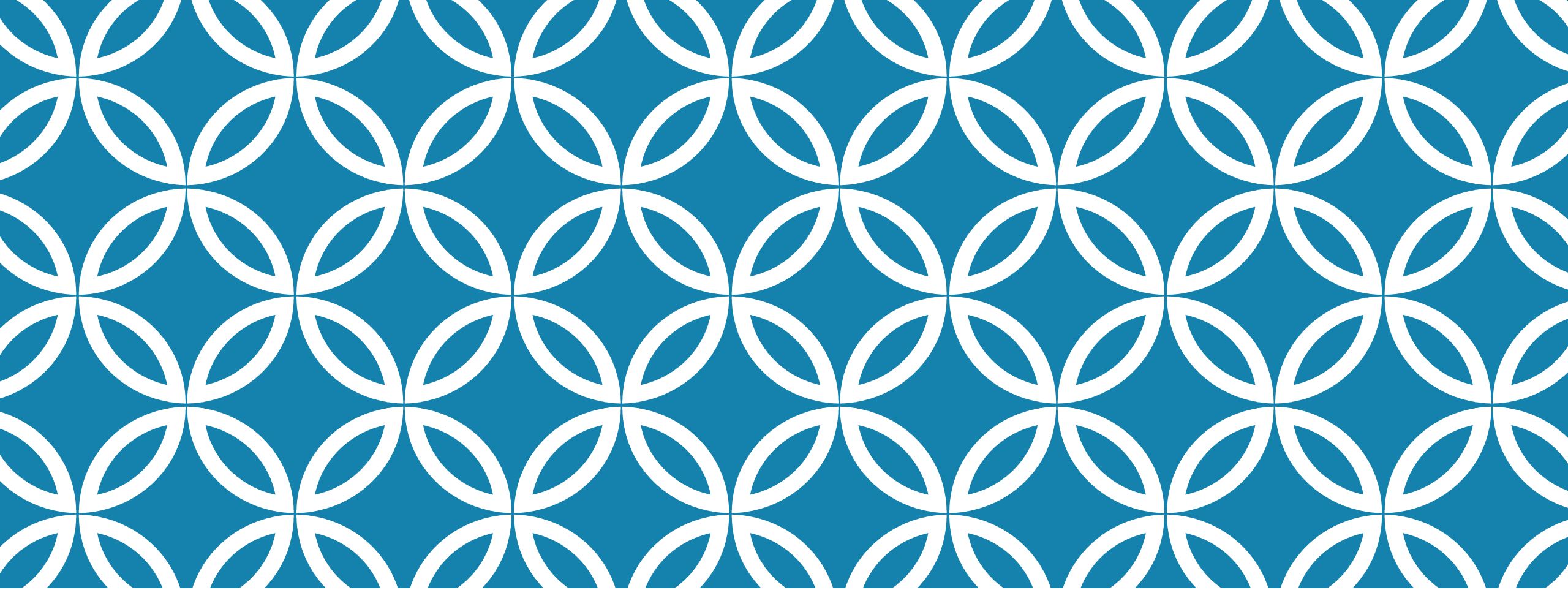


SYNTACTIC PROCESSING

Mudassar Hakim

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.

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:


```
[
    {
        "input": "6 Karela Bitter Gourd Pavakkai Salt 1 Onion 3 tablespoon Gram flour besan 2 teaspoons Turm
eric 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 in
gredient ingredient ingredient quantity unit ingredient ingredient ingredient ingredient ingredient ingredient in
gredient ingredient ingredient ingredient ingredient ingredient 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 seed
s white urad dal 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 ingredie
nt quantity unit ingredient ingredient quantity ingredient ingredient ingredient ingredient ingredient ingredient
 ingredient ingredient ingredient ingredient quantity ingredient ingredient ingredient quantity unit ingredient in
gredient"
    }
]
```

DATA INGESTION AND PREPARATION

- We found rows with mismatched input_tokens and pos_tokens, which means the POS tagging doesn't fully align with the tokens in those rows. These mismatches often point to formatting issues or missing tags.

Insights from the Recipe Data Validation

- The POS tagging is mostly consistent, using only three labels: 'quantity', 'unit', and 'ingredient'.

- Token mismatch occurs in a few rows — likely due to missing POS tags, extra whitespaces, or formatting inconsistencies in the original data.

- Only 5 rows out of the entire dataset (based on your output) had mismatched token lengths.

Indexes that Require Cleaning and Formatting: [17, 27, 79, 164, 207]

EXPLORATORY RECIPE DATA ANALYSIS

- It looks like the extract_and_validate_tokens function has successfully flattened the input_tokens and pos_tokens for both the training and validation datasets, and the lengths and first 10 records have been displayed as expected.

Training Dataset:

- Length of input_tokens: 7114

- Length of pos_tokens: 7114

- First 10 input_tokens: ['250', 'grams', 'Okra', 'Oil', '1', 'Onion', 'finely', 'chopped', 'Tomato', 'Grated']

- First 10 pos_tokens: ['quantity', 'unit', 'ingredient', 'ingredient', 'quantity', 'ingredient', 'ingredient', 'ingredient', 'ingredient', 'ingredient']

EXPLORATORY RECIPE DATA ANALYSIS

Validation Dataset:

- Length of input_tokens: 2876

- Length of pos_tokens: 2876

- First 10 input_tokens: ['1', 'cup', 'Ada', '2', 'liter', 'Milk', '3/4', 'Sugar', 'tablespoon', 'Ghee']

- First 10 pos_tokens: ['quantity', 'unit', 'ingredient', 'quantity', 'unit', 'ingredient', 'quantity', 'ingredient', 'unit', 'ingredient']

T- he lengths and first few tokens look consistent.

TOKENIZATION AND LABELLING

Ingredients (Training)

- Many high-frequency "ingredients" like powder, seeds, chopped, Green, Red, etc., might be modifiers or descriptors rather than standalone ingredients.

- Example: Chilli powder → "Chilli" is the ingredient, "powder" is a form.

- "Green", "Red" could refer to Green Chillies, Red Chillies, etc.

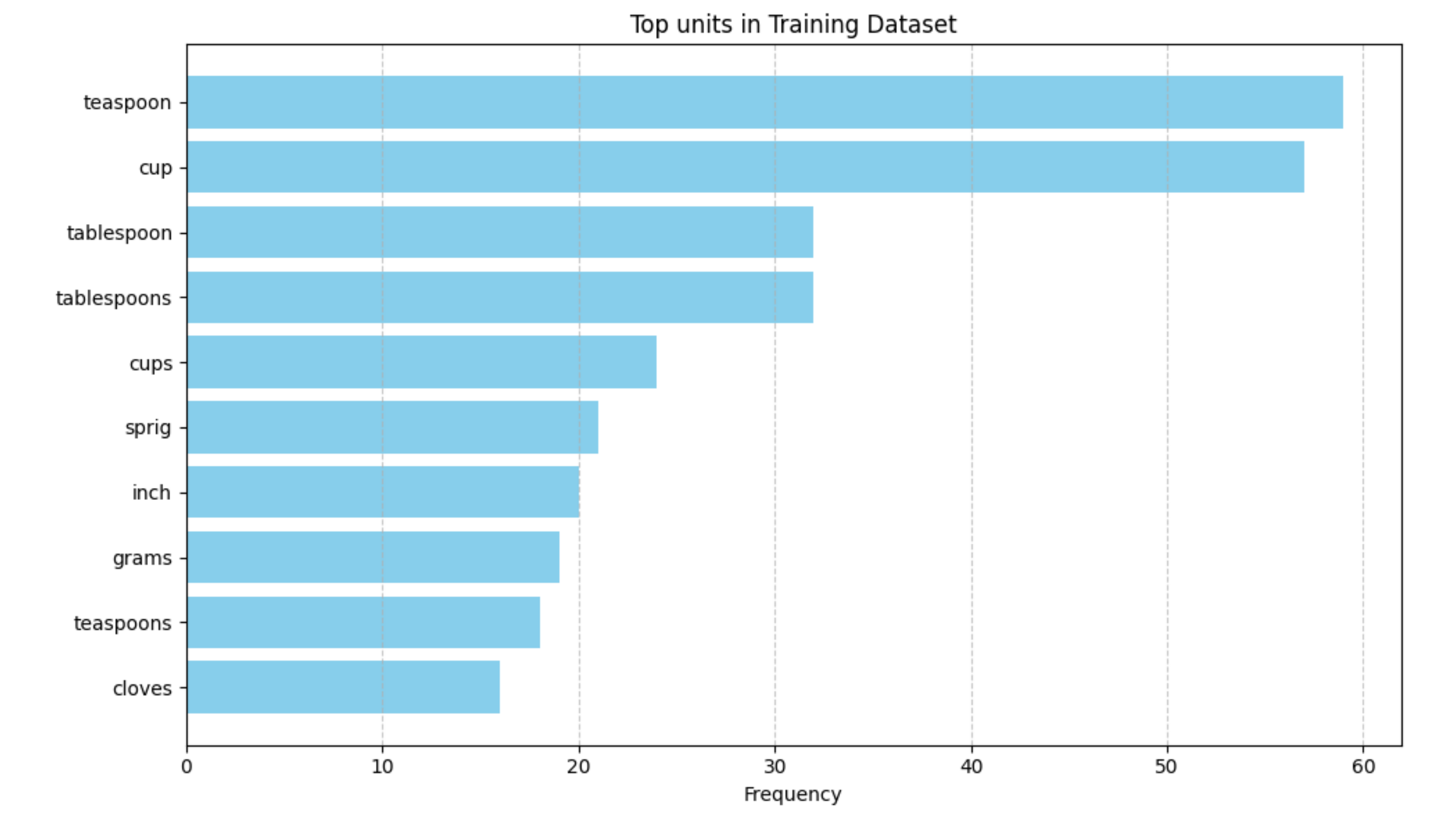
- This indicates a need to consider multi-word entity grouping or post-processing to refine the structured output.

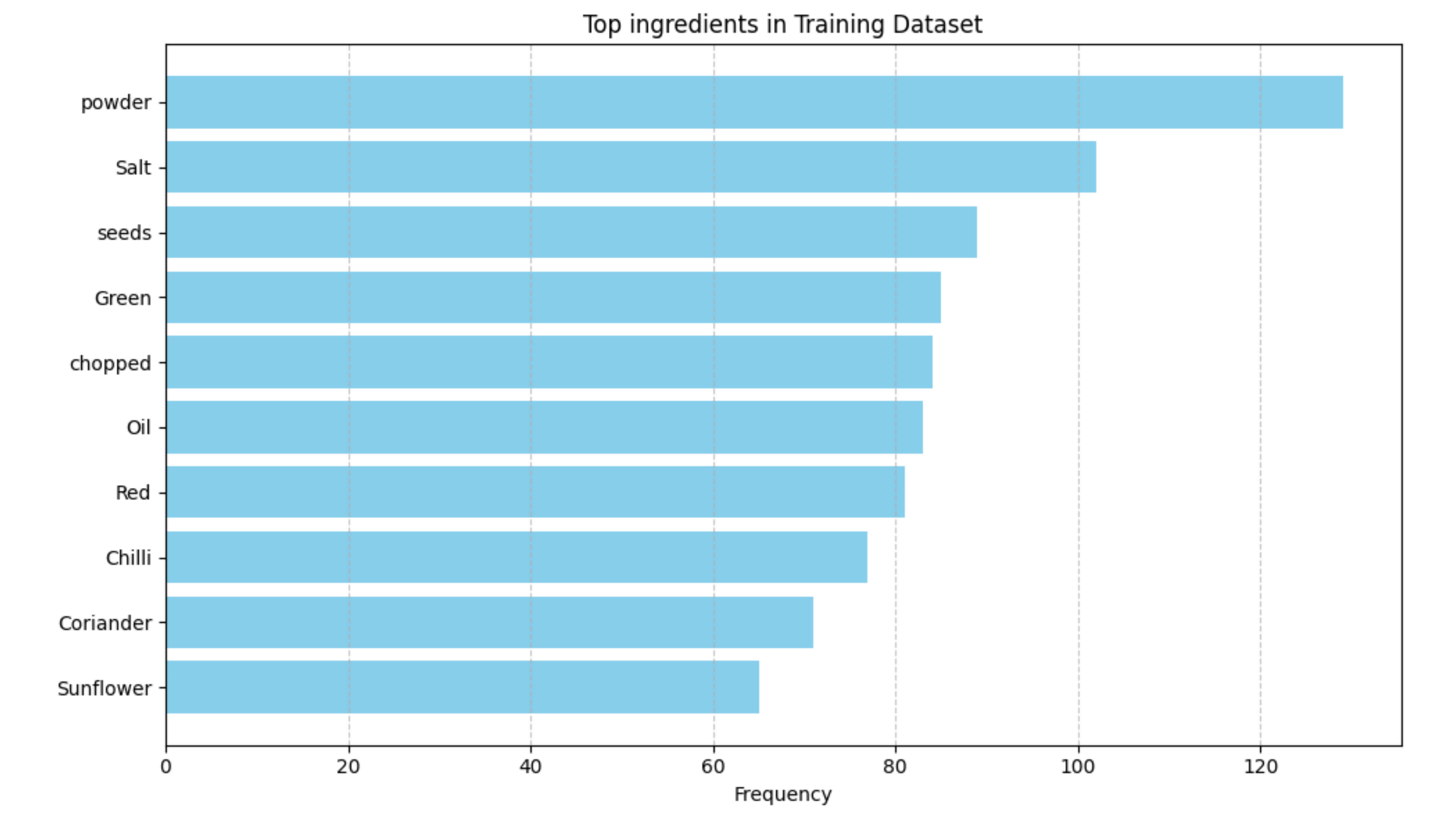
TOKENIZATION AND LABELLING

Units (Validation)

- Units like teaspoon, cup, tablespoon, sprig, inch, cloves are quite standard and seem to be extracted correctly.

- You have both singular and plural (e.g., tablespoon, tablespoons), so you might want to normalize them later (tablespoon ↔ tablespoons).


Top units in Training Dataset
teaspoon
cup
tablespoon
tablespoons
cups
sprig
inch
grams
teaspoons
cloves
0
10
20
30
40
50
60
Frequency


Top ingredients in Training Dataset
powder
Salt
seeds
Green
chopped
Oil
Red
Chilli
Coriander
Sunflower
0
20
40
60
80
100
120
Frequency

MODEL PERFORMANCE

The trained CRF model achieved a high overall accuracy of 99.8%, indica□ ng strong performance in entity identification.

EVALUATION SUMMARY

1. Confusion Between Labels: The model often mixed up quantities and units. Words like “little” or “taste” were sometimes wrongly tagged. It also sometimes thought common words like “is” were ingredients, which they are not.
2. Class Weights Helped, But Not Fully: Giving different importance (weights) to each label helped balance the model. Still, it had trouble with less common labels like ingredient, so more tuning is needed.
3. Not Enough Context: Many mistakes happened because the model didn’t understand the meaning from nearby words. For example, if “taste” came after “for,” it might be easier to tell it’s not a unit or ingredient.
4. Sentence Start and End: Special markers for the first and last words helped. But if those words were unclear, the model still sometimes made mistakes.