Summary Report: Identifying Key Entities in Recipe Data

Problem Statement

The objective is to build a Named Entity Recognition (NER) model that extracts key components from unstructured recipe text. This is crucial for creating structured recipe databases used in dietary apps, recipe managers, or food-related e-commerce.

Softwares used: Jupyter Notebook created using Google Colab. Required libraries imported.

Approach Taken

• **Data Source**: A JSON-formatted (**ingredient_and_quantity.json**) dataset with annotated recipe ingredients is used to load in pandas' data frame for further analysis.

• **NER Labels**: quantity, unit, ingredient.

• **Model Used**: Conditional Random Fields (CRF).

• Pipeline:

1. Data preprocessing (tokenization, POS tagging).

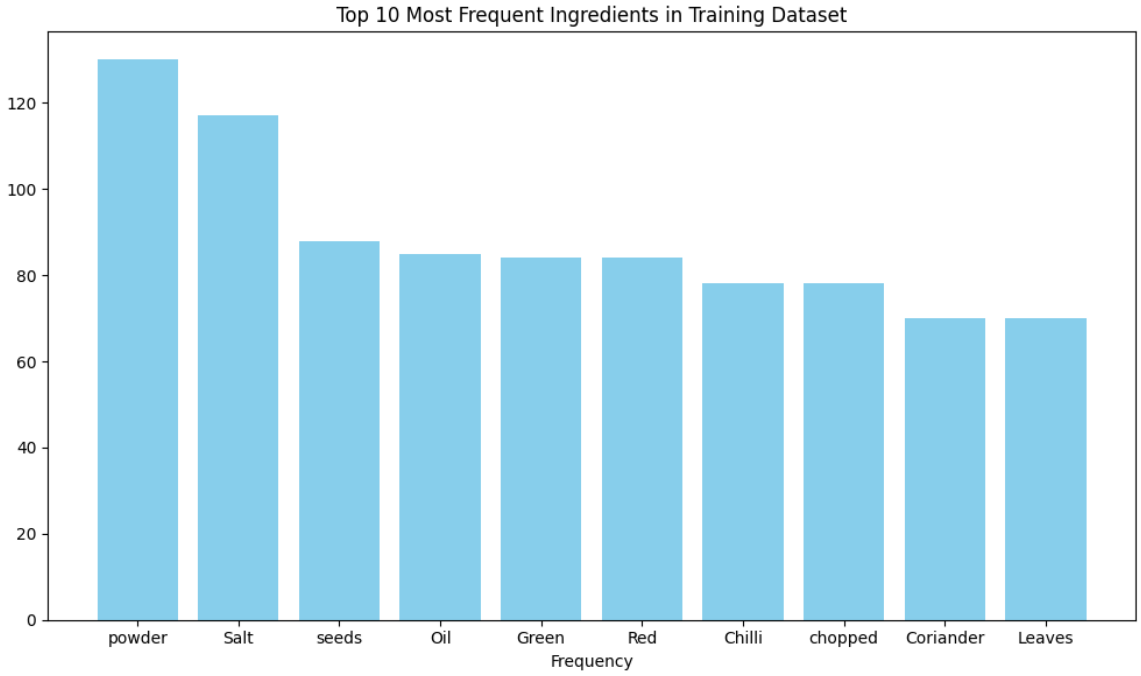
2. Feature extraction (capitalization, POS, word shape, etc.).

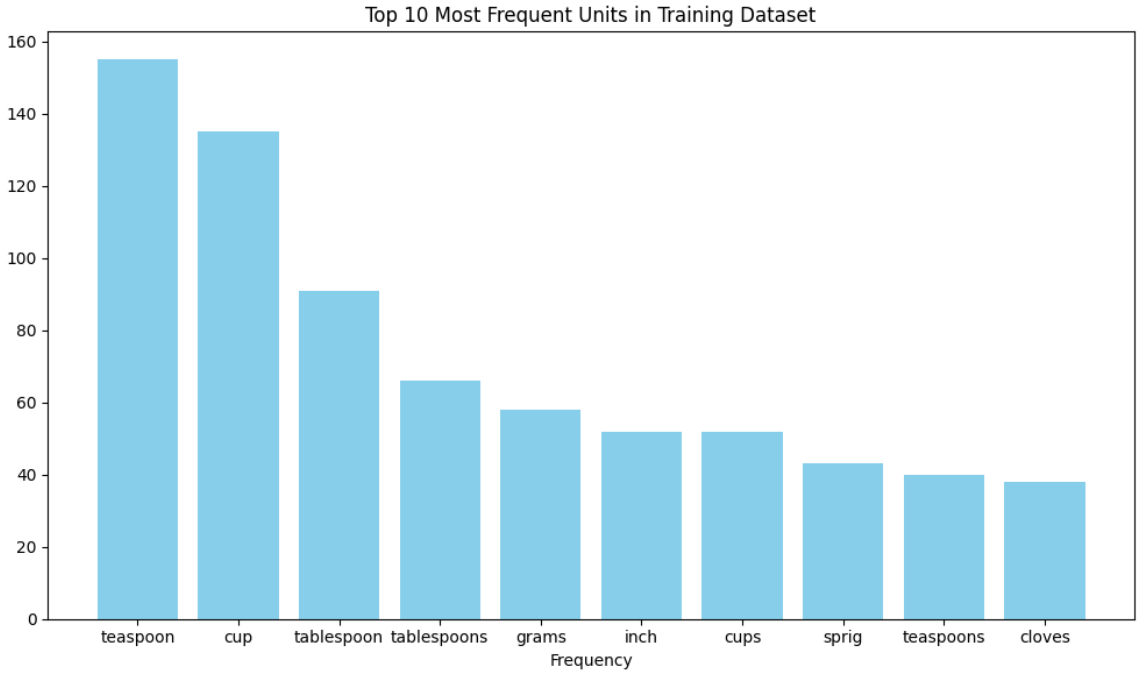
3. Model training using sklearn-crfsuite.

4. Evaluation using metrics like F1-score and classification report.

Visualisations and Explanations

• Label Distribution Chart: A bar graph was used to show how frequently each NER label occurs in the dataset. This helps in understanding class balance.





• Performance Metrics Chart: Visual representations of precision, recall, and F1-score help quickly identify which entity classes are performing well and which need improvement.

• Sample Prediction Examples: Annotated text samples show how the model classifies tokens into quantity, unit, and ingredient categories.

Analysis

The CRF model performed well, particularly in identifying quantities and ingredients. Units were more prone to misclassification, often confused with parts of ingredients. POS tags and contextual features were essential in improving the model's understanding of entity types.

Results

• Precision and recall were high for 'quantity' and 'ingredient' entities.

• Model generalized well on unseen recipe examples.

• Confusion matrix revealed some overlap between 'unit' and 'ingredient' classifications.

Insights

• Recipes are semi-structured and require NLP for reliable parsing.

• Contextual features (e.g., surrounding words, POS tags) significantly improve entity extraction.

• Feature engineering plays a major role in traditional CRF-based systems.

Outcomes

• A CRF-based NER model was developed and evaluated.

• The system effectively extracts quantities, units, and ingredients.

• The model can be integrated into applications to convert free-text recipes into structured format