# Natural Language Processing Tasks Perspectives on Food Data

Maria Eduarda M. de Holanda
*Dep. de Ciência da Computação*
*Universidade de Brasília*
Brasília, DF, Brazil
eduarda.holanda@aluno.unb.br

***Abstract*—Food data analysis is a growing area in Natural Language Processing, considering many different tasks that can be performed in order to assistant health-related professional. In this article, two tasks were taken into consideration: text retrieval and machine learning classification. The proposed method contemplated two embeddings to structure the datasets, Word2Vec and GloVe. After that, a document ranking were performed to analyse the type of dataset selected and a classification task considering two state-of-the-art and recent machine learning approaches, CNN and LSTM. Experiments were conducted to assess the quality of the classification techniques and the results reported a satisfactory metrics evaluation, such as loss reduction and accuracy. Nonetheless, future work can be guided exploring another NLP tasks, such as NER and EL.**

***Index Terms*—food composition data, natural language processing, text retrieval, machine learning classification**

## I. Introduction

Natural Language Processing (NLP) is a sub field of Artificial Intelligence (AI) that includes various tasks and problems that aims to produce useful information or identify interesting patterns about human-produced data [1]. Along the many areas of study that NLP covers, food data analysis is in constant grow due to the challenges the wide diversity of foods brings to dietitians, nutrologist physicians and health-related professionals. In this regard, NLP as well as Machine Learning strategies can support those specialists for knowledge discovery on food data to facilitate or even improve their work.

In the last decades, several researches have explored food data analysis using natural language processing and machine learning algorithms. Some works on literature have proposed new methods for classifying the quality of drinks [2], recognizing entities in recipes [3] and comparing [4] different food data sets to explore its potential. Although the proposed methods were effective to analyse the chosen data, dealing with different data sets can be a challenging assignment since the results produced by machine learning techniques might not be as good as expected.

In this sense, this article mean to explore classification tasks and text retrieval tasks in two different datasets to better understand the type of input we are dealing with and to analyse different machine learning approaches, so that we can obtain better results. Thus, the proposed method also includes data preprocessing step for generating a structured representation of food data instances and an evaluation step employing well-known metrics for assessing the quality of the machine learning classifiers.

This paper is structured as follows. Section II presents the related works on analysing food data and the methods that were applied. Section III details the proposed method and its constituting steps. Section IV describes the experimental results that were conducted on two food datasets and using qualitative evaluation metrics. Section V concludes this paper and discusses possibilities for future work.

## II. Related work

The literature on food data analysis contains several studies that employ data mining strategies to identify relevant patterns and useful information from food datasets [1]. This section describes previous studies that proposed specific tasks, such as food data representation, classification and named entity recognition.

Food computing is a field in computer science that focus on food-related studies and its main goal it to provide useful applications and tools to facilitate human understanding [5]. Hence, collecting food data is the first step in this matter. Anna Wróblewska et al. [3] introduced a new dataset with a total 700 sets of ingredients from online recipes, which were manually annotated so that Named Entity Recognition (NER) models should find various entities helpful to pre-process recipes. The authors tested BERT and LUKE architectures to train the models which presented fine results recognizing food-entities so that this method has potential to be valuable in other food-related tasks.

Furthermore, Matej Petkovi'c et al. [4] designed a AI workflow methodology named *DietHub* to annotate and classify recipes with the food concepts related to them. The proposed workflow is divided into representation learning, which is crucial for efficient performance, and two predictive modeling tasks, classification and hierarchical multi-label classification, to predict the class of a given example. To evaluate the model, recipes of Mediterranean diet were used and compared to other diets considering aspects as health, cooking style and region. Overall, the proposed workflow showed high predictive power and correctly annotate the recipes with semantic tags.

Artificial Neural Networks (ANN) have shown successful results over the years in regard to their use in different fields of applications. Due to its great ability to learn patterns, ANN are extremely useful at classification tasks. For that reason, B Debska et al. [2] utilized an ANN to classify the dataset into good quality of beer or insufficient quality of beer. To design the proper network, intelligent problem solver (IPS) and automatic network designer (AND), automatically selected the most fitting type and architecture. The employed techniques showed a brilliant result with 100% accuracy, indicating great success at classification effort.

As reflected, Natural Language Processing (NLP) can be extremely useful in different fields with a variety of applications [6]. Thus, this article proposes a study about several tasks in NLP that can be helpful to better understand food data related subjects.

## III. Proposed Method

The proposed method is divided into four steps: dataset definition, preprocessing, text retrieval task, classification task and performance comparisons of classifiers as shown in Figure 1. Each step is detailed in the following subsections.

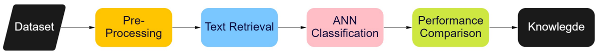


Fig. 1. Flowchart showing the steps illustrating the proposed method.

### A. Dataset

Two food composition datasets presenting similar aspects regarding their attributes were considered to design the proposed method.

*1) Recipe Ingredients Cuisine Dataset:* The Recipe Ingredients Cuisine dataset describes the ingredients used in different dishes from over 20 cuisines. It presents 39774 data instances, described by 2 attributes as shown in Table I. However, only 10000 instances were used, due to its extensive size. This unique dataset was featured in a Kaggle competition [1] for fun and practice.

TABLE I
Description of the Recipe Ingredients Cuisine Dataset's attributes.

| Attributes | Type | Cardinality |
|---|---|---|
| Cuisine Category | Nominal | 1 |
| [Ingredients] | List of Nominals | 1 |

*2) Tradicional Indian Food Dataset:* The Tradicional Indian Food dataset contains informations of a variety of regional and traditional cuisines native to the Indian subcontinent [2]. The dataset presents 255 data instances, which are characterized by 7 attributes. Table II describes the types of all attributes. The diet attribute includes vegetarian or non vegetarian, the flavor profile includes whether the dish is spicy, sweet, bitter or sour and the course meal consists of main course, dessert, snack or starter.

### B. Preprocessing

As the selected food datasets present nominal attributes, a preprocessing is required to generate a new structured representation containing only numerical attributes, called word embedding, to allow their input to the natural language processing tasks. In this regard, two different approaches were chosen: Word2vec and GloVe (Global Vectors for Word Representation).

[1]Recipe Ingredients Cuisine Dataset https://www.kaggle.com/datasets/kaggle/recipe-ingredients-dataset

[2]Tradicional Indian Food Dataset https://www.kaggle.com/datasets/nehaprabhavalkar/indian-food-101

TABLE II
Description of the Tradicional Indian Food Dataset's attributes.

| Attributes | Type | Cardinality |
|---|---|---|
| Food name | Nominal | 1 |
| [Ingredients] | List of Nominals | 1 |
| Diet | Nominal | 1 |
| Flavour Profile | Nominal | 1 |
| Course of Meal | Nominal | 1 |
| State Origin | Nominal | 1 |
| Region | Nominal | 1 |

Word2Vec is a neural network technique proposed in 2013 that generates word embeddings and can utilize two model architectures: Continuous Bag-of-Eords (CBOW) and Skipgram. Those models produces a high dimension vector space, both with its particularity, where each unique word from the input corpus is assigned a corresponding vector in the space [7].

Similarly, GloVe is an unsupervised learning algorithm developed by Stanford for obtaining vector representations for words. However, instead of using local word co-occurrences, it uses aggregated word co-occurrence statistics in the entire dataset [8].

### C. Text Retrieval

Most of the time, analysing and understanding a dataset is not a straightforward task, since it may have multiple instances and attributes. That being said, manually investigating the data is not the right way to address the problem, which is why there are different tasks of text retrieval to resolve this issue.

Text retrieval is the process of analysing a dataset in order to identify meaningful patterns and new insights. There are different methods of text retrieval, but this article adopted the document ranking approach. Thus, this task will sort all documents according to their ingredients so it is possible to determine if a food recipe is more similar to another from a equal or distinct category.

However, since the chosen task will rank a food according to a set of ingredient, it is required a small modification in the preprocessing step so that we have a sentence embedding instead of a word embedding. One approach is to simply return for each sentence the average embedding of the contained words.

### D. Artificial Neural Network Classification

Text classification is a NLP task with a wide range of applications, such as document classification, spam filtering and sentiment analysis. It consists of simply categorizing a collection of sentences into different labels previously defined. Two machine-learning based system were considered in the proposed method: Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM).

A CNN is a type of neural network that involves a mathematical operation called convolution that takes two functions and produces a third one, which expresses how the shape of one is modified by the other. Briefly, a CNN can take an input, assign importance to various aspects in the input and be able to differentiate one from the other [9]. The model architecture was designed with an Embedding Layer, a Convolution Layer, two Dense Layers and finally the Output Layer. In between the Convolution Layer and the first Dense Layer we added a Global Max Pooling Layer to reduce the input dimension and a Dropout Layer to prevent over fitting. Also, it was chosen for every Layer the Rectified Linear Unit (ReLU) as the activation function, apart from the last layer that used a Softmax Function to normalize the output.

Another notorious neural network in NLP is the Recurrent Neural Network (RNN), that has the unique ability of a internal state memory, processing each input and updating the cell state according to the information received. The chosen model LSTM is a special kind RNN, which is able to retain information in a memory cell for long periods of time [10]. The proposed LSTM model was designed with first a Masking Layer to read the input more efficiently, an Embedding Layer, two connected LSTM layers, a Dropout layer to prevent overfitting, a Dense Layer and the Output Layer. For the Dense Layer there was also a ReLu activation function and a Softmax activation function for the last layer.

In order to compile either neural network, a categorical crossentropy loss was chosen combined with a gradient descent optimizer. 100 epochs were performed in every model to obtain the results discussed in section IV.

### E. Evaluation Metrics

The performance of the proposed method depends on how well the machine learning methods can predict the expected output. Thus, the classification techniques are evaluated using 6 state-of the-art metrics to measure the accuracy, the model loss, the precision, the recall, the F1 score and the confusion matrix.

The confusion matrix is an important classification method, since it can be useful to measure all the other mentioned metrics. It can be defined as a combination of predicted and actual labels. Hence, for each actual class, there is a number of times it was correctly predicted to the class and number of times it was wrongly predicted to each of the other classes.

The accuracy simply predicts the percentage of correct output results, or in other words, is defined as the ratio of the number of correct predictions and the total number of predictions. It is a good measure if the predicted classes are balanced, but not a great choice otherwise.

The model loss is a machine learning metric that quantifies the error of the model, evaluates the quality of the prediction after each epoch of training. In other words, indicating how bad or good the model's prediction was after each interaction, so if the loss is zero the model's prediction is perfect, otherwise, the loss is greater.

Furthermore, precision is a metric that evaluate from the positives outcomes, how many of them were correctly predicted. For instance, if we want to tell if the class is "True" and we have 20 "True" outcomes, the precision metric can compute how many of the "True" outcomes are really "True". Similarly, the recall explains how many of the actual positive cases we were able to predict correctly with our model. In this case, we want to tell from all the outcomes that were supposed to be "True", how many were correctly predicted.

Finally, there is the F1-score, which gives the a combined idea about precision and recall metrics. It is defined as the harmonic mean of precision and recall, where we can see in the Equation 1.

$$F1\_Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$

## IV. Experimental results

In this section we performed experiments aiming to: first analyse the text retrieval results from both datasets and second evaluate the quality of the employed neural networks considered in Section III-D. The development of the proposed method was based on Python 3.8 alongside with pandas, sklearn, keras, tensorflow, gensim, glove and plotly libraries for data processing and artificial inteligence techniques.

### A. Text Retrieval Results

As mentioned, two datasets were selected to this task. For the Recipe Ingredients Cuisine Dataset, the Cuisine Category attribute was considered to investigate the dataset and create a rank of similarity between the instances. So, two steps were taken to perform this task: for each embedding we calculated the cosine similarity and sorted from the most similar to the least similar; then we created a top10 rank the tells if the $i$-th most similar instance belongs to the same cuisine category. Similarly, for the Tradicional Indian Food Dataset, the same steps were performed, but 3 different attributes were considered to create the rank: Diet, Flavour Profile and Course of Meal.

The task revealed that Word2Vec presented better results in comparison to GloVe for the Cuisine Dataset. For 10000 instances, 63% from the most similar are categorized with the same Cuisine with Word2Vec, while 56% for GloVe. The Top 10 rank of Word2Vec can be seen in Eq. 2 and the Glove results in Eq. 3.

$$[6335, 5951, 5889, 5800, 5707, 5576, 5623, 5537, 5421, 5465] \tag{2}$$

$$[5615, 5394, 5234, 5142, 4972, 5021, 4911, 4881, 4859, 4779] \tag{3}$$

As for the Indian Food Dataset, the results for both embeddings were quite similar since the size of the dataset is small. But the diet category presented better results with almost 92% use for Word2Vec and GloVe. However, this result may have occurred as this category is unbalanced and 90% of the attributes are categorized as vegetarian and only 10% are non vegetarian. The Top 10 Diet rank of Word2Vec can be seen in Eq. 4 and the Glove results in Eq. 5. Since the Diet Category presented better results, the Classification task will be based on that.

$$[233, 235, 233, 230, 230, 232, 228, 228, 234, 228] \tag{4}$$

$$[236, 240, 233, 235, 236, 231, 234, 231, 230, 227] \tag{5}$$

## B. Evaluation of ANN Classification Performance

In order to evaluate the quality of the selected artificial neural network techniques, we followed three steps: finding the best architecture and the best choices of hyperparameters for each method and then comparing the results with the previous mentioned evaluation metrics. Thus, first different set ups for each model were designed, adding and removing layers in order to obtain better results. After that, the Keras Tuner library from Python was used to select the optimal set of hyperparameters for each classification method and then we compared the evaluation metrics for each technique so that we selected the combination of values that presented the best results.

The results obtained with the Tradicional Food Dataset were not at all satisfying, since it could only predict one class for all models and embeddings. This can be due to the fact that the dataset is very small and unbalanced, and for that the models could not predict the classes right. This way, the results with CNN-Word2Vec for the class "vegetarian" were somewhat high with 87% precision, 100% recall and 93% of f1_score, while the class "non vegetarian" obtained all values 0.

In the other hand, results for the Recipe Ingredients Cuisine Dataset showed more diversity. In Figure 2, we can observe the model loss for both CNN and LSTM architectures using Word2Vec and GloVe embeddings. It is possible to observe that the CNN training obtained a smaller loss with both embeddings, indicating a smaller prediction error. Correspondingly, in Figure 3 the CNN also presented better results of accuracy against the LSTM model.

In Figure 4, we may see this results more clearly since there are more values in the diagonal line in the CNN model than in LSTM model. This indicate that the number of classes correctly predicted are higher in CNN than in LSTM.

Lastly, considering there are many classes predicted, is impracticable to show the other metrics results for every model and embedding. Thus, it is possible to see the CNN-Word2Vec results in table III, for once it presented the best results for every evaluation metric.

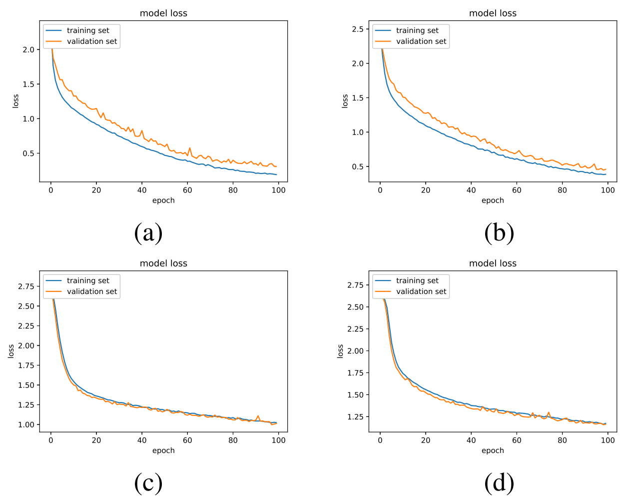


Fig. 2. CNN model loss layout for (a) Word2Vec and (b) GloVe against LSTM model loss layout for (b) Word2Vec and (b) GloVe.

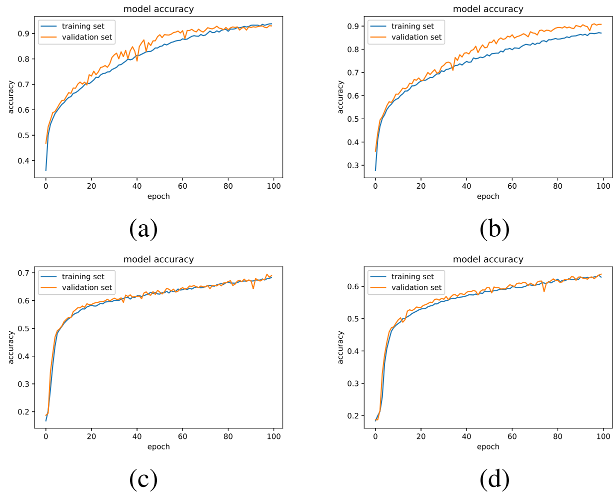


Fig. 3. CNN model accuracy layout for (a) Word2Vec and (b) GloVe against LSTM model accuracy layout for (b) Word2Vec and (b) GloVe.

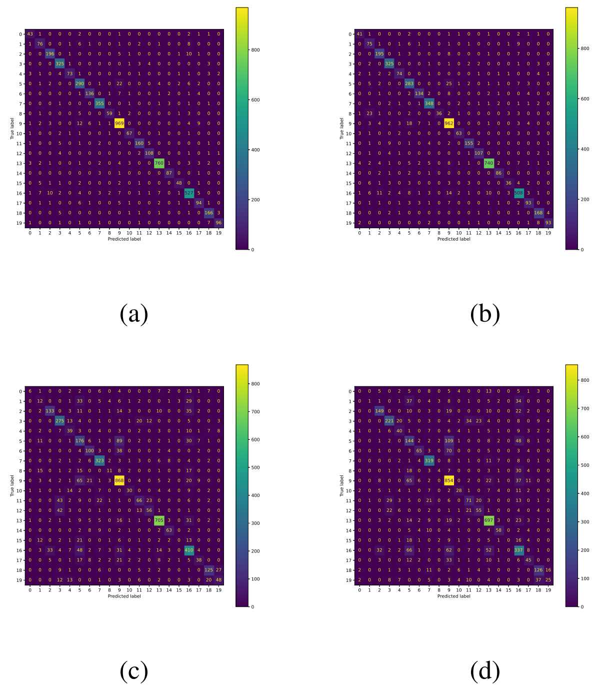


Fig. 4. CNN Confusion Matrix layout for (a) Word2Vec and (b) GloVe against LSTM model Confusion Matrix layout for (b) Word2Vec and (b) GloVe.

## V. Conclusion

This paper explored two tasks of natural language processing that are suitable when considering food-related datasets: text retrieval and classification. For that purpose, the proposed method consisted of pre-processing the data into sentence embeddings or word embeddings depending on the task using either Word2Vec or GloVe. Therefore, a document ranking approach were chosen for the text retrieval task and the a CNN and LSTM were considered for the machine-learning classification problem. As for the experimental results, the CNN along with Word2Vec presented better results for all evaluation metrics considered. Future work can be guided exploring another tasks in natural language process-

TABLE III
EVALUATION METRICS FOR THE RECIPE INGREDIENTS CUISINE DATASET USING CNN AND WORD2VEC

| | ***precision*** | ***recall*** | ***f1_score*** |
|---|---|---|---|
| brazilian | 0.78 | 0.84 | 0.81 |
| british | 0.78 | 0.78 | 0.78 |
| cajun_creole | 0.92 | 0.91 | 0.91 |
| chinese | 0.93 | 0.96 | 0.95 |
| filipino | 0.95 | 0.81 | 0.87 |
| french | 0.87 | 0.88 | 0.87 |
| greek | 0.94 | 0.88 | 0.91 |
| indian | 0.96 | 0.98 | 0.97 |
| irish | 0.92 | 0.82 | 0.87 |
| italian | 0.95 | 0.96 | 0.95 |
| jamaican | 0.86 | 0.89 | 0.88 |
| japanese | 0.93 | 0.91 | 0.92 |
| korean | 0.92 | 0.92 | 0.92 |
| mexican | 0.98 | 0.97 | 0.97 |
| moroccan | 0.94 | 0.95 | 0.94 |
| russian | 0.87 | 0.79 | 0.83 |
| southern_us | 0.92 | 0.91 | 0.92 |
| spanish | 0.78 | 0.82 | 0.80 |
| thai | 0.89 | 0.93 | 0.91 |
| vietnamese | 0.94 | 0.89 | 0.91 |

ing, such as Named-entity recognition (NER) and Entity linking. Furthermore, other models like Transformers can be considered in those tasks, along with other embeddings techniques.

## REFERENCES


[1] M. Hagiwara, *Real-World Natural Language Processing: Practical Applications with Deep Learning*. Manning, 2021. [Online]. Available: https://books.google.com.br/books?id=A92_zQEACAAJ

[2] B. Debska and B. Guzowska-Swider, "Application of artificial neural network in food classification," *Analytica Chimica Acta*, vol. 705, no. 1, pp. 283–291, 2011, a selection of papers presented at the 12th International Conference on Chemometrics in Analytical Chemistry. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003267011008622

[3] A. Wróblewska, A. Kaliska, M. Pawłowski, D. Wiśniewski, W. Sosnowski, and A. Ławrynowicz, "Tasteset – recipe dataset and food entities recognition benchmark," 2022. [Online]. Available: https://arxiv.org/abs/2204.07775

[4] M. Petković, G. Popovski, B. K. Seljak, D. Kocev, and T. Eftimov, "Diethub: Dietary habits analysis through understanding the content of recipes," *Trends in Food Science & Technology*, vol. 107, pp. 183–194, 2021.

[5] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–36, 2019.

[6] K. R. Chowdhary, *Natural Language Processing*. New Delhi: Springer India, 2020, pp. 603–649. [Online]. Available: https://doi.org/10.1007/978-81-322-3972-7_19

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[8] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[9] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *CoRR*, vol. abs/1511.08458, 2015. [Online]. Available: http://arxiv.org/abs/1511.08458

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735