CHeF: Classification of Home Cooked Food

Harsh Vishwakarma(2022205) | Dhairya(2022157)
BTP Advisor: Dr. Ganesh Bagler

Introduction

Food preparation involves a wide range of culinary processes that vary in complexity, technique, and level of transformation applied to raw ingredients. Understanding these processes is essential for modeling recipe behavior, comparing cooking styles, and quantifying the degree of processing involved in food preparation. In this work, we construct a large-scale dataset of unique culinary processes and develop an unsupervised pipeline to analyze, cluster, and quantify their processing intensity. Using Word2Vec embeddings trained on thousands of recipe instructions, we capture the semantic and contextual similarity among 270 unique actions such as *boil*, *knead*, *caramelize*, and *puree*. These embeddings are clustered using K-Means, DBSCAN, Agglomerative and GMM to discover natural groupings of cooking techniques based purely on linguistic and contextual cues.

To evaluate the relative complexity of each process, we introduce a novel **Processing Intensity Score (PIS)** that integrates three dimensions: (i) a lexical-based Word Complexity Score, (ii) an Embedding Dispersion Score measuring semantic variability, and (iii) a Cluster Transition Difficulty Score reflecting how distant a process is from the global cooking-technique space. This normalized, interpretable score enables ranking individual processes as well as comparing clusters on an objective processing scale. Finally, we propose the **CHEF Classification System**, a five-level scale ranging from *minimal* to *highly technical*, which categorizes each process and cluster based on its computed intensity.

Dataset Overview

The dataset used in this project comes from **RecipeDB (Cosylab)**, a large curated repository of global recipes. It combines structured metadata, ingredients, processes, and cooking instructions into a unified format. Only the most relevant columns were used for analysis:

Key Columns (Used in This Project)

- **Recipe ID** – Unique identifier for each recipe.
- **Recipe Title** – Name of the dish.
- **Total Time** – Total preparation and cooking time.
- **Ingredients** – Cleaned list of ingredients used.
- **Instructions / Steps** – Full cooking steps used for extracting processes and training embeddings.
- **Processes** – Extracted cooking actions (e.g., *heat, stir, chop, simmer*).
- **Process Count** – Total number of operations per recipe
- **Region / Sub-Region / Continent** – Geographic origin of the recipe for cultural comparisons.
- **Diet Labels** – Vegetarian, vegan, pescetarian, etc., for nutritional/dietary grouping.
- **Basic Nutrients** – Energy, fat, carbs, protein

Key Insights from data

Top 5 Most Frequent Cooking Processes

- **add** — 187,921
- **heat** — 102,759
- **cook** — 94,106
- **stir** — 83,544
- **place** — 55,651

Top 5 Recipes With the Most Cooking Steps

- **5-Cheese Crab Lasagna** — 121
- **Murgh Makhani (Butter Chicken)** — 90
- **Italian Rum Cake** — 84
- **Vegetable Lasagna** — 83
- **Cod Brandade** — 82

Methodology

Word2Vec Embeddings

A Word2Vec model was trained on 270 standardized cooking process terms to learn **100-dimensional semantic embeddings**.
These embeddings capture similarities in *culinary meaning*: processes used in similar cooking contexts (e.g., "mix", "whisk", "blend") appear close together in the embedding space, while more specialized actions (e.g., "crimp", "deflate") lie farther away.
This vector space becomes the foundation for all clustering experiments.

Model Configuration

- **Architecture:** Skip-Gram (SG = 1)
- **Vector Size:** 100-dimensional embeddings
- **Window Size:** 8 (context window)
- **Minimum Token Count:** 1
- **Training Epochs:** 10
- **Workers:** 4 parallel CPU threads

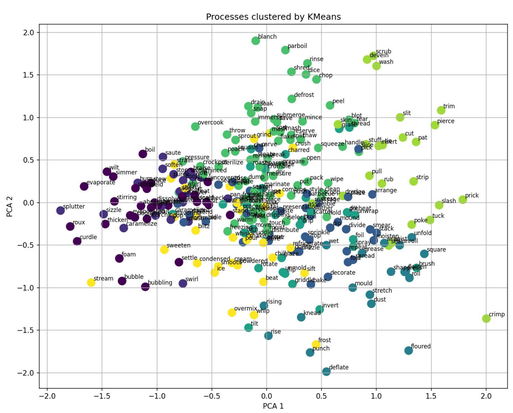
Handling Multi-Word Processes

For multi-token process terms (e.g., *stir-fry*, *pan-sear*),
the final embedding is computed as the **mean of the individual token vectors**.

Clustering Methods Used

1. **K-Means:** Groups process embeddings into a fixed number of clusters by minimizing within-cluster distance.
2. **Agglomerative Clustering:** Builds clusters hierarchically by repeatedly merging the closest groups.
3. **DBSCAN:** Forms clusters based on dense regions in the data and marks low-density points as noise.
4. **Gaussian Mixture Model (GMM):** Creates soft, probabilistic clusters assuming the data is generated from multiple Gaussian distributions.

METHOD	Silhouette Score:	Calinski–Harabasz Index:	Davies–Bouldin Score:
K-Means	0.357	7.48	3.3103
Agglo / MBKMeans	0.216	6.91	3.4477
GMM	0.341	7.45	3.3184
DBSCAN	–	–	–



1. Word Complexity Score (WCS)

Measures linguistic difficulty based on:

- Token length
- Number of tokens
- Rarity of words
- Technical suffixes (*–ize, –ate, –ifying, –ation*)

$$WCS(p) = 0.4 \times \Sigma \text{ len}(t_i) + 0.3 \times n + 0.3 \times (\Sigma [1 / (1 + \text{freq}(t_i))] + \text{suffix\_count}(p))$$

Interpretation:
Higher WCS → more technical / industrial-sounding process.

2. Embedding Dispersion Score (EDS)

Measures how "spread out" a process embedding is:

$$EDS = ||e||^2 + Var(e)$$

Interpretation:
High EDS → semantically intense process
Low EDS → simple, generic actions (e.g., *add, mix*)

3. Cluster Transition Difficulty Score (CTDS)

Distance from the **global embedding centroid**:

$$g = (1 / N) * sum(e\_i \text{ for all processes})$$

$$CTDS = ||e - g||^2$$

Interpretation:
High CTDS → unusual, specialized process
Low CTDS → everyday, common process

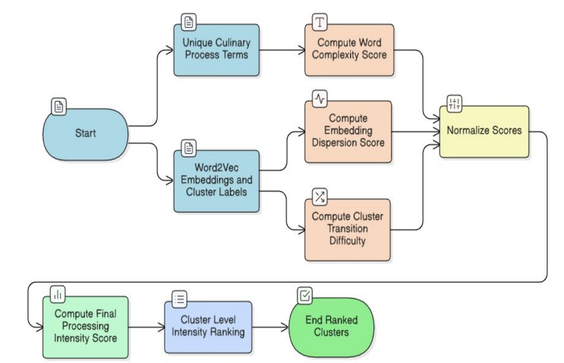
All metrics are min–max normalized to [0,1]:

$$M^* = (M - min(M)) / (max(M) - min(M))$$

$$\text{Final PIS: } PIS = 0.33 * WCS^* + 0.33 * EDS^* + 0.34 * CTDS^*$$

$$\text{Cluster\_Intensity}(k) = \text{average(PIS of all processes in cluster } k)$$

Work Flow



RESULTS

Cluster	Mean PIS	Interpretation
3	0.4575	Most intense — structural shaping. *Examples:* knead, mould, press, stretch, roll, wrap,
4	0.4446	High-heat surface transformations. *Examples:* roast, broil, sear, barbecue, char, glaze
6	0.4359	Precision cutting & modification. *Examples:* trim, pierce, slit, stuff, tie, seal
7	0.4047	Mechanical mixing & blending. *Examples:* blend, puree, grind, whisk, crush, whip
0	0.4044	Thermal processing & slow cooking. *Examples:* boil, simmer, braise, reduce, thicken
1	0.3854	Quick stovetop actions. *Examples:* fry, sauté, scramble, splutter, swirl
2	0.3392	Assembly, finishing, plating. *Examples:* garnish, drizzle, sprinkle, arrange, serve
5	0.3276	Basic everyday kitchen actions. *Examples:* add, chop, wash, peel, melt, remove

TOP 5 PROTOCOLS

Process	PIS
splutter	0.812
deflate	0.747
stream	0.742
crimp	0.742
floured	0.719

BOTTOM 5 PROTOCOLS

Process	PIS
add	0.001
set	0.013
mix	0.018
put	0.020
take	0.070

References

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv:1301.3781. Available at: https://arxiv.org/abs/1301.3781

CoSyLab, IIIT-Delhi (2020). *RecipeDB: A structured compendium of world recipes and their ingredients*. Available at: https://cosylab.iiitd.edu.in/recipedb/ (Accessed: February 2025)

Arora, N., Bhagat, S., & Dhama, R. (2024). *Machine Learning and Natural Language Processing Models to Predict the Extent of Food Processing*. arXiv:2412.17217. Available at: https://arxiv.org/pdf/2412.17217v1

Kaggle (2017). *Recipe Ingredients Dataset*. Available at: https://www.kaggle.com/datasets/kaggle/recipe-ingredients-dataset (Accessed: February 2025)

da Costa Louzada, M. L. & Gabe, K. T. (2025). *Nova food classification system: a contribution from Brazilian epidemiology*. Revista Brasileira de Epidemiologia, 28, e250027. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12129239/

Complete work details

Project GitHub Repositories:

- Home-Cooked Food Protocol Classification
- Numerical Model: Nutrient & NOVA Classification
- Food Pairing Analysis