Objective

The notebook aims to build a Named Entity Recognition (NER) model using **Conditional Random Fields (CRF)** to identify key components in recipe data, such as:

- **Ingredients**
- **Quantities**
- **Units**

Data Description

The given data is in JSON format, representing a **structured recipe ingredient list** with **Named Entity Recognition (NER) labels**. Below is a breakdown of the data fields:

Json Input

```
[
  {
     "input": "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",
     "pos": "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"
  }, …
]
```

Key	Description
input	Contains a raw ingredient list for recipe
pos	Represents the corresponding part-of-speech (POS) tags or NER labels, identifying quantities, ingredients, and units.

Data Ingestion and Preparation

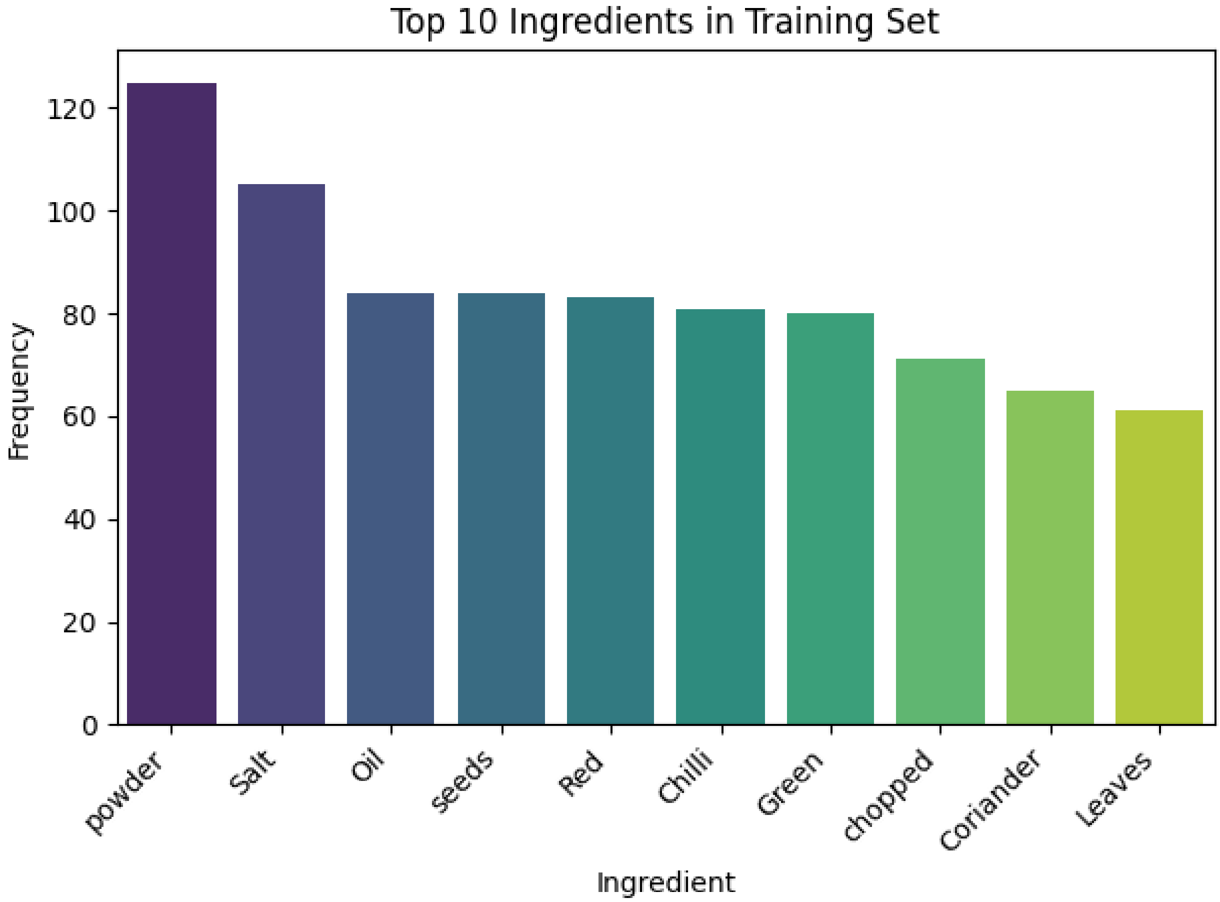
- Given input was read through pandas and a dataframe was created
- The input and pos columns were splitted to input_tokens and pos_tokens
- There were totally **285 input values**
- After cleaning up the data based on the comparison of lengths of input_tokens and pos_tokens, we are left with **280** input values and corresponding pos values.

Train Validation Split

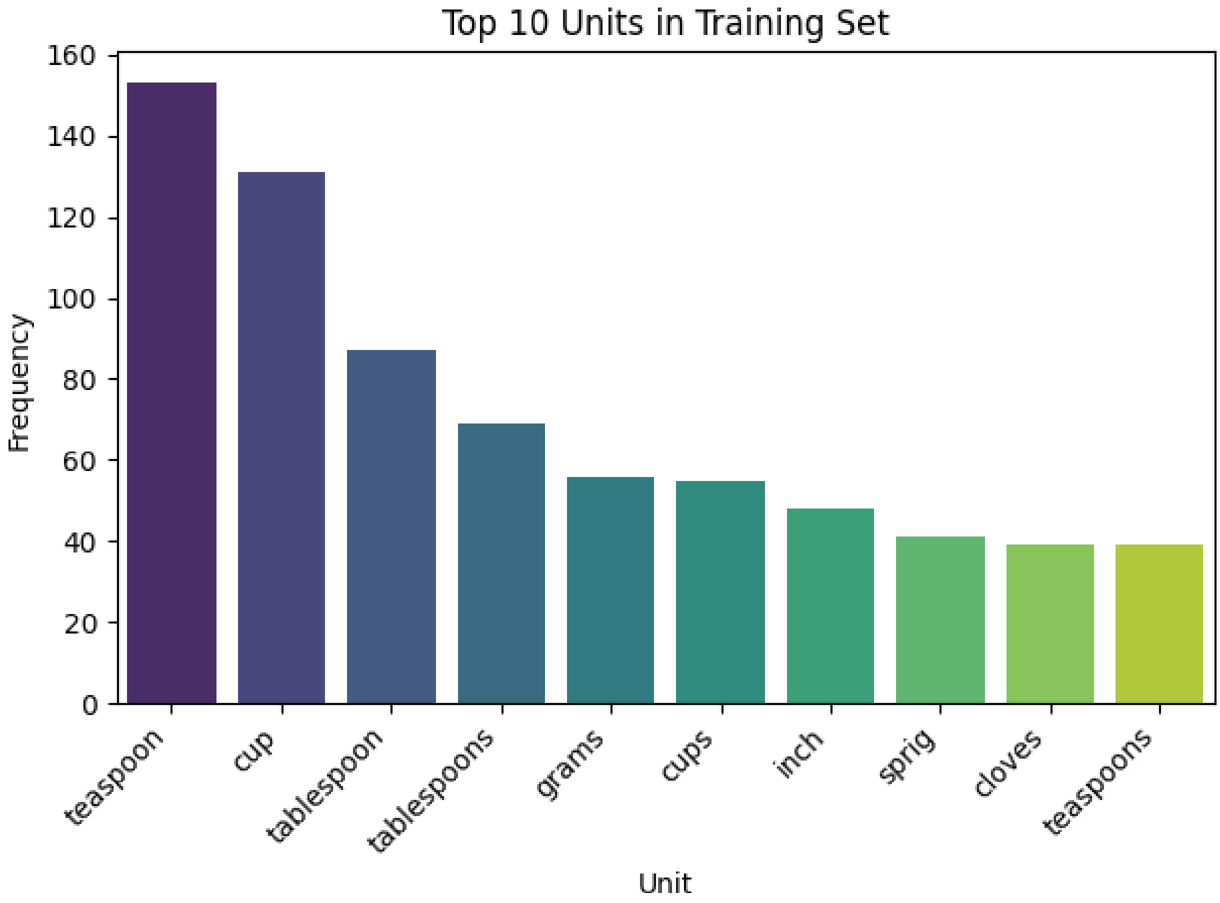
- Train and validation split was performed on the data which resulted in **196 rows of train data** and **84 rows of validation data**.

ERD on Training Data set

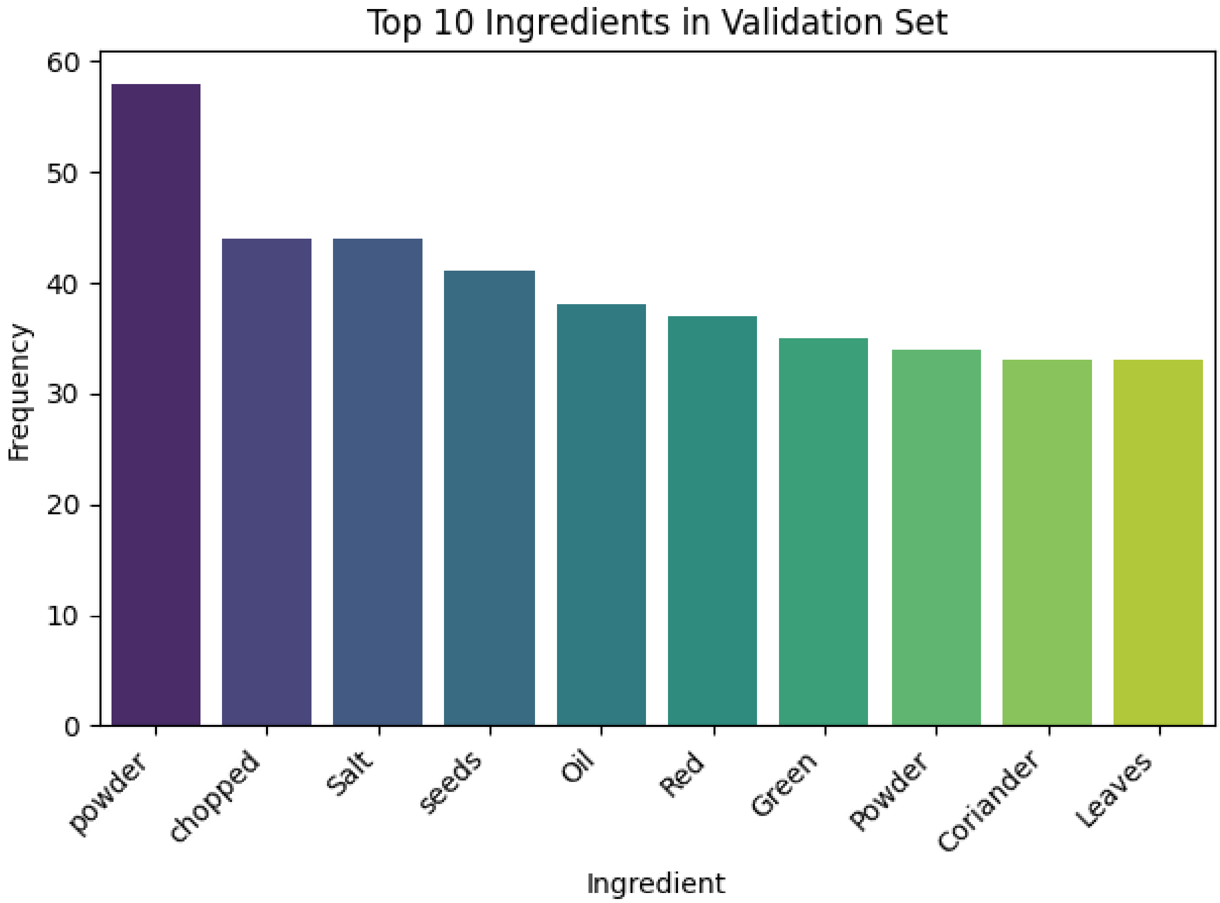
- Flattening the input data, we got totally **6772 tokens** of input and their respective pos tokens
- No of unique labels are identified to be 3 which are **['ingredient' 'quantity' 'unit']**
- Most frequent Ingredients in the train data are found to be

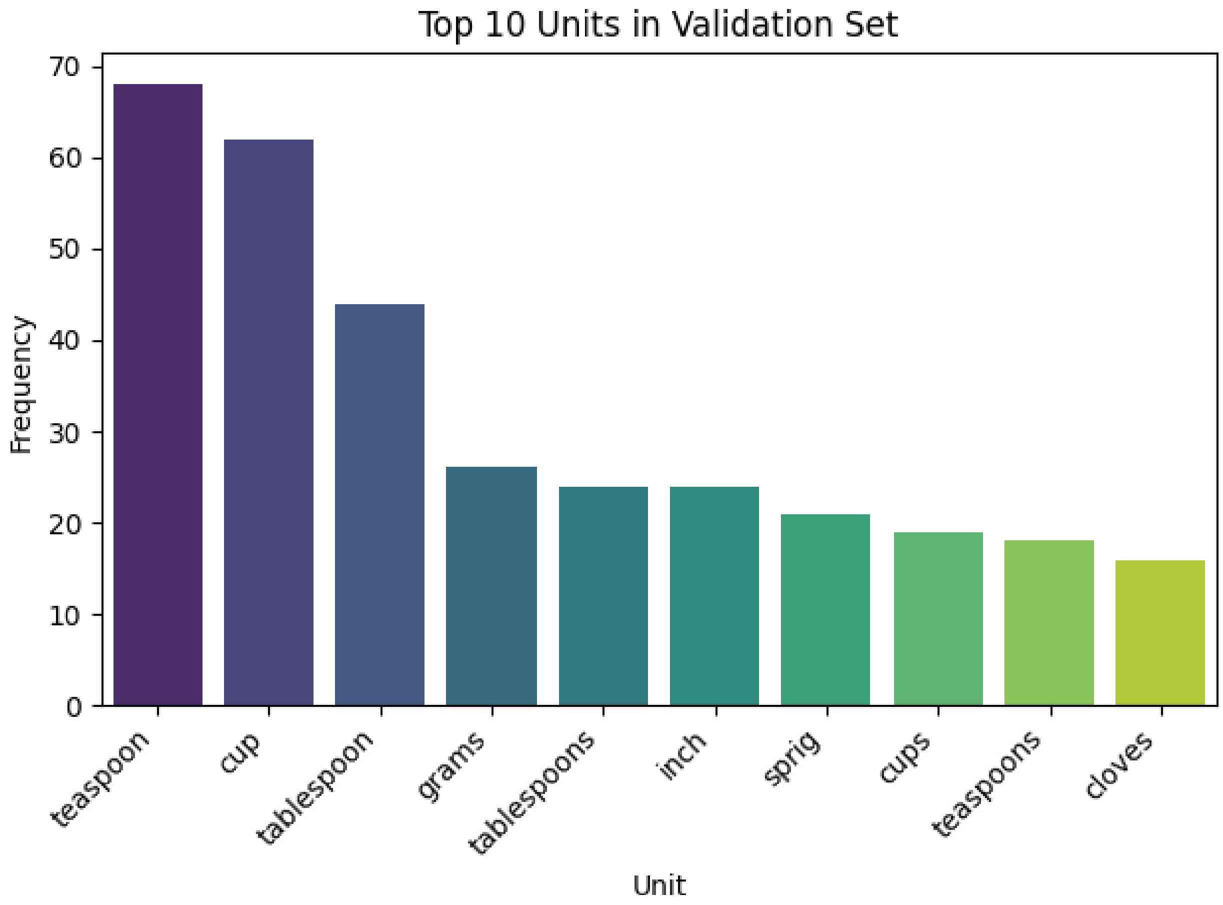


- Most frequent units in the train data are found to be



- Similarly most frequent Ingredients and units in the validation data are found as below


Top 10 Ingredients in Validation Set
Frequency
60
50
40
30
20
10
0
powder
chopped
Salt
seeds
Oil
Red
Green
Powder
Coriander
Leaves
Ingredient



- Both training data and validation data shows similar pattern when visualized for the most frequent tokens.

Insights on the top 10 ingredient tokens in training data

1. "powder" & "Salt" dominate ingredient tokens

- "powder" (125) is highest—reflecting repeated mentions in spice names (e.g. "Turmeric powder").

- "Salt" (105) ranks second, underscoring its universal use.

2. Preparation descriptors are frequent

- "chopped" (71) and "Green" (80) appear as often as many core ingredients, indicating your NER must handle both food items and prep cues.

3. Whole-spice mentions

- "seeds" (84) highlights the frequent use of items like "Cumin seeds" and "Mustard seeds".

Insights on the top 10 unit token in the training data

1. Units skewed to small measures

- "teaspoon" (153) and "cup" (131) together make up almost half of all unit mentions.

- "tablespoon" (87) is a distant third, while larger measures ("inch", "sprig") are much rarer.

2. Plural and Singular forms of tokens

- teaspoon and teaspoons are taken as 2 different tokens

- similarly tablespoon and tablespoons are treated as different tokens

- these forms should be handled through lemmatization which we will do it later.

- Total label counts were found to be
  - **quantity: 959**
  - **unit: 813**
  - **ingredient: 5000**

Model Creation and Evaluation

- CRF model was created using following script

```
crf = sklearn_crfsuite.CRF(
    algorithm='lbfgs',
    c1=0.5,
    c2=0.1,
    max_iterations=100,
    all_possible_transitions=True
)
```

- **F1-Score** calculated for the training set is **0.99** which is much good and is supported with the classification report as below

```
             precision    recall  f1-score   support

  ingredient      0.990     0.997     0.994      5000
    quantity      0.996     0.985     0.991       959
        unit      0.984     0.954     0.969       813

    accuracy                          0.990      6772
   macro avg      0.990     0.979     0.984      6772
weighted avg      0.990     0.990     0.990      6772
```

Insights

1. ingredient (most frequent class):

- Very high recall (0.997): Model captures nearly all true ingredient tokens.

- Precision (0.990): Few false positives.

- F1-score (0.994): Excellent overall — the model is extremely confident and accurate for this class.

2. quantity:

- Precision (0.996) > Recall (0.985): Model is slightly more conservative, making fewer mistakes when it does predict a quantity — but misses a few.

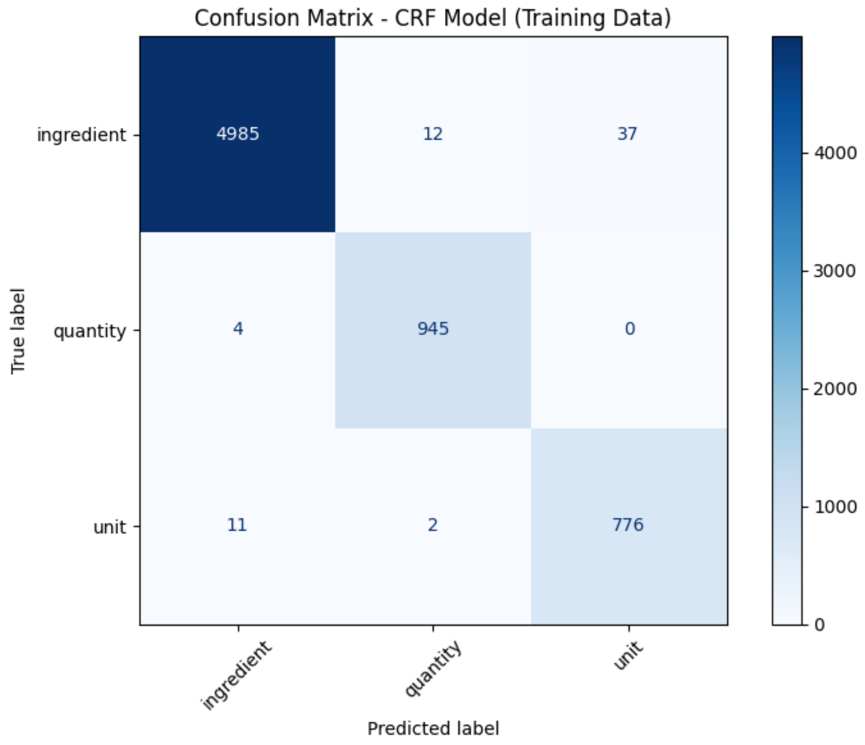
- F1-score (0.991): Still very strong.

3. unit (least frequent class):

- Recall is lowest (0.954): Some actual unit tokens are missed.

- F1-score (0.969): Lower than other classes but still strong, especially considering its relatively small support (813).

Confusion Matrix



Interpretation

- CRF model is highly effective, especially on more frequent classes like ingredient.

- Slight performance drops on the unit class are expected due to lower representation.

- These results suggest the model is well-fit to the training data — possibly even slightly overfitting, as these are training scores.

Evaluation on Validation set

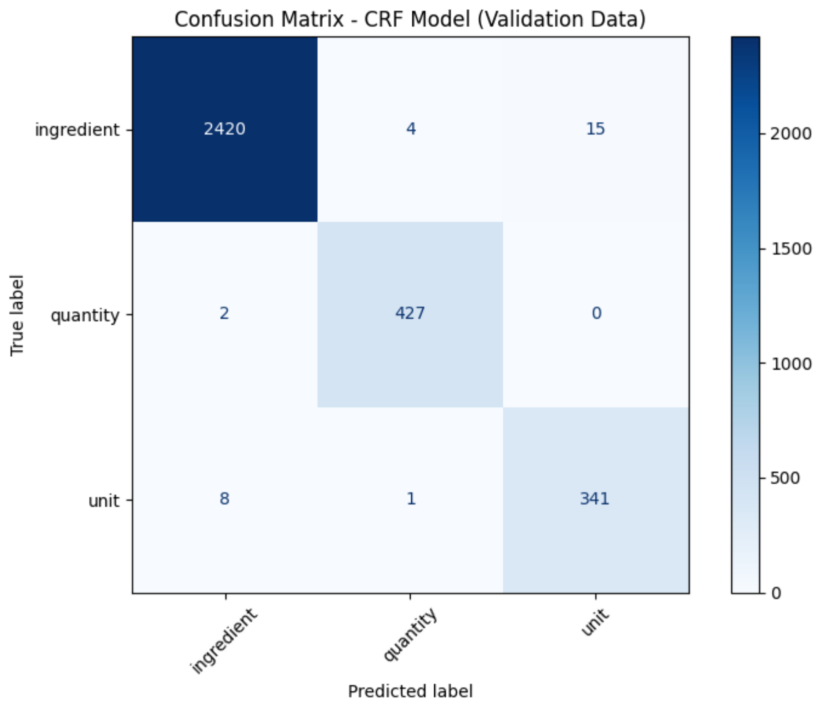
Classification Report

```
              precision    recall  f1-score   support

  ingredient      0.992     0.996     0.994      2430
    quantity      0.995     0.988     0.992       432
        unit      0.974     0.958     0.966       356

    accuracy                          0.991      3218
   macro avg      0.987     0.981     0.984      3218
weighted avg      0.991     0.991     0.991      3218
```

Confusion Matrix



- ➢ Validation Accuracy = 99.07% which is too good.
- ➢ Label wise error analysis was done and is captured in the following table

Error Analysis

```
Label-wise Error Analysis:

        Label  Total  Errors  Accuracy  Class Weight
1        unit    356      15    0.9579        0.1106
2  ingredient   2430      10    0.9959        0.7551
0    quantity    432       5    0.9884        0.1342

Sample Errors with Context:

    token previous_token next_token  true_label predicted_label               context
0   sprig          laung      curry  ingredient            unit      laung sprig curry
1    inch           salt     ginger  ingredient            unit       salt inch ginger
2  little           meat      extra    quantity      ingredient      meat little extra
3     for          honey    glazing    quantity      ingredient      honey for glazing
4   clove        chopped     garlic        unit      ingredient   chopped clove garlic
5  cloves        florets      thyme        unit      ingredient   florets cloves thyme
6     cut        breasts       into        unit      ingredient        breasts cut into
7    into            cut         cm        unit      ingredient            cut into cm
8      cm           into      cubes        unit      ingredient          into cm cubes
9  finely         garlic    chopped        unit      ingredient  garlic finely chopped
```

Transition captured

ingredient -> ingredient 0.716170 ➔ **Highly probable sequence observed by the model**

quantity -> unit 0.203486

unit -> ingredient 0.002671

unit -> quantity -0.028030

unit -> unit -0.151569

ingredient -> unit -0.246678 ➔ **Least probable sequence observed by the model**

Conclusion

Accuracies for both training data and validation data is more than 99% which shows that the model is very good.