Named Entity Recognition for Recipe Data Using Conditional Random Fields

- Rohit Chandel

Objective

- The goal of this assignment is to train a Named Entity Recognition (NER) model using Conditiona 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 structu database of recipes and ingredients that can be used to power advanced features in recipe management systems, dietary tracking apps, or e-commerce platforms.

Data Description

- **Data Format**: JSON with structured recipe ingredient lists
- **Dataset** :
  - Shape – (285,2)
  - input: Raw ingredient list from recipes
  - pos: Corresponding NER labels (quantity, ingredient, unit)
- **Sample Data**:

[{   "input": "6 Karela Bitter Gourd Pavakkai Salt 1 Onion 3 tablespoon Gram flour besan 2 teaspoons Turmeric powder Haldi Red Chill
seeds Jeera Coriander Powder Dhania Amchur Dry Mango Sunflower Oil",

        "pos": "quantity ingredient ingredient ingredient ingredient ingredient quantity ingredient quantity unit ingredient ingredient ingredie
quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingred
ingredient ingredient ingredient ingredient"

   }, {

      "input": "2-1/2 cups rice cooked 3 tomatoes teaspoons BC Belle Bhat powder 1 teaspoon chickpea lentils 1/2 cumin seeds white urad d
mustard green chilli dry red 2 cashew or peanuts 1-1/2 tablespoon oil asafoetida",

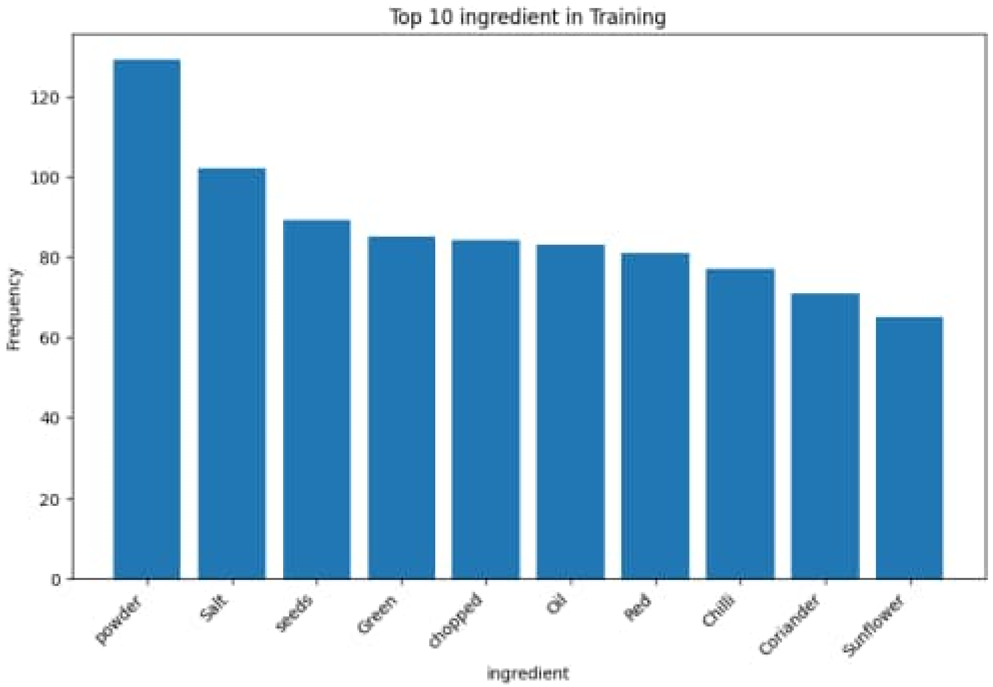
      "pos": "quantity unit ingredient ingredient quantity ingredient unit ingredient ingredient ingredient ingredient quantity unit ingredient i
quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient in
ingredient quantity unit ingredient ingredient"   }]

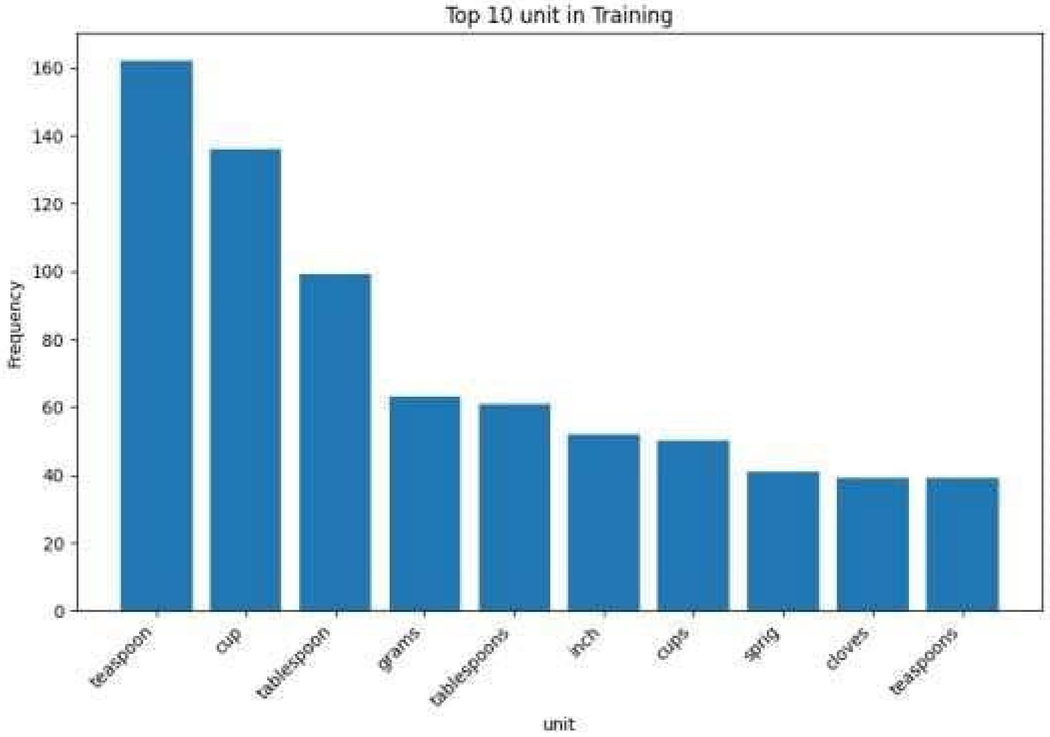
Data Preparation & Cleaning

- **Data Validation**:
  - Checked alignment between input tokens and POS labels
  - Removed misaligned records
  - Final dataset shape after cleaning - (280, 6)
- **Data Split**: 70% training, 30% validation
- **Label Distribution**: Three main entity types
  - Ingredient
  - Quantity
  - unit

Exploratory Data Analysis

- Identified top 10 ingredients and units in the training data



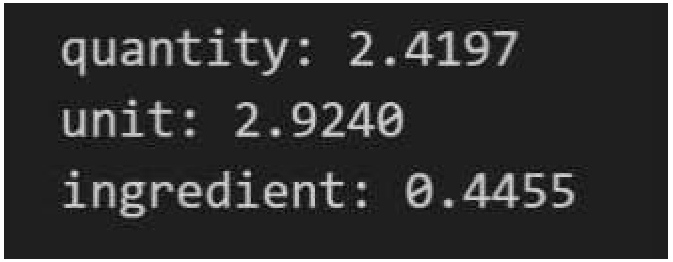


Feature Engineering Strategy

- **Multi-Layered Feature Architecture (27+ Features)**
- **Core Linguistic Features (16)**
  - **spaCy Integration**: Token, lemma, POS tags, dependency parsing
  - **Character Analysis**: Digits, alpha, hyphens, slashes, case patterns
  - **Shape & Structure**: Token shape, punctuation, stopwords
- **Recipe-Specific Features (7)**
  - **Domain Keywords**: 45 units + 62 quantities
  - **Quantity Patterns**: `^\d+$|^\d+\.\d+$|^\d+\/\d+$|^\d+-\d+\/\d+$`
  - **Entity Detection**: is_quantity, is_unit, is_numeric, is_fraction
- **Contextual Features (8)**
  - **Sequential Context**: Previous/next token analysis
  - **Boundary Markers**: Beginning/End of sequence (BOS/EOS)
  - **Neighborhood Intelligence**: Adjacent entity type prediction

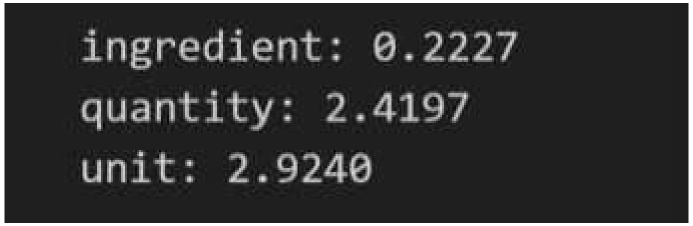
Class Imbalance Handling

- Observed unequal distribution of entity types based on weighted class method



```
quantity: 2.4197
unit: 2.9240
ingredient: 0.4455
```

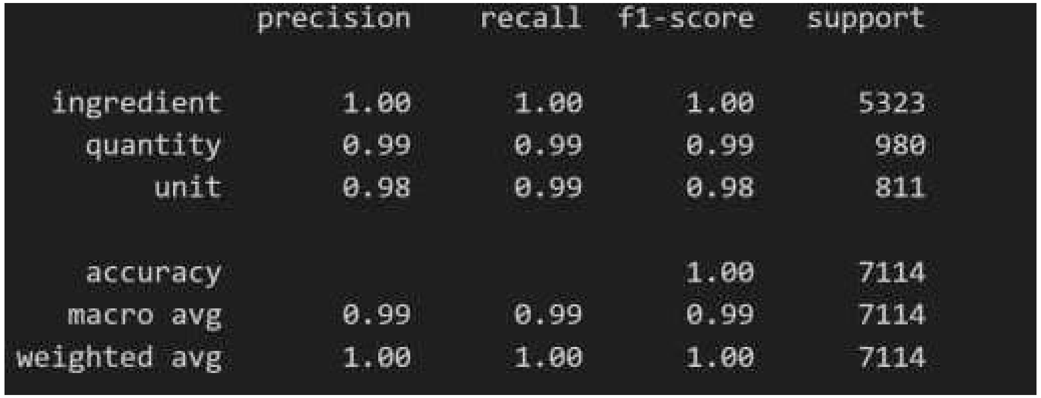
- **Ingredient Penalization**: Applied 0.5x weight to ingredient class



```
ingredient: 0.2227
quantity: 2.4197
unit: 2.9240
```

Model Architecture & Training

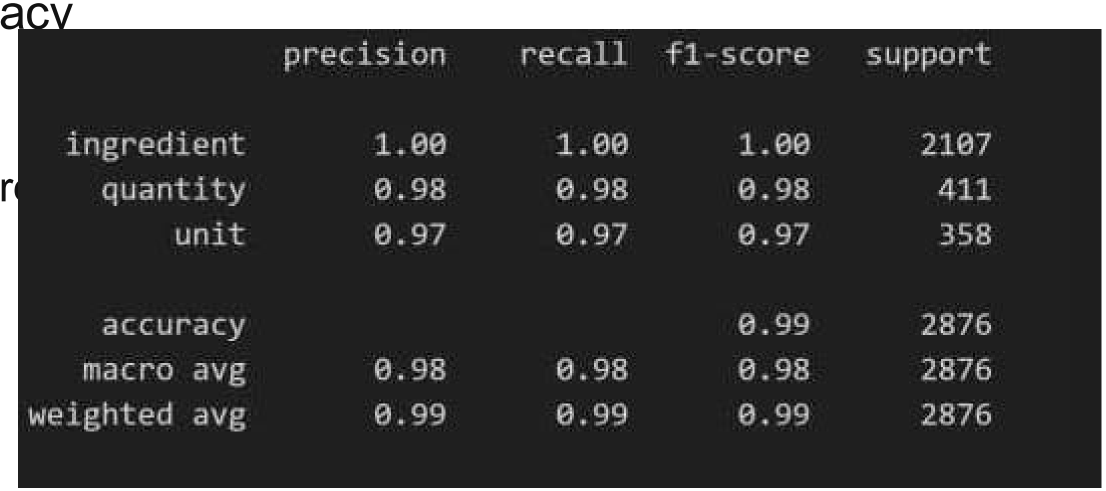
- **Algorithm**: Conditional Random Fields (CRF)
- **Hyperparameters**:
  - Algorithm: L-BFGS
  - L1 regularization (c1): 0.5
  - L2 regularization (c2): 1.0
  - Max iterations: 100
- **Training Strategy**: Weighted feature extraction with class weights
- **Training Accuracy**: High training accuracy observed



```
              precision    recall  f1-score   support

  ingredient       1.00      1.00      1.00      5323
    quantity       0.99      0.99      0.99       980
        unit       0.98      0.99      0.98       811

    accuracy                           1.00      7114
   macro avg       0.99      0.99      0.99      7114
weighted avg       1.00      1.00      1.00      7114
```

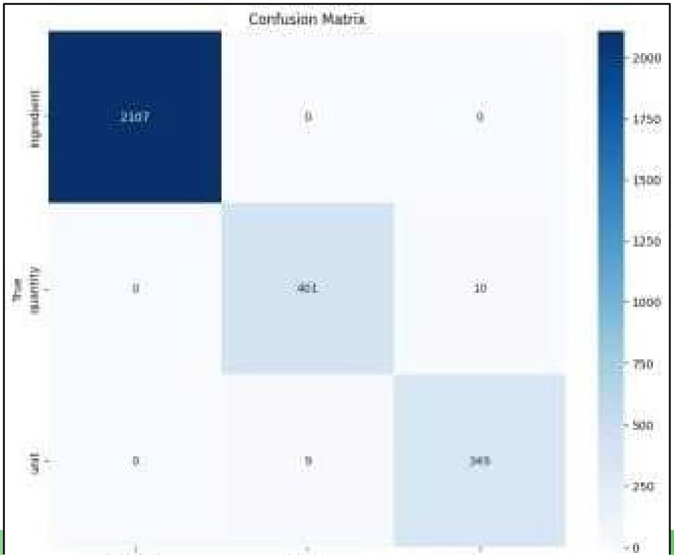
Model Performance - Validation Results

- **Validation Accuracy**: 99% overall accuracy
- **Per-Class Performance**:
  - Ingredients: 100% precision, recall, F1-score
  - Quantities: 98% across all metrics
  - Units: 97% precision, recall, F1-score



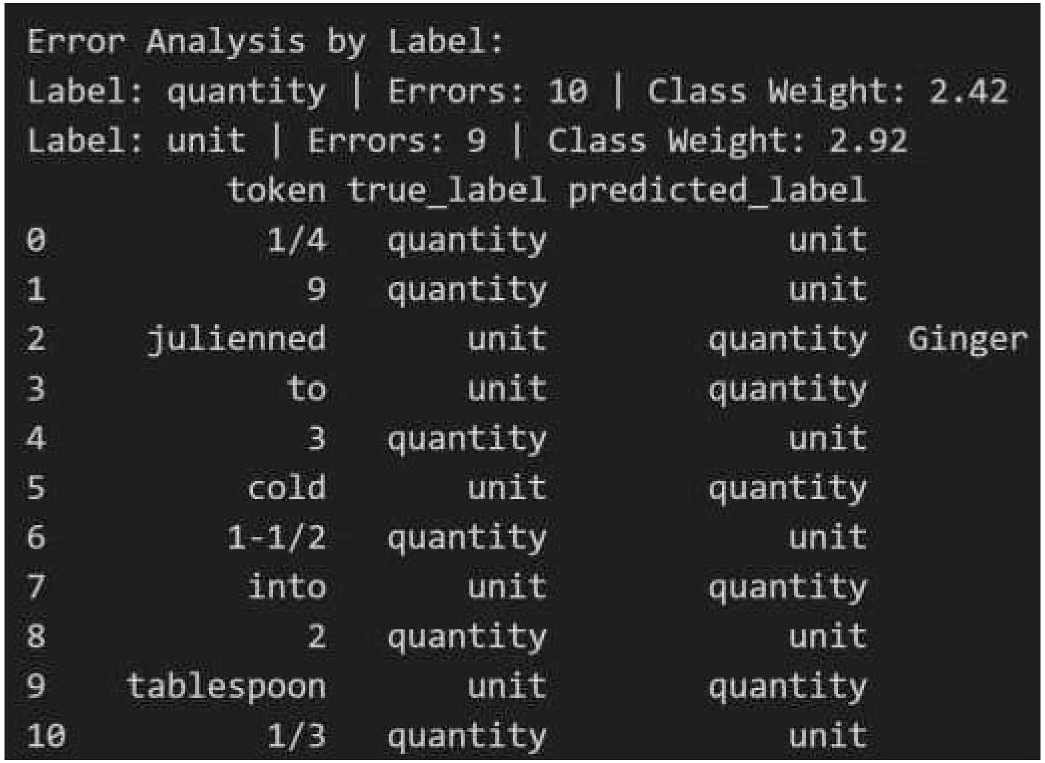
```
              precision    recall  f1-score   support

  ingredient       1.00      1.00      1.00      2107
    quantity       0.98      0.98      0.98       411
        unit       0.97      0.97      0.97       358

    accuracy                           0.99      2876
   macro avg       0.98      0.98      0.98      2876
weighted avg       0.99      0.99      0.99      2876
```

- **Validation Confusion Matrix**:



Error Analysis

- **Error Rate**: Only 1% error rate on validation data
- **Error Distribution by Label and Sample Misclassifications**:
- Total errors 19/280 (~1%)



```
Error Analysis by Label:
Label: quantity | Errors: 10 | Class Weight: 2.42
Label: unit | Errors: 9 | Class Weight: 2.92
          token true_label predicted_label
0           1/4   quantity            unit
1             9   quantity            unit
2     julienned       unit        quantity  Ginger
3            to       unit        quantity
4             3   quantity            unit
5          cold       unit        quantity
6         1-1/2   quantity            unit
7          into       unit        quantity
8             2   quantity            unit
9    tablespoon       unit        quantity
10          1/3   quantity            unit
```

Key Insights & Findings

- **Model Effectiveness**: CRF with comprehensive features works well for recipe NER
- **Feature Importance**: Domain-specific patterns and contextual information crucial
- **Class Weighting Success**: Effective handling of imbalanced data
- **Generalization**: Strong performance across all entity types
- **Success Metrics**: 99% accuracy achieved on validation data
- **Robust Performance**: Consistent results across all entity types