The Impact of Named Entity Recognition on Transformer-Based Multi-Label Dietary Recipe Classification

Kemalcan Bora
Universitat Pompeu Fabra
Barcelona, Spain
kemalcan.bora01@estudiant.upf.edu

Horacio Saggion
Universitat Pompeu Fabra
Barcelona, Spain
horacio.saggion@upf.edu

Abstract

This research explores the impact of Named Entity Recognition (NER) on transformer-based models for multi-label recipe classification by dietary preference. To support this task, we introduce the NutriCuisine Index: a collection of 23,932 recipes annotated across six dietary categories (Healthy, Vegan, Gluten-Free, Low-Carb, High-Protein, Low-Sugar). Using BERT-base-uncased, RoBERTa-base, and DistilBERT-base-uncased, we evaluate how NER-based preprocessing affects the performance (F1-score, Precision, Recall, and Hamming Loss) of Transformer-based multi-label classification models. RoBERTa-base shows significant improvements with NER in F1-score (ΔF1 $= +0.0147, p < 0.001$), Precision, and Recall, while BERT and DistilBERT show no such gains. NER also leads to a slight but statistically significant increase in Hamming Loss across all models. These findings highlight the model dependent impact of NER on classification performance.

1 Introduction

The growth of obesity, diabetes and diet related diseases has significantly intensified the focus on healthy lifestyles (Kupper, 2005) and the necessity for specific nutritional guidelines (James and Gill, 2022; Voigt et al., 2014; Reece et al., 2009; Kessler and Michalsen, 2018; Ley et al., 2014; Moore et al., 2004; Vijan et al., 2005). According to the World Health Organization (WHO), cardiovascular diseases are the leading cause of death from non communicable diseases (NCDs), resulting in approximately 17.9 million deaths each year. This is followed by cancers, which cause about 9.3 million deaths annually. Chronic respiratory diseases account for 4.1 million deaths and diabetes, including deaths from kidney diseases caused by diabetes, accounts for 2.0 million deaths each year [1].

The increasing global focus on healthier eating habits, customized to meet individual health conditions, is manifest in the move towards more health conscious diets (Curtain and Grafenauer, 2019). Studies from various countries find out a considerable demand for personalized nutrition, emphasizing a desire for recipes that satisfy both health and individual dietary preferences (Ge et al., 2015). This trend highlights the need food preparation and recipe development, designed to accommodate the varied dietary needs and health objectives of individuals globally (Curtain and Grafenauer, 2019; Ge et al., 2015). As a result, diets such as Gluten-Free, Vegan, Nut-Free, Low-Sugar are gaining popularity. Studies indicate that these diets tend to be adhered to more consistently compared to other restrictive diets (Cruwys et al., 2020).

However growing interest in health conscious diets, we still face challenges in fully understanding and categorizing the nutritional content and suitability of recipes. Many existing recipe databases lack comprehensive nutritional information or diet types, limiting their utility for individuals with specific dietary needs. To address this gap, we developed the NutriCuisine Index, a database that provides recipe instructions, detailed nutritional content, and classifications into various dietary categories, then conducted various experiments for multi-label classification using this database we developed.

Our work makes three primary contributions: (1) We introduce the NutriCuisine Index, a comprehensive recipe database comprising 23,932 recipes categorized across 22 recipe categories and we provide expert dietitian annotations for six key dietary labels (Healthy, Vegan, Low-Carb, Gluten-

[1] https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases accessed July 26, 2025

Proceedings of Recent Advances in Natural Language Processing,pages 184–193
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-022

Free, High-Protein, and Low-Sugar) within this dataset. (2) We perform an investigation into the impact of Named Entity Recognition (NER) on the performance of three transformer-based language models (BERT-base-uncased, RoBERTa-base and DistilBERT-base-uncased) for multi-label dietary classification from recipe ingredients, focusing on the six expert-annotated dietary labels. (3) We deliver detailed performance analyses, revealing that incorporating NER yields statistically significant F1-score improvements (ΔF1 $= +0.0147, p < 0.001$), precision, and recall for RoBERTa-base on this task. In contrast, BERT-base-uncased and DistilBERT-base-uncased do not realize similar gains, with all models showing a slight increase in Hamming Loss.

In this paper, we first explore related research in Section 2 to contextualize our work. Section 3 details our structured methodology for evaluating the influence of NER preprocessing on transformer-based models for multi-label classification. We then introduce the NutriCuisine Index, our novel dataset, in Section 4, discussing its creation, content and significance. Section 5 presents our experimental findings, followed by a discussion of these results. Finally, Section 6 summarizes our contributions, key findings, and potential directions for future research. Through this work, we aim to contribute to a more comprehensive understanding of recipe nutrition and dietary suitability, potentially improving recipe recommendations and dietary planning tools.

2 Related Work

This section reviews literature related to our research, focusing on two primary domains: computational methods for classifying recipes and ingredients with an emphasis on dietary characteristics, and the application of NER in understanding food-related textual data.

Early research into recipe and ingredient analysis often focused on tasks like identifying cuisines based on ingredient combinations. For instance, Guria et al. (2023) investigated the relationship between cuisines and ingredients, demonstrating that techniques like TF-IDF of ingredient lists coupled with traditional machine learning classifiers (e.g., Naive Bayes) could distinguish between major cuisine types.

Similarly, in the domain of ingredient identification, DeepFood (Pan et al., 2017) showcased the utility of Convolutional Neural Networks (CNNs) for classifying ingredients from images, often combining deep learning extracted features with classifiers like Support Vector Machines (SVMs). While these studies are foundational for food informatics, their primary aim was often ingredient recognition or general cuisine classification, rather than multi-label dietary attribute classification (e.g., 'Vegan', 'Low-Carb', 'Gluten-Free') that forms the core of our investigation. Our work extends these efforts by employing modern Transformer-based architectures for this specific dietary categorization task.

Research has also leveraged computational methods to link food to health outcomes and standardize food data.

Campese and Pozza (2021) utilized Natural Language Processing (NLP) techniques, comparing shallow learning models (e.g., SVMs with lexical features) against deep learning approaches (e.g., Recurrent Neural Networks) to classify foods based on their potential association with inflammation.

While their goal of identifying health-related food properties shares conceptual similarities with our objective, their focus was on a specific health marker (inflammation) rather than comprehensive multi-label dietary preferences. Furthermore, their NLP methodologies differ from our approach, which specifically examines the impact of NER as a preprocessing step for Transformer models.

Concurrently, other efforts have concentrated on classifying foods by processing levels, such as the system proposed by Monteiro et al. (2010) for Brazilian diets and more recently by Dickie et al. (2023) for the Australian food supply.

These frameworks are valuable for public health policy but address a different facet of food characterization processing level rather than the recipe level ingredient-based dietary classification we target.

In parallel, significant advancements have occurred in food data standardization. The FoodEx2 framework, along with systems like StandFood that implement it (Eftimov et al., 2017; European Food Safety Authority, 2011), aims to establish a comprehensive and hierarchical system for describing food items across diverse food safety and consumption survey domains. While such standardization is critical for data interoperability and large-scale epidemiological studies, these frameworks do not typically provide the granular, expert-annotated dietary labels (e.g., 'Healthy,' 'High-Protein,' 'Low-

Sugar') for individual recipes that our NutriCuisine Index specifically offers. This distinction is key for facilitating research in personalized nutrition and developing tools for dietary guidance.

NER has also emerged as a valuable tool for extracting structured information from unstructured recipe text. Several datasets have been developed to train and evaluate NER models in this domain. The TASTEset dataset (Wróblewska et al., 2022), for example, comprises 700 recipes with over 13,000 annotated entities. As detailed in Table 1, TASTEset targets a wide array of culinary concepts, including 'Food,' 'Quantity,' 'Unit,' 'Process,' 'Physical Quality,' 'Color,' and 'Taste.'

Entity	Description
Physical Quality	Texture, state (e.g., liquid, solid)
Food	Specific food items or ingredients
Quantity	Amount of each food item used
Unit	Measurement units (e.g., cups, grams)
Color	Color of food items
Taste	Flavor profile (e.g., sweet, bitter)
Part	Specific part of ingredient (e.g., chicken breast)
Process	Preparation methods (e.g., ground, chopped)

Table 1: TASTEset Entity Types

Another notable resource is the Food Ingredient Named Entity Recognition (FINER) dataset, introduced by Komariah et al. (2023) through their Semi-Supervised Multi-Model Prediction Technique (SMPT).

FINER focuses predominantly on food ingredient entities, labeled from a substantial corpus of online recipes. While both TASTEset and FINER are pivotal for advancing general culinary NER, their annotated entity sets are not explicitly tailored for, nor have they been extensively benchmarked on, the task of multi-label dietary preference classification. For instance, identifying 'Food' entities (e.g., "flour," "sugar," "chicken") is a fundamental first step provided by these NER systems. However, determining if a recipe qualifies as 'Gluten-Free,' 'Vegan,' or 'Low-Sugar' often requires more nuanced interpretation, consideration of implicit knowledge (e.g., that 'butter' is not vegan, or that 'soy sauce' often contains gluten unless specified), or the aggregation of information across multiple ingredients. Our research investigates whether general-purpose NER, by explicitly highlighting ingredient entities and potentially other tags like quantities or units, can augment the ability of Transformer-based language models to perform such dietary classifications when trained on a dataset like our NutriCuisine Index, which contains explicit expert-provided dietary labels.

Finally, large-scale recipe datasets have significantly fueled research in culinary NLP, enabling tasks from recipe generation to food image analysis. Key examples include:

- **YouCook2 (Zhou et al., 2018):** This dataset, with 2000 long, untrimmed videos from 89 cooking recipes, is instrumental for research in instructional video understanding.
- **Recipe1M/Recipe1M+ (Salvador et al., 2019):** Comprising over one million recipes paired with approximately 13 million food images, this collection is a cornerstone for multimodal food research, particularly in image-to-recipe retrieval and generation.
- **RecipeNLG (Bień et al., 2020):** An extension of Recipe1M, RecipeNLG offers over two million unique recipes and incorporates NER tags for food names, facilitating more targeted ingredient-level analyses and innovations in recipe generation.

Despite their scale (e.g., YouCook2 (Zhou et al., 2018), Recipe1M/1M+ (Salvador et al., 2019), RecipeNLG (Bień et al., 2020)), existing large recipe datasets often lack comprehensive, expert-verified annotations for specific dietary attributes (e.g., 'Vegan', 'Low-Carb') or detailed nutritional information. This limits their utility for dietary categorization and personalized recommendation. Our NutriCuisine Index fills this gap by providing 23,932 recipes annotated by expert dietitians with six key dietary labels (Healthy, Vegan, Low-Carb, Gluten-Free, High-Protein and Low-Sugar; see Section 4). This resource supports a investigation of multi-label dietary classification based on recipe text, including the specific impact of incorporating NER explored in this study.

3 Proposed Methodology

The methodology proposed herein is designed to evaluate the impact of NER on the efficacy of Language Models (LMs) in multi-label classification of recipes according to dietary categories. This process integrates the development of a dietary database, the training of a NER model tailored

to culinary contexts, and the application of transformer based LMs for classification, with a comparative analysis of performance with and without NER preprocessing. The approach comprises three principal stages, detailed as follows.

i. **Database Development**: The initial phase entails the construction of the NutriCuisine Index, a structured repository of dietary-specific recipes. Utilizing the Scrapy framework[2], we collected 23,932 recipes from reputable online sources, including BBC Good Food, Heart UK, and Delish. These data were stored in Elasticsearch[3], selected for its NoSQL architecture and proficiency in text processing. To ensure reliability, two expert dietitians annotated and validated the dietary labels. For the classification experiments, recipes were categorized into six primary classes based on the **Healthy, Vegan, Low-Carb, Gluten-Free, High-Protein** and **Low-Sugar**. This step establishes a quality dataset, important for classification tasks.

ii. **Named Entity Recognition Model Training and Application**: A `bert-base-cased` model was fine-tuned for NER on the TASTEset database, which was annotated using the IOB2 tagging scheme (Ramshaw and Marcus, 1999; Tjong Kim Sang and Buchholz, 2000). In IOB2 format, each token is labeled as either the Beginning (B-) or Inside (I-) of a named entity, or Outside (O) if it does not belong to any entity. This scheme allows precise identification of entity spans such as `B-FOOD`, `I-FOOD`, `B-QUANTITY`, etc.

   We employed 5-fold cross-validation, training each fold for 7 epochs using a learning rate of 2×10^{-5}, batch size 16 (train/eval), weight decay 0.01, and a maximum sequence length of 128 tokens. Sub-word tokens produced by the tokenizer were assigned a label of -100 to exclude them from loss computation. For each fold, the model checkpoint with the lowest evaluation loss was selected and designated as the NERPredictor.

   For the multi-label classification task, this NERPredictor processed ingredient texts from our NutriCuisine Index. Identified entities (e.g., FOOD, QUANTITY) were used to reformat ingredient strings by inserting special tokens (e.g., `[B-QUANTITY]1 [B-UNIT]tablespoon [B-FOOD]oil [ING_END]`). The set of special tokens, including `[B-COLOR]`, `[B-FOOD]`, `[B-QUANTITY]`, `[B-UNIT]`, `[I-FOOD]`, `[I-QUANTITY]`, `[O]`, and our custom delimiter `[ING_END]`, were added to the vocabulary of the subsequent transformer-based classification models.

iii. **Multi-Label Classification**: The final stage employs three transformer-based models: BERT-base-uncased, RoBERTa-base and DistilBERT-base-uncased to classify recipes into multiple dietary categories. For each model, the tokenizer was augmented with the special NER related tokens (detailed in Stage **ii.**), and the model's token embeddings were resized accordingly to accommodate these new tokens. Ingredient lists from the NutriCuisine Index were processed in two configurations: (i) with NER tags incorporated and (ii) without NER preprocessing (using outputs from the `processed_ingredients` field, which result from applying the custom text preprocessing pipeline). Data preparation included appropriate tokenization for each model and a label binarization process, where each of the six dietary categories was represented as an independent binary label for multi-label classification. The dataset was split into training (70%), validation (15%), and test (15%) sets. Models were trained for up to 6 epochs with a batch size of 16, employing the AdamW optimizer (lr=2×10^{-5}) and Binary Cross-Entropy with Logits Loss. Early stopping with a patience of 3 epochs and a minimum delta of 0.001 was used to prevent overfitting, based on validation loss. An approximate randomization test (1000 repetitions) assessed the statistical significance of NER's impact on classification performance metrics (F1-score, Precision, Recall, Accuracy, and Hamming Loss).

This structured methodology aims to clarify the differential effects of NER on transformer model performance, providing a foundation for optimizing dietary classification systems.

[2] https://scrapy.org/
[3] https://www.elastic.co/elasticsearch

4 NutriCuisine Index: Tailored Recipes for Every Diet

We introduce the NutriCuisine Index, a new database of **23,932** recipes scraped from public sources like BBC Good Food[4], Heart UK[5], and Delish[6]. The data, confirmed usable for research per General Data Protection Regulation (GDPR), includes ingredients, preparation steps, nutritional content, and diet types. Addressing a gap in existing resources like RecipeNLG, which lack detailed dietary categorizations, NutriCuisine provides multi-label dietary classifications across 22 categories, of which 6 are key health-relevant dietary labels validated by expert dietitians: **Healthy, Vegan, Low-Carb, Gluten-Free, High-Protein**, and **Low-Sugar**. These labels were initially scraped and then reviewed by two certified dietitians, who conducted manual annotation based on ingredient analysis and nutritional assessment. Each expert annotated overlapping subsets of the recipe corpus to allow for consistency checks. To quantify annotation reliability, inter-annotator agreement (IAA) was calculated using Cohen's kappa (Artstein, 2017) across 1,000 sampled recipes, achieving an agreement score of 0.82. The experts used a shared labelling guide and resolved any disagreements by discussing the recipe together. The final distribution of the validated dietary types is shown in Table 2, while other non-exclusive recipe categories are shown in Table 3.

Dietary Types	Count	Dietary Types	Count
Vegan	20438	Low-Carb	1469
Gluten-Free	11353	Low-Sugar	1556
Healthy	9416	High-Protein	864

Table 2: Distribution of recipes with validated dietary types.

Other Categories	Count	Other Categories	Count
Freezable	4096	Slow-Cooker	63
Easily-Doubled	597	Whole-30	74
Easily-Halved	447	Soup	60
30-Minute-Meals	129	Lunch	58
Appetizers	115	Kid-Friendly	99
Dinner	158	Salads	61
One-Pot-Meals	86	Pressure-Cooker	45

Table 3: Distribution of recipes across other general categories. Counts are non-exclusive.

[4] https://www.bbcgoodfood.com/
[5] https://www.heartuk.org.uk/
[6] https://www.delish.com/

The NutriCuisine database schema includes detailed recipe information, including various dietary requirements. Table 4 outlines the key fields that provide a complete recipe overview. Listing 1 presents a JSON example of the data format.

Field	Description
Title	Recipe name
Serve	Number of servings
Link	URL to original recipe source
Ingredients	List of ingredients with quantities
Directions	Step by step cooking instructions
Nutrition	Nutritional content per serving
Diets	Suitable diet types for the recipe

Table 4: NutriCuisine Database Schema

```
{
  "title": "Creamy Courgette Lasagne",
  "serve": "4",
  "link": "https://www.bbcgoodfood.com
      /...",
  "ingredients": ['9 dried lasagne
      sheets', '1 tbsp sunflower oil',
      '1 onion, finely chopped'..],
"directions": ['Heat oven to 220C/fan,
    Boil lasagne sheets for 5 mins,
    rinse in cold water, and drizzle
    with oil.', 'Fry onion in a large
    pan. After 3 mins, add courgettes
    ....],
  "nutrition": {"kcal": "405",
                "fat": "21g",
                "carbs": "38g",
                "protein": "18g"},
  "diets": ["Low-Carb", "Vegetarian"]}
```

Listing 1: NutriCuisine Index Sample

5 Experiments & Results

This section details the experimental setup and outcomes for two core components of our study: first, the development and evaluation of our NER model on the TASTEset dataset (Wróblewska et al., 2022), and second, the multi-label classification of recipes from the NutriCuisine Index into dietary categories, evaluating the impact of NER integration.

5.1 Named Entity Recognition Model

An NER model was developed to identify and structure culinary entities within ingredient texts, forming a preparatory step for the dietary classification task. We trained and evaluated a `bert-base-cased` model for token classification using the TASTEset dataset.

5.1.1 Dataset and Training Protocol

The NER model development followed the protocol described in Section 3, Stage **ii.** Briefly,

the TASTEset dataset, with words and their corresponding IOB2 labels, was used. A 5-fold cross-validation approach was employed to train the `bert-base-cased` model for 7 epochs per fold, with a learning rate of 2×10^{-5} and a batch size of 16. The `NERPredictor` component, subsequently used for annotating ingredients for the multi-label classification task, is based on the models trained through this cross-validation process.

5.1.2 Data Preparation

A data preprocessing protocol was implemented to enhance the quality and consistency of input data for the NER experiments. This process encompassed text normalization, including case uniformity and character encoding standardization, as well as meticulous cleaning procedures. Numerical data underwent conversion to decimal format, with careful attention paid to mixed fractions and multiplicative expressions. These measures aim to optimize model performance by standardizing and simplifying the dataset for more accurate dietary classification.

5.1.3 Training Process

NER task utilized the `bert-base-cased` (Bidirectional Encoder Representations from Transformers) architecture. Our choice of this model was motivated by the work of Wróblewska et al. (2022) (Wróblewska et al., 2022), who showed good performance with BERT-based models in similar culinary NER tasks, providing a relevant benchmark for our approach.

As established in Section 3 (Stage **ii.**) and recapped in Section 5.1.1, the NER model was trained using a 5-fold cross-validation strategy. Within each fold, training was conducted for 7 epochs. This epoch count was determined to adequately balance model learning on the TASTEset data against computational resources and to prevent potential overfitting that might occur with more extended training. The learning rate was set to $2 \cdot 10^{-5}$ to ensure stable convergence during fine-tuning. Batch sizes for both training and evaluation were 16, a common choice that balances memory constraints with effective gradient estimation. For regularization, a weight decay of 0.01 and a dropout rate of 0.1 (applied to specified layers) were employed. The AdamW optimizer (Loshchilov and Hutter, 2019) was used, and a random seed of 42 was set to ensure the reproducibility of the training experiments. Furthermore, the tokenizer was configured with a `max_length` of 128. During training, labels were aligned with tokenized inputs, and subword tokens were assigned a label of -100 to be ignored by the loss function. The model was also configured to output hidden states and attention weights, which can be valuable for deeper model analysis, although these were not directly utilized in the main classification pipeline.

5.1.4 NER Performance

The performance of the NER model, aggregated across the 5 folds of cross-validation, is presented in Table 5. The model was tasked with identifying entity types such as FOOD, QUANTITY, UNIT, and COLOR, each distinguished by IOB2 tags (e.g., B-FOOD for the beginning of a food entity, I-FOOD for tokens inside a food entity, and O for tokens outside any entity).

Entity Tag	Precision	Recall	F1-Score	Support
B-COLOR	0.901	0.928	0.914	332
B-FOOD	0.939	0.957	0.948	5481
B-QUANTITY	0.954	0.989	0.971	5202
B-UNIT	0.958	0.985	0.972	4334
I-FOOD	0.889	0.915	0.902	2767
I-QUANTITY	0.948	0.882	0.914	618
O	0.962	0.907	0.933	8096
Overall Accuracy			0.946	26830
Macro Avg.	0.936	0.938	0.936	26830
Weighted Avg.	0.947	0.946	0.946	26830

Table 5: Overall NER Model Performance (5-Fold Cross-Validation on TASTEset)

The average training loss across the 5 folds was 0.360, and the average evaluation loss was 0.172. The NER model demonstrated strong performance, particularly in identifying B-UNIT (F1: 0.972), B-QUANTITY (F1: 0.971), and B-FOOD (F1: 0.948) tags. The I-FOOD and I-QUANTITY tags also achieved F1-scores of 0.902 and 0.914, respectively. These results confirm the model's capability to accurately recognize relevant culinary entities, providing a reliable basis for feature enrichment in the subsequent dietary classification experiments. The entity tags used to generate special tokens for the multi-label classification task were `[B-COLOR]`, `[B-FOOD]`, `[B-QUANTITY]`, `[B-UNIT]`, `[I-FOOD]`, `[I-QUANTITY]` and `[O]`.

5.1.5 Results and Analysis

In comparison to Wróblewska et al.'s work, we will compare our results with BERT-based models, specifically BERT-base-uncased. Table 6 compares the performance of our BERT-base-uncased imple-

mentation, incorporating the aforementioned preprocessing techniques and hyperparameters, with the F1 scores reported for BERT-base-uncased models in Wróblewska et al.'s work.

Entity Type	Our Model (BERT-base-uncased)	Wróblewska et al. (2022) (BERT-base-uncased)
FOOD	**0.933**	0.889
QUANTITY	0.965	**0.983**
UNIT	0.972	**0.979**
COLOR	**0.914**	0.873

Table 6: NER Performance Comparison: F1-Scores for Entity Classes. Our results are weighted averages of B/I tags from 5-fold CV.

These results show that our NER approach works well for identifying important parts of food ingredients. Using what we learned from NER, we then moved on to recipe classification. The next part of our work involves creating and testing a method for multi-label classification of recipes.

5.2 Multi-Label Dietary Classification

This subsection details the experiments and results for classifying recipes from the NutriCuisine Index into six dietary categories: Healthy, Vegan, Low-Carb, Gluten-Free, High-Protein, and Low-Sugar. We evaluate how features extracted through NER impact the performance of three transformer-based models. Notably, we excluded Dairy-Free and Nut-Free from this classification task because they are allergy-related dietary (Vlieg-Boerstra et al., 2023; Novak and Leung, 2005; Fälth-Magnusson et al., 1989).

5.2.1 Experimental Setup and Evaluation Metrics

The experimental setup followed the protocol outlined in Section 3, Stage **iii.** Briefly, three models BERT-base-uncased, RoBERTa-base and DistilBERT-base-uncased were trained and evaluated in two configurations: (i) with ingredient texts augmented by NER tags and (ii) using only preprocessed ingredient texts (without NER tags). The dataset was split into 70% training, 15% validation, and 15% test sets. Models were trained for up to 6 epochs with a batch size of 16, using the AdamW optimizer (lr=2×10^{-5}), `BCEWithLogitsLoss`, and early stopping criteria (patience=3, min_delta=0.001).

Performance was primarily evaluated on the test set using F1-score (F1), Precision (P), Recall (R), exact match Accuracy (Acc.), and Hamming Loss (HL). An approximate randomization test (1000 repetitions) was used to determine the statistical significance of performance differences between the NER and no-NER configurations for each model. We report p-values for these key metrics.

5.2.2 Results and Analysis

Table 7 summarizes test set performance with and without NER, while Table 8 shows the statistical significance of the differences.

Model	F1	P	R	Acc.	HL
With NER					
RoBERTa-base	0.788	0.784	0.797	0.922	0.0157
DistilBERT-base-uncased	0.783	0.783	0.786	0.931	0.0133
BERT-base-uncased	0.791	0.790	0.795	0.943	0.0114
Without NER					
RoBERTa-base	0.773	0.772	0.778	0.922	0.0154
DistilBERT-base-uncased	0.786	0.786	0.790	0.942	0.0113
BERT-base-uncased	0.794	0.793	0.797	0.951	0.0097

Table 7: Overall Test Set Performance on the NutriCuisine Index for multi-label dietary classification.

Model	Metric	Δ	p-value
RoBERTa-base	F1	+0.0147	**0.001**
RoBERTa-base	Precision	+0.0116	**0.001**
RoBERTa-base	Recall	+0.0196	**0.001**
RoBERTa-base	Accuracy	0.0000	**0.001**[a]
RoBERTa-base	Hamming Loss	+0.0003	**0.001**
DistilBERT-base-uncased	F1	-0.0035	0.522
DistilBERT-base-uncased	Precision	-0.0030	0.497
DistilBERT-base-uncased	Recall	-0.0032	0.498
DistilBERT-base-uncased	Accuracy	-0.0106	**0.001**
DistilBERT-base-uncased	Hamming Loss	+0.0019	**0.001**
BERT-base-uncased	F1	-0.0025	0.523
BERT-base-uncased	Precision	-0.0028	0.504
BERT-base-uncased	Recall	-0.0014	0.466
BERT-base-uncased	Accuracy	-0.0084	**0.001**
BERT-base-uncased	Hamming Loss	+0.0017	**0.001**

Table 8: Significance of NER integration on the NutriCuisine Index (NutriIndex), where $\Delta = \text{Score}_{\text{NER}} - \text{Score}_{\text{NoNER}}$. Bold p-values indicate statistical significance ($p < 0.05$). For Hamming Loss (HL), positive Δ indicates worse performance. [a]RoBERTa accuracy: rounded $\Delta = 0$, but $p < 0.05$ from unrounded values.

As indicated in Table 7 and corroborated by the statistical analysis in Table 8, RoBERTa-base demonstrated a statistically significant improvement in F1-score ($\Delta = +0.0147, p < 0.001$), Precision ($\Delta = +0.0116, p < 0.001$), and Recall ($\Delta = +0.0196, p < 0.001$) when NER features were incorporated. In contrast, for DistilBERT-base-uncased and BERT-base-uncased, the integration of NER did not yield significant improvements in these sample-averaged F1, precision, or recall

metrics. These models generally performed numerically better without NER, though these differences were not statistically significant for F1S, PS, and RS.

Regarding overall label prediction accuracy (exact match) and Hamming Loss, NER integration showed a different trend. For DistilBERT-base-uncased and BERT-base-uncased, NER led to a statistically significant decrease in accuracy (DistilBERT: $\Delta = -0.0106, p < 0.001$; BERT: $\Delta = -0.0084, p < 0.001$) and a significant increase (i.e., worsening) in Hamming Loss (DistilBERT: $\Delta = +0.0019, p < 0.001$; BERT: $\Delta = +0.0017, p < 0.001$). RoBERTa-base also exhibited a statistically significant, albeit smaller, increase in Hamming Loss ($\Delta = +0.0003, p < 0.001$) with NER. Its accuracy, while appearing unchanged after rounding (Table 7), showed a statistically significant effect in the randomization test (Table 8), suggesting minor, consistent variations detrimental to exact match accuracy due to NER. While RoBERTa-base benefits from NER in terms of average per-sample F1, precision, and recall, this advantage does not extend to DistilBERT-base-uncased or BERT-base-uncased for these metrics. Furthermore, the introduction of NER tags appears to have a generally detrimental effect on metrics sensitive to exact label set predictions (Accuracy) or cumulative individual label errors (Hamming Loss).

To further investigate the impact on RoBERTa-base, Table 9 presents its per-class F1-scores.

Dietary Class	With NER	Without NER	Δ
Healthy	0.964	0.968	-0.004
Vegan	0.959	0.949	+0.010
Low-Carb	0.800	0.884	-0.084
Gluten-Free	1.000	0.997	+0.003
High-Protein	0.182	0.625	-0.443
Low-Sugar	0.987	0.980	+0.007
Macro Avg. F1	0.815	0.900	-0.085

Table 9: RoBERTa-base per class F1-Scores with and without NER on the Test Set.

The per class analysis for RoBERTa-base (Table 9) reveals a mixed impact of NER. F1-scores improved for 'Vegan' (+0.010), 'Gluten-Free' (+0.003), and 'Low-Sugar' (+0.007). However, performance degraded for 'Healthy' (-0.004), 'Low-Carb' (-0.084), and substantially for 'High-Protein' (-0.443). This disparate per-class effect explains why the Macro Average F1-score for RoBERTa-base decreased with NER (from 0.900 to 0.815), despite the improvement in F1-score. It suggests that while NER enhances the signal for certain classes or on average across samples, it may introduce noise or less effective structural cues for other classes, particularly those with smaller representation (like 'High-Protein') or where the entities identified by NER are less discriminative for that specific dietary type.

6 Conclusion and Future Work

We investigated the impact of NER on transformer-based multi-label dietary recipe classification, introducing the NutriCuisine Index a dataset of 23,932 recipes annotated for six dietary types and a custom NER model trained on TASTEset. Experiments with BERT-base-uncased, RoBERTa-base, and DistilBERT-base-uncased showed that while transformers classify dietary labels effectively (e.g., BERT-base-uncased Macro F1 0.966 without NER), NER's impact varies by model. RoBERTa-base benefited significantly in F1-score ($\Delta = +0.0147$, $p < 0.001$), Precision, and Recall, while others showed marginal or no improvement. NER also slightly increased Hamming Loss and reduced exact match accuracy in some cases.

These results suggest that NER can enhance specific architectures like RoBERTa, though its benefits are not universal. Future directions include refining NER with contextual disambiguation (e.g., "egg white" as ingredient vs. color), investigating RoBERTa-base's unique gains through attention analysis, and exploring alternative NER feature representations (e.g., embeddings or adaptive integration). Our datasets and code[7] [8] are publicly available to support further work in personalized nutrition and recipe analysis.

Acknowledgments

We acknowledge support from the Spanish State Research Agency through the María de Maeztu Units of Excellence Program (CEX2021-001195-M), and from the Departament de Recerca i Universitats de la Generalitat de Catalunya via the SGR-Cat 2021 grants. We also thank the anonymous reviewers for their comments and constructive feedback, which helped improve the quality of this work.

[7] https://github.com/NutriCuisine/NERonLLM
[8] https://github.com/NutriCuisine/database

References

Ron Artstein. 2017. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.

Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. Recipenlg: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28.

Stefano Campese and Davide Pozza. 2021. Food classification for inflammation recognition through ingredient label analysis: A real nlp case study. In *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 2*, pages 172–181. Springer.

Tegan Cruwys, Niklas K Steffens, S Alexander Haslam, Catherine Haslam, Matthew J Hornsey, Craig McGarty, and Daniel P Skorich. 2020. Predictors of social identification in group therapy. *Psychotherapy Research*, 30(3):348–361.

Felicity Curtain and Sara Grafenauer. 2019. Plant-based meat substitutes in the flexitarian age: an audit of products on supermarket shelves. *Nutrients*, 11(11):2603.

Sarah Dickie, Julie Woods, Priscila Machado, and Mark Lawrence. 2023. A novel food processing-based nutrition classification scheme for guiding policy actions applied to the australian food supply. *Frontiers in Nutrition*, 10:1071356.

Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. 2017. Standfood: standardization of foods using a semi-automatic system for classifying and describing foods according to foodex2. *Nutrients*, 9(6):542.

European Food Safety Authority. 2011. Evaluation of the foodex, the food classification system applied to the development of the efsa comprehensive european food consumption database. *EFSA Journal*, 9(3):1970.

K Fälth-Magnusson, N-IM Kjellman, and K-E MAGNUSSON. 1989. Effects of various types of diets on food allergy in the infant. *Acta Pædiatrica*, 78:53–56.

Mouzhi Ge, Francesco Ricci, and David Massimo. 2015. Health-aware food recommender system. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 333–334.

Sudipa Guria, Sudarshan Banerjee, Suddhasattwa Bhattacharjee, Swarup Paul, Sobhan Halder, and Priyanka Das. 2023. Classification of foods based on ingredients. In *2023 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE.

W Philip T James and Tim Gill. 2022. Obesity–introduction: history and the scale of the problem worldwide. *Clinical Obesity in Adults and Children*, pages 1–16.

Christian Kessler and Andreas Michalsen. 2018. Vegmed:«vegmed–scientific congress for plant-based nutrition and medicine».

Kokoy Siti Komariah, Ariana Tulus Purnomo, Ardianto Satriawan, Muhammad Ogin Hasanuddin, Casi Setianingsih, and Bong-Kee Sin. 2023. Smpt: A semi-supervised multi-model prediction technique for food ingredient named entity recognition (finer) dataset construction. In *Informatics*, 1, page 10. MDPI.

Cynthia Kupper. 2005. Dietary guidelines and implementation for celiac disease. *Gastroenterology*, 128(4):S121–S127.

Sylvia H Ley, Osama Hamdy, Viswanathan Mohan, and Frank B Hu. 2014. Prevention and management of type 2 diabetes: dietary components and nutritional strategies. *The Lancet*, 383(9933):1999–2007.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Carlos Augusto Monteiro, Renata Bertazzi Levy, Rafael Moreira Claro, Inês Rugani Ribeiro de Castro, and Geoffrey Cannon. 2010. A new classification of foods based on the extent and purpose of their processing. *Cadernos de saude publica*, 26:2039–2049.

Helen Moore, Carolyn D Summerbell, Lee Hooper, Kennedy Cruickshank, Avni Vyas, Paul Johnstone, Vicki Ashton, Peter Kopelman, and J Kennedy Cruickshank. 2004. Dietary advice for treatment of type 2 diabetes mellitus in adults. *Cochrane Database of Systematic Reviews*, (2).

Natalija Novak and Donald YM Leung. 2005. Diet and allergy: you are what you eat? *Journal of allergy and clinical immunology*, 115(6):1235–1237.

Lili Pan, Samira Pouyanfar, Hao Chen, Jiaohua Qin, and Shu-Ching Chen. 2017. Deepfood: Automatic multi-class classification of food ingredients using deep learning. In *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)*, pages 181–189. IEEE.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

E Albert Reece, Gustavo Leguizamón, and Arnon Wiznitzer. 2009. Gestational diabetes: the need for a common ground. *The Lancet*, 373(9677):1789–1797.

Amaia Salvador, Michal Drozdzal, Xavier Giró-i Nieto, and Adriana Romero. 2019. Inverse cooking: Recipe generation from food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10453–10462.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Sandeep Vijan, NS Stuart, J Thomas Fitzgerald, David L Ronis, Rodney A Hayward, S Slater, and Timothy P Hofer. 2005. Barriers to following dietary recommendations in type 2 diabetes. *Diabetic medicine*, 22(1):32–38.

Berber Vlieg-Boerstra, Marion Groetch, Emilia Vassilopoulou, Rosan Meyer, Kirsi Laitinen, Anne Swain, Raquel Durban, Olga Benjamin, Rachelle Bottse, Kate Grimshaw, et al. 2023. The immune-supportive diet in allergy management: a narrative review and proposal. *Allergy*, 78(6):1441–1458.

Kristin Voigt, Stuart G Nicholls, and Garrath Williams. 2014. *Childhood obesity: ethical and policy issues*. Oxford University Press, USA.

Ania Wróblewska, Agnieszka Kaliska, Maciej Pawłowski, Dawid Wiśniewski, Witold Sosnowski, and Agnieszka Ławrynowicz. 2022. Tasteset–recipe dataset and food entities recognition benchmark. *arXiv preprint arXiv:2204.07775*.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.