Personalized Recipe Recommendation System

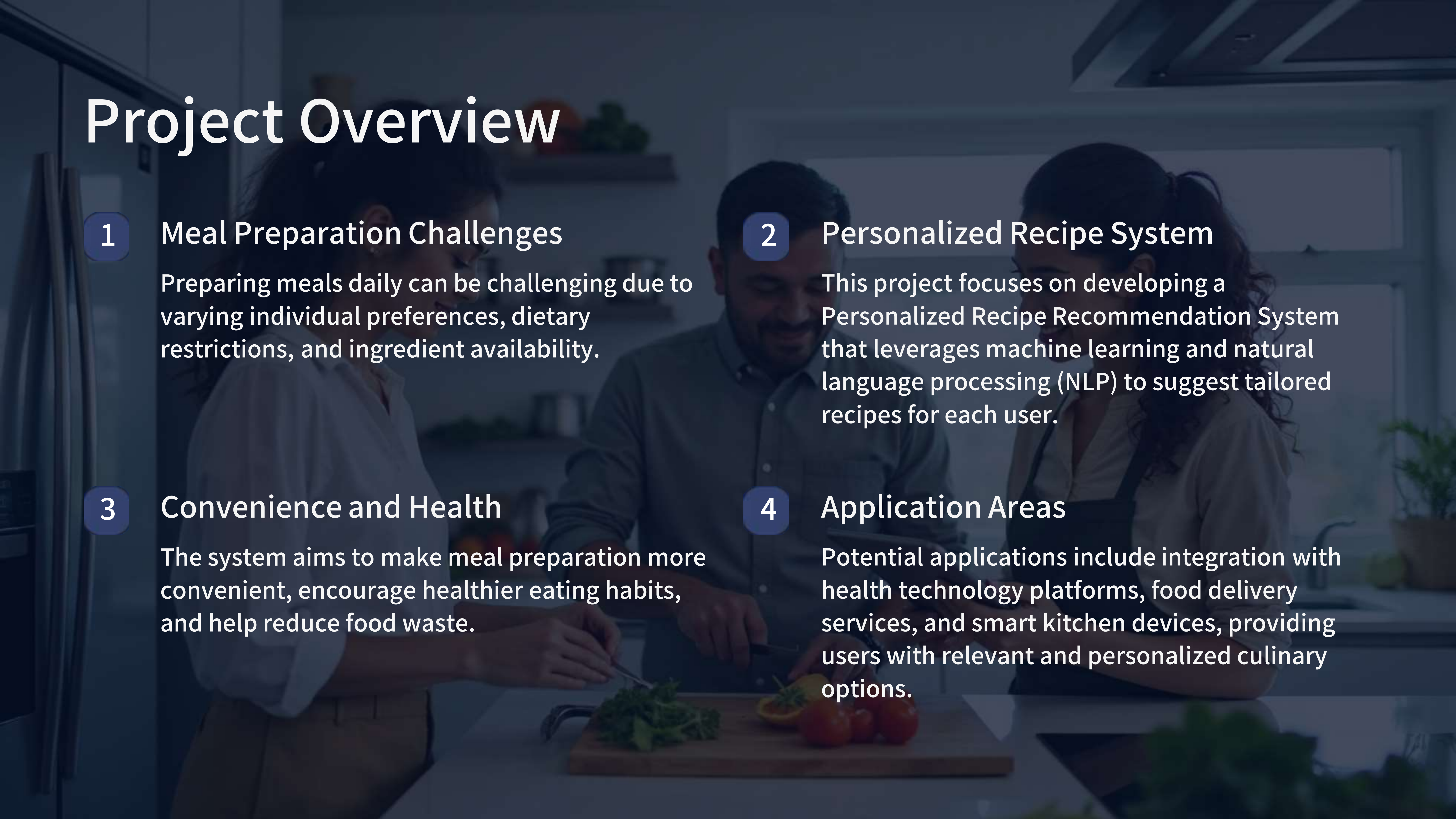
Susan Mworia

Neema Gatonye

Maureen Maina

Valentine Gacheri

George Ikuro



Project Overview

1 Meal Preparation Challenges

Preparing meals daily can be challenging due to varying individual preferences, dietary restrictions, and ingredient availability.

2 Personalized Recipe System

This project focuses on developing a Personalized Recipe Recommendation System that leverages machine learning and natural language processing (NLP) to suggest tailored recipes for each user.

3 Convenience and Health

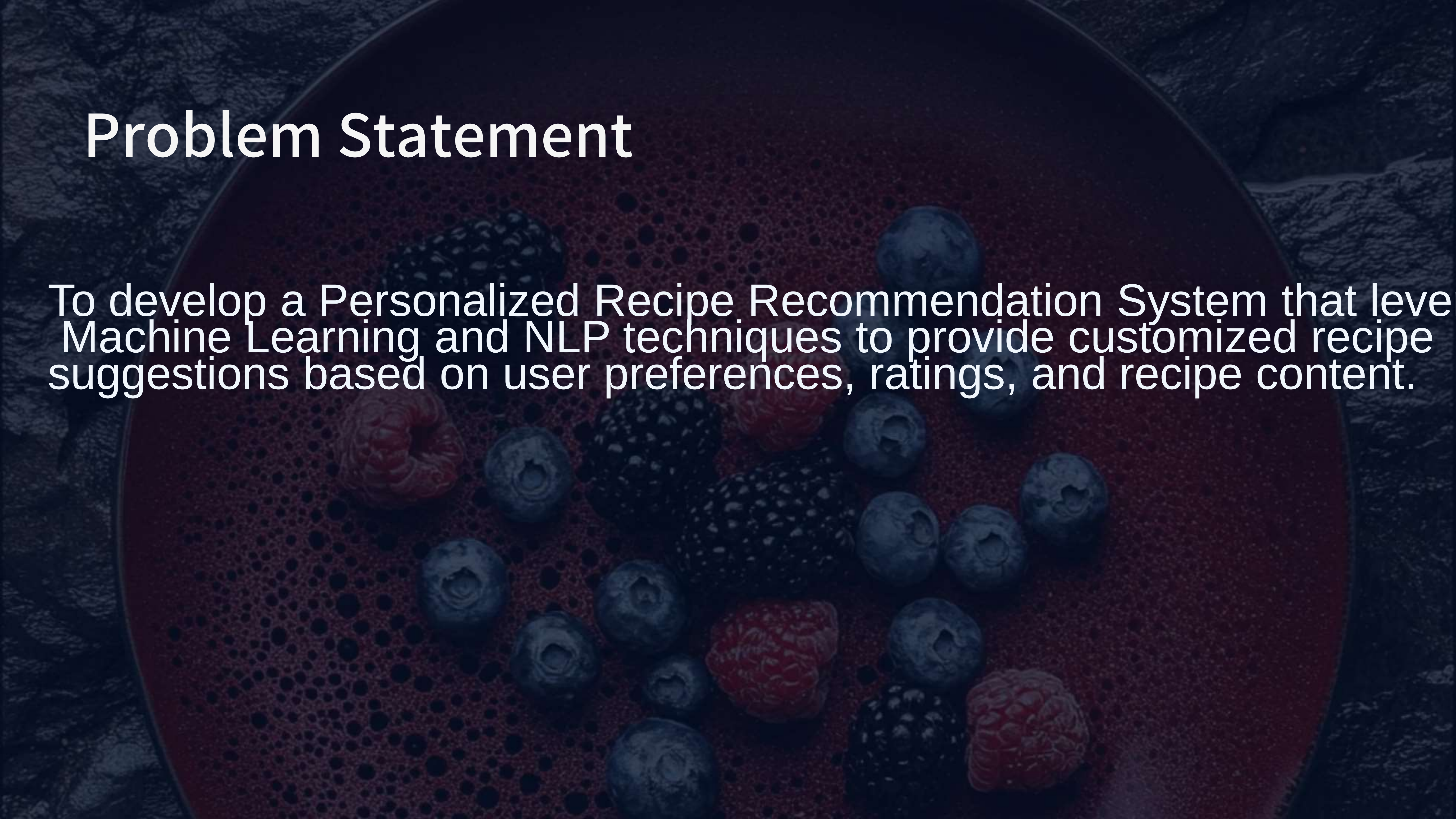
The system aims to make meal preparation more convenient, encourage healthier eating habits, and help reduce food waste.

4 Application Areas

Potential applications include integration with health technology platforms, food delivery services, and smart kitchen devices, providing users with relevant and personalized culinary options.

Problem Statement

To develop a Personalized Recipe Recommendation System that leve Machine Learning and NLP techniques to provide customized recipe suggestions based on user preferences, ratings, and recipe content.



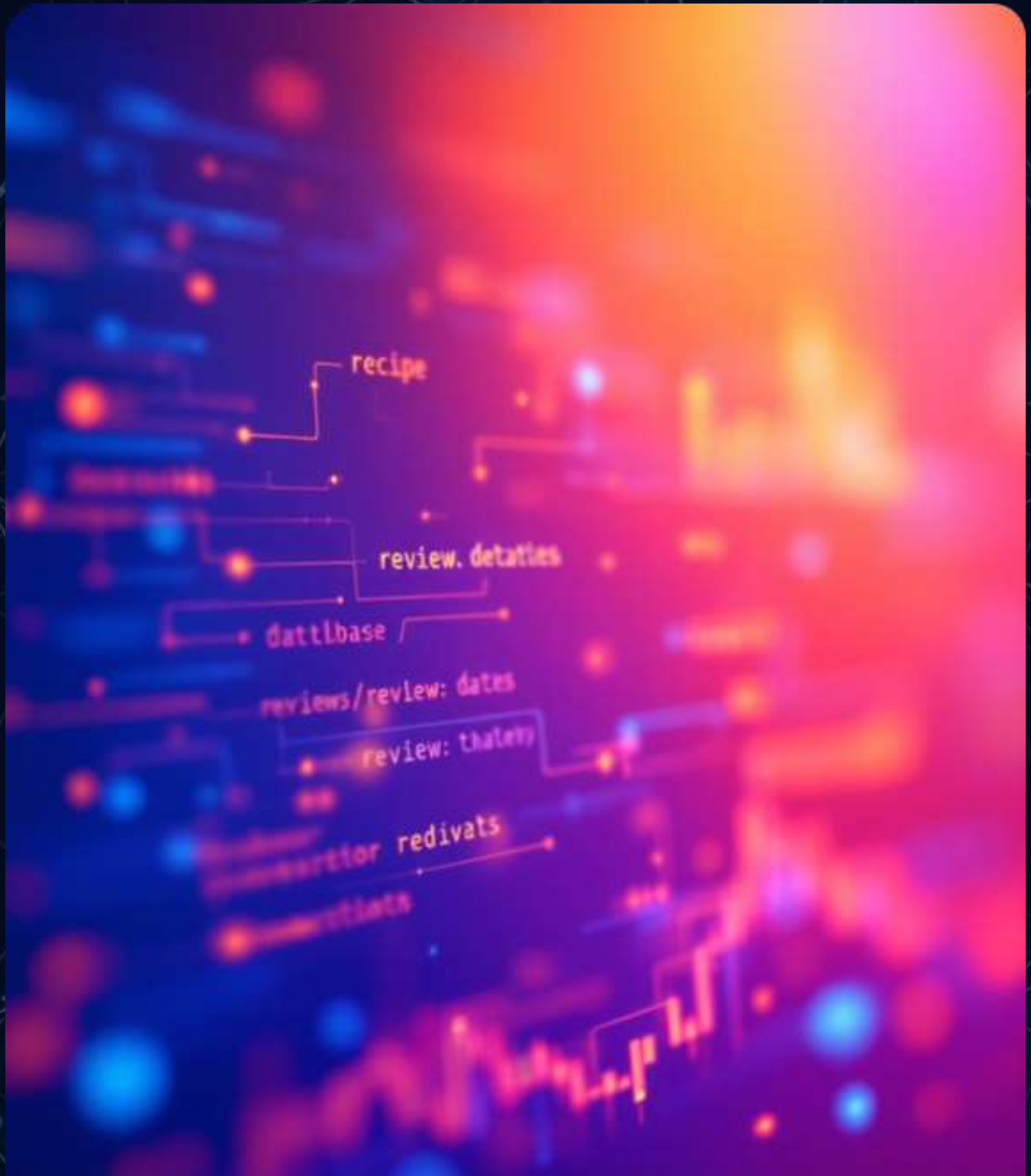
Project Objectives

Develop a model using Natural Language Processing (NLP) to recommend recipes based on ingredients, instructions, and keywords, focusing on recipe content similarity.

Use user interaction data such as ratings and reviews to identify patterns and recommend recipes liked by similar users through matrix factorization techniques.

Recommendation Method	Description
Content-Based Model	Focus on recipe content similarity.
Collaborative Filtering Model	Leverage user interaction data for recommendations.
Hybrid Recommendation System	Combine strengths of both methods for better accuracy.

Data Overview



Data Understanding

The recipes dataset contains 522,517 entries with 28 attributes including recipe IDs, names, author details, cooking times, nutritional info, and instructions.

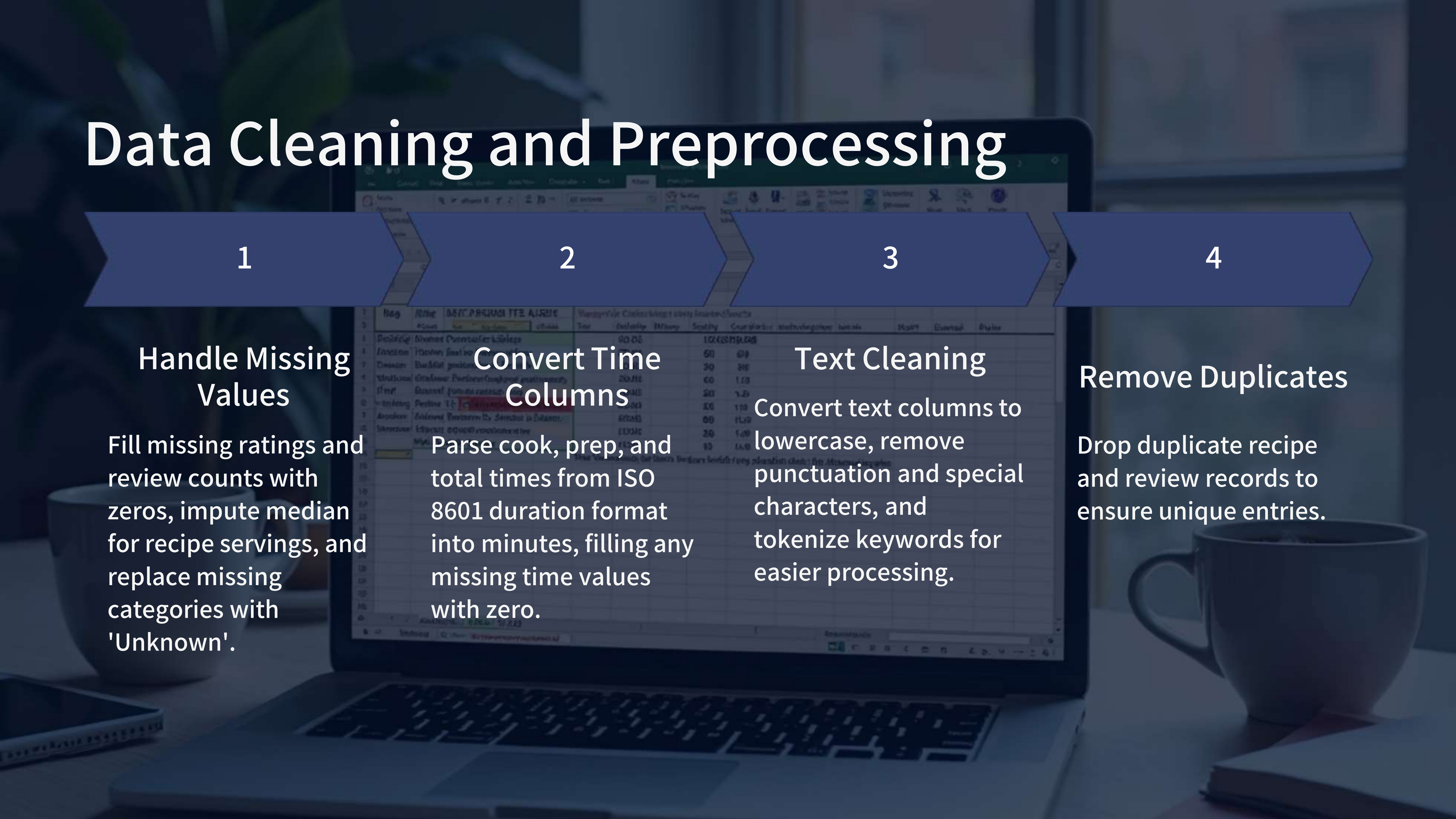
- The reviews dataset includes 1,401,982 entries with 8 columns such as review IDs, recipe references, author info, ratings, and review texts.

Key Recipe Attributes

- Each recipe entry includes detailed information like cook time, prep time, total time, ingredient quantities and parts, nutritional facts (calories, fat, protein, sodium, etc.), aggregated ratings, and review counts, enabling rich content-based analysis.

Key Review Attributes

Data Cleaning and Preprocessing



1

Handle Missing Values

Fill missing ratings and review counts with zeros, impute median for recipe servings, and replace missing categories with 'Unknown'.

2

Convert Time Columns

Parse cook, prep, and total times from ISO 8601 duration format into minutes, filling any missing time values with zero.

3

Text Cleaning

Convert text columns to lowercase, remove punctuation and special characters, and tokenize keywords for easier processing.

4

Remove Duplicates

Drop duplicate recipe and review records to ensure unique entries.

Feature Engineering & Category Mapping

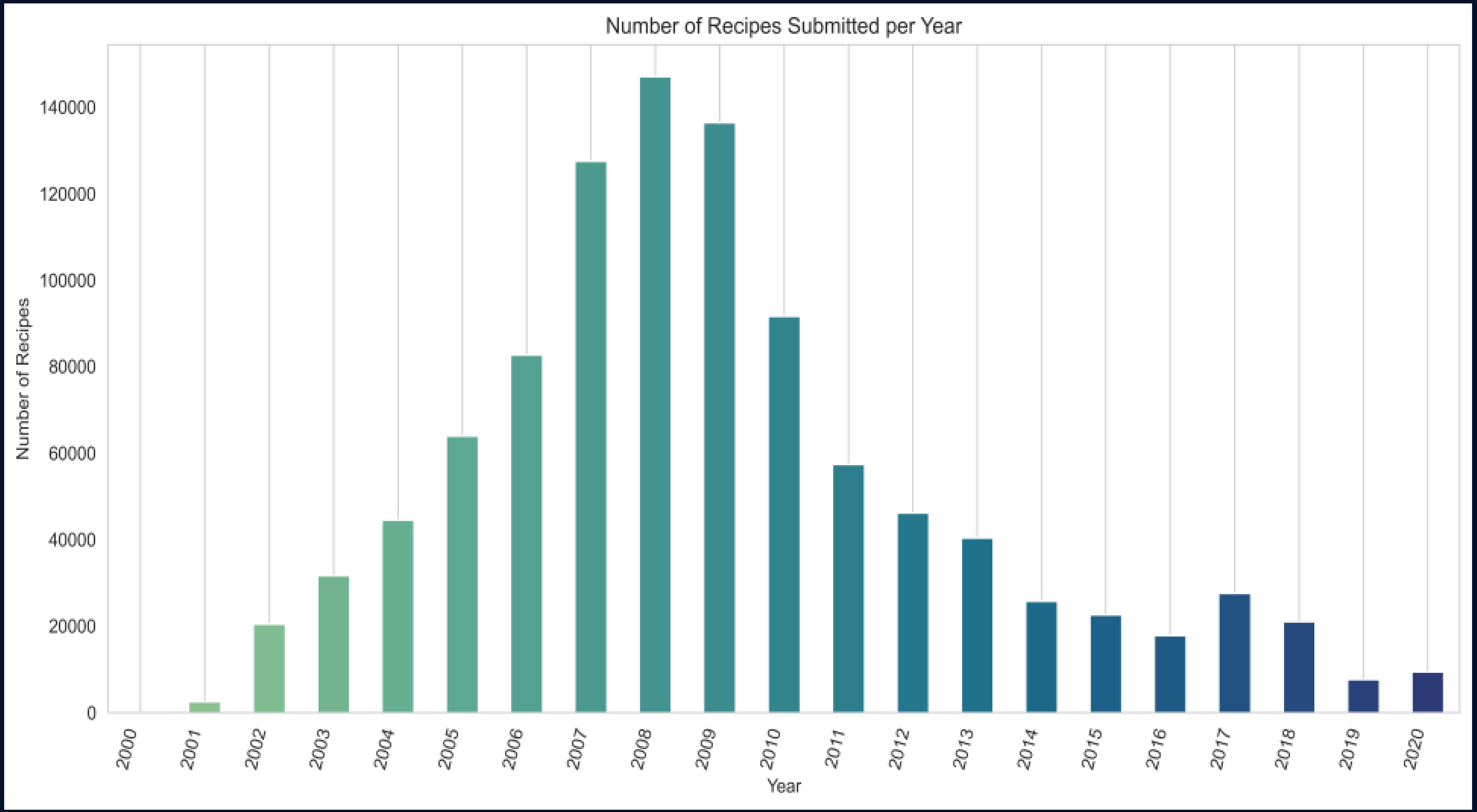
Feature Engineering

- We created a dictionary linking specific keywords from recipe categories to broader, meaningful groups to standardize diverse category labels.
- A flattened lookup dictionary was generated from the mapping to efficiently match recipe categories to their broader groups during preprocessing.

Category Mapping Logic

- A function scans each recipe's category, converts it to lowercase, and assigns it to a broad group based on keyword matches or defaults to 'others' if no match is found.
- The mapping was applied to the dataset, creating a new column with standardized category groups, enabling cleaner data analysis and more accurate recommendations.

Exploratory Data Analysis (EDA) - Time Trends



Peak Average Rating

average rating in 2006

Highest Review Count

review count in 2008

Max Recipes Submitted

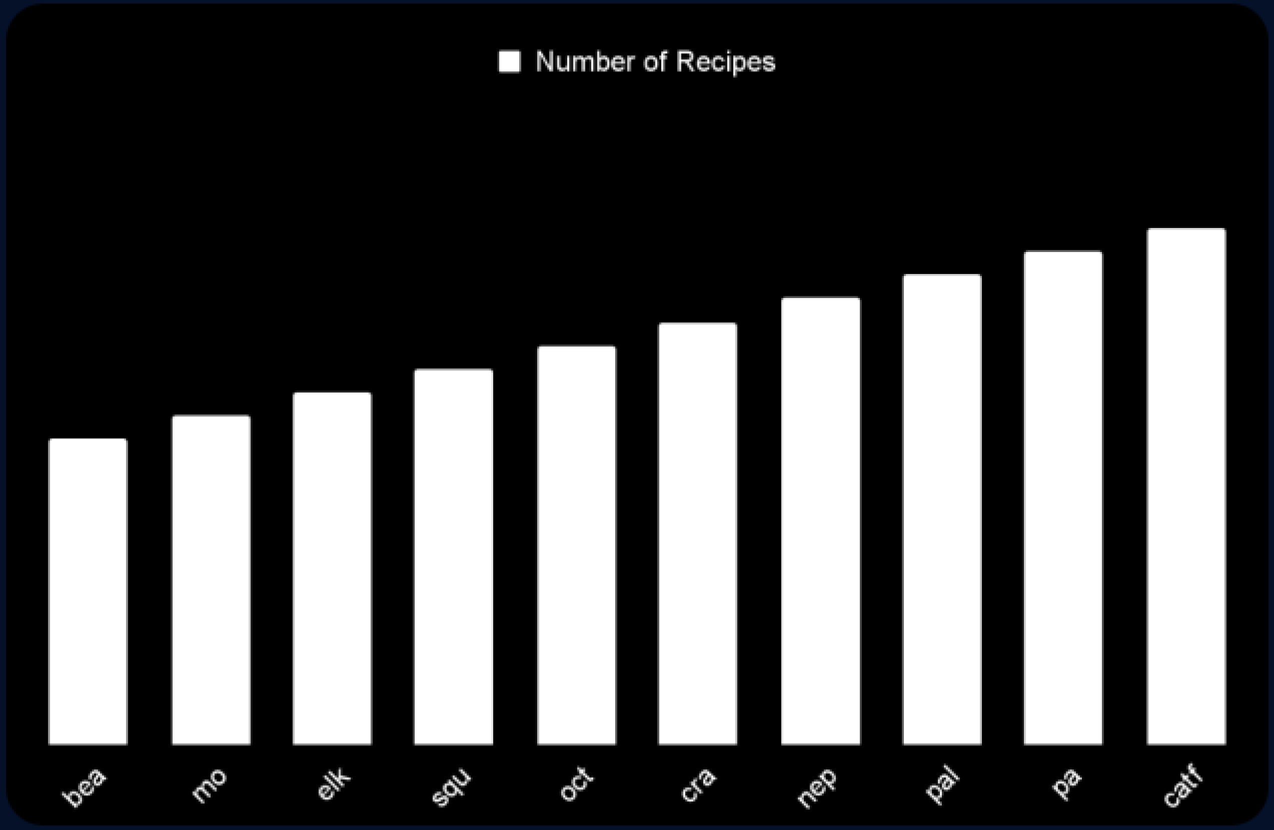
recipes submitted in 2008

Review Distribution and Recipe Popularity Data Insights

Histogram reveals that the majority of recipe reviews are 5-star, indicating widespread user satisfaction across different recipe categories.

Bar charts display the top 10 most reviewed recipes, demonstrating their popularity, alongside the 10 least reviewed, highlighting less common or niche recipes.

Recipe Category Popularity



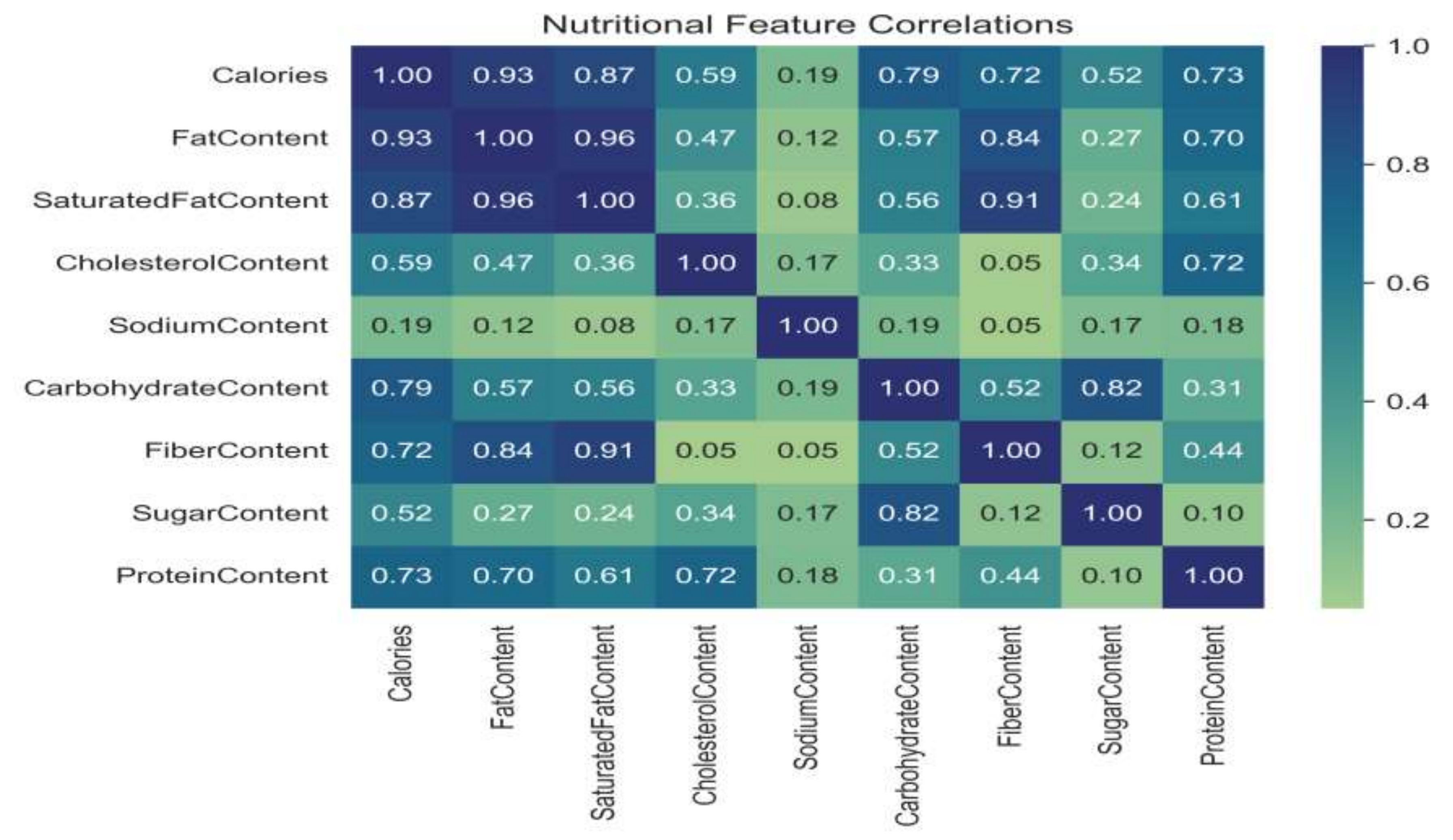
Poultry, desserts, and vegetable-based recipes lead in popularity, reflecting common dietary preferences and cooking trends.

Categories like exotic meats, niche diets, and uncommon ingredients show the lowest recipe counts, indicating specialized user interest.

Nutritional Feature Correlations

EDA

Correlation Heatmap of Nutritional Features



	Calories	FatContent	SaturatedFatContent	CholesterolContent	SodiumContent	CarbohydrateContent	FiberContent	SugarContent	ProteinContent
Calories	1.00	0.93	0.87	0.59	0.19	0.79	0.72	0.52	0.73
FatContent	0.93	1.00	0.96	0.47	0.12	0.57	0.84	0.27	0.70
SaturatedFatContent	0.87	0.96	1.00	0.36	0.08	0.56	0.91	0.24	0.61
CholesterolContent	0.59	0.47	0.36	1.00	0.17	0.33	0.05	0.34	0.72
SodiumContent	0.19	0.12	0.08	0.17	1.00	0.19	0.05	0.17	0.18
CarbohydrateContent	0.79	0.57	0.56	0.33	0.19	1.00	0.52	0.82	0.31
FiberContent	0.72	0.84	0.91	0.05	0.05	0.52	1.00	0.12	0.44
SugarContent	0.52	0.27	0.24	0.34	0.17	0.82	0.12	1.00	0.10
ProteinContent	0.73	0.70	0.61	0.72	0.18	0.31	0.44	0.10	1.00

Text Preprocessing for NLP

Preprocessing Steps

1

Tokenization

Split user reviews and ingredient text into individual words or tokens to enable detailed text analysis and feature extraction.

2

Stopword Removal

Remove common, non-informative words (stopwords) from the tokens to reduce noise and focus on meaningful content.

3

Stemming

Reduce words to their root form using Snowball Stemmer to group similar terms and improve text matching and model generalization.

4

Text Cleaning

Lowercase text, remove numbers, punctuation, mentions, hashtags, and URLs to standardize data and eliminate irrelevant symbols.

NLP Analysis - Top Words by Category

Analysis of Recipe Categories

1 Chicken Recipes

Top words include 'chicken', 'cook', 'easy', 'delicious', highlighting common cooking actions and positive feedback for chicken dishes.

2 Vegetarian Recipes

Frequent words such as 'vegetable', 'fresh', 'healthy', and 'simple' emphasize health-conscious and fresh ingredient themes in vegetarian recipes.

3 Dessert Recipes

Words like 'sweet', 'sugar', 'chocolate', and 'easy' dominate, reflecting the focus on sweetness and simplicity in dessert reviews.

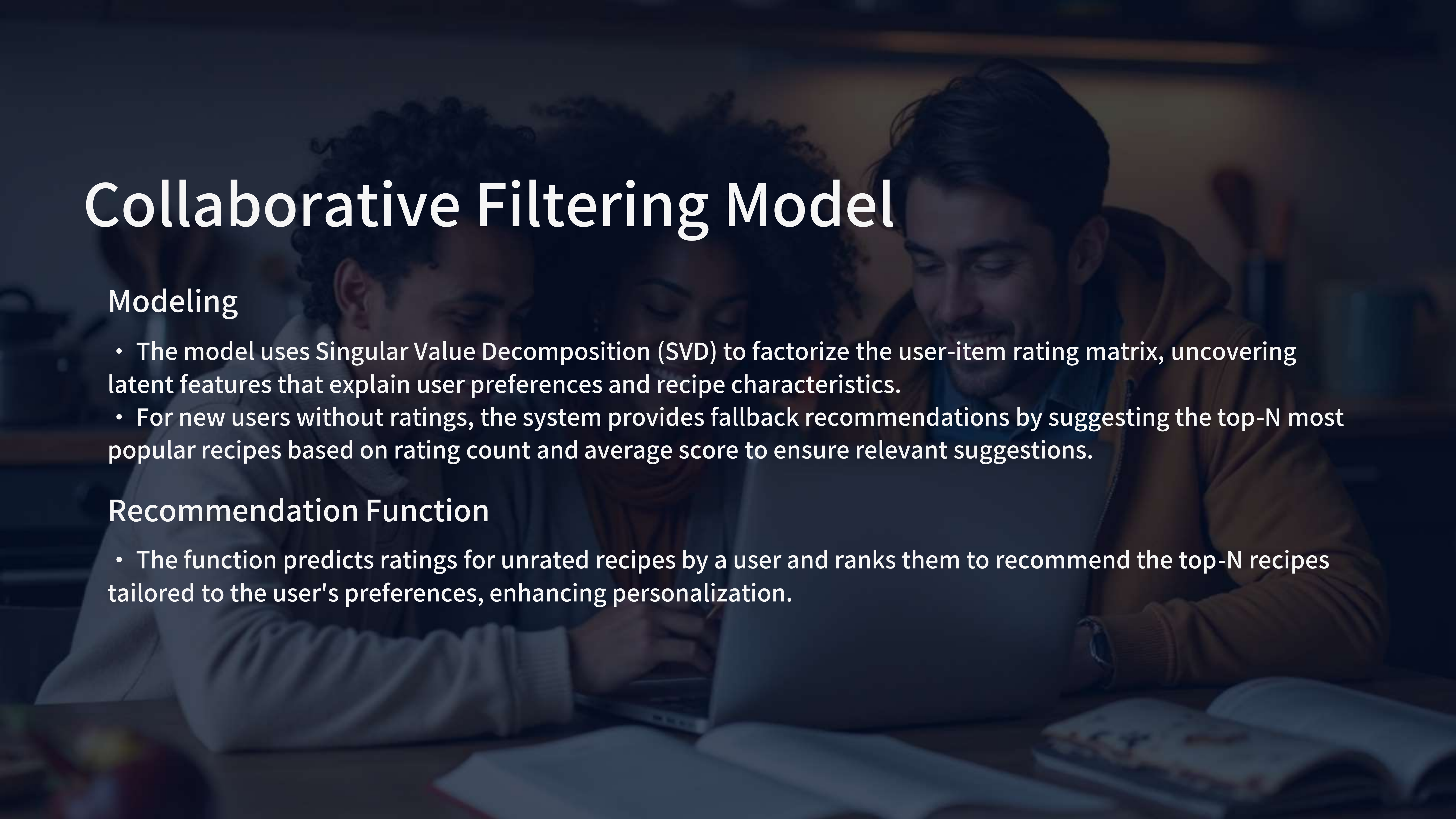
Ingredient and Keyword Analysis

Top 10 Ingredients

- The most frequent ingredients include salt, butter, egg, sugar, and flour, reflecting common recipe staples across categories.

Top 10 Keywords

- Keywords such as 'min', 'easi', 'low', and 'hour' dominate, indicating common themes in recipes like cooking time, simplicity, and dietary preferences.



Collaborative Filtering Model

Modeling

- The model uses Singular Value Decomposition (SVD) to factorize the user-item rating matrix, uncovering latent features that explain user preferences and recipe characteristics.
- For new users without ratings, the system provides fallback recommendations by suggesting the top-N most popular recipes based on rating count and average score to ensure relevant suggestions.

Recommendation Function

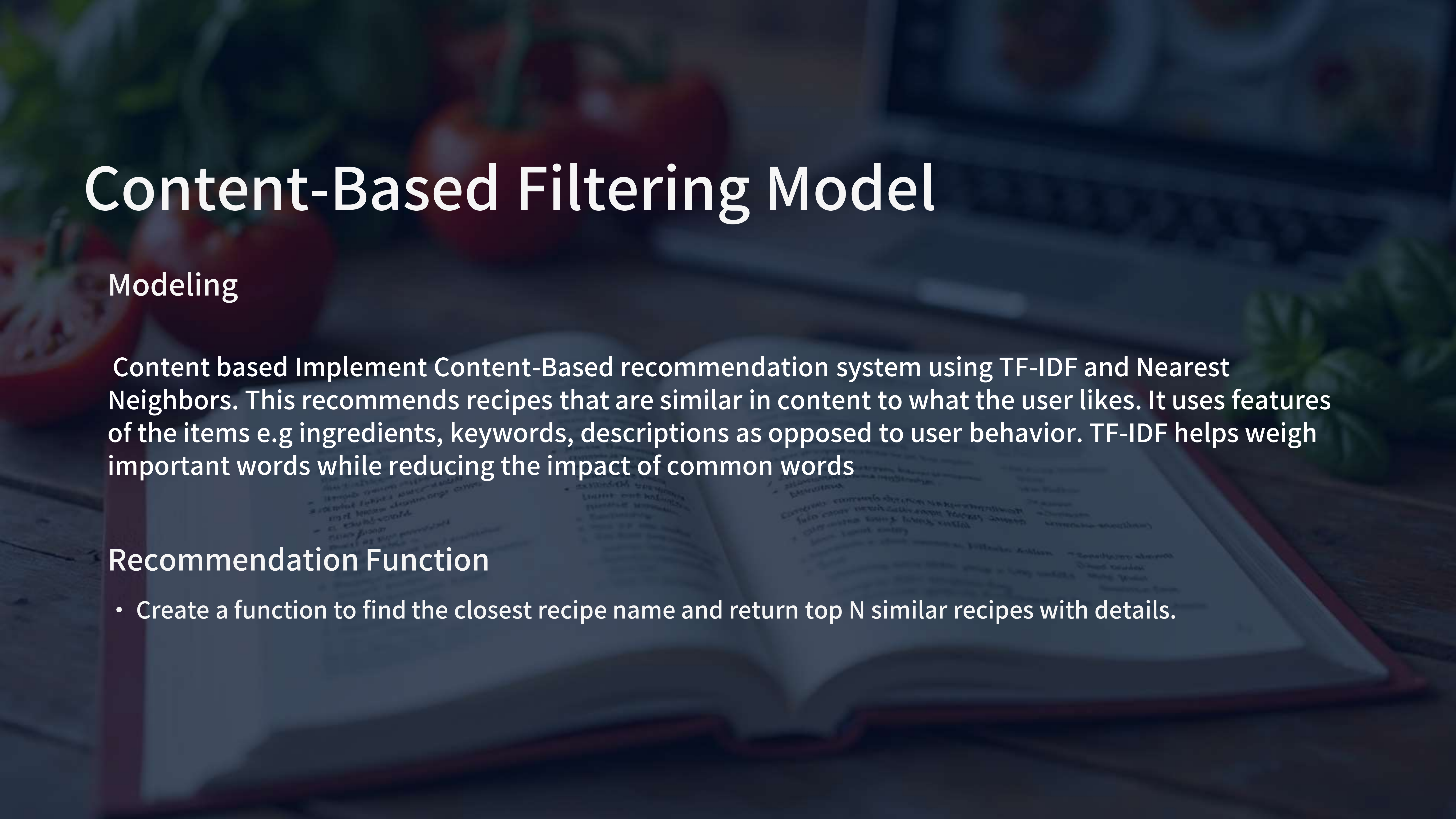
- The function predicts ratings for unrated recipes by a user and ranks them to recommend the top-N recipes tailored to the user's preferences, enhancing personalization.

Collaborative Filtering Evaluation

Evaluation Metrics

1.9881 Root Mean Squared Error (RMSE)

1.6032 Mean Absolute Error (MAE)



Content-Based Filtering Model

Modeling

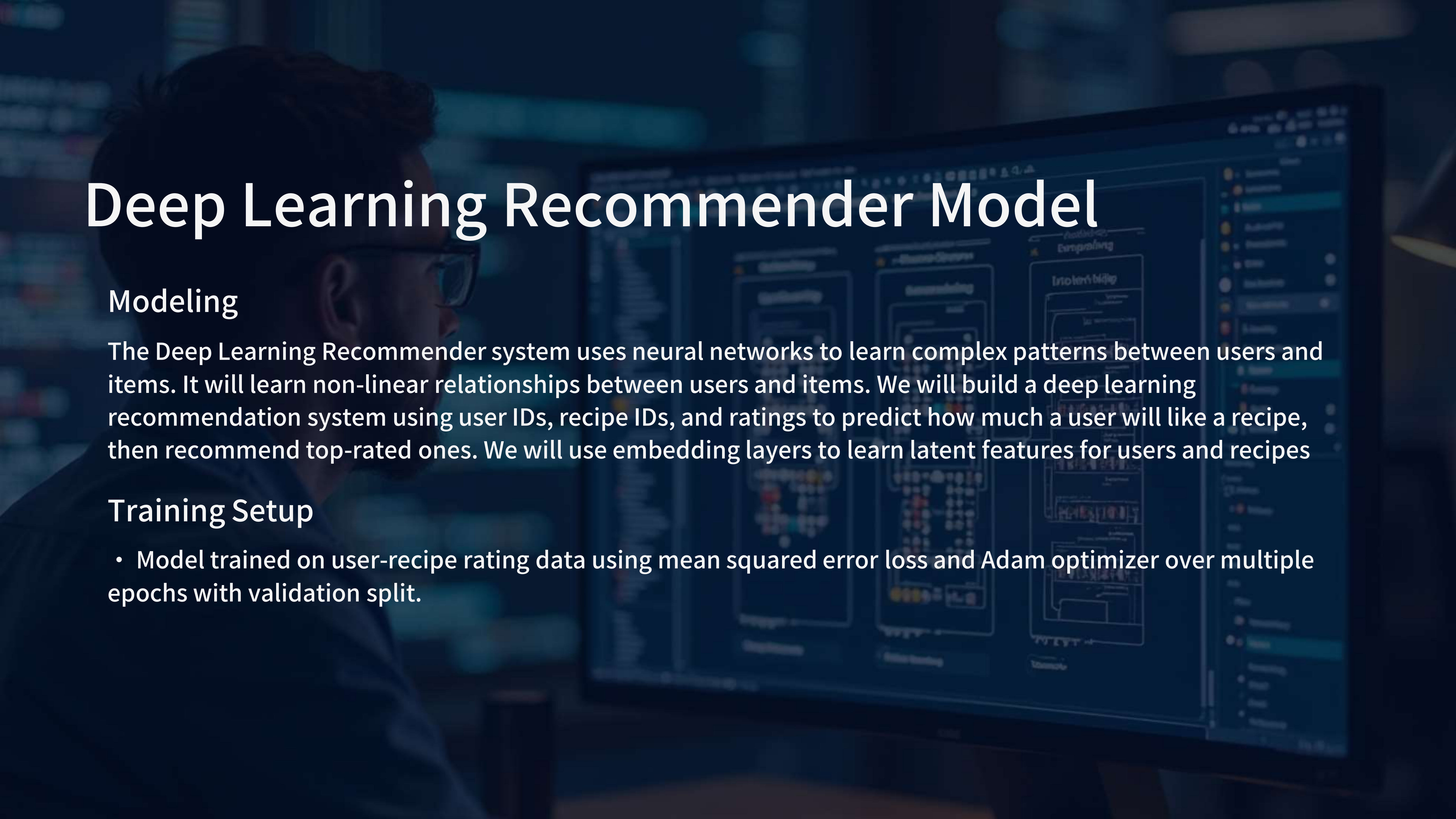
Content based Implement Content-Based recommendation system using TF-IDF and Nearest Neighbors. This recommends recipes that are similar in content to what the user likes. It uses features of the items e.g ingredients, keywords, descriptions as opposed to user behavior. TF-IDF helps weigh important words while reducing the impact of common words

Recommendation Function

- Create a function to find the closest recipe name and return top N similar recipes with details.

Content-Based Filtering Evaluation Results

The Content-Based Filtering model achieved a Precision@5 score of 0.6971, indicating strong relevance in recommendations.



Deep Learning Recommender Model

Modeling

The Deep Learning Recommender system uses neural networks to learn complex patterns between users and items. It will learn non-linear relationships between users and items. We will build a deep learning recommendation system using user IDs, recipe IDs, and ratings to predict how much a user will like a recipe, then recommend top-rated ones. We will use embedding layers to learn latent features for users and recipes

Training Setup

- Model trained on user-recipe rating data using mean squared error loss and Adam optimizer over multiple epochs with validation split.

Deep Learning Model Evaluation Metrics

RMSE

2.1537

Epochs

5 trained

Embedding size

50

Loss Type

Mean Squared Error

Hybrid Model Architecture

Modeling Steps

1

Data Preparation

- Encode user and recipe IDs as categorical variables and scale the preparation time feature for integration into the model.

2

Embedding Layers

- Create embedding layers for users and recipes to learn latent features.

3

Feature Concatenation

- Concatenate the flattened user and recipe embeddings with the scaled preparation time as an additional continuous input feature.

4

Dense Neural Network

- Feed the combined features into fully connected dense layers with activation functions to model complex interactions and predict ratings.

Hybrid Model Training and Evaluation

Hybrid Model Performance

The hybrid model showed rapid loss convergence within 5 epochs, reducing both training and validation loss consistently.

Feature Incorporation

Incorporation of prep time as a content feature enhanced the model's ability to generalize beyond user-item interactions alone.

Rating Predictions Accuracy

The RMSE of 0.3440 on test data signifies highly accurate rating predictions on a 1-5 scale, outperforming other models.

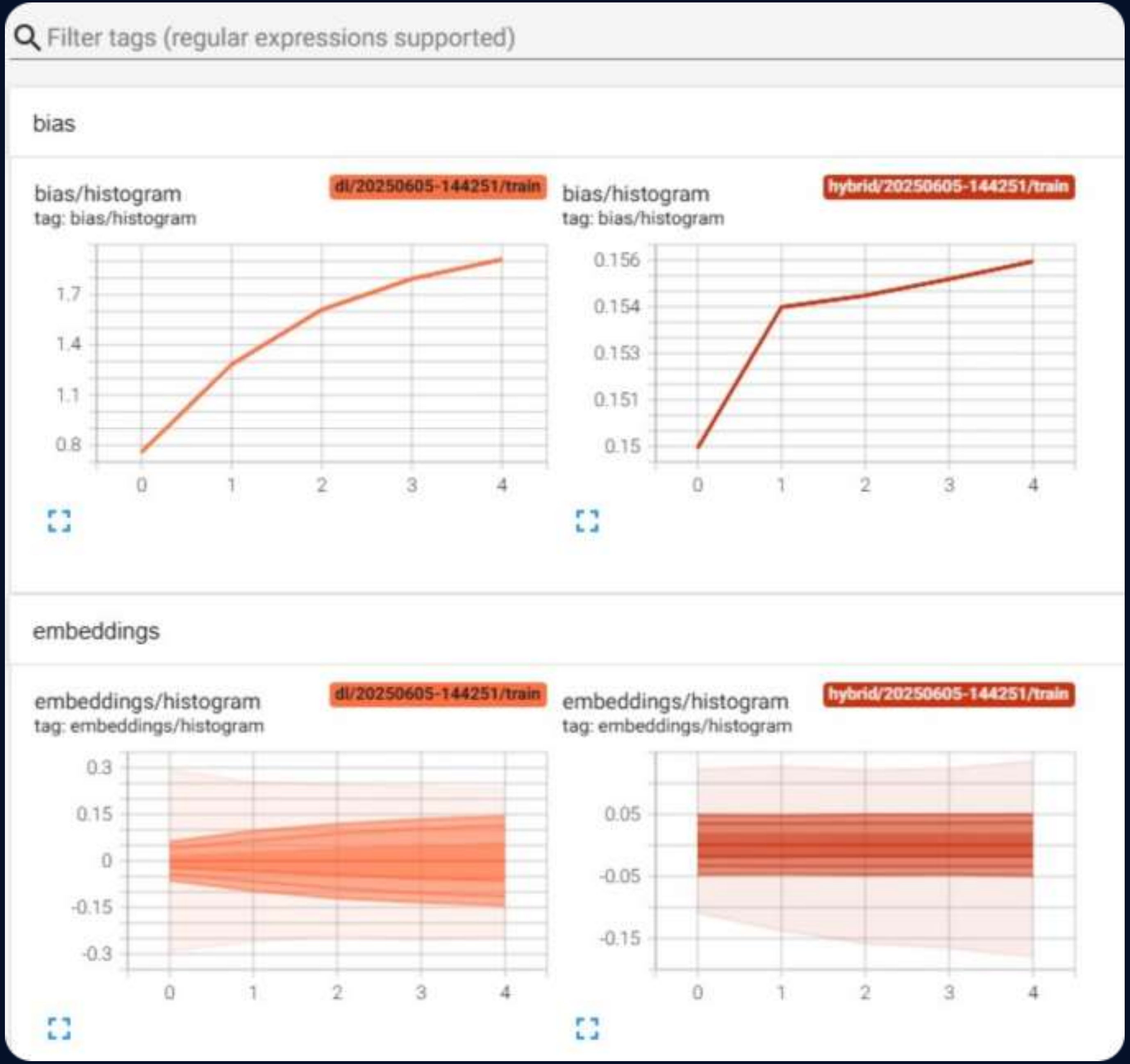
Training Dynamics

TensorBoard logs confirmed stable training dynamics and minimal overfitting, supporting robust recommendations.

Model Evaluation

Assess the performance of models using metrics like RMSE, MAE, and Precision@K to ensure the system recommends relevant and accurate recipes.

TensorBoard Insights



Loss Curves Comparison

- Hybrid model shows rapid and steady loss reduction over epochs.
- Deep learning model starts with higher loss and converges more slowly.
- Lower validation loss in hybrid model indicates better generalization.

Deployment

Conclusion and Key Findings

Recommendation Approaches

This project implemented and evaluated four recommendation approaches: collaborative filtering (SVD), content-based filtering, a deep learning model, and a hybrid model.

Model Performance

Among these, the hybrid model outperformed all others, achieving the lowest RMSE of 0.3757, indicating high accuracy and strong generalization on unseen data.

TensorBoard Visualizations

TensorBoard visualizations confirmed its stable loss convergence and balanced parameter distribution, contrasting with the overfitting tendencies of the deep learning model alone.

Collaborative Filtering Insights

Integration of Methodologies