Scientific Report CSE 455/555 Term Project

Jacob Szczudlik

Hanwen Dong

7/27/25

Part 2: NLP Engineer, Hanwen Dong. BERT-based Recipe Recommendation Model

Data Analysis: I leveraged the RAW_recipes.csv corpus of 231637 recipes (and optionally RAW_interactions. csv for ratings) to fine-tune a semantic retrieval model. I preprocessed each record by concatenating its tags and ingredients into a single input sequence in the following **format:**

[CLS] tag1 tag2 ... [SEP] ingredient1 ingredient2 ... [SEP]

This step produced the file bert_recipe_inputs.json, which served as my training corpus.

Model Setup: I chose bert-base-uncased from HuggingFace as the foundation. Tokenization and model weights were loaded from the pretrained checkpoint. For fine-tuning, I configured the hyperparameters as follows:

Batch size: 8

Learning rate: 2×10^-5

Epochs: 3

Training was executed on a Google Colab A100 GPU through pro, with both model and data tensors moved to CUDA. I optimized the **[CLS]** embedding representations without defining a supervised classification loss, focusing solely on refining the semantic vectors for retrieval.

Evaluation: After training completed, I ran several multi-tag queries to assess retrieval quality. Below are representative **results:**

Top 5:

Query: 'banana', 'bread']

beat this banana bread (0.927)

spicy banana bread (0.915)

say what banana sandwich (0.903)

the best banana bread or muffins (0.892)

bananas 4 ice cream pie (0.879)

Query: ['avocado', 'dip']

sour cream avocado dip vegan (0.912)

a fun and frolic kind of avocado bacon and tomato wrap yippee (0.887)

a 1 steak puffs with corn and avocado salsa a1 (0.865)

a 1 butter deconstructed pinwheel steak with avocado sauce a1 (0.854)

a b l c avocado bacon lettuce crab (0.841)

Query: ['carrot', 'soup', 'ginger']

jiffy extra moist carrot cake (0.919)

the carrot cake (0.906)

german barbecued carrots (0.894)

s wonderful carrot and raisin salad (0.882)

sweet carrots (0.871)

{
  "name": "arriba   baked winter squash mexican style",
  "input_text": "[CLS] ['60-minutes-or-less' 'time-to-make' 'course' 'main-ingredient'
},
{
  "name": "a bit different breakfast pizza",
  "input_text": "[CLS] ['30-minutes-or-less' 'time-to-make' 'course' 'main-ingredient'
},
{
  "name": "all in the kitchen chili",
  "input_text": "[CLS] ['time-to-make' 'course' 'preparation' 'main-dish' 'chili' 'cro
},
{
  "name": "alouette potatoes",
  "input_text": "[CLS] ['60-minutes-or-less' 'time-to-make' 'course' 'main-ingredient'
},
{
  "name": "amish tomato ketchup for canning",
  "input_text": "[CLS] ['weeknight' 'time-to-make' 'course' 'main-ingredient' 'cuisine'
},
{
  "name": "apple a day milk shake",
  "input_text": "[CLS] ['15-minutes-or-less' 'time-to-make' 'course' 'main-ingredient'
},
{
  "name": "aww marinated olives",
  "input_text": "[CLS] ['15-minutes-or-less' 'time-to-make' 'course' 'main-ingredient'
},
{
  "name": "backyard style barbecued ribs",
  "input_text": "[CLS] ['weeknight' 'time-to-make' 'course' 'main-ingredient' 'cuisine'
},
{
  "name": "bananas 4 ice cream pie",
  "input_text": "[CLS] ['weeknight' 'time-to-make' 'course' 'main-ingredient' 'preparat
},
{
  "name": "beat this banana bread",
  "input_text": "[CLS] ['weeknight' 'time-to-make' 'course' 'main-ingredient' 'preparat
},
{
  "name": "berry good sandwich spread",
  "input_text": "[CLS] ['15-minutes-or-less' 'time-to-make' 'course' 'preparation' 'low
},
{
  "name": "better than sex strawberries",
  "input_text": "[CLS] ['weeknight' 'time-to-make' 'course' 'main-ingredient' 'preparat
},
{
  "name": "better then bush s baked beans",

Query: ['pineapple', 'grilled', 'dessert']

golden circle pineapple upside down cake (0.933)

2 ingredient pineapple cake with pineapple sauce (0.915)

pink stuff cherry pie filling pineapple dessert (0.903)

5 points plus pineapple and honey glazed chicken (0.888)

2blueberry pineapple ring (0.876)

Failure Case:

Query: ['banana', 'gluten-free']

Returned recipes:

beat this banana bread (0.750)

say what banana sandwich (0.742)

jiffy extra moist banana muffins (0.735)

These results show that the 'gluten-free' descriptor is under-represented in the dataset and often overshadowed by popular banana recipes, indicating a need for better ingredient filtering or explicit dietary labels.

Conclusion:

I successfully developed and fine-tuned a bert-based semantic search engine over 231637 recipes focused on avocado , banana, carrot, and pineapple. The model efficiently returns relevant multi?tag recommendations and clearly identifies areas for enhancement, tehn embedding caching for faster queries, input enrichment with titles and steps, and rating based ranking to refine results.