term-frequency models to find exact or close keyword matches.

### B. Semantic retrieval (Sentence-BERT embeddings)

Uses dense vector embeddings (e.g., all-MiniLM-L6-v2) to capture meaning similarity even when wording differs.

## 4. Evaluation

Model effectiveness is assessed using common IR metrics implemented in the repository:

<i>Metric</i>	<i>Description</i>
Precision@K	Fraction of top K retrieved documents that are relevant
Recall@K	Fraction of all relevant documents retrieved
Mean Average Precision (MAP)	Average precision over recall positions for each query

## 5. System Overview

Raw JSONL Reviews + Metadata



[1] Preprocessing → Cleaned & merged text corpus



[2] Retrieval → BM25 / TF-IDF / Semantic models



[3] Evaluation → Precision@K, Recall@K, MAP

## 6. Project Structure

```
project_root/
├── 1. preprocessing/
│   ├── load_raw_data.py           # Load & merge reviews + metadata
│   ├── text_cleaning.py          # Clean, normalize, lemmatize text
│   └── exploratory_data_analysis.ipynb # EDA notebooks (visual insights)
├── 2. retrieval/
│   ├── bm25_retrieval.py         # BM25 and TF-IDF retrieval
│   └── semantic_retrieval.py     # Transformer-based semantic model
├── 3. evaluation/
│   ├── retrieval_evaluation.py    # Benchmark BM25, TF-IDF, semantic models
│   └── evaluation_plot.ipynb    # Visualization of evaluation metrics
├── data/
│   ├── raw/                     # Amazon JSONL input files
│   ├── processed/               # Cleaned CSVs & cached embeddings
│   └── groundtruth/              # Gold relevance + evaluation outputs
├── requirements.txt              # Library dependencies
└── README.md                    # Project description
```

## 7. Setup

### a. Clone and Enter Folder

```
git clone https://github.com/jasonccfok/amazon-electronics-troubleshooting-retrieval.git
cd amazon-electronics-troubleshooting-retrieval
```

### b. Create Virtual Environment

```
python-m venv venv
source venv/bin/activate    # macOS/Linux
venv\Scripts\activate      # Windows
```

### c. Install Dependencies

```
pip install-r requirements.txt
```

### d. Prepare Data

Category	#User	#Item	#Rating	#R_Token	#M_Token	Download
Cell_Phones_and_Accessories	11.6M	1.3M	20.8M	935.4M	1.3B	<a href="#">review, meta</a>
Clothing-Shoes_and_Jewelry	22.6M	7.2M	66.0M	2.6B	5.9B	<a href="#">review, meta</a>
Digital_Music	101.0K	70.5K	130.4K	11.4M	22.3M	<a href="#">review, meta</a>
Electronics	18.3M	1.6M	43.9M	2.7B	1.7B	<a href="#">review, meta</a>
Gift_Cards	132.7K	1.1K	152.4K	3.6M	630.0K	<a href="#">review, meta</a>
Grocery_and_Gourmet_Food	7.0M	603.2K	14.3M	579.5M	462.8M	<a href="#">review, meta</a>

Download from <https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023> and place Amazon Electronics data in:

```
data/raw/Electronics.jsonl.gz
data/raw/meta_Electronics.jsonl.gz
```

## 8. Usage Guide

### a. Step 1 – Load and Merge Raw Data

```
python "1. preprocessing/load_raw_data.py" ^  
--review "data/raw/Electronics.jsonl.gz" ^  
--meta "data/raw/meta_Electronics.jsonl.gz" ^  
--output "data/processed/electronics_merged.csv" ^  
--limit 50000
```

Output → data/processed/electronics\_merged.csv

### b. Step 2 – Clean and Normalize Text

```
python "1. preprocessing/text_cleaning.py" ^  
--input "data/processed/electronics_merged.csv" ^  
--output "data/processed/reviews_clean.csv"
```

### c. Step 3 – BM25 / TF-IDF Retrieval

```
python "2. retrieval/bm25_retrieval.py" ^  
--data "data/processed/reviews_clean.csv" ^  
--query "tripod screw loose problem" ^  
--topk 5
```

Sample terminal output:

```

===== QUERY RESULTS =====
Query: tripod screw loose problem

----- BM25 Top Results -----
1. parent_asin: B0BZB56D2J | Score: 16.0709
   Review (raw): Loose screw inside, Bulky but very useful! The design is very bulky but it's very useful feature to clamp into place. quality feels nice. Mine has 1 problem which does worries me a little: when I shake it a little you can tell there is a screw loose inside. this is a potential cause of a short circuit. But because once is set in place I don't move it anymore it doesn't worry me as much. Tripp Lite 3 Outlet Surge Protector Power Strip with Desk Clamp, 10ft. Cord, 510 Joules, 2 USB Charging Ports, Black, $20K Insurance & (TLP310USBC)

2. parent_asin: B01F5SIVBQ | Score: 15.5608
   Review (raw): Good but could be improved It doesn't come with any sort of quick release for the attachment of the camera, so very time you want to use it you have to screw the camera into the entire tripod. I believe they sell an attachment that you can use for quick release, but I wish it was integrated. The tripod itself is really easy to use and holds my dslr sturdily with no problem. The height is not really adjustable so it's not as flexible that way. Tamrac TR404 ZipShot® Mini Tripod - Black

3. parent_asin: B00430AI9M | Score: 15.3906
   Review (raw): Heavy duty! This tripod is heavy duty and very sturdy. The only thing that wasn't mentioned in any reviews, is the fact that the tripod sits at an angle and there isn't anything you can do about it. If you want your antenna to point straight ahead, like I did, then you got a problem. I will just get some longer screws to adjust the 5' mast that I ordered, because the one that came with it, same shipment I might add, was too short. Like I said, it's heavy duty, easily mounts to the roof, doesn't come with pitch pads or screws, but all in all it serves its purpose. Four stars, only for the fact it doesn't sit straight up and down, rather at an angle. 3 feet Satellite Tripod Mount with 2-Inch OD Mast

----- TF-IDF Top Results -----
1. parent_asin: B0BZB56D2J | Score: 0.1722
   Review (raw): Loose screw inside, Bulky but very useful! The design is very bulky but it's very useful feature to clamp into place. quality feels nice. Mine has 1 problem which does worries me a little: when I shake it a little you can tell there is a screw loose inside. this is a potential cause of a short circuit. But because once is set in place I don't move it anymore it doesn't worry me as much. Tripp Lite 3 Outlet Surge Protector Power Strip with Desk Clamp, 10ft. Cord, 510 Joules, 2 USB Charging Ports, Black, $20K Insurance & (TLP310USBC)

2. parent_asin: B013FUZJ2U | Score: 0.1578
   Review (raw): Very Handy! Very handy to have these adapters! EXMAX® 17 in 1 Kit 1/4" 3/8" Threaded Male Female Screw Adapter D-Shaft D-Ring Tripod Screw Hot Shoe Mount Compatible with Canon Nikon Camera Flash Stand Bracket Holder Monopod

3. parent_asin: B07L3TN59S | Score: 0.1465
   Review (raw): Lensball suction cup for tripod heads Exactly as pictured in description..  
Would be nice to have a step up adapter for tripod legs included.. as is you will have to use tripod head. Suction seems adequate with original 80mm lensball. Would prefer suction cup to be just a bit bigger for increased grip insurance Original Lensball Stand | Suction Mount + Flat Base + Tripod Screw Thread

```

#### d. Step 4 – Semantic Retrieval (Transformer Embeddings)

```

python "2. retrieval/semantic_retrieval.py" ^
--data "data/processed/reviews_clean.csv" ^
--query "phone adapter screw problem" ^
--topk 5 ^
--model "all-MiniLM-L6-v2"

```

Sample terminal output:



```

===== QUERY RESULTS =====
Query: phone adapter screw problem
----- Semantic (SentenceTransformer) Top Results -----
1. parent_asin: B006C13X4Q | Score: 0.6141
   Review (raw): wont work with adapter I bought this as a cute gag gift for my dad for Christmas. He doesn't have an iPhone so I bought an adapter to use with it and it didn't work for him AT ALL! It was so upsetting on Christmas day. It did however work on the iPhone, just not on his with an adapter. SANODY Retro Cell Phone Handset
2. parent_asin: B07P94PB7Z | Score: 0.5454
   Review (raw): Adaptor VERY poorly constructed This adaptor simply does not fit an iPhone. Period. It is not useable. The plug that fits into the iPhone is too small so it is loose when inserted into the iPhone and falls right out. I am baffled that such a poor device would be sold by Amazon. All I can do now is throw it away. It is an absolute waste of money. USB 3 Camera Adapter, 3 in 1 USB Female OTG Adapter with Charging and 3.5mm Headphone Audio Jack Splitter for iPhone/iPad, Support USB Flash Drive, MIDI Keyboard
3. parent_asin: B078SQWBDY | Score: 0.5348
   Review (raw): Adapter fell apart The Adapter broke and fell right off in my computer I was so lucky I could pull it out. I was able to return for a full credit IPEAK USB C to USB Adapter with Keychain 4Pack, Type C Male to USB A Female OTG Connector for MacBook Pro 2017/2016 15&13inch MacBook 12inch Samsung Galaxy S9 S8 A8 Plus Note 8 Chrome Book Pixelbook

```

## e. Step 5 – Quantitative Evaluation of Models

Prepare a gold\_set.csv file such as:

<i>query</i>	<i>parent_asin</i>	<i>relevance</i>
bluetooth disconnecting	B001234ABC	1
bluetooth disconnecting	B00XYZ567	0

Run:

```

python "3. evaluation/retrieval_evaluation.py" ^
--data "data/processed/reviews_clean.csv" ^
--gold "data/groundtruth/gold_set.csv" ^
--topk 5 ^
--embeddings "data/processed/review_embeddings.pkl"

```

Sample terminal output:

```

=== SUMMARY RESULTS (avg across queries) ===
bm25_p@k    0.400000
bm25_r@k    0.139125
bm25_map    0.149821
tfidf_p@k   0.100000
tfidf_r@k   0.029380
tfidf_map   0.050740
sem_p@k     0.600000
sem_r@k     0.313692
sem_map     0.295665
dtype: float64
Detailed results saved → data/groundtruth/retrieval_evaluation_results.csv

```



## 9. Technologies

<i>Library</i>	<i>Role</i>
pandas, numpy	Data manipulation, math operations
nltk	Tokenization, stopwords lists, lemmatization
rank-bm25	BM25 ranking algorithm
scikit-learn	TF-IDF vectorization, cosine similarity, metrics
sentence-transformers	Transformer-based sentence embeddings
torch	GPU support for embedding computation
tqdm	Iteration progress monitoring
matplotlib, seaborn, wordcloud	Visualization support
faiss-cpu (optional)	Fast nearest-neighbor search for large embedding sets