Identifying Key Entities in Recipe Data

Rudragada Sunil Kumar

Business Objective

The goal of this assignment is to train a Named Entity Recognition (NER) model using Conditional Random Fields (CRF) to extract key entities from recipe data. The model will classify words into predefined categories such as ingredients, quantities and units, enabling the creation of a structured database of recipes and ingredients that can be used to power advanced features in recipe management systems, dietary tracking apps, or e-commerce platforms.

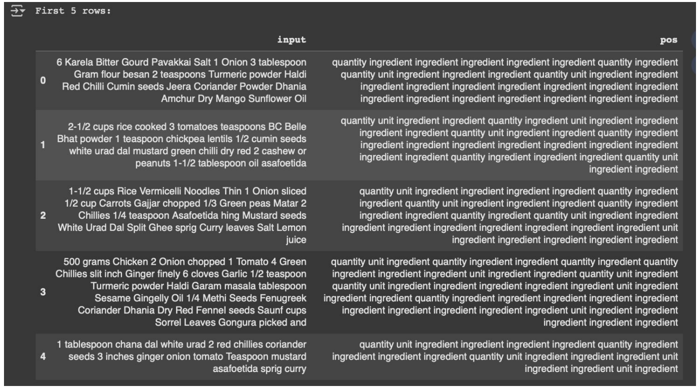
Dataset Overview

The dataset contains **285 rows** and **2 columns**:

- `input`: Raw recipe instruction text
- `pos`: Corresponding POS tags for each token (e.g., `quantity`, `unit`, `ingredient`)

The data was provided in **JSON format** and loaded using `pandas.read_json()` into a DataFrame for preprocessing.
Each row represents a complete ingredient phrase with space-separated tokens and their respective labels

Example (first 5 rows) shown in the figure below 👇:

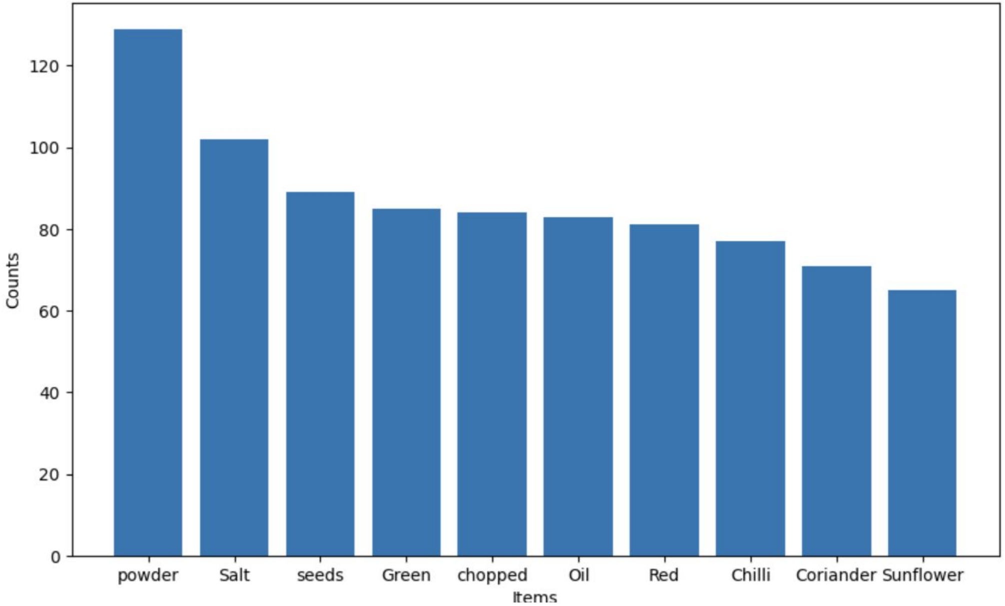


First 5 rows:

	input	pos
0	6 Karela Bitter Gourd Pavakkai Salt 1 Onion 3 tablespoon Gram flour besan 2 teaspoons Turmeric powder Haldi Red Chilli Cumin seeds Jeera Coriander Powder Dhania Amchur Dry Mango Sunflower Oil	quantity ingredient ingredient ingredient ingredient ingredient quantity ingredient quantity unit ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient
1	2-1/2 cups rice cooked 3 tomatoes teaspoons BC Belle Bhat powder 1 teaspoon chickpea lentils 1/2 cumin seeds white urad dal mustard green chilli dry red 2 cashew or peanuts 1-1/2 tablespoon oil asafoetida	quantity unit ingredient ingredient quantity ingredient unit ingredient ingredient ingredient ingredient quantity unit ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient quantity unit ingredient ingredient
2	1-1/2 cups Rice Vermicelli Noodles Thin 1 Onion sliced 1/2 cup Carrots Gajjar chopped 1/3 Green peas Matar 2 Chillies 1/4 teaspoon Asafoetida hing Mustard seeds White Urad Dal Split Ghee sprig Curry leaves Salt Lemon juice	quantity unit ingredient ingredient ingredient ingredient quantity ingredient ingredient quantity unit ingredient ingredient ingredient quantity ingredient ingredient ingredient quantity ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient
3	500 grams Chicken 2 Onion chopped 1 Tomato 4 Green Chillies slit inch Ginger finely 6 cloves Garlic 1/2 teaspoon Turmeric powder Haldi Garam masala tablespoon Sesame Gingelly Oil 1/4 Methi Seeds Fenugreek Coriander Dhania Dry Red Fennel seeds Saunf cups Sorrel Leaves Gongura picked and	quantity unit ingredient quantity ingredient ingredient quantity ingredient quantity ingredient ingredient ingredient unit ingredient ingredient quantity unit ingredient quantity unit ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient
4	1 tablespoon chana dal white urad 2 red chillies coriander seeds 3 inches ginger onion tomato Teaspoon mustard asafoetida sprig curry	quantity unit ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient unit ingredient ingredient unit ingredient

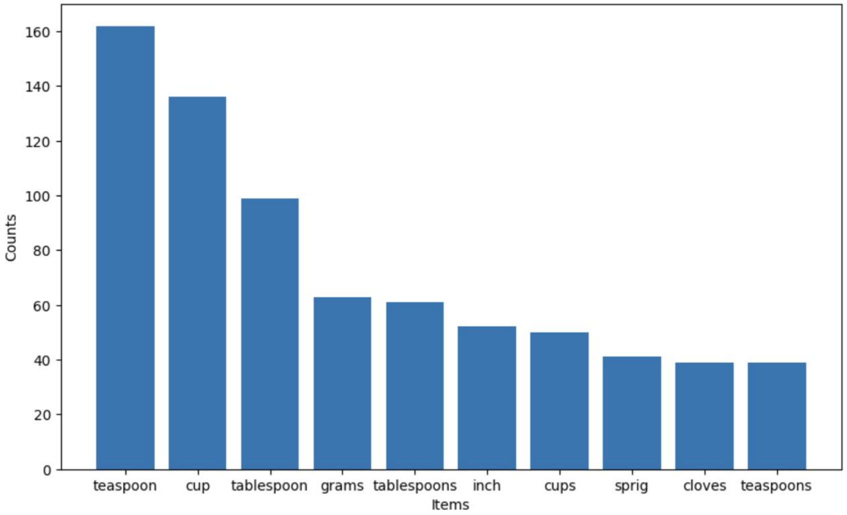
Exploratory Data Analysis (EDA On Train Ingredients)

- **"Powder"** is the most common token, likely due to frequently used ingredients like *chili powder*, *garam masala*, etc.
- High-frequency items like **"Salt"**, **"Oil"**, and **"seeds"** indicate staple ingredients commonly found in recipes.
- Tokens like **"chopped"**, **"Green"**, and **"Red"** show that **descriptive words are often labeled as ingredients**, making context features crucial for correct classification.



Exploratory Data Analysis (EDA On Train Units)

- **"teaspoon"**, **"cup"**, and **"tablespoon"** are the most commonly used measurement units, highlighting the importance of **volume-based units** in recipes.
- Both **singular and plural forms** (e.g., *tablespoon* vs. *tablespoons*, *cup* vs. *cups*) appear frequently, emphasizing the need for normalization during preprocessing.
- Units like **"inch"**, **"sprig"**, and **"cloves"** indicate that recipes also include **non-standard or descriptive units**, which require contextual understanding for accurate classification.



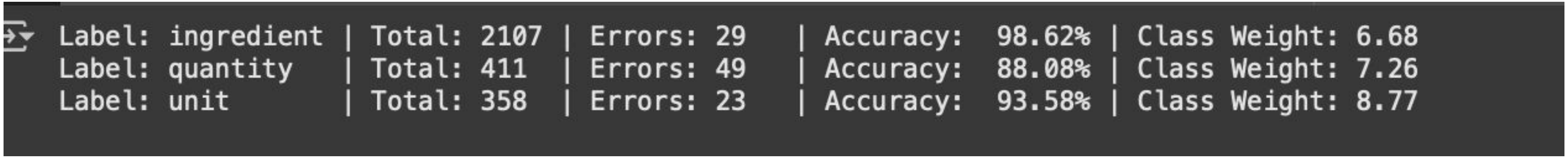
Feature Engineering

- Added token-level features: `lemma`, `pos`, `shape`, `is_digit`, etc.
- Contextual features: `prev_token`, `next_token`
- Custom features: `is_unit`, `is_quantity` using regex & keyword lists

Model Building & Evaluation

- Model used: Conditional Random Field (CRF)
- Weighted features using inverse frequency
- Validation Accuracy: **96.49%**

Below Image showing accuracy and class weight for each label

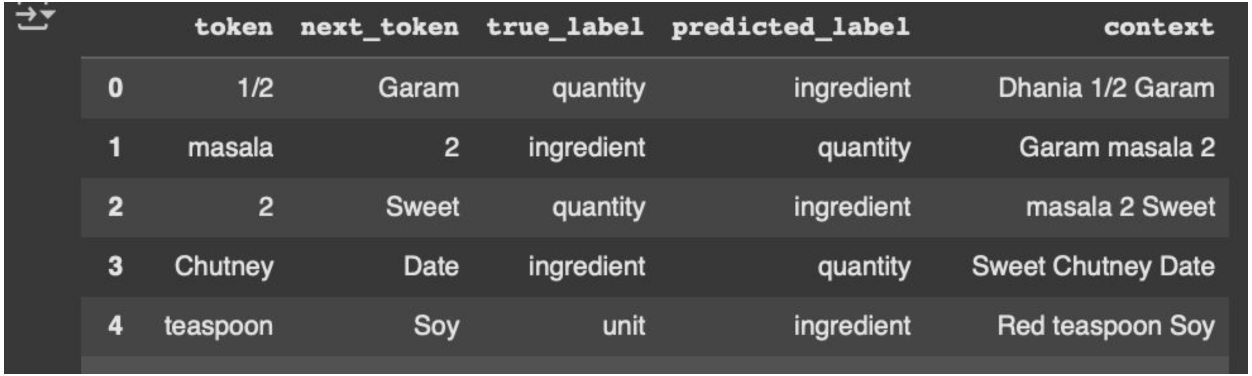


```
Label: ingredient | Total: 2107 | Errors: 29   | Accuracy:  98.62% | Class Weight: 6.68
Label: quantity   | Total: 411  | Errors: 49   | Accuracy:  88.08% | Class Weight: 7.26
Label: unit       | Total: 358  | Errors: 23   | Accuracy:  93.58% | Class Weight: 8.77
```

Error Analysis & Insights

- Most confusion: `quantity ↔ ingredient`
- Context is critical for disambiguation
- Model handles common and rare units well
- There are **101 tokens** that were **misclassified** in the validation dataset

Sample error data frame image



	token	next_token	true_label	predicted_label	context
0	1/2	Garam	quantity	ingredient	Dhania 1/2 Garam
1	masala	2	ingredient	quantity	Garam masala 2
2	2	Sweet	quantity	ingredient	masala 2 Sweet
3	Chutney	Date	ingredient	quantity	Sweet Chutney Date
4	teaspoon	Soy	unit	ingredient	Red teaspoon Soy

Conclusion

- CRF model is effective for structured NER on recipe data
- Domain-specific features and class weighting improved accuracy
- Future improvements: Handle rare units/quantities, expand dataset