DATA*6100

FINAL PROJECT PROPOSAL

Recipe Classification Dataset: What’s Cooking?

University of Guelph

Submitted by: Kshitiz Pokhrel, Santosh Kumar Satapathy

Under the guidance of: Dr. Mihai Nica

PROJECT AIM

Our primary goal in this project is to develop and fine-tune a neural network model that can navigate the text data of the "What's Cooking" dataset and make accurate predictions about cuisine types. This involves exploring and applying various text processing techniques, like tokenization, vectorization, and neural network architectures adept at handling textual information. The challenge lies in effectively interpreting the variances in ingredient lists and utilizing these insights to make accurate predictions. A successful completion of this project would help us advance our understanding and capabilities in text-based data processing and predictive modeling in the context of an engaging and practical dataset.

INTRODUCTION TO THE DATASET

The "What's Cooking" dataset from Kaggle is a compilation of various cuisines, each assigned to one of 20 diverse cuisine types. These range from more popular cuisines like Italian and Chinese to more unique ones like Brazilian and Russian. The dataset's key feature is the list of ingredients commonly used in each cuisine type, provided as text. The dataset presents a unique challenge due to the varying number of ingredients, the recurrence of similar ingredients across various cuisines, and the way these ingredients are described textually. It serves as an excellent opportunity to test our skills in creating predictive models using text-based data.

METHODOLOGY

In this project, we are aiming to build a neural network model to predict cuisines based on the ingredients list and evaluate our model using various metrics and comparing it with other ML models. The methodologies that will be used across different sections of the project are:

Data Preprocessing:

- We will begin by vectorizing the ingredients list, converting the textual data into a format that can be provided as input to machine learning models.
- The categorical data representing the cuisine types will be encoded to facilitate the model's ability to classify the recipes.

Neural Networks:

- **Convolutional Neural Networks (CNNs)** will be employed for their proficiency in pattern recognition, which is expected to be useful in identifying patterns within ingredient lists.
- **Recurrent Neural Networks (RNNs)** will be utilized to process the sequential nature of the data found in recipes.

Feature Extraction and Selection:

Techniques such as TF-IDF will be used to determine the significance of each ingredient in relation to the cuisine types.

Model Optimization and Validation:

- To optimize the model, we will use methods like cross-validation and grid search to fine-tune the hyperparameters.
- The performance of the model will be evaluated on metrics such as accuracy, precision, recall, and F1 score to ensure reliability.

Comparative Analysis with Other ML Techniques:

We will also explore alternative machine learning models such as Random Forests and Support Vector Machines (SVMs) to compare their effectiveness against our neural network model.