Multi-label Classification of Over-The-Counter healthcare products (A real-world use case)

Course: AI in Healthcare

Final Project

Fall 2024

Multi-label Classification of Over-the-Counter Healthcare Products

- Using ELECTRA Transformer Model
- Using MultiLabelBinarizer
- Accuracy up to 92% on unseen data

Problem Statement
The Challenge in Healthcare Product Classification

- Manual classification is time-consuming and error-prone
- Products often span multiple categories
- Critical for:
  - Inventory management
  - Regulatory compliance
  - Customer experience
  - Search optimization

Data Overview
Dataset Characteristics

- 65,996 product descriptions
- Three classification dimensions
  - 9 Segments (broad categories)
  - 50 Sub-segments (specific types)
  - 4 Target Customers
- Have to correctly classify Segment, Sub-segment and Target Customer based on long-form product description.
- Hierarchical department structure
- Split: 65% training, 25% validation, 10% test (unseen data)

Example 1

Segment	Sub-Segment
Allergy	Allergy INS
	Oral Antihistamine
	Other Allergy
	Topical Itch

Example 2

Segment	Sub-Segment
Vitamins, Minerals & Supplements	Herbal & Natural (H&N)
	Minerals
	Multivitamins
	Pregnancy Vitamins
	Probiotics
	Single Vitamins
	Supplements

Target Group
Child
Adult
Infant
Unknown

System Architecture

Product Text
- Description
- Department

Preprocessing
- Tokenization
- Cleaning

ELECTRA Encoder (12 Layers)

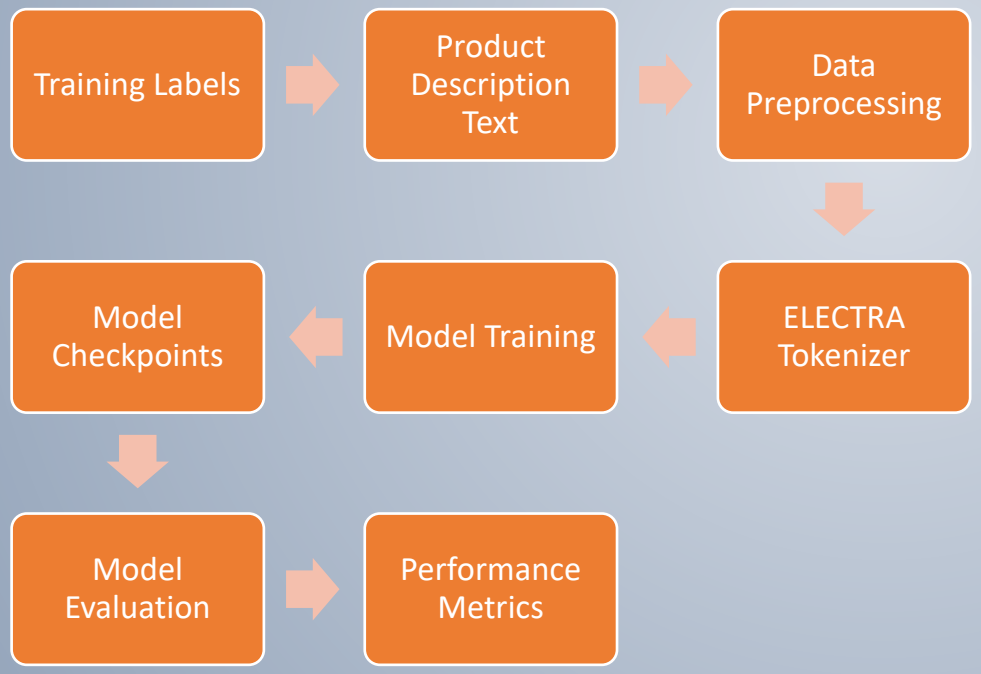
Hidden States (768-dim)

Classification Heads (3)

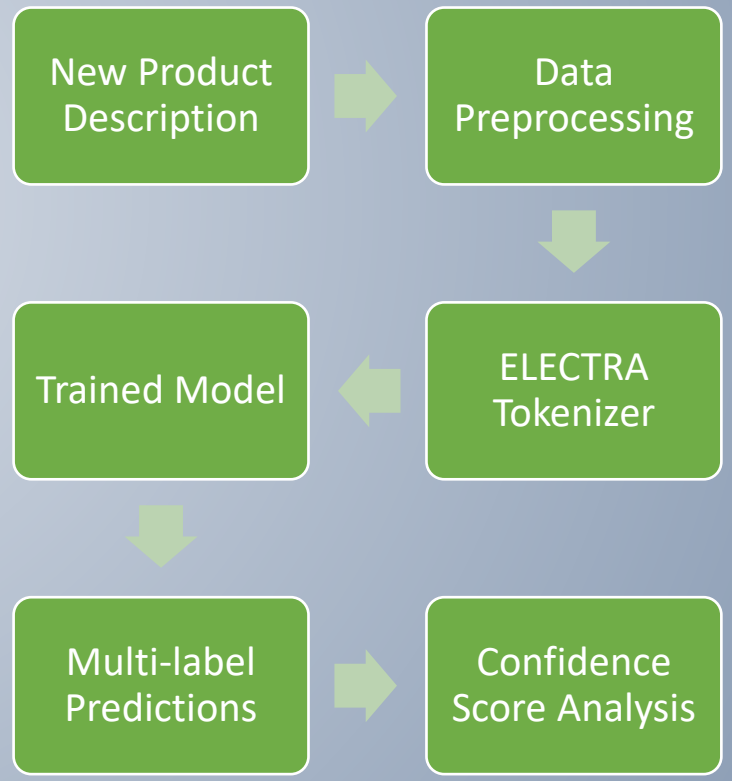
Predictions with Confidence

Data Flow Architecture

TRAINING DATA PIPELINE



INFERENCE PIPELINE



Detailed Model Architecture ELECTRA-base Model

Input Text

ELECTRA Encoder (12 Layers)

Hidden States (768 dimensions)

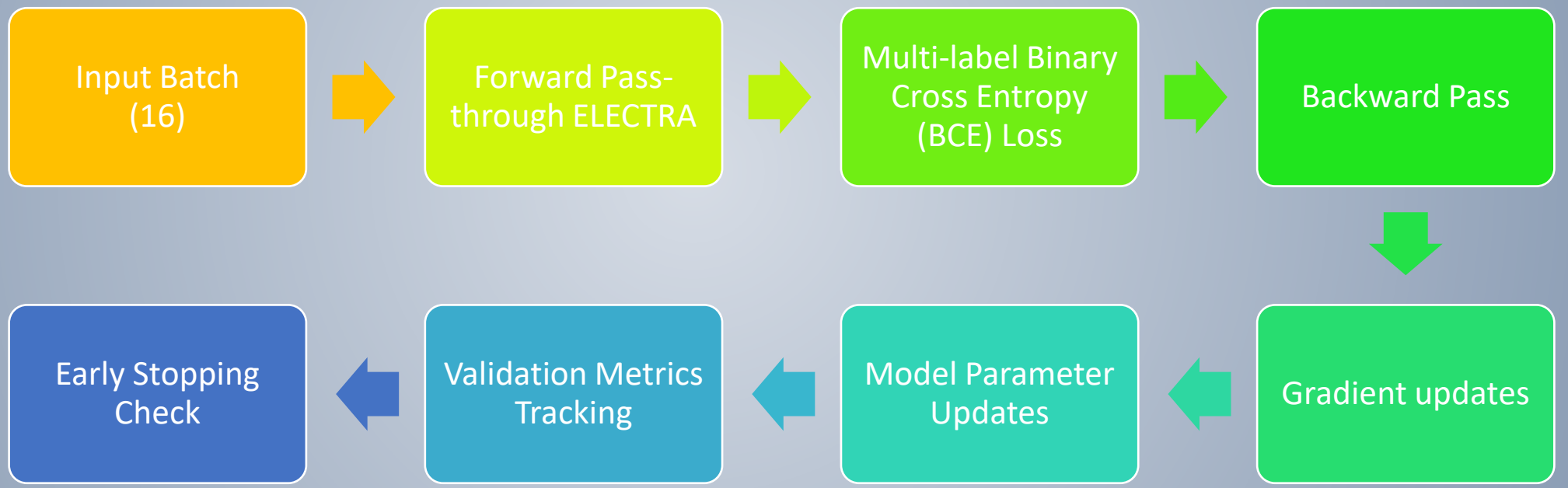
Classification Heads

- Segment Classifier (9 labels)
- Sub-Segment Classifier (50 labels)
- Consumer Classifier (4 labels)

Sigmoid Activation

Confidence Scores & Prediction

Training Process
Training Workflow

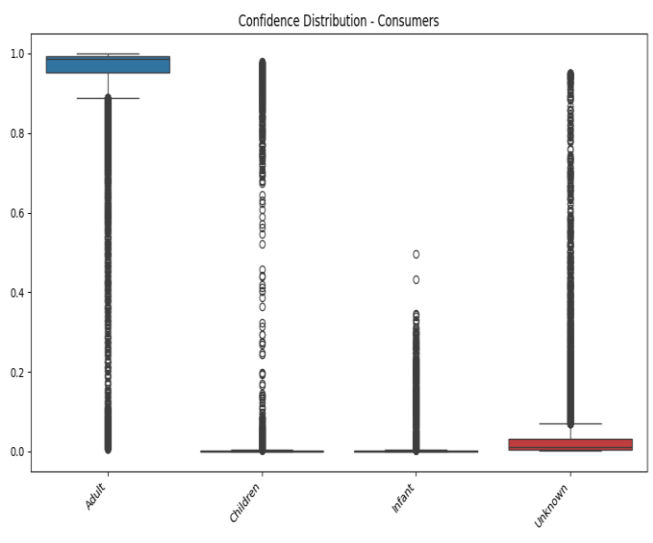


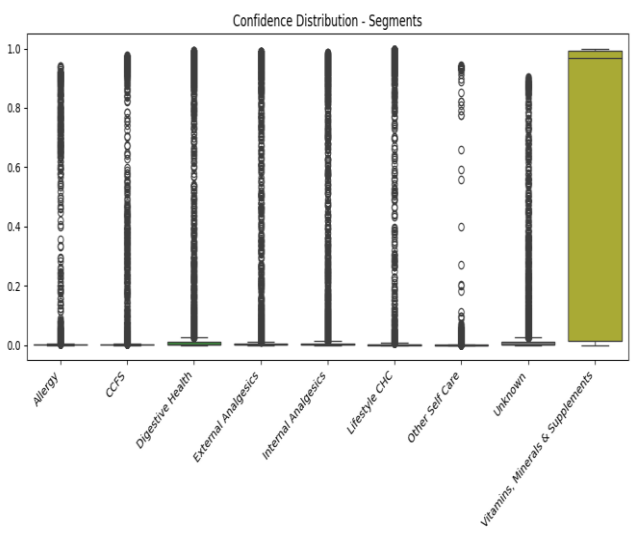
Results Overview
Performance Metrics on Unseen Data

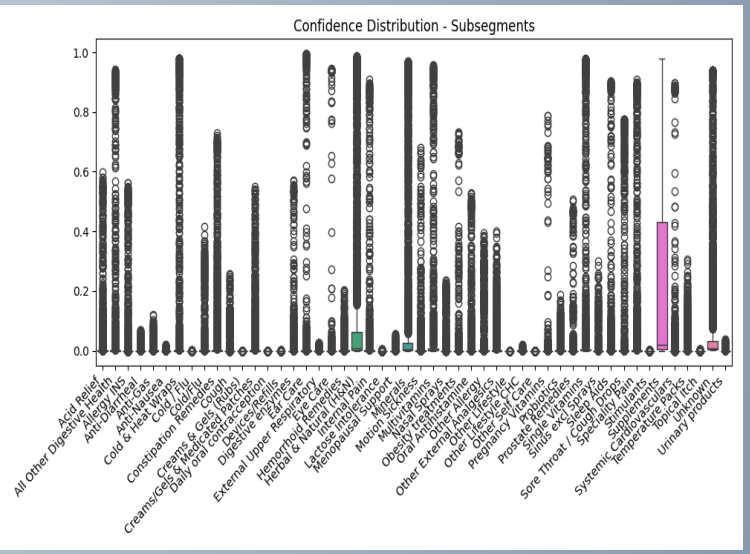
	Segments	Sub-segments	Consumers
Accuracy	89.33%	67.35%	92.27%
F1 score	89.21%	66.22%	90.62%

Confidence Analysis

Consumer Classification	Segment Confidence	Sub-Segment
High confidence	Strong performance in Vitamins, Minerals & Supplements	Variable confidence across sub-segments







Error Analysis
Common Error Patterns

- Correct Segment, wrong sub-segment
- Example: Beet Root Supplement
  - Correct: Vitamin, Minerals & Supplements
  - Missed: Minerals sub-segment
- Higher Accuracy in well-defined categories

Future Directions
Next Steps

- Enhanced sub-segment classification
- Integration of product images
- Active learning implementation
- Regulatory compliance features
- API development for real-time classification