LabelSense

An AI-driven Food Label Analyzer

Team Members:
Khushboo Harsh Patel (kp3329@drexel.edu)
Shobhit Dixit (sd3733@drexel.edu)
Neel Rakeshbhai Patel (np928@drexel.edu)

April 28, 2025

Introduction

Consumers frequently encounter difficulties interpreting complex food labels, especially when managing dietary preferences and allergen sensitivities. Current applications such as **Yuka** and **Fooducate** predominantly rely on barcode scanning, significantly limiting their effectiveness only to products present within their databases. LabelSense proposes overcoming this barrier by directly analyzing images of ingredient lists, employing advanced Natural Language Processing (NLP) techniques, thus enabling personalized and globally accessible dietary insights.

Objective

LabelSense will address critical limitations of current products by clearly:

- Extracting and analyzing ingredient lists directly from images.
- Identifying products' compliance with vegetarian and vegan dietary requirements.
- Offering personalized allergen detection based on user profiles.
- Providing intuitive, simplified definitions of complex ingredients.
- Supporting multiple languages to ensure global applicability.

Existing Limitations

Existing solutions (**Yuka**, **Fooducate**) have critical shortcomings:

- **Barcode Dependency:** Limits scope strictly to database-listed products, excluding local or unregistered products.
- **Dietary Analysis:** Insufficient explicit detection of vegan or vegetarian ingredients.
- **Personalization Deficiency:** Generalized allergen detection, failing to adapt to individual user needs.
- **Complex Ingredient Definitions:** Missing or overly technical definitions, confusing users.

Incremental Improvements in LabelSense

LabelSense addresses these limitations through:

- **Direct Image-based Ingredient Analysis:** Utilizing OCR and NLP without barcode constraints.
- **Diet-specific Ingredient Detection:** Explicit NLP categorization of vegan/vegetarian ingredients.
- **Personalized Allergen Detection:** Tailored allergen alerts based on individual profiles.
- **Clear Ingredient Definitions:** NLP-driven intuitive explanations derived from trusted knowledge bases.
- **Multilingual Support:** Analyzing labels in multiple languages using multilingual NLP models.

Data Requirements

- Primary data from **Open Food Facts** (comprehensive ingredient and allergen data).
- Supplementary ingredient definitions from Wikipedia API.
- Ingredient label images from publicly available datasets for OCR training.

Evaluation Dataset Strategy

For effective end-to-end evaluation, **LabelSense** will utilize the *Open Food Facts (OFF)* dataset, which provides both **ingredient label images** and structured annotations such as allergens, dietary labels (e.g., vegan, vegetarian), and multilingual entries. This allows us to efficiently use a single dataset for both development and evaluation, ensuring consistency while avoiding the need to source additional data.

To simulate real-world usage—where users input images of ingredient lists—we will assess the complete pipeline, from OCR-based text extraction to NLP-driven analysis. This includes evaluating:

- The reliability of **OCR** across diverse image qualities, languages, and label formats.
- The accuracy of:
  - **Personalized allergen detection**.
  - **Dietary compliance verification**.
  - **Simplified ingredient explanations**.

A curated subset of approximately **100 samples** will be selected from OFF to ensure diversity across languages, product categories, and inclusion of challenging cases such as ambiguous or region-specific ingredient expressions.

Since the same dataset will work for evaluation, introduced Evaluation section for more information

Baseline Methods

- **OCR/Text Extraction**: Tesseract or Google Cloud Vision API for extracting ingredient lists from label images.
- **Ingredient Parsing (NLP)**: Initially, we plan to use pretrained NLP models such as `spaCy` or HuggingFace NER models to identify key entities within extracted text.
- **Dietary & Allergen Classification**: Rule-based ingredient matching against authoritative allergen and dietary compliance lists to detect potential allergens and verify vegan/vegetarian status.

Alternative Approach:

In addition to the above methods, we will explore leveraging **Large Language Models (LLMs)**, such as GPT-4, for parsing and classification tasks. This approach offers greater flexibility, particularly in handling:

- Ambiguous or region-specific ingredient expressions (e.g., variations in labeling gelatin across countries).
- Multilingual ingredient lists.
- Uncertain allergen cases where predefined rules may be insufficient.

Updated section based on feedback regarding LLM usage.

Future Potential

- **Personalized Allergen Prediction:** Transformer-based NLP (e.g., DistilBERT) trained incrementally on user-provided data.
- **Multilingual NLP Pipeline:** Using multilingual transformer models (e.g., XLM-Roberta).
- **Proactive Ingredient Health Risks (Stretch Goal):** Automatically summarizing recent research via Semantic Scholar API to highlight ingredient concerns.

Conclusion

LabelSense distinctly addresses existing gaps by eliminating barcode limitations, providing precise dietary analyses, personalized allergen detection, simplified ingredient explanations, and multilingual support. Its incremental advancements set LabelSense apart from current market solutions, confirming its novelty, feasibility, and potential for global impact within the proposed timeframe.