Analysis Report: Recipe Entity Extraction using CRF

This report presents the complete analysis, visualizations, insights, and outcomes extracted from the Jupyter Notebook titled 'Identifying_Key_Entities_Recipe.ipynb'.

 # Identifying Key Entities in Recipe Data

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.

Diagrams, Insights, Findings and Conclusion:

Based on the analysis of the Conditional Random Fields (CRF) model trained for recipe ingredient line parsing, the following findings and conclusions can be drawn:

Findings:

1. **Overall Performance:** The model achieved a high overall token-level accuracy of approximately 97.95% on the validation dataset. While this indicates a generally strong performance, the misclassification analysis reveals areas for improvement.
2. **Label-Specific Performance:** The flat classification report and the analysis by true label highlight significant differences in performance across different entity types (labels):
    - The 'unit' label has the lowest accuracy (recall) at around 89.39%, accounting for a large portion of the misclassification errors (38 out of 59 errors). This suggests the model struggles more with correctly identifying units compared to quantities or ingredients.
    - 'Quantity' has a higher accuracy (~98.30%),
    - and 'ingredient' has the highest accuracy (~99.34%).
    - The higher misclassification count for 'unit', despite having a moderate frequency, indicates that differentiating units from other tokens, particularly ingredients, is a key challenge.
3. **Common Error Patterns:**
    - The most frequent misclassification is predicting 'ingredient' when the true label is 'unit' (38 occurrences). This suggests that some tokens, particularly those that can function as units or ingredient names (e.g., "cloves", "tsp"), or units that appear in less standard contexts, are being incorrectly tagged as ingredients.
    - Errors also occur in the reverse direction ('ingredient' -> 'unit', 11 occurrences), and between 'quantity' and 'ingredient' ('quantity' -> 'ingredient', 5 occurrences).

    - Specific tokens like 'cloves', 'tsp', and 'few' are frequently involved in misclassifications, likely due to their potential ambiguity or context-dependent labeling.

4. **Influence of Class Weights:** The analysis table includes the class weights applied during training. While 'ingredient' has the highest frequency and the lowest weight (due to penalization), it still achieves the highest accuracy. 'unit' has a moderate frequency and a higher weight, yet its accuracy is the lowest. This suggests that while weighting helps, it may not fully compensate for ambiguities or complexities inherent in certain labels or tokens. The features themselves might be insufficient to distinguish 'unit' tokens reliably in all contexts.
5. **Types of Errors:** Observed error types align with common challenges in sequence labeling: boundary errors (misclassifying tokens at the transition points between entities), ambiguity of tokens, errors on less frequent labels or tokens, and difficulty with complex or unusual phrasing.

Conclusion:

The trained CRF model is a solid starting point for recipe ingredient line parsing, demonstrating high overall accuracy. However, the error analysis reveals that the model's performance is significantly impacted by the challenge of accurately identifying 'unit' tokens. Misclassifications between 'unit' and 'ingredient' are the most prominent issue.

To improve the model's performance, especially for the 'unit' label, future work should focus on:

1. **Enhanced Feature Engineering:** Develop features that better capture the context distinguishing units from ingredients, potentially including more specific patterns, n-grams, or dictionary lookups for units.
2. **Data Quality and Quantity:** Review and potentially augment the training data for 'unit' labels, ensuring comprehensive coverage of unit variations and contexts, including edge cases.
3. **Model Exploration:** While CRF is effective, exploring models like Bi-LSTM-CRF might help in capturing longer-range dependencies and more complex contextual patterns that could improve performance on challenging labels like 'unit'.

The error analysis provides a clear roadmap for targeted improvements, allowing for focused effort on the labels and tokens where the model currently struggles the most.

Insights from the Validation Dataset Error Analysis ---

- The model had an overall token-level accuracy of 0.9795 on the validation set, misclassifying 59 out of 2876 tokens.
- The misclassification analysis per true label (shown in the table above) highlights that certain labels are more prone to errors than others.

Labels with the lowest accuracy (most errors relative to their frequency): true_label total_in_validation misclassification_count accuracy_for_label class_weight 1 unit 358 38 0.893855 8.771887 0 quantity 411 7 0.982968 7.259184 2 ingredient 2107 14 0.993355 0.334116

- Common misclassification patterns (True Label -> Predicted Label) include: 'unit'->'ingredient' (38), 'ingredient'->'unit' (11), 'quantity'->'ingredient' (5).
- Tokens most frequently involved in misclassifications are: 'cloves' (7), 'tsp' (4), 'few' (3).

Specific types of errors observed (based on sample inspection and common patterns):

- Ambiguity: Tokens that can have different meanings depending on context (e.g., 'powder', 'extract', numbers used as names).
- Boundary Errors: Misclassifying the first or last token of a sequence (e.g., mislabeling the unit or the start of the ingredient).
- Rare Labels/Tokens: Labels or tokens that appear infrequently in the training data might be harder to predict accurately.
- Complex Phrases: Multi-word ingredients or descriptors can sometimes cause the model to incorrectly tag individual words.
- Insufficient Contextual Features: The current features might not fully capture complex dependencies between tokens or long-range relationships within the recipe line.

Recommendations for Improvement:

- Feature Engineering: Explore additional features, such as surrounding tokens (beyond immediate previous/next), n-grams, or external lexical resources/dictionaries.
- Data Augmentation: Increase the diversity and quantity of training data, especially for less frequent labels or challenging phrasing.
- Model Hyperparameter Tuning: Experiment with different CRF hyperparameters (c1, c2, max_iterations) to potentially improve convergence and generalization.
- Explore Different Models: Consider alternative sequence labeling models like Bi-LSTM-CRF which can better capture long-range dependencies.
- Error-Specific Handling: If certain error patterns are very frequent, investigate specific rules or features to address them.

Insights from the Validation Dataset Error Analysis ---

- The model had an overall token-level accuracy of 0.9795 on the validation set, misclassifying 59 out of 2876 tokens.

- The misclassification analysis per true label (shown in the table above) highlights that certain labels are more prone to errors than others.

Labels with the lowest accuracy (most errors relative to their frequency):

	true_label	total_in_validation	misclassification_count	accuracy_for_label	class_weight
1	unit	358	38	0.893855	8.771887
0	quantity	411	7	0.982968	7.259184
2	ingredient	2107	14	0.993355	0.334116

- Common misclassification patterns (True Label -> Predicted Label) include: 'unit'->'ingredient' (38), 'ingredient'->'unit' (11), 'quantity'->'ingredient' (5).

- Tokens most frequently involved in misclassifications are: 'cloves' (7), 'tsp' (4), 'few' (3).

Specific types of errors observed (based on sample inspection and common patterns):

- Ambiguity: Tokens that can have different meanings depending on context (e.g., 'powder', 'extract', numbers used as names).

- Boundary Errors: Misclassifying the first or last token of a sequence (e.g., mislabeling the unit or the start of the ingredient).

- Rare Labels/Tokens: Labels or tokens that appear infrequently in the training data might be harder to predict accurately.

- Complex Phrases: Multi-word ingredients or descriptors can sometimes cause the model to incorrectly tag individual words.

- Insufficient Contextual Features: The current features might not fully capture complex dependencies between tokens or long-range relationships within the recipe line.

Recommendations for Improvement:

- Feature Engineering: Explore additional features, such as surrounding tokens (beyond immediate previous/next), n-grams, or external lexical resources/dictionaries.

- Data Augmentation: Increase the diversity and quantity of training data, especially for less frequent labels or challenging phrasing.

- Model Hyperparameter Tuning: Experiment with different CRF hyperparameters (c1, c2, max_iterations) to potentially improve convergence and generalization.

- Explore Different Models: Consider alternative sequence labeling models like Bi-LSTM-CRF which can better capture long-range dependencies.

- Error-Specific Handling: If certain error patterns are very frequent, investigate specific rules or features to address them.

Making predictions on the training dataset...

Predictions on training dataset completed.

Flat Classification Report on Training Dataset:

	precision	recall	f1-score	support
quantity	0.997	0.980	0.988	980
unit	0.981	0.957	0.969	811
ingredient	0.990	0.997	0.993	5323
accuracy			0.990	7114
macro avg	0.989	0.978	0.983	7114
weighted avg	0.990	0.990	0.990	7114

Confusion Matrix on Training Dataset:

	ingredient	quantity	unit
ingredient	5305	3	15
quantity	20	960	0
unit	35	0	776

Generating Flat Classification Report for the TRAINING dataset:

	precision	recall	f1-score	support
ingredient	0.990	0.997	0.993	5323
quantity	0.997	0.980	0.988	980
unit	0.981	0.957	0.969	811
accuracy			0.990	7114
macro avg	0.989	0.978	0.983	7114
weighted avg	0.990	0.990	0.990	7114

Flat Classification Report for the TRAINING dataset has been generated.

Making predictions on the training dataset...

Predictions on training dataset completed.

Computing Confusion Matrix on Training Dataset...

Confusion Matrix computation completed.

Confusion Matrix (Training Dataset):

	ingredient	quantity	unit
ingredient	5305	3	15
quantity	20	960	0
unit	35	0	776

Making predictions on the validation dataset...

Predictions on validation dataset completed.

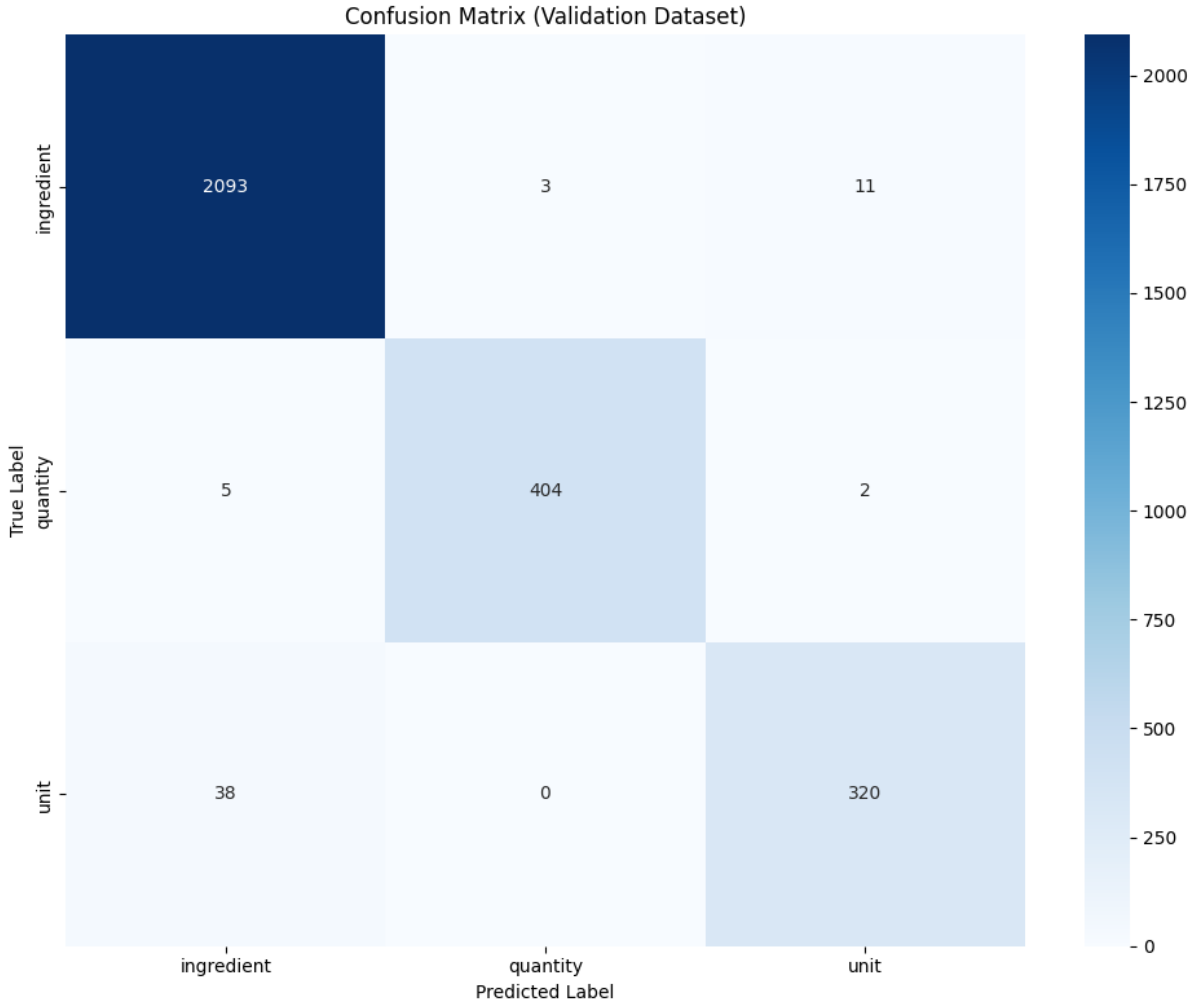
Flat Classification Report on Validation Dataset:

	precision	recall	f1-score	support
ingredient	0.980	0.993	0.987	2107
quantity	0.993	0.983	0.988	411
unit	0.961	0.894	0.926	358
accuracy			0.979	2876
macro avg	0.978	0.957	0.967	2876
weighted avg	0.979	0.979	0.979	2876

Confusion Matrix on Validation Dataset:

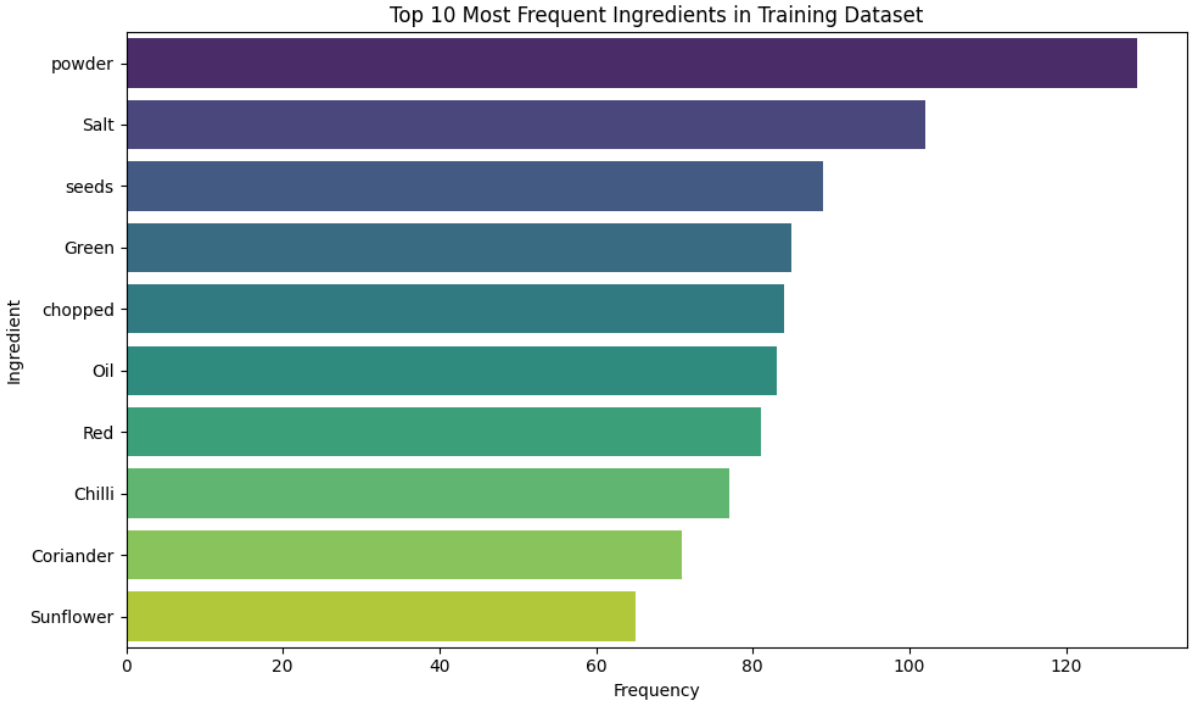
	ingredient	quantity	unit
ingredient	2093	3	11
quantity	5	404	2
unit	38	0	320

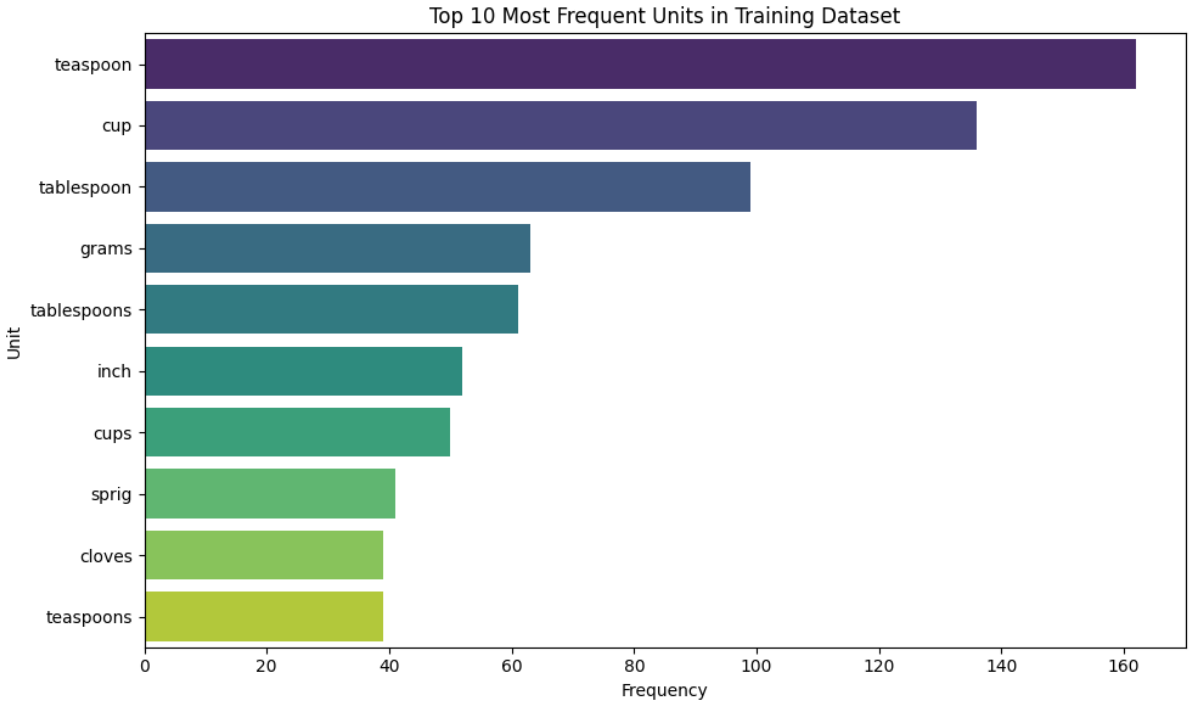
Plotting Confusion Matrix for Validation Dataset...

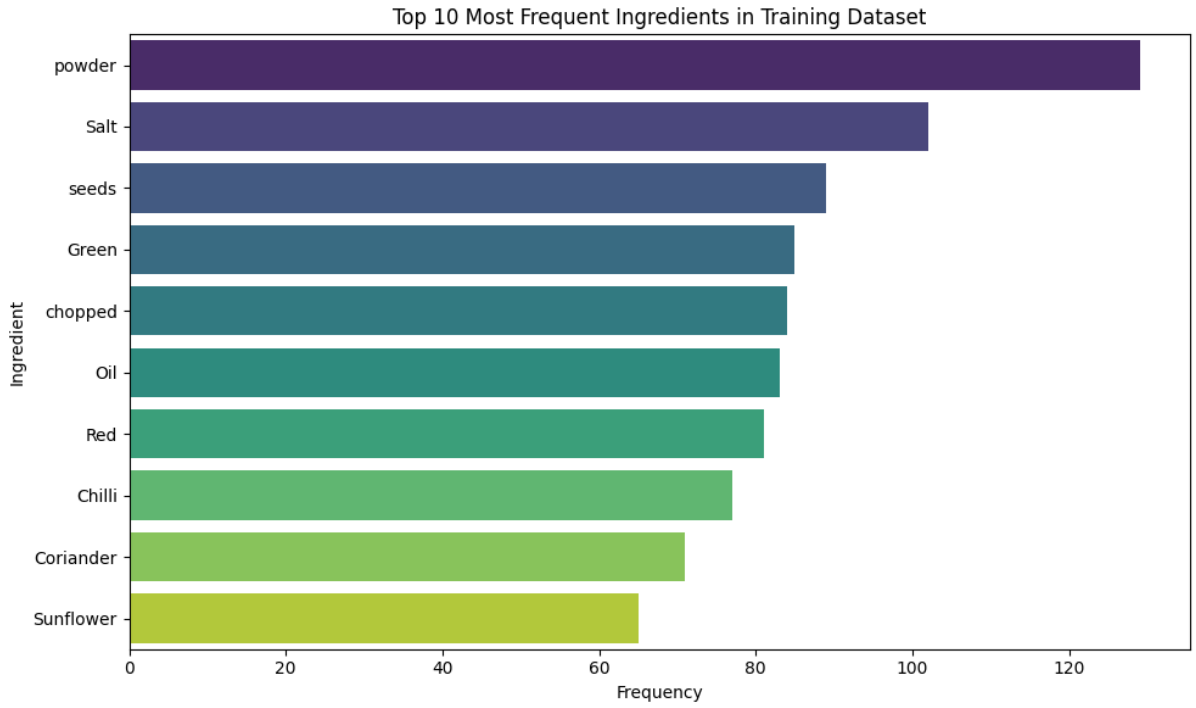


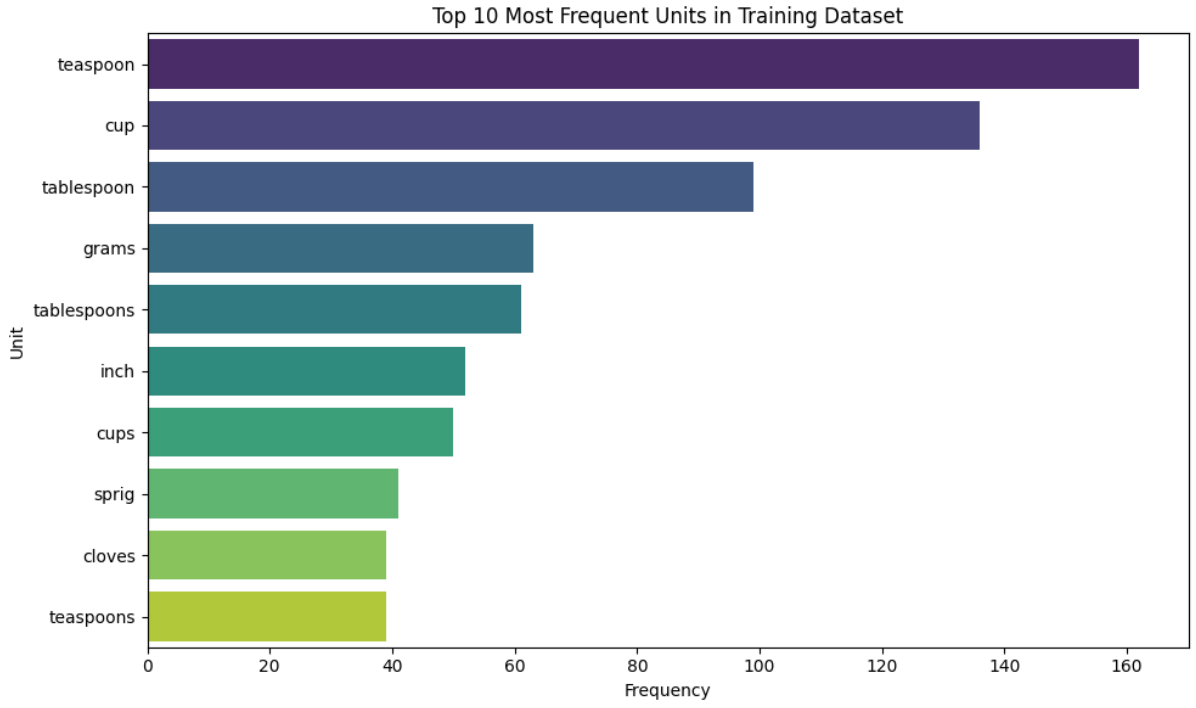
Identifying misclassified samples in the validation dataset...

Found 29 misclassified recipe samples out of 84.


Top 10 Most Frequent Ingredients in Training Dataset
powder
Salt
seeds
Green
chopped
Oil
Red
Chilli
Coriander
Sunflower
Ingredient
0
20
40
60
80
100
120
Frequency


Top 10 Most Frequent Units in Training Dataset
teaspoon
cup
tablespoon
grams
tablespoons
inch
cups
sprig
cloves
teaspoons
Unit
0
20
40
60
80
100
120
140
160
Frequency





Insights from the EDA on the Training Dataset:

- Ingredients: The plots show that 'powder', 'salt', 'oil', and 'water' are among the most frequently mentioned ingredients.

This indicates that these are common staple ingredients found across a wide variety of recipes in the dataset.

- Units: The most frequent units include 'teaspoon(s)', 'cup(s)', and 'tablespoon(s)'.

This suggests that volumes are the most commonly used measurement types in these recipes.

Other units like 'inch(es)', 'piece(s)', and weight units ('gram(s)', 'kg') appear less frequently in the top 10.

- Both plots highlight the distribution of frequency, with a few items being significantly more common than others.

This suggests that while there's a diverse set of ingredients and units, a core set dominates the data.

- Understanding these frequently occurring entities is crucial for building a robust NER model,

as the model will encounter these tokens more often during training.

It might also inform feature engineering, such as creating features for common units or ingredient types.

The lower frequency of other entities suggests that the model might struggle with less common ingredients or units if not adequately represented or if specific features aren't designed to handle them.

Top 10 Most Frequent Units in Validation Dataset ---

teaspoon: 59

cup: 57

tablespoon: 32

tablespoons: 32

cups: 24

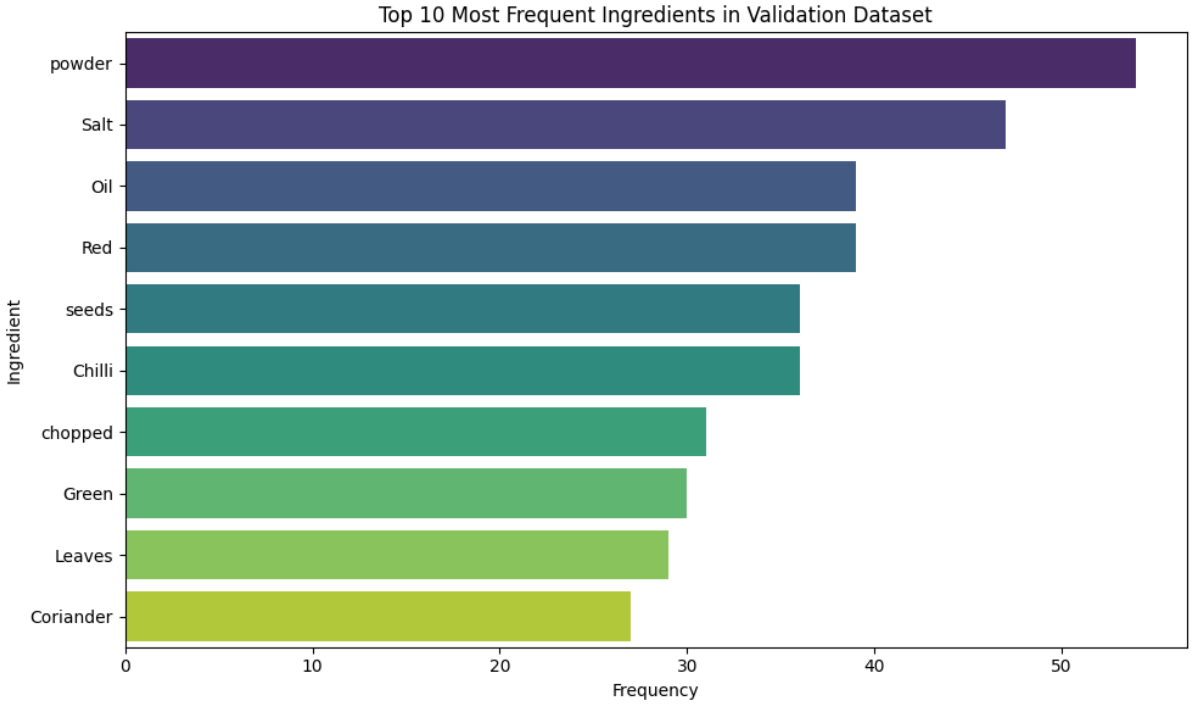
sprig: 21

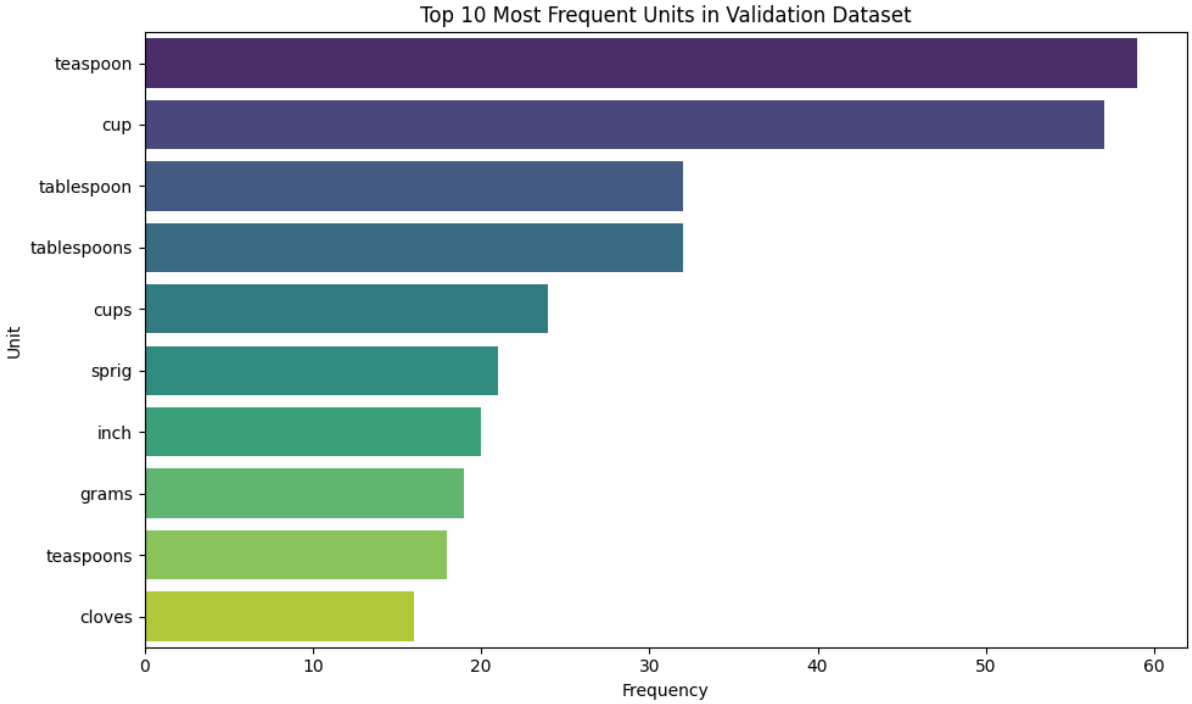
inch: 20

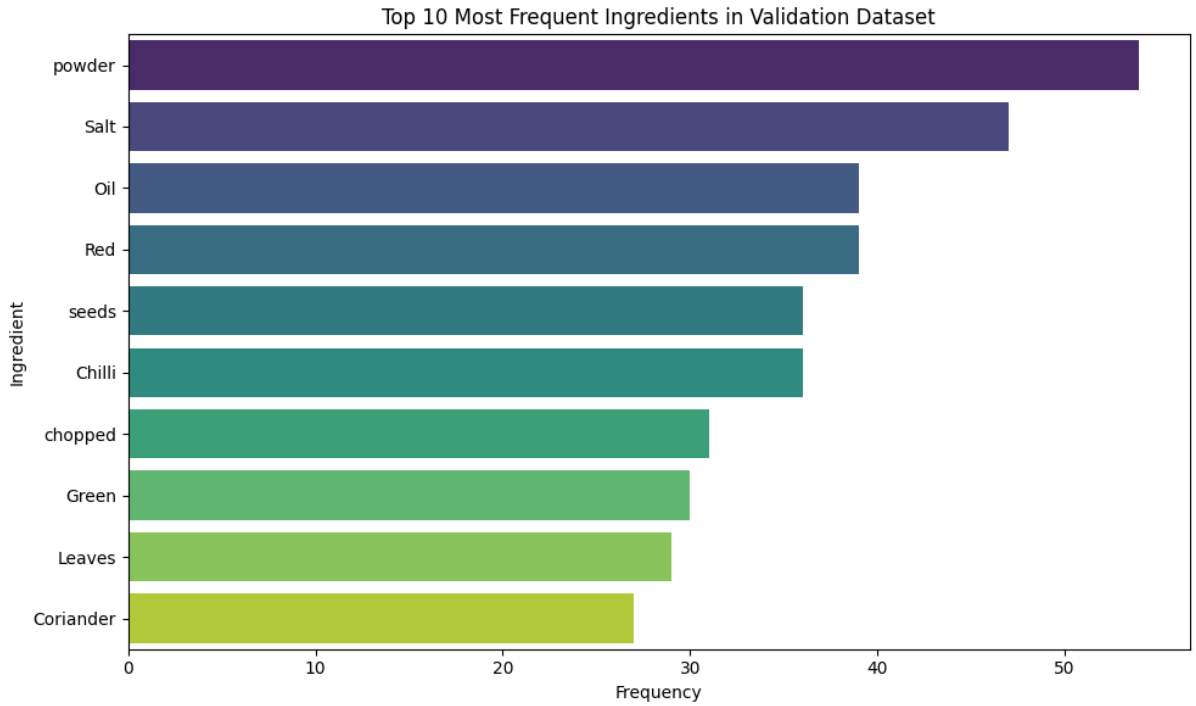
grams: 19

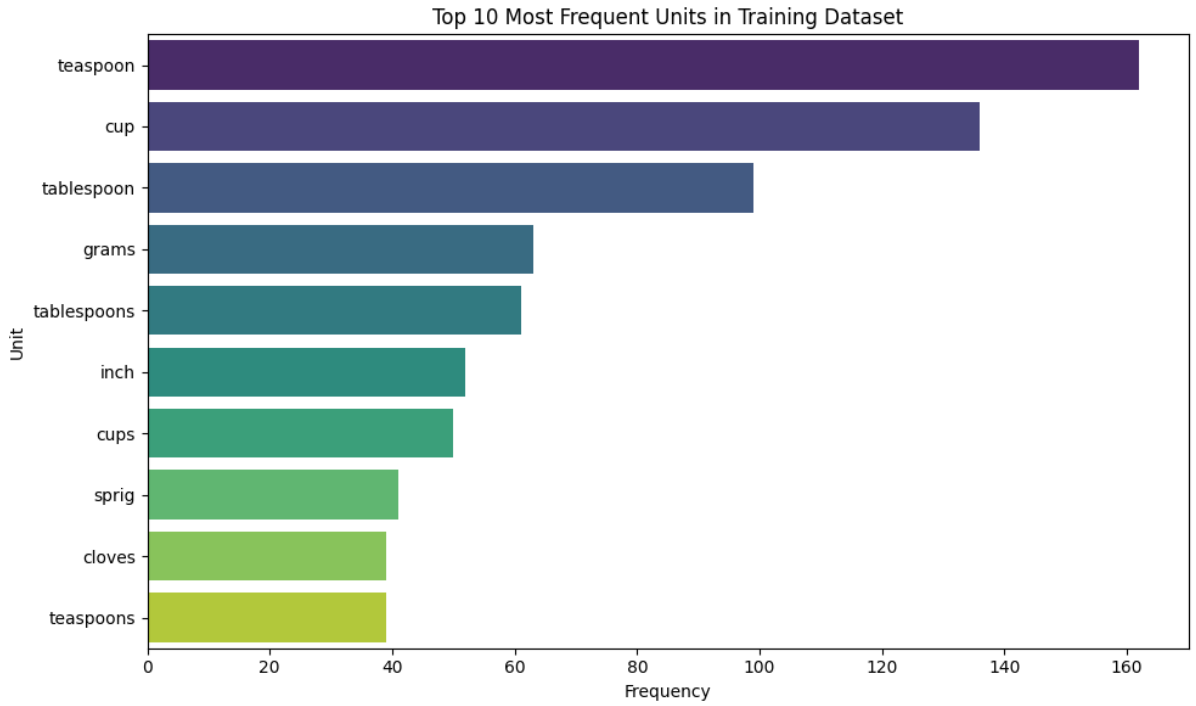
teaspoons: 18

cloves: 16


Top 10 Most Frequent Ingredients in Validation Dataset
powder
Salt
Oil
Red
seeds
Chilli
chopped
Green
Leaves
Coriander
Ingredient
0
10
20
30
40
50
Frequency


Top 10 Most Frequent Units in Validation Dataset
teaspoon
cup
tablespoon
tablespoons
cups
sprig
inch
grams
teaspoons
cloves
Unit
0
10
20
30
40
50
60
Frequency





Investigating details of misclassified samples:

--- Misclassified Sample 1 (Original Index: 2) ---

Tokens: 1 tablespoon Sunflower Oil 3 Potato Aloo Ginger paste Green Chilli chopped 1-1/12 tablespoons Sesame seeds Til teaspoon Red powder Cumin Jeera Coriander Powder Dhania 1/2 Garam masala 2 Sweet Chutney Date Tamarind Leaves few

True Labels: quantity unit ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient

Predicted Labels: quantity unit ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient quantity

Differences: 1 (quantity) tablespoon (unit) Sunflower (ingredient) Oil (ingredient) 3 (quantity) Potato (ingredient) Aloo (ingredient) Ginger (ingredient) paste (ingredient) Green (ingredient) Chilli (ingredient) chopped (ingredient) 1-1/12 (quantity) tablespoons (unit) Sesame (ingredient) seeds (ingredient) Til (ingredient) teaspoon (unit) Red (ingredient) powder (ingredient) Cumin (ingredient) Jeera (ingredient) Coriander (ingredient) Powder (ingredient) Dhania (ingredient) 1/2 (quantity) Garam (ingredient) masala (ingredient) 2 (quantity) Sweet (ingredient) Chutney (ingredient) Date (ingredient) Tamarind (ingredient) Leaves (ingredient) few (ingredient -> quantity)

--- Misclassified Sample 2 (Original Index: 3) ---

Tokens: 1 cup green peas gram flour 1/2 cheese tsp ginger 2 chillies turmeric powder cumin teaspoon salt oil

True Labels: quantity unit ingredient ingredient ingredient ingredient quantity ingredient unit ingredient quantity ingredient ingredient ingredient ingredient unit ingredient ingredient

Predicted Labels: quantity unit ingredient ingredient unit ingredient quantity ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient unit ingredient ingredient

Differences: 1 (quantity) cup (unit) green (ingredient) peas (ingredient) gram (ingredient -> unit) flour (ingredient) 1/2 (quantity) cheese (ingredient) tsp (unit -> ingredient) ginger (ingredient) 2 (quantity) chillies (ingredient) turmeric (ingredient) powder (ingredient) cumin (ingredient) teaspoon (unit) salt (ingredient) oil (ingredient)

--- Misclassified Sample 3 (Original Index: 5) ---

Tokens: 1 cup cabbage leaves 3/4 tomatoes 18 grams tamarind 2 tablespoons white urad dal 4 red chillies 3 cloves garlic big Spoon oil teaspoon Rye 1/2 Cumin seeds sprig Curry

True Labels: quantity unit ingredient ingredient quantity ingredient quantity unit ingredient quantity unit ingredient ingredient ingredient quantity ingredient ingredient quantity ingredient ingredient ingredient unit ingredient unit ingredient quantity ingredient ingredient unit ingredient

Predicted Labels: quantity unit ingredient ingredient quantity ingredient quantity unit ingredient quantity unit ingredient ingredient ingredient quantity ingredient ingredient quantity unit ingredient ingredient ingredient ingredient unit ingredient quantity ingredient ingredient unit ingredient

Differences: 1 (quantity) cup (unit) cabbage (ingredient) leaves (ingredient) 3/4 (quantity) tomatoes (ingredient) 18 (quantity) grams (unit) tamarind (ingredient) 2 (quantity) tablespoons (unit) white (ingredient) urad (ingredient) dal (ingredient) 4 (quantity) red (ingredient) chillies (ingredient) 3 (quantity) cloves (ingredient -> unit) garlic (ingredient) big (ingredient) Spoon (unit -> ingredient) oil (ingredient) teaspoon (unit) Rye (ingredient) 1/2 (quantity) Cumin (ingredient) seeds (ingredient) sprig (unit) Curry (ingredient)

--- Misclassified Sample 4 (Original Index: 6) ---

Tokens: 2 teaspoons oil 1 teaspoon cumin seeds cloves garlic grated onions finely chopped red chilli powder 1/2 turmeric cup coconut milk vegetable Stock tablespoons Dijon Mustard carrots cut round thinly 5 green beans into small pieces 1/4 peas steam potatoes boiled salt

True Labels: quantity unit ingredient quantity unit ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient unit ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient

Predicted Labels: quantity unit ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient unit ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient unit quantity ingredient ingredient ingredient ingredient ingredient

Differences: 2 (quantity) teaspoons (unit) oil (ingredient) 1 (quantity) teaspoon (unit) cumin (ingredient) seeds (ingredient) cloves (unit -> ingredient) garlic (ingredient) grated (ingredient) onions (ingredient) finely (ingredient) chopped (ingredient) red (ingredient) chilli (ingredient) powder (ingredient) 1/2 (quantity) turmeric (ingredient) cup (unit) coconut (ingredient) milk (ingredient) vegetable (ingredient) Stock (ingredient) tablespoons (unit) Dijon (ingredient) Mustard (ingredient) carrots (ingredient) cut (ingredient) round (ingredient) thinly (ingredient) 5 (quantity) green (ingredient) beans (ingredient) into (ingredient) small (ingredient) pieces (ingredient -> unit) 1/4

(quantity) peas (ingredient) steam (ingredient) potatoes (ingredient) boiled (ingredient) salt (ingredient)

--- Misclassified Sample 5 (Original Index: 13) ---

Tokens: 18 Pani Pur is 2 Potato Aloo boiled 1/4 cup Green Moong Sprouts 1 teaspoon Cumin powder Jeera Chaat Masala Powder 1/2 Red Chilli Mango Raw 10 Mint Leaves Pudina Black Salt Kala Namak pepper tablespoons Sugar

True Labels: quantity ingredient ingredient quantity quantity ingredient ingredient ingredient quantity unit ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient

Predicted Labels: quantity ingredient ingredient ingredient quantity ingredient ingredient ingredient quantity unit ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient

Differences: 18 (quantity) Pani (ingredient) Pur (ingredient) is (quantity -> ingredient) 2 (quantity) Potato (ingredient) Aloo (ingredient) boiled (ingredient) 1/4 (quantity) cup (unit) Green (ingredient) Moong (ingredient) Sprouts (ingredient) 1 (quantity) teaspoon (unit) Cumin (ingredient) powder (ingredient) Jeera (ingredient) Chaat (ingredient) Masala (ingredient) Powder (ingredient) 1/2 (quantity) Red (ingredient) Chilli (ingredient) Mango (ingredient) Raw (ingredient) 10 (quantity) Mint (ingredient) Leaves (ingredient) Pudina (ingredient) Black (ingredient) Salt (ingredient) Kala (ingredient) Namak (ingredient) pepper (ingredient) tablespoons (unit) Sugar (ingredient)

--- Misclassified Sample 6 (Original Index: 15) ---

Tokens: 1 cup Quinoa 2 cups Water tablespoons Extra Virgin Olive Oil teaspoon Mustard seeds 1/2 Cumin Jeera White Urad Dal Split Chana dal Bengal Gram 6 Curry leaves Green Chillies finely chopped Shallot Tomato 4 Carrot Gajjar 1/4 Del Monte Whole Corn Kernels peas Matar beans French Beans Coriander Powder Dhania Garam masala powder Salt Leaves few

True Labels: quantity unit ingredient quantity unit ingredient unit ingredient ingredient ingredient ingredient unit ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient

Predicted Labels: quantity unit ingredient quantity unit ingredient unit ingredient ingredient ingredient ingredient unit ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity

Differences: 1 (quantity) cup (unit) Quinoa (ingredient) 2 (quantity) cups (unit) Water (ingredient) tablespoons (unit) Extra (ingredient) Virgin (ingredient) Olive (ingredient) Oil (ingredient) teaspoon (unit) Mustard (ingredient) seeds (ingredient) 1/2 (quantity) Cumin (ingredient) Jeera (ingredient) White (ingredient) Urad (ingredient) Dal (ingredient) Split (ingredient) Chana (ingredient) dal (ingredient) Bengal (ingredient) Gram (ingredient) 6 (quantity) Curry (ingredient) leaves (ingredient) Green (ingredient) Chillies (ingredient) finely (ingredient) chopped (ingredient) Shallot (ingredient) Tomato (ingredient) 4 (quantity) Carrot (ingredient) Gajjar (ingredient) 1/4 (quantity) Del (ingredient) Monte (ingredient) Whole (ingredient) Corn (ingredient) Kernels (ingredient) peas (ingredient) Matar (ingredient) beans (ingredient) French (ingredient) Beans (ingredient) Coriander (ingredient) Powder (ingredient) Dhania (ingredient) Garam (ingredient) masala (ingredient) powder (ingredient) Salt (ingredient) Leaves (ingredient) few (ingredient -> quantity)

--- Misclassified Sample 7 (Original Index: 16) ---

Tokens: 2 cups Tomatoes chopped 1/2 Onion finely cup Red Wine Vinaigrette Dried oregano cloves Garlic minced Black pepper powder Dijon Mustard 3 tablespoon Cane sugar to tablespoons Extra Virgin Olive Oil Salt

True Labels: quantity unit ingredient ingredient quantity ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient

Predicted Labels: quantity unit ingredient ingredient quantity ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient

Differences: 2 (quantity) cups (unit) Tomatoes (ingredient) chopped (ingredient) 1/2 (quantity) Onion (ingredient) finely (ingredient) cup (unit) Red (ingredient) Wine (ingredient) Vinaigrette (ingredient) Dried (ingredient) oregano (ingredient) cloves (unit) Garlic (ingredient) minced (ingredient) Black (ingredient) pepper (ingredient) powder (ingredient) Dijon (ingredient) Mustard (ingredient) 3 (quantity) tablespoon (unit) Cane (ingredient) sugar (ingredient) to (quantity -> ingredient) tablespoons (unit) Extra (ingredient) Virgin (ingredient) Olive (ingredient) Oil (ingredient) Salt (ingredient)

--- Misclassified Sample 8 (Original Index: 19) ---

Tokens: 1 cup Whole Wheat Flour 1/4 All Purpose Maida Sooji Semolina Rava 2 tablespoon Curd Dahi Yogurt teaspoon Turmeric powder Haldi Salt a pinch Sunflower Oil for kneading 4 Potatoes Aloo boiled and mashed 1/2 Cumin seeds Jeera Onion finely chopped cloves Garlic crushed inch Ginger Green Chillies 5 Curry leaves sprig Coriander Dhania Leaves Red Chilli

True Labels: quantity unit ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient quantity unit ingredient ingredient quantity ingredient quantity ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient unit ingredient ingredient ingredient quantity ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient

Predicted Labels: quantity unit ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient unit ingredient ingredient unit ingredient ingredient ingredient quantity ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient

Differences: 1 (quantity) cup (unit) Whole (ingredient) Wheat (ingredient) Flour (ingredient) 1/4 (quantity) All (ingredient) Purpose (ingredient) Maida (ingredient) Sooji (ingredient) Semolina (ingredient) Rava (ingredient) 2 (quantity) tablespoon (unit) Curd (ingredient) Dahi (ingredient) Yogurt (ingredient) teaspoon (unit) Turmeric (ingredient) powder (ingredient) Haldi (ingredient) Salt (ingredient) a (quantity -> ingredient) pinch (unit) Sunflower (ingredient) Oil (ingredient) for (quantity -> ingredient) kneading (ingredient) 4 (quantity) Potatoes (ingredient) Aloo (ingredient) boiled (ingredient) and (ingredient) mashed (ingredient) 1/2 (quantity) Cumin (ingredient) seeds (ingredient) Jeera (ingredient) Onion (ingredient) finely (ingredient) chopped (ingredient) cloves

(unit) Garlic (ingredient) crushed (ingredient) inch (unit) Ginger (ingredient) Green (ingredient) Chillies (ingredient) 5 (quantity) Curry (ingredient) leaves (ingredient) sprig (unit) Coriander (ingredient) Dhania (ingredient) Leaves (ingredient) Red (ingredient) Chilli (ingredient)

--- Misclassified Sample 9 (Original Index: 21) ---

Tokens: 1 cup Black Eyed Beans Lobia Onion chopped 3 cloves Garlic minced Red Yellow Green Bell Pepper Capsicum finely 2 Tomatoes blanched inch Ginger julienned tablespoon Extra Virgin Olive Oil teaspoon Cumin powder Jeera Chilli or red chilli flakes 4 sprig Coriander Dhania Leaves Lemon juice adjustable

True Labels: quantity unit ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient unit ingredient unit unit ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient

Predicted Labels: quantity unit ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity ingredient ingredient unit ingredient ingredient unit ingredient ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient

Differences: 1 (quantity) cup (unit) Black (ingredient) Eyed (ingredient) Beans (ingredient) Lobia (ingredient) Onion (ingredient) chopped (ingredient) 3 (quantity) cloves (unit) Garlic (ingredient) minced (ingredient) Red (ingredient) Yellow (ingredient) Green (ingredient) Bell (ingredient) Pepper (ingredient) Capsicum (ingredient) finely (ingredient) 2 (quantity) Tomatoes (ingredient) blanched (ingredient) inch (unit) Ginger (ingredient) julienned (unit -> ingredient) tablespoon (unit) Extra (ingredient) Virgin (ingredient) Olive (ingredient) Oil (ingredient) teaspoon (unit) Cumin (ingredient) powder (ingredient) Jeera (ingredient) Chilli (ingredient) or (ingredient) red (ingredient) chilli (ingredient) flakes (ingredient) 4 (quantity) sprig (unit) Coriander (ingredient) Dhania (ingredient) Leaves (ingredient) Lemon (ingredient) juice (ingredient) adjustable (ingredient)

--- Misclassified Sample 10 (Original Index: 26) ---

Tokens: 8 Mooli Mullangi Radish purple 2 cups Water 1/4 teaspoon Garam masala powder 1/2 Salt Black pepper slices Coriander Dhania Leaves chopped

True Labels: quantity ingredient ingredient ingredient ingredient quantity unit ingredient quantity unit ingredient ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient ingredient ingredient

Predicted Labels: quantity ingredient ingredient ingredient ingredient quantity unit ingredient quantity unit ingredient ingredient ingredient quantity ingredient ingredient ingredient unit ingredient ingredient ingredient ingredient

Differences: 8 (quantity) Mooli (ingredient) Mullangi (ingredient) Radish (ingredient) purple (ingredient) 2 (quantity) cups (unit) Water (ingredient) 1/4 (quantity) teaspoon (unit) Garam (ingredient) masala (ingredient) powder (ingredient) 1/2 (quantity) Salt (ingredient) Black (ingredient) pepper (ingredient) slices (ingredient -> unit) Coriander (ingredient) Dhania (ingredient) Leaves (ingredient) chopped (ingredient)

--- Insights from Investigating Misclassified Samples ---

- A total of 29 out of 84 validation recipes were misclassified.

- Many errors seem to occur at the boundaries between different entities (e.g., quantity ending and unit beginning, or unit ending and ingredient beginning).

- The model sometimes confuses labels for less frequent entities or specific phrasing it hasn't seen sufficiently during training.

- Cases where a word could belong to multiple categories based on context (e.g., 'powder' could be 'ingredient' or part of a method instruction if present in the full recipe text) might be challenging.

- Numbers or units that appear in unusual positions or formats occasionally cause misclassifications.

- Misclassifications might highlight limitations in the current feature set or the model's ability to capture complex contextual cues.

- Analyzing specific examples (like those printed above) is crucial to pinpoint exact error patterns and potential areas for improvement (e.g., adding more features, expanding the dictionary, exploring different models).

Most Common Misclassification Errors (True -> Predicted):

- unit -> ingredient: 38 times

- ingredient -> unit: 11 times
- quantity -> ingredient: 5 times
- ingredient -> quantity: 3 times
- quantity -> unit: 2 times

Analysis of Misclassified Tokens:
- Token: 'cloves'
  - ingredient->unit: 4 times
  - unit->ingredient: 2 times
  - quantity->unit: 1 times
- Token: 'tsp'
  - unit->ingredient: 4 times
- Token: 'few'
  - ingredient->quantity: 3 times
- Token: 'a'
  - quantity->ingredient: 2 times
  - unit->ingredient: 1 times
- Token: 'tbsp'
  - unit->ingredient: 3 times
- Token: 'gram'
  - ingredient->unit: 2 times
- Token: 'pieces'
  - ingredient->unit: 2 times
- Token: 'to'
  - quantity->ingredient: 1 times
  - unit->ingredient: 1 times
- Token: 'cut'
  - unit->ingredient: 2 times
- Token: 'into'
  - unit->ingredient: 2 times

Error analysis completed.

Flattened true validation labels (y_val_labels_flat) with length: 2876

Flattened predicted validation labels (y_val_pred_flat) with length: 2876

Initialized empty error_data list.

Verification successful: Flattened lists have consistent lengths.

Length of flattened true labels: 2876

Length of flattened predicted labels: 2876

Initialized empty error_data list.

Length of flattened X_val with indices: 2876

Length of flattened true validation labels: 2876

Length of flattened predicted validation labels: 2876

Lengths of flattened data are consistent.

Iterating through validation data to collect error information...

Collected information for 59 misclassified tokens.

Error analysis data is stored in the `error_data` list.

First 5 misclassified tokens details:

--- Error 1 ---

Recipe Index: 2, Token Index: 34

Token: 'few'

True Label: 'ingredient'

Predicted Label: 'quantity'

Context: Previous='Leaves', Next='None'

True Label Weight: 0.33

Predicted Label Weight: 7.26

--- Error 2 ---

Recipe Index: 3, Token Index: 4

Token: 'gram'

True Label: 'ingredient'

Predicted Label: 'unit'

Context: Previous='peas', Next='flour'

True Label Weight: 0.33

Predicted Label Weight: 8.77

--- Error 3 ---

Recipe Index: 3, Token Index: 8

Token: 'tsp'

True Label: 'unit'

Predicted Label: 'ingredient'

Context: Previous='cheese', Next='ginger'

True Label Weight: 8.77

Predicted Label Weight: 0.33

--- Error 4 ---

Recipe Index: 5, Token Index: 18

Token: 'cloves'

True Label: 'ingredient'

Predicted Label: 'unit'

Context: Previous='3', Next='garlic'

True Label Weight: 0.33

Predicted Label Weight: 8.77

--- Error 5 ---

Recipe Index: 5, Token Index: 21

Token: 'Spoon'

True Label: 'unit'

Predicted Label: 'ingredient'

Context: Previous='big', Next='oil'

True Label Weight: 8.77

Predicted Label Weight: 0.33

Most Common Misclassified Label Pairs:

- 'unit' -> 'ingredient': 38 times
- 'ingredient' -> 'unit': 11 times
- 'quantity' -> 'ingredient': 5 times
- 'ingredient' -> 'quantity': 3 times
- 'quantity' -> 'unit': 2 times

Most Common Tokens Involved in Misclassifications:

- 'cloves': 7 times
- 'tsp': 4 times
- 'few': 3 times
- 'a': 3 times
- 'tbsp': 3 times
- 'gram': 2 times
- 'pieces': 2 times
- 'to': 2 times
- 'cut': 2 times
- 'into': 2 times

Misclassification Counts per True Label:

- 'unit': 38 times
- 'ingredient': 14 times
- 'quantity': 7 times

Misclassification Counts per Predicted Label:

- 'ingredient': 43 times

- 'unit': 13 times

- 'quantity': 3 times

DataFrame shape: (59, 10)

Total tokens in validation data: 2876

Number of misclassified tokens: 59

Number of correctly classified tokens: 2817

Overall Accuracy on Validation Data: 0.9795

Error Analysis: Misclassified Tokens DataFrame ---

Displaying DataFrame of misclassified tokens with context and label weights:

	token	previous_token	next_token	true_label	predicted_label	true_label_weight	predicted_label_weight
0	few	Leaves	None	ingredient	quantity	0.334116	7.259184
1	gram	peas	flour	ingredient	unit	0.334116	8.771887
2	tsp	cheese	ginger	unit	ingredient	8.771887	0.334116
3	cloves	3	garlic	ingredient	unit	0.334116	8.771887
4	Spoon	big	oil	unit	ingredient	8.771887	0.334116
5	cloves	seeds	garlic	unit	ingredient	8.771887	0.334116
6	pieces	small	1/4	ingredient	unit	0.334116	8.771887
7	is	Pur	2	quantity	ingredient	7.259184	0.334116

	token	previous_token	next_token	true_label	predicted_label	true_label_weight	predicted_label_weight
8	few	Leaves	None	ingredient	quantity	0.334116	7.259184
9	to	sugar	tablespoons	quantity	ingredient	7.259184	0.334116

Total misclassified tokens found: 59

--- Error Analysis: By True Label ---

true_label	total_in_validation	misclassification_count	correctly_classified_count	accuracy_for_label	class_weight
ingredient	2107	14	2093	0.9934	0.3341
quantity	411	7	404	0.9830	7.2592
unit	358	38	320	0.8939	8.7719

--- Overall Accuracy ---

Overall Accuracy on Validation Data: 0.9795

Error analysis by label and overall accuracy displayed.

Computed weight_dict (Inverse Frequency Weights, 'ingredient' penalized):

{'quantity': 7.259183673469388,

'unit': 8.771886559802713,

'ingredient': 0.6682321998872816}

Sorted weight_dict:

{'ingredient': 0.6682321998872816,

'quantity': 7.259183673469388,

'unit': 8.771886559802713}

weight_dict after penalising 'ingredient' label:

{'quantity': 7.259183673469388,

'unit': 8.771886559802713,

'ingredient': 0.3341160999436408}

Defined extract_features_with_class_weights function.

This function extracts features using dataset2features and returns the features, labels, and the global weight_dict.

The weight_dict is intended to be used as a parameter for the CRFSuite trainer.

Computed weight_dict (Inverse Frequency Weights, 'ingredient' penalized):

{'quantity': 7.259183673469388,

'unit': 8.771886559802713,

'ingredient': 0.6682321998872816}

Sorted weight_dict:

{'ingredient': 0.6682321998872816,

'quantity': 7.259183673469388,

'unit': 8.771886559802713}

Label counts in y_train_flat:

Counter({'quantity': 980, 'unit': 811, 'ingredient': 5323})

Total number of samples (tokens) in the training set: 7114

Label Counts:

- ingredient: 5323

- quantity: 980

- unit: 811

Length of original y_train_labels (list of lists): 196

Length of flattened y_train_flat: 7114

First 10 elements of y_train_flat:

['quantity',

'unit',

'ingredient',

'ingredient',

'quantity',

'ingredient',

'ingredient',

'ingredient',

'ingredient',

'ingredient']

Length of validation features (X_val_features): 84

Length of validation labels (y_val_labels): 84

Length validation successful: Number of sequences and tokens per sequence match for validation set.

Length of training features (X_train_features): 196

Length of training labels (y_train_labels): 196

Length of validation features (X_val_features): 84

Length of validation labels (y_val_labels): 84

Length validation successful: Number of sequences and tokens per sequence match for train and validation sets.

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	[2-1/2, 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	[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, ingredient, quantity, unit, ingredient, ingredient, ingredient, ingredient, ingredient, ingredient,

		ingredient		ingredient,
		ingredient		ingredient,
		ingredient unit		ingredient, unit,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient]
		quantity unit		[quantity, unit,
		ingredient quantity		ingredient, quantity,
		ingredient		ingredient,
		ingredient quantity		ingredient, quantity,
		ingredient quantity		ingredient, quantity,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient unit		ingredient, unit,
		ingredient		ingredient,
	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	ingredient quantity	[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]	ingredient, quantity,
		unit ingredient		unit, ingredient,
		quantity unit		quantity, unit,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient unit		ingredient, unit,
3		ingredient		ingredient,
		ingredient		ingredient,
		ingredient quantity		ingredient, quantity,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient unit		ingredient, unit,
		ingredient		ingredient,
		ingredient		ingredient,
		ingredient		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	[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]

Length of flattened input_tokens for Training dataset: 7114

Length of flattened pos_tokens for Training dataset: 7114

----------Thank You----------