Recipe Review Analysis and Sentiment-Driven Recommendation System

1. Introduction

This project analyzes **18,182 recipe reviews** to predict user ratings, uncover engagement patterns, and derive actionable insights for improving recipe offerings. By leveraging machine learning (ML), natural language processing (NLP), and interactive dashboards, we address critical business questions:

- Which recipes are most/least popular?
- How has user satisfaction evolved over time?
- What themes dominate user reviews?
- How can recipe quality and user experience be enhanced?

The analysis combines predictive modeling, topic mining, and visualization to guide data-driven decision-making.

---

2. Data Overview & Preprocessing

Dataset Description :UCI Repository (Download)

- **Source:** Recipe_Reviews.csv (15 columns, including recipe_name, stars, text, and user metadata).
- **Key Issues Addressed:**
    - **Missing Values:** 2 rows with missing text dropped; thumbs_up/thumbs_down filled with zeros.
    - **Class Imbalance:** 70% of reviews were 5-star ratings.
    - **Text Noise:** Special characters, numbers, and stopwords removed.

Preprocessing Steps

1. **Text Cleaning:**
    - Standardized text by removing non-alphabetic characters, converting to lowercase, and eliminating stopwords.
2. **Balancing Data:**
    - Upsampled minority classes (1–4-star ratings) to match the 5-star majority using resample.
3. **Feature Engineering:**
    - Transformed text into 5,000 TF-IDF features for ML modeling.

---

3. Exploratory Data Analysis (EDA)

Star Rating Distribution

- **Original Data:** Severe imbalance, with 5-star ratings dominating (≈12,000 instances).
- **Balanced Data:** Uniform distribution post-resampling (≈14,000 instances per class).

User Engagement

- **Light Engagement:** 80% of users contributed 1–5 reviews.
- **Power Users:** A small subset (20+ reviews) drove 30% of interactions.

Temporal Trends

- **Decline in Satisfaction:** Average ratings dropped from **4.3 (2021)** to **3.89 (2022)**, signaling potential quality or expectation mismatches

---

4. Model Development & Evaluation

Logistic Regression

- **Accuracy:** 72%
- **Weakness:** Poor recall for minority classes (e.g., 41% for 2-star ratings).

Random Forest Classifier

- **Accuracy:** 95%
- **Strengths:**
    - High precision/recall across classes (e.g., 99% precision for 3-star ratings).
    - Robust to overfitting (**cross-validation score: 95.3%**).
- **Confusion Matrix Insights:**
    - Strong 5-star prediction (2,720 correct).
    - Minor confusion between adjacent classes (e.g., 4-star vs. 5-star).

---

5. Topic Modeling & Recommendations

LDA-Driven Themes

Five key topics emerged from reviews:

1. **Family-Friendly Recipes** ("family," "love," "favorite").
2. **Baking Recipes** ("cake," "sugar," "bread").
3. **Savory Dishes** ("chicken," "soup," "cheese").
4. **Positive Sentiments** ("delicious," "tasty," "yummy").
5. **Dessert Feedback** ("pie," "apple," "sugar").

Actionable Recommendations

For Recipe Developers:

1. **Discontinue Low-Rated Recipes:**
    - **Bottom 5:** Pineapple Orange Cake (2.61 ⭐), Fluffy Key Lime Pie (2.94 ⭐).
2. **Promote Top Recipes:**
    - **Top 5:** Rustic Italian Tortellini Soup (4.73 ⭐), Corn Pudding (4.71 ⭐).
3. **Refine Recipes Using Themes:**
    - Enhance savory dishes (e.g., improve "chicken" or "sauce" components).

For Platform Managers:

1. **Reverse Satisfaction Decline:** Investigate 2021–2022 rating drop (4.3 → 3.89).
2. **Boost Engagement:** Reward users for high-quality reviews.

---

6. Interactive Dashboard

A **Streamlit** dashboard enables real-time exploration:

- **Filters:** Select recipes by name.
- **Key Features:**
    - **LDA Topic Summaries:** Top words for themes like "baking" or "desserts."
    - **Sentiment Distribution:** Star rating trends for selected recipes.
    - **Word Clouds & Top Reviews:** Highlight frequent terms and popular feedback.
    - **Popularity Rankings:** Tables of top/bottom recipes (e.g., Pumpkin Bread vs. Caramel Heavenlies).

---

7. Conclusion & Next Steps

This project delivers:

- A **95%-accurate Random Forest model** for rating prediction.
- Data-driven insights to optimize recipes and user experience.
- An **interactive dashboard** for stakeholder decision-making.

Next Steps:

1. Expand the dataset with recent reviews.
2. Deploy the dashboard for real-time use.
3. A/B test recipe improvements (e.g., tweak "sauce" in chicken dishes).