**Identifying Key Entities in Recipe Data**

**1. Objective**

This project focuses on building a system that can recognize important components in cooking recipes using a Named Entity Recognition (NER) approach. Specifically, it identifies entities such as:

* Ingredients
* Quantities
* Measurement units

By extracting these elements from unstructured text, recipes can be converted into a structured format. This structured data can then support use cases like nutritional tracking, automated grocery lists, and interactive cooking apps.

**2. Methodology**

**Data Preparation**

* The recipe text comes pre-tagged using the IOB format, which marks each word as the beginning (B), inside (I), or outside (O) of an entity.
* This format is commonly used for labeling sequences in NER tasks.

**Feature Engineering**

Each token (word) is enriched with several features to improve prediction, including:

* The lowercase version of the word
* Part-of-speech (POS) information
* Prefixes and suffixes (e.g., first/last few characters)
* Whether it’s a digit or capitalized
* Word position and other contextual cues

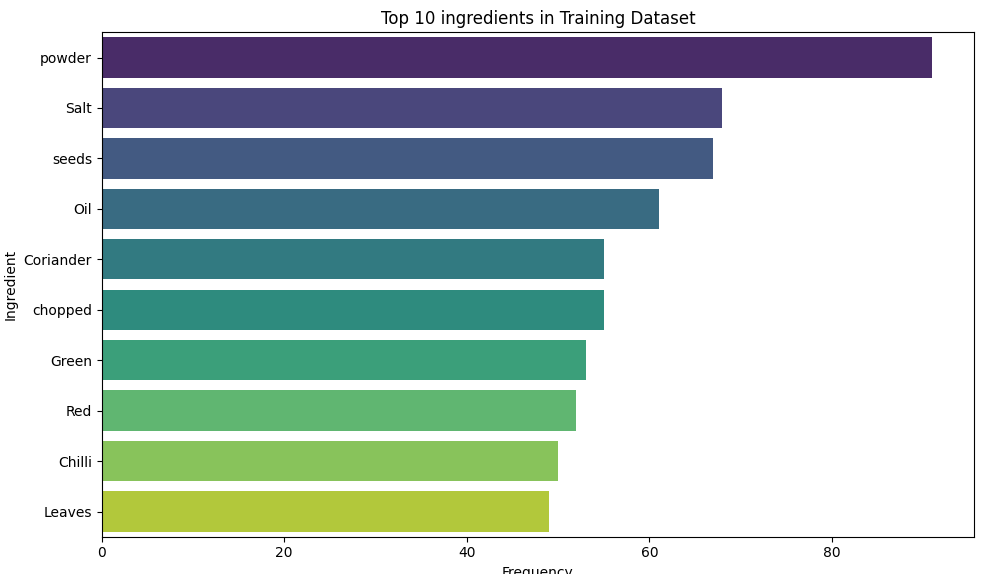
**Model Development**

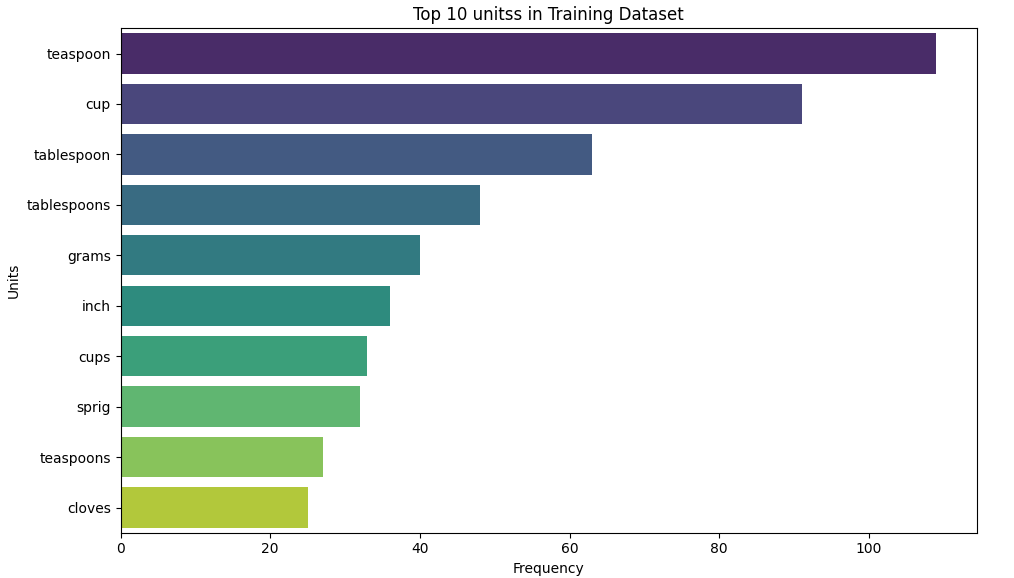
* The model is built using the sklearn-crfsuite implementation of Conditional Random Fields (CRF), which is well-suited for identifying patterns in sequential data.
* It was evaluated based on standard classification metrics: precision, recall, and F1-score.

**3. Visualizations and Insights**

**Entity Frequency**

Visual analysis showed that common ingredients like **powder**, **salt**, and **seeds** were among the most frequently mentioned. Similarly, **teaspoon**, **cup**, and **tablespoon** were the top units identified in the dataset.





* Ingredients are the most frequently tagged entity.
* Units and quantities follow consistent patterns, making them easier to recognize.

**Model Performance**

| **Entity** | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| Ingredient | 0.99 | 0.99 | 0.99 | 1611 |
| Quantity | 0.99 | 0.97 | 0.98 | 294 |
| Unit | 0.96 | 0.95 | 0.95 | 244 |
| **Overall Accuracy** |  |  | **0.98** | 2149 |

Insights:

* Ingredient detection is highly accurate, thanks to effective feature design.
* Quantities and units have slightly lower recall, possibly due to irregular expressions like "a pinch" or "half".
* The model is well-balanced, with macro and weighted F1-scores around 0.97–0.98.

**4. Assumptions Made**

* The dataset is accurately annotated using the IOB format.
* Each sentence or recipe line is treated independently.
* The model does not rely on pretrained embeddings or language models.
* All features are manually engineered.

**5. Conclusion**

The CRF-based NER model achieves strong performance with 98% overall accuracy in identifying entities in recipe data. Its success demonstrates the effectiveness of traditional feature-based models in structured text domains where patterns are predictable and consistent.