Natural language processing for allergen identification on food labels: an application in the Brazilian context

Processamento de linguagem natural na identificação de alérgenos em rótulos alimentares: uma aplicação no contexto brasileiro

Procesamiento del lenguaje natural en la identificación de alérgenos en etiquetas de alimentos: una aplicación en el contexto brasileño

Giovanna Alves Gadelha[1], Renan Augusto Pereira[2], Flávia Magalhães Guedes[2], Ana Trindade Winck[3]

1 Bel., Federal University of Health Sciences of Porto Alegre - UFCSPA, Porto Alegre (RS), Brazil.
2 Me., Federal University of Health Sciences of Porto Alegre - UFCSPA, Porto Alegre (RS), Brazil.
3 Dra., Federal University of Health Sciences of Porto Alegre - UFCSPA, Porto Alegre (RS), Brazil.

Corresponding author: Profa. Dra. Ana Trindade Winck
E-mail: anatw@ufcspa.edu.br

Resumo

Objetivo: Alergias alimentares impactam parte significativa da população, apresentando desafios para a saúde pública. A abordagem para o manejo dessas alergias exige a eliminação de alimentos que as desencadeiam. Entretanto, ler e interpretar rótulos de alimentos é desafiador devido à nomenclatura variada e inconsistente, bem como à falta de regulamentação adequada. Para o contexto brasileiro, propomos uma solução de Processamento de Linguagem Natural, que será integrada a um aplicativo móvel dedicado. Método: Para o reconhecimento das diversas nomenclaturas associadas aos quatro principais alérgenos, foi desenvolvido um banco de dados de alérgenos e um modelo de reconhecimento de entidades nomeadas, além de funções de pré-processamento de texto. Resultados: a avaliação dos modelos obteve uma precisão média de 96.50. Conclusão: Esta solução apoia a promoção de práticas alimentares mais seguras para indivíduos com alergias alimentares, fornecendo suporte tecnológico para obter informações sobre a presença de alérgenos em alimentos.

Descritores: Rotulagem de Alimentos; Hipersensibilidade Alimentar;Processamento de Linguagem Natural

Abstract

Objective: Food allergies impact a significant portion of the population, presenting challenges to public health. The approach to managing these allergies is by the elimination of specific trigger foods. However, reading and interpreting food labels is challenging due to diverse and inconsistent nomenclature, as well as inadequate regulations. For the Brazilian context, we propose a Natural Language Processing solution, which will be integrated into a dedicated mobile application. Method: To recognize the diverse nomenclatures associated with allergens focusing on Portuguese terms, we developed an allergen database and a named entity recognition model, as well as text preprocessing functions. Results. The evaluation of the models achieved an average precision of 96.50. Conclusion: This solution supports safer dietary practices for individuals with food allergies, providing technological support in obtaining information about the presence of allergens in products.

Keywords: Food Labeling; Food Hypersensitivity; Natural Language Processing

Resumen

Objetivo: Las alergias alimentarias presentan desafíos para la salud pública. El manejo de estas alergias exige la eliminación de los alimentos que las desencadenan. Todavía, leer e interpretar las etiquetas de los alimentos es desafiante debido a la nomenclatura variada, así como a la falta de regulación adecuada. Para el contexto brasileño, proponemos una solución de Procesamiento de Lenguaje Natural, que se integrará en una aplicación móvil dedicada. Método: Para el reconocimiento de las diversas nomenclaturas asociadas a los cuatro principales alérgenos, se desarrolló una base de datos de alérgenos y un modelo de reconocimiento de entidades nombradas, además de funciones de preprocesamiento de texto. Resultados: la evaluación de los modelos obtuvo una precisión media de 96.50. Conclusión: Esta solución apoya la promoción de prácticas alimentarias más seguras para individuos con alergias alimentarias, proporcionando soporte tecnológico para obtener información sobre la presencia de alérgenos en los alimentos.

Descriptores: Etiquetado de Alimentos; Hipersensibilidad a los Alimentos; Procesamiento de Lenguaje Natural

Introduction

Food allergy is a condition that has been increasingly prevalent in Brazil and globally, posing a significant challenge to public health. It is characterized by an adverse and hypersensitive response of the immune system to antigens found in certain foods, affecting a considerable portion of the population. Recent estimates suggest that approximately 6 to 8% of children and 2 to 3% of adults in Brazil are affected by some form of food allergy(1). This issue not only significantly impacts the quality of life but also carries economic implications and strains healthcare systems, leading to an increasing demand for specialized services and appropriate interventions. Presently, the primary available treatment involves a restrictive diet, entailing the elimination of foods that trigger the allergic reaction.

In this context, the ability to accurately read food labels and identify potential allergenic substances is crucial for individuals affected by food allergies. Resolution RDC No. 26/2015 of the Brazilian National Health Surveillance Agency (ANVISA) mandates that major allergenic foods, which are described in Table 1, must be clearly labeled(2). However, studies indicate that, despite regulations, many food products do not comply, either due to a lack of allergen declaration or spelling errors(3,4). Moreover, a fundamental challenge in reading food labels accurately is the existence of different nomenclatures associated with the same allergen, often unknown to the general population(5) .

Addressing this challenge involves the utilization of tools that assist in identifying allergens on food labels. In today's landscape of continuous technological advancements and their substantial impact on the healthcare sector, the use of technologies appears as a viable approach, particularly through the utilization of tools like mobile applications. Within this context, natural language processing can be utilized to confront the issue of diverse nomenclatures found on food labels, employing techniques such as Named Entity Recognition (NER). These approaches can lead to significant progress in understanding and accurately identifying allergens, thereby actively contributing to public health by promoting food safety.

Table 1 – Main foods causing food allergies[6].

Group	Food
1	Wheat, rye, barley, oats, and their hybridized strains
2	Shellfish
3	Eggs
4	Fish
5	Peanuts
6	Soy
7	Milks of all mammalian species
8	Almond (Prunus dulcis, syn.: Prunus amygdalus, Amygdalus communis L.)
9	Hazelnuts (Corylus spp.)
10	Cashew (Anacardium occidentale)
11	Brazil nut or Para nut (Bertholletia excelsa)
12	Macadamias (Macadamia spp.)
13	Walnuts (Juglans spp.)
14	Pecans (Carya spp.)
15	Pistachios (Pistacia spp.)
16	Pine nuts (Pinus spp.)
17	Chestnuts (Castanea spp.)
18	Natural látex

In this context, Artificial Intelligence (AI) techniques, particularly in healthcare, have shown promising capabilities in identifying and comprehending allergens[7]. Named entity recognition through natural language processing enables the identification of naming patterns in various texts, allowing for the extraction of pertinent information.

As highlighted in a study led by Gorjan Popovski and team, there is a scarcity of Natural Language Processing (NLP) approaches focused on extracting food-related information. To address this gap, they developed FoodIE, a named entity recognition method based on rules for extracting unstructured text information related to food. FoodIE utilizes a limited set of rules derived from computational linguistics and semantic information to identify food entities. The method's evaluation was conducted on two distinct datasets, revealing highly promising results[8] .

Another study by Stefano Campese and Davide Pozza focused on food intolerances and utilized NLP strategies as a promising solution to classify foods into appropriate intolerance groups based on their ingredients. This study assessed and compared various machine learning techniques, along with different feature extraction methods, conducted on actual commercial products[9].

A recent study by Alessandra Amato and Giovanni Cozzolino developed a system using AI techniques to extract and analyze recipe ingredients, providing users with alerts about potential allergens. Although promising in detecting allergens in specific recipes, this study represents the closest account found in literature concerning the use of NLP for allergen analysis, yet not contextualized within food labels, which poses greater challenges due to nomenclature[10].

Additionally, it is essential to note that electronic health records serve as a primary source for NLP studies in the health sector. For instance, a study by L. Bilaver and colleagues focused on predicting food allergy diagnosis by allergists from textual notes from general pediatrician visits. Applying various machine learning algorithms to analyze textual data collected from general pediatric care encounters, the study illustrated that general notes could be useful in identifying patients with food allergies, enabling timely referrals to allergists[11].

Although explicit applications employing natural language processing for allergen recognition are scarce, we can highlight applications aiming to read and provide information on food product labels and nutritional tables. An example is the "Desrotulando" app, developed by nutritionists based on the Brazilian Ministry of Health's Dietary Guidelines. This app assists users in decision-making by reading food barcodes and connecting to a platform to retrieve nutritional information, including ingredient lists, obviating the need for text processing as it directly connects to product databases. Furthermore, the recent "Loomos" app stands out by scanning food labels via Optical Character Recognition (OCR) and Augmented Reality, highlighting ingredients based on user food preferences. This solution is highly relevant for aiding diets and identifying food allergies and intolerances. However, it's important to note that label information is compared to an exclusive internal app database, yielding information returns based on user queries without explicitly mentioning text processing.

Despite these studies, there is still a gap in specialized literature in applying natural language processing within the context of food labels. Therefore, with a focus on the Brazilian context and its Portuguese terms, we present a REST API to be integrated into an application for reading and identifying allergens on food labels. Its main functionality is to recognize the different nomenclatures used for the same allergen from food labels. This development is based on natural language processing techniques, with a primary focus on a named entity recognition model. Additionally, it is important to highlight that this study focused on recognizing the four main proteins: milk, wheat, soy, and eggs. This scope is driven by the greater variety of nomenclature associated with them[12].

Methods

Modeling of the allergen database

In this work, we propose the creation of a database containing allergens and their respective corresponding terms, aiming to serve as a testbase for developing and training a NER model. This model is intended to identify allergens present in food labels. Currently, the data are structured around four major allergy-triggering proteins: milk, eggs, soy, and wheat. Each of these proteins is considered an entity and has its associated nomenclatures.

In this context, we chose to employ the MongoDB database, a NoSQL solution, to store project information. Additionally, we utilized the Python library Pymongo to connect with the API functions. Unlike traditional relational databases, the chosen option features a flexible and scalable structure capable of accommodating variations in data structure, such as different input formats, optional fields, and additional properties. This flexibility enables us to meet the specific project requirements without imposing a rigid table structure or schema. Moreover, it allows for the addition of new fields or properties as needed without the necessity of altering the entire database structure.

Development of the NER model

This study aims to identify the various nomenclatures associated with the same allergen to facilitate the recognition of potential allergens on food labels. To achieve this goal, we chose NER as the primary natural language processing technique. We developed a customized NER model using the SpaCy framework in its 3.0 version, which integrates a pre-designed CNN architecture into the training process, enabling us to personalize this neural network architecture. An example of the feature extraction method utilized is tok2vec.

To train the model, we utilized a Brazilian database of publicly available food labels[13]. The initial step involved cleaning this dataset to ease its processing, involving the removal of unnecessary columns and standardizing the text to lowercase. Subsequently, we selected only the column containing the product ingredients.

To train a dataset using SpaCy, it is necessary to format them according to annotations. This involves an array containing the text and another array containing the start and end of the entity, along with the entity type, following the format:

'document text', {'entities': [(start, end, type), ..., (start, end, type)]

However, since the database used for training did not contain such annotations, we employed SpaCy's Entity Ruler. This component allows the addition of custom entities to the model. The Entity Ruler comprises the entity and its patterns. In this context, we utilized the allergen database to construct it, defining each protein as an entity and assigning the associated allergens as patterns of that entity.

This enabled us to generate an annotated dataset without the need for manual annotation. We utilized 153 data points for training and 52 for testing. Subsequently, we converted the training data from JSON format into SpaCy Doc objects and configured the training files, adjusting the hyperparameters.

Development of text processing functions

In addition to the customized NER model, text preprocessing functions were implemented, using Python, to be applied to food labels before entity recognition by the model. This is due to the fact that preprocessed texts generally yield more effective recognition. Furthermore, by employing the annotation approach with the Entity Ruler, it is possible to annotate the data exactly in the same way as defined in this standard. As a result, it is natural that some entities may not be correctly identified or even lost. In recognition of this limitation, four specialized functions for distinct contexts were implemented utilizing tokenization techniques and various Python libraries, including pandas, sklearn, sys, re, and tqdm:

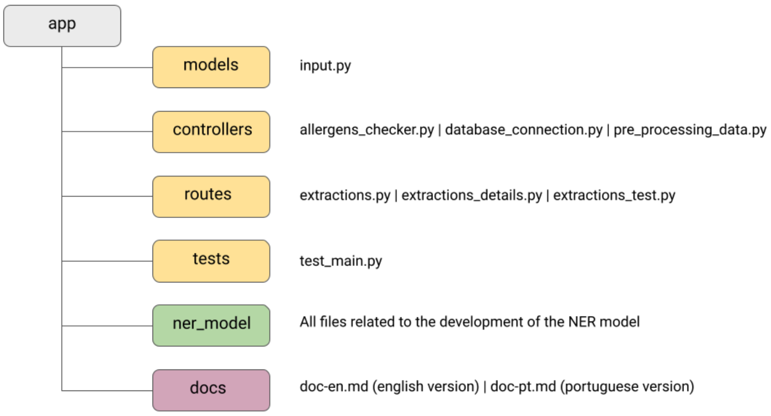
- The function *no_contains_identifier* deletes the token after the phrase 'no contains' following the label convention. This is because the model recognizes only the word and was not trained considering the mentioned context.

- The function *may_contains_identifier* removes subsequent tokens following the expression 'may contains' according to the label standard. This occurs because the scope of the work does not involve identifying traces related to cross-contamination.
- Lastly, the function *label_cleaning*, working in conjunction with the *preposition_identifier* function, checks subsequent prepositions for unique tokens. This ensures that the model does not classify cases such as 'coconut milk' or 'cocoa butter' as belonging to the milk entity.

Structure of the REST API

At the end of the process, a REST API was implemented using the FAST API framework in Python to encapsulate the developed functions and load the built model. The choice was based on simplicity and speed, reflecting an architecture in accordance with the examples from the tool's documentation. In Figure 1, it is possible to identify the layers of the developed API. These layers include the control layer with functions dedicated to handling requests, the model layer containing data schemas, routes aimed at the endpoints, a testing layer for unit tests, documentation, and a folder with files related to the development of the NER model.

Figura 1 – Folder structure, made by the author.


app
models
input.py
controllers
allergens_checker.py | database_connection.py | pre_processing_data.py
routes
extractions.py | extractions_details.py | extractions_test.py
tests
test_main.py
ner_model
All files related to the development of the NER model
docs
doc-en.md (english version) | doc-pt.md (portuguese version)

Altogether, three routes are available: two focused on allergen recognition and one intended for testing. In the request body, a JSON is sent containing an entry for the label and another for the list of allergens the user wishes to identify. If this list is empty, the return will be all allergens found in the text, among the four initially established. In the test route, an additional input is used to compare the allergens truly present in the text with the results identified by the model. Although all returns are in JSON format, it is the specifics of each route that differentiate their functionalities, as described in the instructions below:

- The extractions route returns only the original proteins found in the label.
- The extractions details route returns both the original protein found and the associated allergen.
- The extractions test route presents the label, the validated original proteins, those identified by the model, as well as errors and accuracies.

Upon completing the development of the REST API, it is important to integrate it into the application for allergen identification on food labels. In this context, it was necessary to establish the communication architecture among the project components. The project encompasses three distinct APIs, each dedicated to a specific functionality. Consequently, we have the current Natural Language Processing (NLP) API with the objective of recognizing different nomenclatures used for the same allergen. Additionally, although outside the scope of this work, there is an Optical Character Recognition (OCR) API responsible for reading the product label image and extracting its text. In this scenario, the OCR API receives an image from the application and returns the extracted text. Subsequently, the application will send this text to the NER API to process it and return the result, indicating whether the product in question contains the initially selected four allergens in its composition.

Documentation and Testing

Upon concluding the development of the proposed REST API, documentation was made available in both English and Portuguese for the project team, aiming to provide guidance on its utilization. These instructions encompass details regarding the installation of utilized libraries and frameworks, along with guidelines on how to run the application locally. Additionally, the available routes and their request and response protocols are outlined.

Regarding testing, the API has integrated unit tests using the pytest library to assess the quality of received requests. Furthermore, an additional test route has been implemented to verify the quality of entity recognition. This information is recorded in the database to assist in future endeavors by identifying potential issues and gaps, as well as capturing new variations of nomenclatures that can be added to the allergen database.

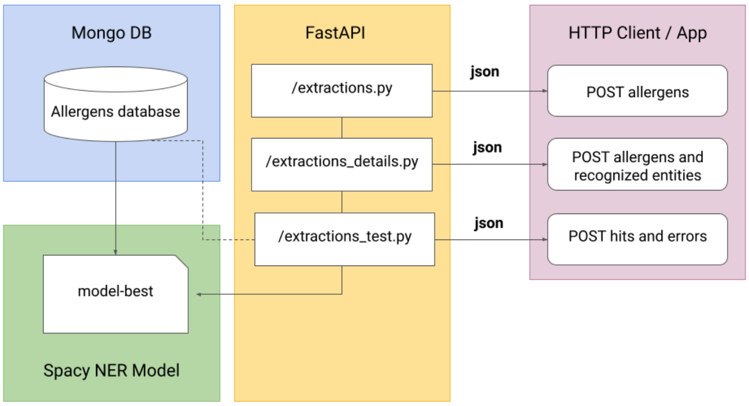
Ethical Aspects

All project stages were conducted using public databases and texts available on the labels of industrialized products. Therefore, in accordance with Resolution No. 510 of April 7, 2016, from the National Health Council, this work does not need to be submitted to an ethics committee.

Results and discussion

As a result of this work, we have developed a REST API capable of recognizing allergens associated with proteins from milk, wheat, eggs, and soy, as illustrated in Figure 2. This API has been integrated into an application specifically developed for allergen identification through the parsing of food product labels. It is important to highlight that the present work did not encompass the traces of foods that may be indicated on the label preceded by the phrase 'may contain', since we are focused solely on allergens directly present in the product.

Figure 2 – REST API architecture



Furthermore, the encapsulation of these functionalities within a REST API not only ensures scalability but also facilitates the maintenance of the architecture, allowing for the effortless addition of new source proteins. Also, the inclusion of a testing route enables the identification of new nomenclatures, rendering the allergen database adaptable to evolving requirements. In order to fulfill the primary function of the API, we constructed a tailored Named Entity Recognition model, specifically calibrated for this context. T in Table 3.

Table 3 – Evaluation metrics of a custom NER model.

Protein	Precision	Recall	F1-Score
Wheat	100	100	100
Soy	100	100	100
Eggs	90.91	100	95.24
Milk	95.08	96.67	95.87

The results of an NER model are directly related to the quantity and quality of the training and annotation data. Therefore, one of the main challenges of this work was the scarcity of data on food labels, along with the lack of proper formatting for use in training models. The absence of annotated data significantly impacts the model's learning capacity and also highlights the limitations that the SpaCy framework presents for languages other than English.

In response to these challenges, we developed preprocessing functions aimed at overcoming these limitations, ensuring the accurate recognition of entities, as detailed in Table 4. All developed functions are tailored to the context of Brazilian food labels, and consequently, variations in the Portuguese language adhere to regulatory standards. While the metrics show satisfactory results, it is essential to acknowledge the constraints arising from the limited training data and non-manual annotation.

Table 4 – Pre-processing Functions of Food Labels

Function	Description
data_cleaning	This function performs text normalization to lowercase and removes characters and punctuation.
no_contains_identifier	This function checks if the allergen is preceded by the phrase 'does not contain' as per the standard and removes it from the sentence, if true, to prevent the model from recognizing it.
may_contains_identifier	This function checks if the allergen is preceded by the phrase 'may contain' as per the standard and removes it from the sentence, if true, to prevent the model from recognizing it. **NOTE:** This function was designed with the project scope in mind, which does not aim to recognize traces.
preposition_identifier	This function is responsible for recognizing preposition tokens.
Wheat label_cleaning	This function is responsible for checking if unique tokens are followed by a preposition and removes them from the sentence if true, preventing the model from recognizing them. **NOTE:** This function was designed to address a limitation of the model that mistakenly identified unique tokens in certain portuguese contexts.

We must warn regarding the potential overfitting of the model to the training data or bias in classifying specific examples, notably evident in the results for wheat and soy, boasting 100% accuracy. Therefore, expanding the dataset and conducting manual review of annotations, even though demanding additional time and effort, are imperative steps toward enhancing the quality of allergen recognition for a more precise analysis. It is also important to mention that we haven't yet managed to test food labels extracted from the OCR API, hence considering the limitations imposed by the product packaging, such as variations in text positioning.

Moreover, we have developed an extensive allergen database encompassing various nomenclatures associated with the four target origin proteins in this study. The structure of this database facilitates the inclusion of new proteins and the expansion of nomenclatures, providing the model with an increasing repository of examples to continually enhance its learning capabilities over time. Table 5 offers a comprehensive overview of the number of documents contained within the database for each entity.

Table 5 – Proteins and quantity of corresponding terms, mapped by the project's multidisciplinary team

Protein	Quantity of corresponding terms
Eggs	19
Wheat	21
Soy	43
Milk	50

Conclusion and future work

Food allergy is a condition that demands efforts from both the affected individual and the involved family members, caregivers, and healthcare professionals. Thus, it becomes a public health issue requiring specialized services and appropriate interventions. As part of the primary treatment, food labels play a role for those affected, providing information about the contents of the product and promoting food safety for the population. However, from the perspective of the Brazilian context, there is a need for a closer inspection of the quality of food labels and the development of tools to assist the public in choosing allergenic foods.

This importance goes beyond simply providing accurate information; it directly impacts the health of the population, as food allergies can trigger critical situations for those affected. Additionally, technology can be employed as an accessible and comprehensive means to include the lay population, lacking the necessary guidance on the various nomenclatures associated with the same allergen. In this context, the development of applications contributes to enhancing food safety, allowing caregivers and family members to benefit from these solutions.

Thus, this study aims to provide more security to the allergic population in selecting and consuming food through a technological solution in the field of dietary health, with the goal of significantly improving the quality of life. Leveraging a multidisciplinary team, the project adopted a collaborative methodology that allows for the expansion and improvement of the main functionalities: label reading and allergen recognition. Furthermore, the study highlighted the scarcity of datasets in the Portuguese language for training natural language processing models, as well as the limitation of named entity recognition libraries for languages other than English.

Finally, as future work, it is essential to dedicate efforts to creating datasets related to food labels and their manual annotation to enhance the developed model, making it less dependent on context-adaptive functions. Additionally, continuous monitoring of allergen databases can lead to substantial improvements and can be leveraged in various applications in the field of dietary health.

References

1. ASBAI, "Alergia alimentar é o tema central da Semana Mundial," https://asbai.org.br/alergia-alimentar-e-o-tema-central-da-semana-mundial/, 2019.
2. ANVISA, "Perguntas e respostas sobre rotulagem de alergênicos," https://www.gov.br/anvisa/pt-br/centraisdeconteudo/publicacoes/alimentos/perguntas-e-respostas-arquivos/rotulagem-nutricional_2a-edicao.pdf/view, 2023.
3. M. Oliveira Andrade, D. Alves, W. Nascimento. "AVALIAÇÃO DA ROTULAGEM DE ALIMENTOS E DA CONFORMIDADE QUANTO A DECLARAÇÃO OBRIGATÓRIA DE ALERGÊNICOS," Alimentos: Ciência, Tecnologia e Meio Ambiente, vol. 3, no. 1, pp. 14–25, 2022.
4. Martins, L. "Rotulagem de alimentos alergênicos: análise das informaçoes". Hig. aliment 2023:e1113–e1113.

a de Alergia Alimentar," https://www.poenorotulo.com.br/, 2014.

6. Brasil, "RDC nº 26, de 02 de julho de 2015," , 2015.
7. Bilaver, L., et al. "P107 FOOD ALLERGY AND INFORMATICS: USING NATURAL LANGUAGE PROCESSING TO IDENTIFY CLINICAL PREDICTORS IN PROGRESS NOTES," in Annals of Allergy, Asthma & Immunology, vol. 127, no. 5, pp. S41–S42, 2021.
8. Popovski, G., et al. "FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction.," in ICPRAM, vol. 12, pp. 915, 2019.
9. S. Campese, D. Pozza, "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, 2021, pp. 172–181.
10. A. Amato, G. Cozzolino, "SafeEat: extraction of information about the presence of food allergens in recipes," in Advances in Intelligent Networking and Collaborative Systems: The 12th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2020) 12, 2021, pp. 194–203.
11. Bilaver, L., et al. "Natural language processing of pediatric progress notes for the identification of food allergy," in The Journal of Allergy and Clinical Immunology: In Practice, 2023.
12. Caimmi, D., et al. "Food allergy in primary care," in Acta Bio Medica: Atenei Parmensis, vol. 92, no. Suppl 7, 2021.
13. Banco Brasileiro de Rotulos de Alimentos, "Banco Brasileiro de Rotulos de Alimentos," https://five.epicollect.net/project/banco-brasileiro-de-rotulos-de-alimentos/, 2021.