Identifying Key Entities in Recipe Data

CREATED BY-

BIKASH SARKAR

Problem Statement

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.

Python Code

GITHUB LINK-
HTTPS://GITHUB.COM/BIKASH3/NLPIDENTIFYINGKEYENTITIES

Project Summary

- **Domain-Specific NER Application**: Tailoring Named Entity Recognition to the culinary domain addresses a niche use case, enabling structured understanding of informal and varied recipe language. **CRF for Context-Aware Sequence Labeling**: Leveraging Conditional Random Fields ensures that predictions factor in surrounding context, improving accuracy in identifying interdependent entities like ingredients and measurements.
- **Enhanced Data Utility**: Unstructured recipe text is transformed into structured datasets, facilitating advanced search, recommendation, and personalization features.
- **Real-World AI Impact**: Demonstrates practical machine learning application in domains beyond standard NLP benchmarks, particularly in the growing **food-tech** and **wellness** sectors.
- **Interdisciplinary Value**: Bridges culinary knowledge, linguistic processing, and machine learning—showcasing the potential of AI in creative and everyday life scenarios.

Methodology

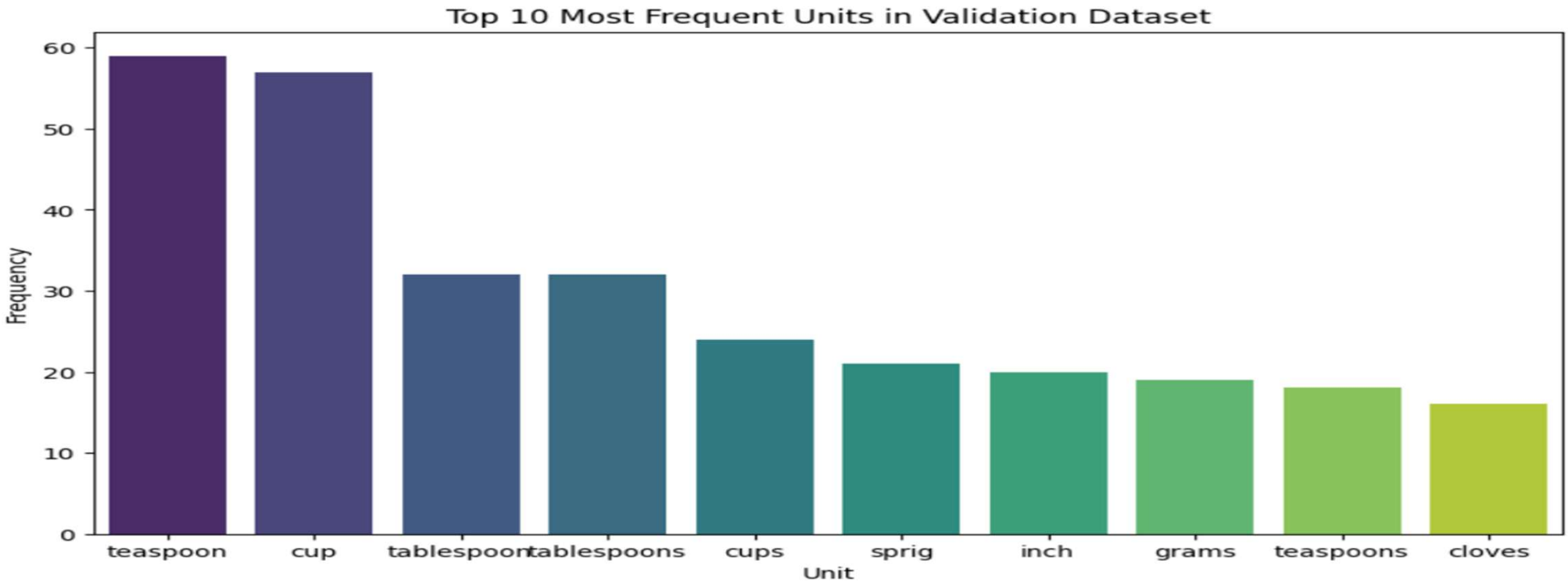
The development of the NER system using Conditional Random Fields (CRF) followed these sequential steps:

- **Library Import and Setup**
  - Essential Python libraries for data manipulation, visualization, natural language processing, and CRF modeling were imported and configured for seamless development.
- **Data Ingestion and Preparation**
  - Recipe datasets were loaded from reliable sources. The text data was cleaned and standardized to ensure uniformity in formatting, casing, and punctuation.
- **Train-Validation Split**
  - The annotated dataset was divided into training and validation sets to facilitate model training and unbiased performance evaluation.
- **Exploratory Data Analysis (EDA) – Training Set**
  - A detailed analysis was conducted on the training data to uncover patterns in entity distribution, token frequency, and common structures in recipe texts.
- **Exploratory Data Analysis (EDA) – Validation Set**
  - Similar exploratory techniques were applied to the validation set to compare distributions and ensure representativeness across subsets.
- **Feature Extraction for CRF Modeling**
  - Token-level features such as word morphology, part-of-speech tags, word casing, and neighboring word context were extracted to enable informed sequence modeling.
- **CRF Model Construction and Training**
  - A CRF model was instantiated and trained on the labeled training data. Hyperparameters were fine-tuned using cross-validation techniques to optimize performance.
- **Prediction and Evaluation**
  - The trained model was applied to the validation set. Performance was measured using metrics like Precision, Recall, and F1-score across each entity class.
- **Error Analysis**
  - Misclassified entities from the validation set were analyzed to uncover common sources of error, providing guidance for potential improvements in annotation and feature engineering.

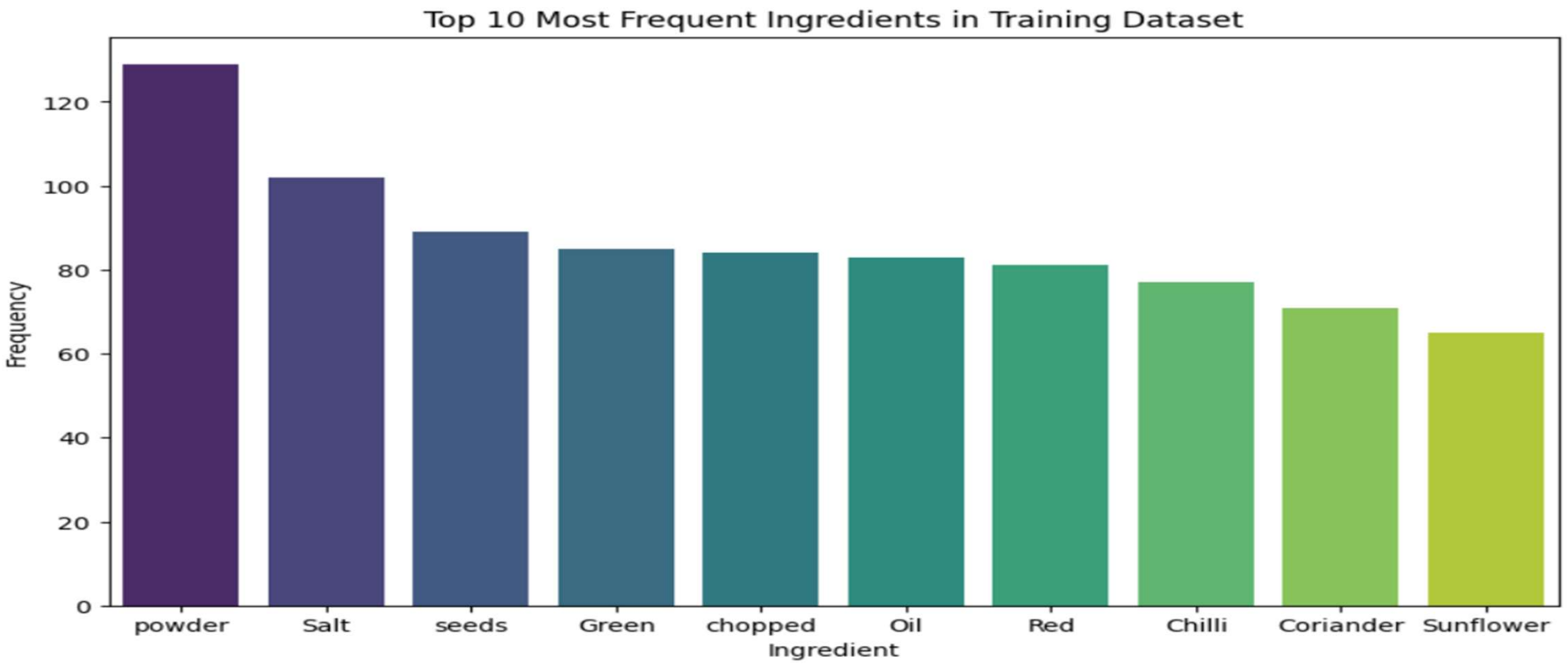
Techniques

- **Library Import and Setup**
  - Utilized essential Python libraries such as pandas, numpy, matplotlib, seaborn, sklearn, and sklearn-crfsuite.
- **Data Handling & Preparation**
  - Used libraries like pandas and numpy for data management.
  - Cleaned and normalized recipe text for consistency and accuracy.
- **Data Splitting & EDA**
  - Split data using train_test_split() to ensure fair evaluation.
  - Conducted visual and statistical exploratory data analysis (EDA) on both training and validation sets.
- **Feature Engineering**
  - Extracted token-level features such as word shape, POS tags, and contextual tokens.
  - Leveraged linguistic and morphological traits relevant to recipes.
- **Model Training with CRF**
  - Implemented and trained the CRF model using sklearn-crfsuite.
  - Applied hyperparameter tuning (e.g., L1/L2 regularization) for optimization.
- **Evaluation & Error Analysis**
  - Used metrics like Precision, Recall, and F1-score to assess performance.
  - Analyzed errors using confusion matrices and visualization techniques to uncover misclassifications.

Visualisations



Top frequent units in training data



Top frequent ingredients in training data

Key Insights

- **Domain-Specific NER Application:**
  - Tailoring Named Entity Recognition to the culinary domain addresses a niche use case, enabling structured understanding of informal and varied recipe language.
- **CRF for Context-Aware Sequence Labeling:**
  - Leveraging Conditional Random Fields ensures that predictions factor in surrounding context, improving accuracy in identifying interdependent entities like ingredients and measurements.
- **Enhanced Data Utility:**
  - Unstructured recipe text is transformed into structured datasets, facilitating advanced search, recommendation, and personalization features.
- **Real-World AI Impact:**
  - Demonstrates practical machine learning application in domains beyond standard NLP benchmarks, particularly in the growing food-tech and wellness sectors.
- **Interdisciplinary Value:**
  - Bridges culinary knowledge, linguistic processing, and machine learning—showcasing the potential of AI in creative and everyday life scenarios.

Conclusion

This project successfully demonstrates the development of a domain-specific Named Entity Recognition (NER) system for culinary data using Conditional Random Fields (CRF). By accurately identifying and categorizing ingredients, quantities, and units from unstructured recipe texts, the model transforms scattered textual data into structured insights. The use of CRF enables contextual understanding of sequential terms, enhancing both precision and reliability in entity extraction.

Beyond its technical merit, the project underscores how machine learning can meaningfully impact real-world applications—empowering food-tech platforms, dietary management tools, and intelligent shopping assistants. It highlights the interdisciplinary synergy between natural language processing and everyday life, paving the way for smarter and more intuitive culinary experiences powered by AI.

Thank You