IR Final Project

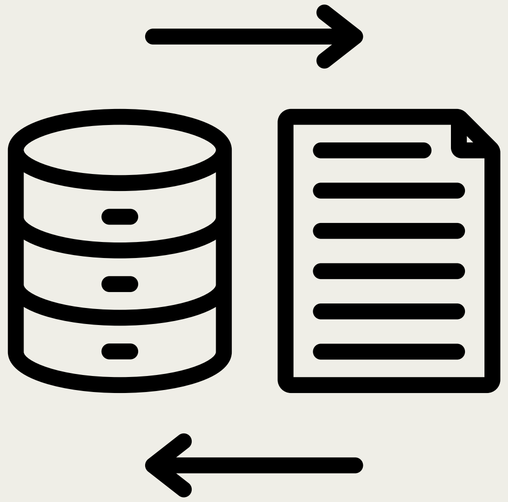
MARKETWATCH: E-COMMERCE COMPETITOR INTELLIGENCE SUITE

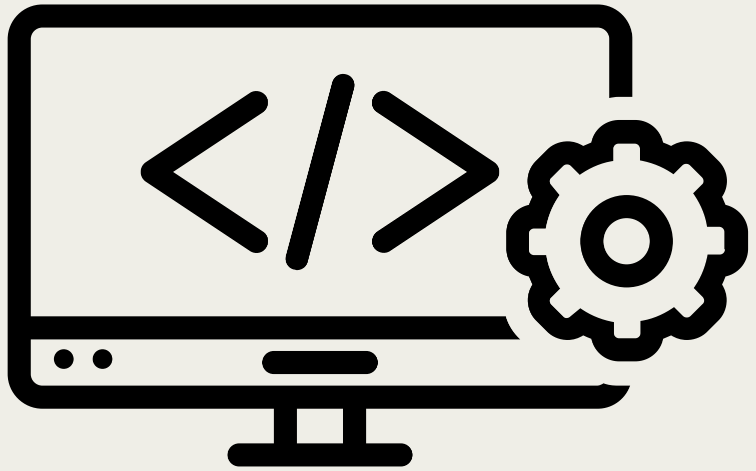
Members:
Aditya Nangia 2020168
Akshat Saini 2020018
Jayan Pahuja 2020071
Nakul Thureja 2020528
Saksham Bhupal 2020573

Problem Statement

The goal is to develop an automated system that can assist e-commerce sellers in optimizing their product listings and pricing strategies by leveraging web scraping, deep learning models, and information retrieval techniques to gather and analyze competitor data and customer feedback.

The system should automate the process of data collection through web scraping, perform feature extraction and analysis using deep learning models and techniques like TF-IDF, and provide sellers with valuable insights to make informed decisions regarding pricing, product listing optimization, and addressing customer needs.





Input and Output

Input:

- Seller’s product listing. This means it can be a product which the seller is selling on different platforms such as iPhone 15, Laptop Stands, etc.
- Product URL

Output:

- Insights and recommendations for optimal pricing strategies based on competitor pricing trends and customer price sensitivity.
- Analysis of customer feedback and sentiment to identify areas for product listing optimization
- Actionable insights to help sellers enhance their competitiveness in the e-commerce market
- A review bot which helps in providing suggestive improvements to the listings.
- Listing strength and performance prediction

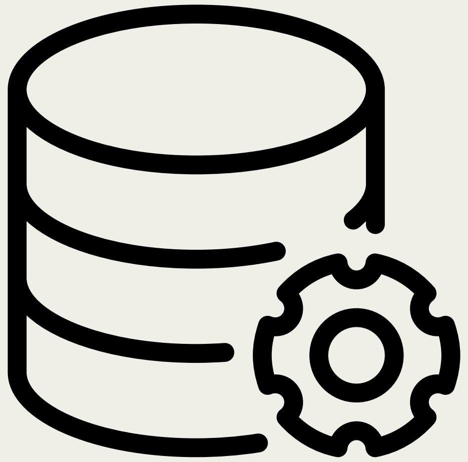
Motivation

In the realm of Information Retrieval, the prevailing emphasis predominantly revolves around consumer-centric solutions, leaving a noticeable gap in support for producers and sellers. Our objective is to address this by offering a tailored solution for sellers operating on platforms such as Amazon, Flipkart, and others. Sellers often face the arduous task of managing product listings across multiple platforms, rendering manual analysis inefficient and error-prone. The sheer magnitude of data available on e-commerce platforms underscores the necessity for further exploration of this issue.

In the fiercely competitive online market, comprehending market dynamics, evaluating competitor products, and formulating effective pricing and marketing strategies are paramount for success. Moreover, extracting meaningful insights from customer feedback and sentiments expressed in product reviews presents an additional challenge for sellers aiming to enhance their product offerings. A solution based on Information Retrieval is essential to aggregate, analyze, and present actionable intelligence derived from competitor data and customer reviews, empowering sellers to make informed decisions, enhance their offerings, and elevate their competitiveness.

Literature Review

Paper - 1: Increased information retrieval capabilities on e-commerce websites using scraping techniques



This study employs web crawling across three e-commerce websites, consolidating data into a database for streamlined retrieval. By scraping HTML tags and storing data systematically, the process ensures efficiency with a 100% recall rate and 93.9% precision rate, enabling rapid and accurate information retrieval. These techniques can be helpful and be further worked upon for our use case.

Paper - 2: A Review on Web Scrapping and its Applications

This review dives into web scraping, exploring its applications and the technology behind it. It examines the reasons why web scraping is valuable, discusses its advantages and limitations, and explores the tools and libraries commonly used for scraping tasks. Additionally, the review highlights the various applications of web scraping across different fields.

Literature Review

Paper - 3: Opinion mining and sentiment analysis on online customer review

Opinion mining, crucial in e-commerce, sees rising importance with the surge in online transactions and user-generated content. Reviews on platforms like Amazon express customer sentiments, offering valuable insights. This study focuses on mining Amazon reviews, utilizing algorithms like Naïve Bayes, Logistic Regression, and SentiWordNet for sentiment analysis. The goal is to automate sentiment recognition and enhance understanding of user emotions.

Paper - 4: Sentiment analysis: A literature review

This paper offers a survey of recent advancements in sentiment analysis, a technique that extracts emotional tones from text data. Due to the surge of subjective content online, particularly in reviews, sentiment analysis has become a hot research area. The paper delves into the core methods used in sentiment analysis research, including framework and lexicon development, feature extraction, and polarity classification (positive, negative, or neutral). It highlights the current methodologies, explores existing limitations, and provides an in-depth look at applications in business and blog analysis. Finally, the paper discusses potential future directions for sentiment analysis research.

Literature Review

Paper - 5: On Application of Learning to Rank for E-Commerce Search

E-commerce search is a burgeoning application of information retrieval, with Learning to Rank (LETOR) emerging as a pivotal strategy. While LETOR is extensively studied for web searches, its application to e-commerce searches remains unexplored. This paper addresses practical challenges in implementing LETOR for e-commerce search, including feature representation, obtaining reliable relevance judgments, and leveraging multiple user feedback signals. Experiments on industry datasets reveal insights: popularity-based features enhance relevance-based ones, reducing query attribute sparsity is beneficial, and order rate proves the most robust training objective, followed by click rate, while add-to-cart ratio is less reliable

Paper-6: Classification of Customer Reviews based on Sentiment Analysis

The paper proposes a system that performs the classification of customer reviews of hotels by means of a sentiment analysis. They extract a domain-specific lexicon of semantically relevant words based on a given corpus, which backs the sentiment analysis for generating a classification of the reviews.

Proposed Solution

The proposed solution is a comprehensive system that combines web scraping, deep learning, and information retrieval techniques to gather and analyze data from various e-commerce platforms. The system will automate the process of collecting competitor product listings, pricing data, and customer reviews through ethical web scraping practices. It will then leverage pre-trained deep learning models, such as ResNet, to extract visual features from product images and employ techniques like TF-IDF to analyze textual data, including product descriptions and customer reviews.

The extracted features will be utilized to identify the top competitors for the seller's products by computing similarity metrics. Additionally, the system will perform sentiment analysis on customer reviews using Natural Language Processing (NLP) tools like VADER Sentiment Analysis and NLTK. This analysis will provide insights into customer satisfaction, areas for improvement, and help optimize product listings. The consolidated competitor analysis will involve feature normalization and weighted scoring to combine data from multiple e-commerce platforms, ensuring comprehensive and relevant insights. Furthermore, the development of an AI chatbot is proposed to answer customer questions based on the analyzed review data, enhancing customer engagement and support.

Methodology

Step - 1: Web Scraping with Ethical Considerations:

a. Technology: Leverages libraries like Beautiful Soup to extract data from product pages on e-commerce websites.

b. Challenge Addressed: Automates data collection, eliminating manual work and ensuring access to the latest competitor information.

c. Ethical Considerations:

i. Respect robots.txt files that instruct web crawlers on restricted areas.

ii. Adhere to website terms of service to avoid copyright infringement.

iii. Be mindful of scraping frequency to avoid overloading website servers.

Step - 2: Feature Extraction for Detailed Competitor Comparison:

a. Technology:

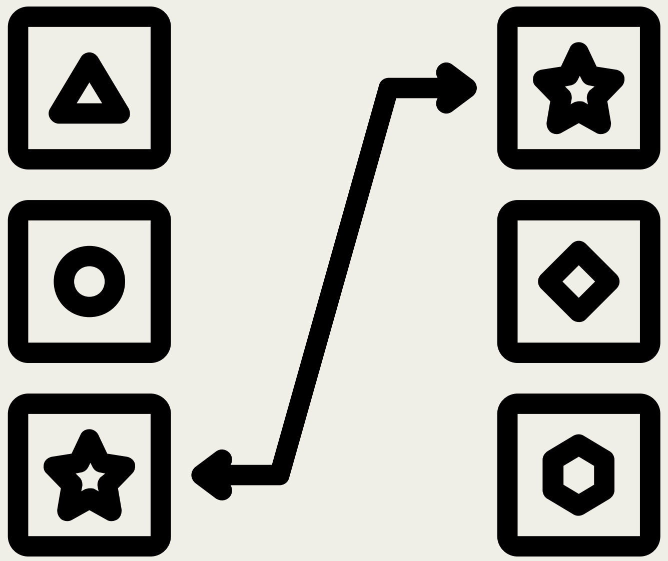
i. Image Features: Uses a pre-trained deep learning model like ResNet to extract visual features from competitor product images (e.g., color, shape, texture).

ii. Text Features: Employs TF-IDF (Term Frequency-Inverse Document Frequency) to analyse product descriptions and identify keywords that highlight product features and benefits.

b. Challenge Addressed: Enables a more nuanced comparison than just product titles. This allows sellers to identify similar products even with slight variations in names or descriptions

Methodology

Step - 3: Identifying Key Competitors Through Feature Matching:
a. Technology: Utilizes a similarity metric (e.g., cosine similarity) based on the extracted image and text features from the seller's product and competitor's products.
b. Challenge Addressed: Provides a targeted set of the most relevant competitors for analysis rather than an overwhelming list from the entire marketplace. This allows sellers to focus their energy on the most impactful competition.



Step - 4: Sentiment Analysis of Customer Reviews:
a. Technology: Employs Natural Language Processing (NLP) tools such as VADER Sentiment Analysis, NLTK, and Beautiful Soup to analyse customer reviews for the seller's product and identify positive, negative, or neutral sentiment.
b. Challenge Addressed: Provides sellers with insights into customer satisfaction and helps them understand areas for improvement in their own product or listing.

Methodology

Step - 5: Consolidated Competitive Analysis

a. Technology:

- Weighted Scoring: Develop a scoring system that assigns weights to data from different websites based on factors like website popularity, relevance to the seller's target audience, or trustworthiness of the source.

b. Challenges Addressed:

- Inconsistent Data Formats: By normalizing features extracted from different websites, the system can handle inconsistencies in data formats and ensure fair comparisons across platforms.
- Varying Website Relevance: The weighted scoring approach addresses the challenge of varying relevance and importance of different e-commerce platforms, allowing the system to prioritize data from more pertinent sources.

Step - 6: AI Chatbot for Customer Reviews:

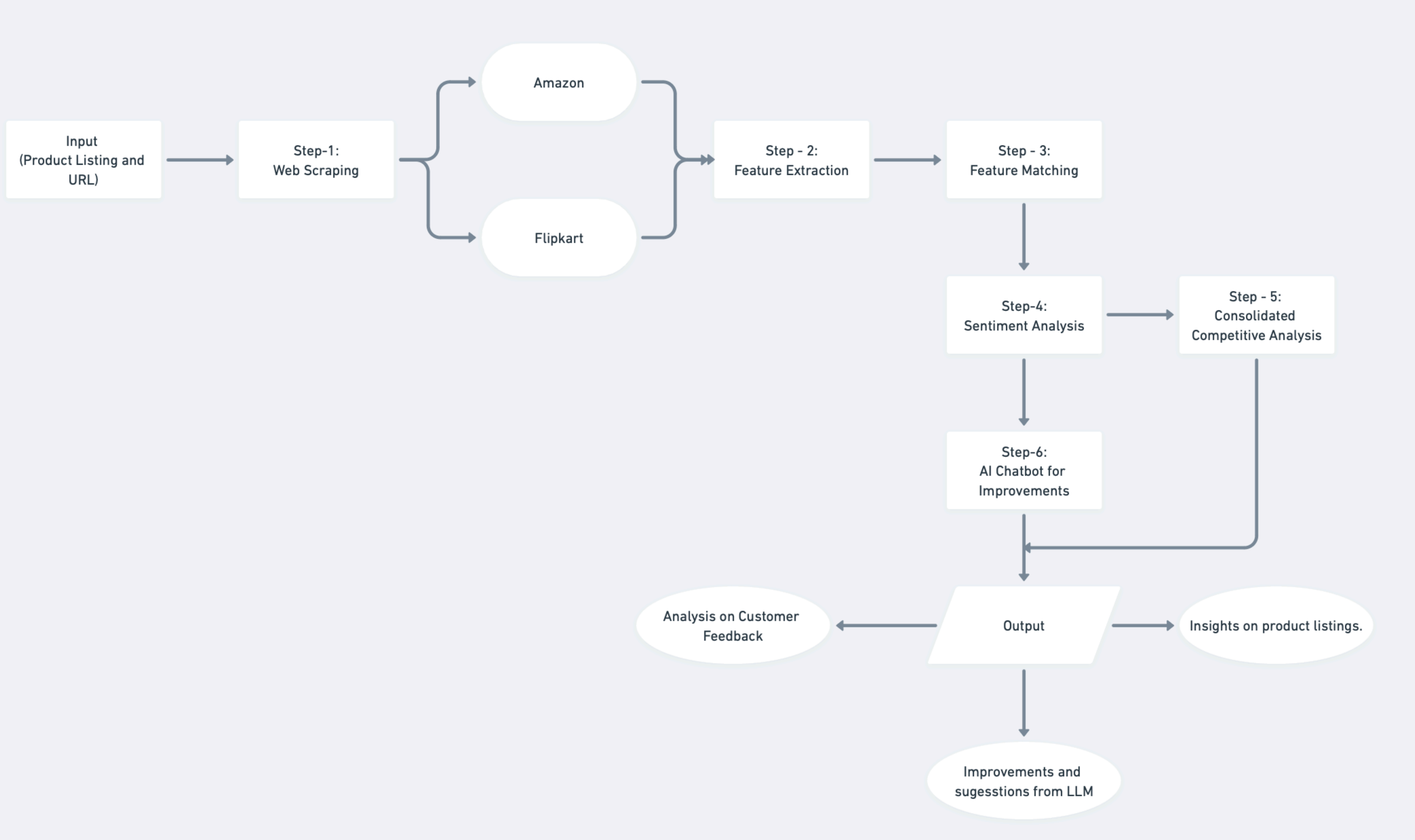
a. Technology:

- Answer Generation: Train the chatbot using natural language generation (NLG) models like GPT-3 to generate clear, concise, and informative answers to user questions using the retrieved information from customer reviews.

b. Challenges Addressed:

- Question Understanding: By employing question classification techniques, the chatbot can understand the context and intent behind user questions, enabling it to provide relevant and targeted responses.

Pipeline



Input
(Product Listing and URL)
Step-1:
Web Scraping
Amazon
Flipkart
Step - 2:
Feature Extraction
Step - 3:
Feature Matching
Step-4:
Sentiment Analysis
Step - 5:
Consolidated Competitive Analysis
Step-6:
AI Chatbot for Improvements
Analysis on Customer Feedback
Output
Insights on product listings.
Improvements and sugesstions from LLM

Novelty

In our initial endeavor to create a comprehensive "MultiModal Listing Strength Predictor," our primary goal was to engineer a model capable of evaluating the overall efficacy of a product listing within an e-commerce environment. We aimed to integrate various data modalities and features to construct a holistic assessment framework. By leveraging this predictor, we sought to provide e-commerce platforms with a powerful tool to gauge the effectiveness of their product listings, considering factors such as image quality, descriptive text, pricing, and other relevant attributes. Through this endeavor, we aimed to enhance the visibility and conversion rates of product listings, ultimately optimizing the shopping experience for both consumers and sellers in the online marketplace.

Evaluation

We employed two distinct evaluation metrics to assess the effectiveness of our system. The first metric gauges the performance of the sentiment analysis component, specifically its ability to accurately classify sentiment (e.g., positive, negative, neutral) within the data. The second metric evaluates the performance of the implemented similarity matching algorithm, focusing on its precision in identifying similar data points.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	0
2	1.00	0.89	0.94	18
accuracy			0.89	18
macro avg	0.50	0.44	0.47	18
weighted avg	1.00	0.89	0.94	18

Evaluation

Evaluating Sentiment Analysis:
We adopted a lexicon-based approach to evaluate the sentiment analysis algorithm's performance. Positive (2), neutral (1), and negative (0) labels were assigned as ground truth based on the average product rating on e-commerce websites. A five-point star rating system was employed, with ratings from 0 to 2.5 mapped to the negative label, 2.5 to 3.5 mapped to neutral, and 3.5 to 5 mapped to positive. For the predicted sentiment scores, a similar discretization strategy was implemented. Scores ranging from -1 to -0.25 were assigned the negative label, -0.25 to 0.25 were assigned neutral, and 0.25 to 1 were assigned positive. Subsequently, the performance evaluation mirrored a standard multi-class classification task. We computed precision, recall, F1-score, and accuracy to assess the model's effectiveness in classifying sentiment.

Evaluating Similarity Matching:
The evaluation of the similarity matching algorithm employed two distinct metrics tailored to the data type: ROUGE score for textual similarity and Root Mean Squared Error (RMSE) for image similarity. A combined similarity score was constructed by dividing the ROUGE score by the RMSE. This score served as the basis for retrieving the top 10 most similar products which served as the ground truth. The retrieved product set was then compared against set of top 10 most similar products generated by the algorithm. The number of matching products served as the evaluation metric for the overall effectiveness of the similarity matching process.

Evaluation

Listing Strength Predictor:

The performance of the listing strength prediction model was assessed using established machine learning evaluation metrics, including precision, recall, and F1-score. These metrics were calculated for both the averaging rating predictor and the average sentiment review predictor.

	precision	recall	f1-score	support
3	0.00	0.00	0.00	6
4	0.00	0.00	0.00	60
5	0.40	1.00	0.57	44
accuracy			0.40	110
macro avg	0.13	0.33	0.19	110
weighted avg	0.16	0.40	0.23	110

test Accuracy: 0.4
test Loss: 45.720574378967285

	precision	recall	f1-score	support
2	0.00	0.00	0.00	25
3	0.88	1.00	0.93	288
4	0.00	0.00	0.00	16
accuracy			0.88	329
macro avg	0.29	0.33	0.31	329
weighted avg	0.77	0.88	0.82	329

test Accuracy: 0.8753799392097265
test Loss: 85.35110169649124

Results

Relevant Demo and Screenshots for Website: