"Doesn't This Look Familiar?"

A Multimodal Approach to Finding Duplicate Products

Macy's Team: Rohan Talati, Albert Egido De Poy, Jay Malepati
Faculty Advisor: Professor Alexandre Jacquillat


Jeremy Michael


Yolanda Wang

Project Background

macy's has
- 519 stores across the USA and online E-commerce;
- Launched a third-party marketplace in 2022, contributing approximately **4.2%** to the total revenue of $24 billion in 2023

Problem Statement


Macy's E-commerce → **Macy's Products**, **Marketplace Products**

Macy's expansion into the Marketplace risks potential **revenue loss** from including third-party products that **closely resemble** existing Macy's products.

Our solution **identifies these overlaps**, ensuring a unique and optimized product selection.

→ **Increased Revenue**
Improved Customer Experience

Challenges

Product nuances:
- silhouette, color, pattern, key elements
- brand, price, size, material


vs.


Diverse categories:
Phase 1: Women's swimwear
Phase 2: Women's dresses
⋮
Phase N: **ALL** product types
↓ Style diversity

Attribute variability:
Duplicates vary in specifications, titles, and images, making simple matching unreliable

Data Limitation:
Only five labeled pairs are provided, insufficient for building a supervised classification model

Data Sources

1.8M	20%	425
Unique Products	from Marketplace	Product Categories

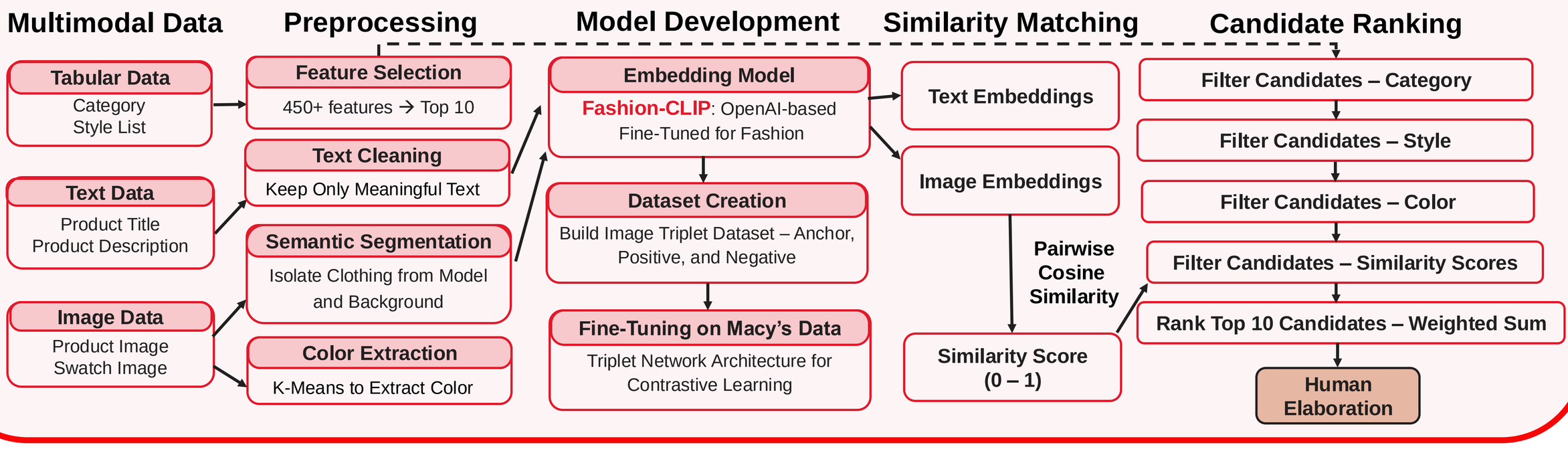
Tabular
Category, color, style pattern

Text
Product titles, descriptions

Image
Front view, close-up details

Miscellaneous
Inventory, fabric, occasion

Methodology



Model Results

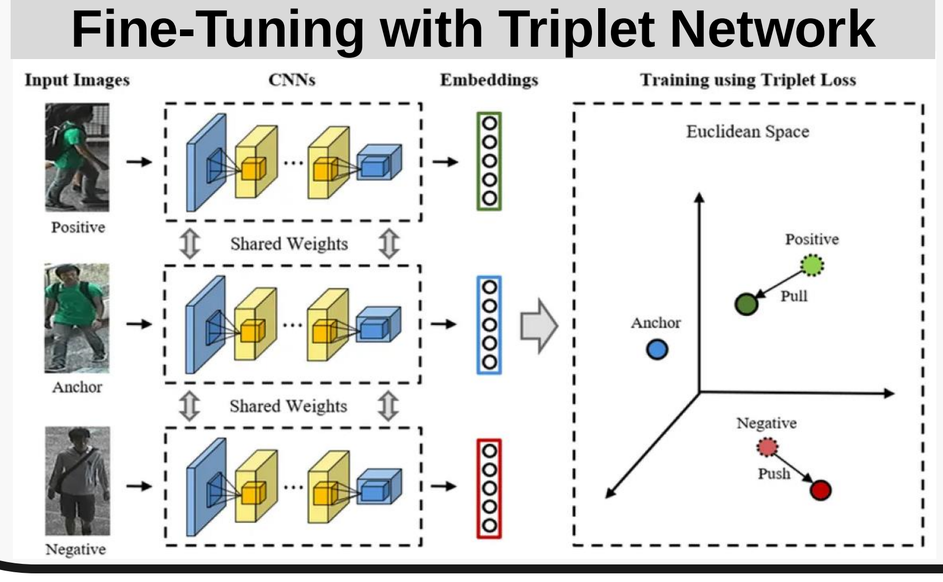
Precision@10
5% Baseline ↑ 93% Final Model

Identified Duplicates
↑ 550 %

Precision@10: Number of top 10 items matching at least 3 of silhouette, color, pattern, or key elements

Baseline: Filter candidates only by tabular features and ranked randomly

Fine-Tuning with Triplet Network



Example Similarity Search

The model returns the true duplicate product with the highest similarity score.

Business Impact

Multiple Use Cases
- Group existing similar products together and categorize new products into clusters of similar items.
- Store ranked matches with similarity scores in Macy's cloud and build a dashboard for labeling exact duplicates and non-duplicates.
- Integrate model into product recommendation systems for enhanced relevance and diversity of suggestions.

→ By eliminating duplicate products across all categories, the estimated cost savings is up to **$3.9M**.

→ Boosted the current number of identified duplicates by **550%**.

Acknowledgements: *We would like to thank the Macy's Inc. EDA team (especially Albert Egido De Poy, Jay Malepati, and Rohan Talati) for their time and support, and our faculty advisor Professor Alexandre Jacquillat for his insights and guidance. Together they have helped us develop the work presented in this poster.*