CachacaNER: a Dataset for Named Entity Recognition in Texts about the Cachaça Drink

Priscilla Silva[1*], Arthur Franco[1], Thiago Santos[1], Mozar Brito[2] and Denilson Pereira[1*]

[1*]Department of Computer Science, Federal University of Lavras, P.O. Box 3037, Lavras, 37.200-900, MG, Brazil.
[2]Department of Agroindustrial Management, Federal University of Lavras, P.O. Box 3037, Lavras, 37.200-900, MG, Brazil.

*Corresponding author(s). E-mail(s): priscilla.silva2@estudante.ufla.br; denilsonpereira@ufla.br;
Contributing authors: arthur.franco@estudante.ufla.br; thiago.santos6@estudante.ufla.br; mozarjdb@ufla.br;

Abstract

Named Entity Recognition (NER) is the task of identifying and classifying tokens in texts corresponding to a set of pre-defined categories, such as names of people, organizations and locations. Datasets labeled for this task are essential for training supervised machine learning models. Although there are many datasets labeled with texts for English, in the Portuguese language they are scarcer. This work contributes to the creation and evaluation of a manually labeled dataset for the NER task, with texts in Brazilian Portuguese, in the specific domain of the beverage called Cachaça. This is a popular drink in Brazil, and of great economic importance. This is the first NER dataset in the beverage domain, and can be useful for other types of beverages with similar entity categories, such as wine and beer. We describe the process of data collection, creation of the dataset and its experimental evaluation. As a result, we created a dataset containing over 180,000 tokens labeled in 17 entity categories. The labeling obtained a high value of agreement among the labelers, according to the Fleiss' Kappa

metric. The size of the dataset, as well as the result of its experimental evaluation, are comparable to other datasets in the Portuguese language, even though ours has a greater number of entity categories.

Keywords: NER, Named Entity Recognition, Dataset, Labeled data, Cachaça

1 Introduction

Named Entity Recognition (NER) (Nadeau & Sekine, 2007) is a Natural Language Processing (NLP) tagging task that recognizes relevant concepts (or entities) in texts and classifies them according to a set of pre-defined semantic categories, such as names of people, organizations, locations, dates and times. In a specific context, the categories can be, for example, names of genes, proteins, drugs and diseases in the biomedical domain. An entity can be any word or sequence of words that refer to the same concept. For example, in the sentence "Rodriguinha cachaça has been produced in Capitólio for over a hundred years", there are the entities: "Rodriguinha" (name of the drink), "Capitólio" (location) and "a hundred years" (time).

Named Entity Recognition plays a key role in Information Extraction and NLP applications such as information retrieval, question answering, machine translation, automatic text summarization, and event/product monitoring. The NER task is important to identify relevant entities and disambiguate the context of a textual content. For example, recognizing the entity "São Paulo" as a location (Brazilian city name) in a sentence can be important to detect where a particular event occurred, and to differentiate it from an entity of the saint name category in the religious context (Saint Paul, in portuguese, São Paulo).

State-of-the-art NER research is based on machine learning techniques (Goyal, Gupta, & Kumar, 2018; Li, Sun, Han, & Li, 2022; Yadav & Bethard, 2019). Most approaches are supervised, which require labeled data to train a learning model. The best quality corpora for training a learning model are those manually labeled by humans. However, labeling is an expensive, tedious and time-consuming process, and requires the engagement of domain experts for quality annotation. Although there are several corpora manually annotated in the English language, in Portuguese they are rarer. Among them are the first and second HAREM (Freitas, Mota, Santos, Oliveira, & Carvalho, 2010; Santos, Seco, Cardoso, & Vilela, 2006), Paramopama (Mendonça Jr., Barbosa, Macedo, & São Cristóvão, 2015) and LeNER-Br (de Araujo et al., 2018). The latter in the specific field of the judiciary.

This work presents the process of creating and evaluating a new NER dataset in Portuguese, manually annotated, in the specific domain of the Brazilian drink called Cachaça. The dataset was named CachacaNER. Cachaça is an alcoholic beverage produced from sugar cane. It is used in the preparation of the cocktail known worldwide as "caipirinha". Its origin dates back to the

beginning of Portuguese colonization in Brazil. It is a very versatile drink, and can be consumed pure, chilled or mixed with other drinks (Instituto Brasileiro da Cachaça, 2022).

Cachaça also has its economic importance. According to Instituto Brasileiro da Cachaça (2022), cachaça is the second most consumed alcoholic beverage in Brazil, after beer, and represents 72% of the spirits market in the country, being considered one of the four most consumed spirits in the world. The country has more than 1,000 registered producers, which produce more than 8 million liters per year and generate more than 600 thousand jobs. Cachaça is exported to 77 countries, including the USA, Germany, France and Paraguay (ExpoCachaça, 2022).

To the best of our knowledge, there is no specific annotated NER corpus available for the beverage domain. Cachaça shares many common characteristics with other types of drinks, such as wine, beer and coffee. Some works in the literature have applied text mining techniques to extract characteristics described by wine experts from product review texts (Katumullage, Yang, Barth, & Cao, 2022; Lefever, Hendrickx, Croijmans, van den Bosch, & Majid, 2018; Palmer & Chen, 2018). Cachaça also shares the same sensory characteristics of wine, such as aroma, flavor, consistency and color, as well as properties such as storage container, place of origin and price. Thus, a NER dataset on the cachaça drink can also be useful for extracting information about other types of beverages.

To illustrate, the following is an excerpt from a text extracted from the CachacaNER dataset containing the analysis of a *Cachacier* (the same as a wine sommelier) on a specific cachaça:

> "De cor amarelo-palha, possui uma mescla de aromas de frutas cítricas, mel e baunilha. No paladar, traz um gosto doce, mas que aguça as papilas salgadas. Além disso, causa uma sensação picante e um pouco alcoólica na boca. Retrogosto agradável e moderado."
> (In English: "Straw yellow in color, it has a mixture of citrus, honey and vanilla aromas. On the palate, it brings a sweet taste, but that sharpens the salty taste buds. In addition, it causes a spicy and slightly alcoholic sensation in the mouth. Pleasant and moderate aftertaste.")

In this work, we developed and evaluated a NER dataset containing seventeen categories of entities in the cachaça beverage domain. The main contributions are: (i) identification of the main categories of entities described in texts about cachaça, (ii) data collection and labeling for the NER task, and (iii) experimental evaluation of the created dataset. As a result, a dataset containing 183,019 tokens labeled by three labelers was generated, with an agreement coefficient of 0,857 in the Kappa metric, which is considered an almost perfect agreement. In the experimental evaluation using a NER model trained in the spaCy tool, a value of 88.9% of F1-measure was obtained in

the test set. The dataset is publicly available at https://github.com/LabRI-Information-Retrieval-Lab/CachacaNER.

The rest of this document is organized as follows. Section 2 describes the main works in the literature on creating datasets for the NER task. Section 3 describes details about the creation and characteristics of the CachacaNER dataset. Section 4 presents the experimental evaluation and the results obtained with the created dataset. And Section 5 presents the conclusions and future work.

2 Related Work

NER is typically a sequence-labeling task, and the main techniques applied are rule-based, learning-based, and hybrid approaches (Goyal et al., 2018). Rule-based approaches were used in early NER systems, and are still used in hybrid approaches combined with learning-based techniques. The most common models used in learning-based approaches are Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), Conditional Random Field (CRF) and Support Vector Machines (SVM). And more recently, approaches based on deep learning models have had good results (Li et al., 2022; Yadav & Bethard, 2019). Supervised learning-based approaches require labeled datasets for training the models.

In English, there are several datasets for evaluating NER models (see Table 1 of Li et al. (2022)). One of the most popular is CoNLL (Sang, 2002; Sang & De Meulder, 2003), which also includes other languages. It is labeled with the categories names of persons, organizations, locations and miscellaneous names (entities that do not belong to the previous three groups). Another, HYENA (Yosef, Bauer, Hoffart, Spaniol, & Weikum, 2012) has 505 entity categories, taken from Wikipedia texts. In specific domains there are also datasets for Biomedicine (Kim, Ohta, Tateisi, & Tsujii, 2003), Agriculture (Malarkodi, Lex, & Devi, 2016), unstructured handwritten document images (Adak, Chaudhuri, & Blumenstein, 2016), among others.

This work contributes with a specific domain dataset for the Portuguese language. In this language, there are few public datasets available. The most popular is HAREM (Santos & Cardoso, 2006; Santos et al., 2006), which consists of a golden collection of texts manually annotated for the first evaluation contest for NER in Portuguese. The first collection was increased with the completion of the Second HAREM edtion (Freitas et al., 2010). The HAREM[1] collections for NER are available in three different versions: FirstHarem, MiniHarem and SecondHarem. The texts are annotated with the following ten entity categories: person (pessoa), organization (organização), location (local), time (tempo), value (valor), abstraction (abstração), event (acontecimento), thing (objects which have names) (coisa), title (works of art, man made things) (obra), and other (outra).

[1] http://www.linguateca.pt/HAREM/

Manually annotated gold-standard corpora are highly expensive to produce, especially for many languages. An alternative is to annotate data automatically. Even though they are of lower quality, automatically generated annotations are important for training NER systems, especially for resource-scarse languages. To meet this demand, Nothman, Ringland, Radford, Murphy, and Curran (2013) created a multilingual corpora annotated for NER, including Portuguese, called WikiNER. The data was labeled by exploiting the text and structure of Wikipedia. The authors present a strategy to label the text for each outgoing link with the entity category of the target article. The method for classifying target articles takes advantage of Wikipedia's multilingual structure. The texts are annotated with the following entity categories: person, organization, location and other miscellaneous entities.

Mendonça Jr. et al. (2015) revised a subset of the Brazilian Portuguese version of the WikiNER corpus, manually making corrections to some of its incorrect labeling. According to the authors, WikiNER has some limitations as it does not consider the context of sentences. They also added new sentences from the news domain, increasing the presence of instances of the organization category, which is deficient in the WikiNER corpus. In addition, they added the time category to the class set. The new corpus was named Paramopama.

WikiANN (Pan et al., 2017) is another silver-standard corpus, i.e., a large amount of automatically labeled text, also from Wikipedia. It is a multilingual NER dataset, which includes Portuguese, annotated with person, organization and location tags. SESAME (Menezes, Savarese, & Milidiú, 2019) is also a silver-standard corpus, which exploites the structure of DBpedia and Wikipedia, annotated with person, organization and location tags. Silva, Silva, Dutra, and Araujo (2021) also created a silver-standard dataset, from journalistic texts. The dataset was automatically labeled by a methodology that uses grammatical classes annotated in a corpus to mark tokens in the categories person, organization and location.

It is well known that NLP tools work better on formal texts. To handle the NER task in informal, noisy and short texts, Peres, Esteves, and Maheshwari (2017) created a gold-standard dataset of tweets in Portuguese, annotated with person, organization and location tags.

We found only two datasets for specific domains, both created from Brazilian legal texts, containing specific tags for law and legal case entities. One of them is the LeNER-Br dataset (de Araujo et al., 2018), which was manually labeled with the generic categories person, organization, location and time, and with two legal categories: legislation and jurisprudence (legal cases). The other dataset in the legal domain is UlyssesNER-Br (Albuquerque et al., 2022). It was manually labeled with seven categories of entities, five being generic (person, organization, location, event and time) and two specific to the legislative domain (law foundation and law product). Some categories have also types, totalizing eighteen types of entities.

Most of the datasets available for the Portuguese language were labeled only with the categories person, organization and location, except for HAREM,

which has ten generic domain categories, and the two datasets of the legal domain, LeNER-Br and UlyssesNER-Br . This work contributes to the creation of the CachacaNER dataset in the specific domain of beverages. It has the largest number of entity categories, seventeen, and is described in detail in the next section.

3 The CachacaNER Dataset

This section describes the methodology for creating the dataset called CachacaNER, as well as some statistics about its data. The flow in Figure 1 presents the steps for creating the dataset, which are described in more detail in the following subsections.

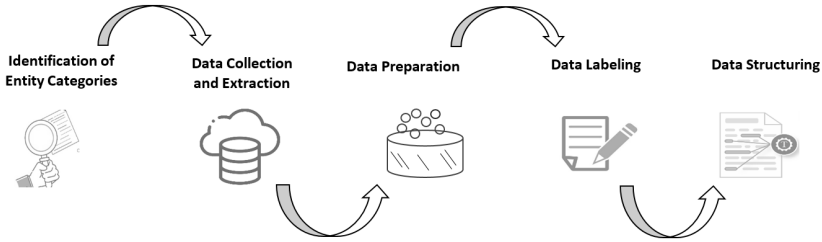

Fig. 1: Steps for creating the CachacaNER dataset

3.1 Named Entity Categories

In order to identify the categories of entities, we carried out a study on the cachaça drink from scientific articles, e-commerce sites, blogs and social networks, among others, to raise the characteristics, attributes and components that best describe the drink. As a result, we identified eleven categories specific to the beverage domain and six generic categories, which are in texts about drinks, but can also be found in texts on different domains. Table 1 describes the categories and presents examples. The last six are the generic categories.

3.2 Data Collection and Extraction

To create the CachacaNER dataset, we collected data from 24 e-commerce sites for the cachaça product. To collect and extract data from web pages, we develop web scraping scripts using the Beautiful Soup[2] and Selenium[3] APIs. In total, we collected data from 3,381 web pages in HTML format.

Data from the pages collected were extracted, structured and stored in a database. Some examples of the extracted metadata are: beverage brand name, price, storage time, sensory analysis, descriptive information, cachacier

[2]https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[3]https ://www.selenium.dev/

Table 1: Named Entity Categories with Examples

Category	Description	Example
Drink Name	Product trade name	Do Chefe, Prazer de Minas, 51
Alcoholic Graduation	Alcoholic strength	42%, 27GL
Drink Classification	Classification according to some manufacturing characteristics and blends	Clássica, Tradicional, Prata, Ouro, Premium, Extra Premium, Reserva Especial, Blend
Distillation Equipment	Type of equipment/device used in the distillation process	Alambique, Coluna
Storage Time	Amount of time the beverage is stored for aging	8 anos, 12 anos
Storage Container	Container in which the beverage is stored before being marketed	barril, dorna, tonel, secular de paróis, tanque de aço inoxidável
Wood Type	Name of the wood used to make the container where the beverage is stored for aging	Amburana, Bálsamo, Jequitibá, Ariribá, Carvalho Europeu, Castanheira, Cumaru
Sensory Characteristic - Color	Drink coloring	amarela, clara, translucida, branca, brilhante
Sensory Characteristic - Aroma	Aroma or smell given off by the drink	canela, toque de especiarias, floral
Sensory Characteristic - Flavor	Taste felt in the mouth when the drink is ingested	doce, azedo, ácido, amargo, salgado, adstringente
Sensory Characteristic - Consistency	Texture or perceived mouthfeel in relation to the drink	aveludado, macio, viscoso, cremoso, oleoso, licoroso, encorpado, pesado
Person Name	Name or nickname referring to a human being	Pedro, Maria S. Silva, Ronaldinho Gaúcho
Organization Name	Name of entity that has its own administration	Fazenda do Cantagalo, Grupo Gouveia Brasil
Location Name	Geographic location	Lavras, Minas Gerais, Brasil
Volume	Occupied space inside a container	250ml, 200 litros
Time	Time representation	10/10/2021, agosto de 2021, 10/02/20 às 10h33
Price	Monetary value	R$120,00, 20 reais

analysis, company or beverage history, awards, and ingredients. The sites are heterogeneous, some did not contain all the data described. The metadata does not directly correspond to the dataset's entity categories. The beverage brand name contains entities referring to the categories drink name, drink classification, wood type and volume, for example, "'Cachaça Vale Verde Tradicional Extra Premium Carvalho Escocês 700ml". Sensory analysis and cachacier analysis contains texts describing the four sensory characteristics of the dataset.

3.3 Data Preparation

The data preparation stage consisted of: (i) preprocessing the data (ii) completing it with metadata (iii) dividing the texts into sentences and (iv) selecting a subset of the data for manual labeling.

The first step was to carry out the preprocessing of the data, which consisted of removing irrelevant attributes for the generation of the dataset, standardizing the character encoding, removing spacing and replicated end-of-line character.

The next step consisted of inserting the text corresponding to the metadata in front of the respective text containing its value. For example, "Price: R$55.00", insertion of the "Price" metadata before the value "R$55.00". This is useful for adding context to data, especially those composed of a single token. They are easily interpreted by a person viewing a web page, however lose meaning when viewed in isolation.

The third step consisted of dividing the texts into sentences, which is the traditional format used in NER datasets. The texts were broken by punctuation marks (.,?,!).

Finally, in the last step, we selected 1,000 documents to be manually labeled. We call a document all the text corresponding to a product (a specific cachaça) contained in a collected web page. The limitation on the amount of documents is due to the high cost of manual labeling. To ensure diversity among the texts that make up the dataset, we selected all documents from the sites with less than 69 pages, and from the others, we selected 69 documents.

3.4 Data Labeling

The dataset labeling was done manually by three students (two undergraduates and one graduate), using the Doccano[4] tool.

To guide the process, we have created a labeling guidelines document. The identification of entities related to sensory characteristics are the ones that most depend on the labeler's perception. The orientation for labeling these entities was based on the Cachaça Sensory Wheel, proposed by Bortoletto (2016), and on the survey of sensory attributes developed in the work of Pinheiro (2010).

Initially, each labeler took annotations on a sample of 30 documents. Afterwards, the labelers had a discussion about the annotations to clarify issues and solve the problems found. From this, the guidelines document was adjusted and each labeler, individually, made their annotations in all documents.

After labeling, the results were automatically compared to identify discrepancies between labelers. The annotations in which there was agreement among the three labelers and those in which two of them agreed were directly inserted into the final dataset, using the majority vote strategy. For only 403 entity annotations, there was disagreement between the three labelers. For these,

[4] https://github.com/doccano/doccano

the labelers jointly reanalyzed them and defined the final annotation that was inserted into the dataset.

We used the Fleiss' Kappa coefficient of agreement, proposed by Fleiss (1971), to assess the consistency of the labels. This metric measures the agreement among three or more raters when attributing categorical classifications to the same data set. As a result, the overall agreement among our three labelers was 0.857, which, according to the interpretation proposed by Landis and Koch (1977), is considered almost perfect. Analyzing the categories of entities individually, the one in which we obtained the lowest agreement result was the Sensory Characteristic - Flavor category, with the value 0.723, which is interpreted as a strong agreement.

3.5 Dataset Structuring

The CachacaNER dataset is available in IOB2 format, a variant of the original IOB schema (Ramshaw & Marcus, 1995). In this scheme, a label starting with "B-" indicates that the token is the beginning of a named entity, "I-" indicates that the token is inside a named entity, and "O" indicates that the token does not pertain to any named entity. Named entities are assumed to be non-overlapping and not spanning more than one sentence. Table 2 presents an excerpt from the dataset.

Table 2: An excerpt from the dataset

Token	Label
A	O
Cachaça	O
Serra	B-NOME_BEBIDA
Limpa	I-NOME_BEBIDA
355ml	B-VOLUME
é	O
uma	O
bebida	O
armazenada	O
por	O
seis	B-TEMPO_ARMAZENAMENTO
meses	I-TEMPO_ARMAZENAMENTO
em	O
tonéis	B-RECIPIENTE_ARMAZENAMENTO
de	I-RECIPIENTE_ARMAZENAMENTO
inox	I-RECIPIENTE_ARMAZENAMENTO
.	O

In addition to the token and label, the dataset also has the following attributes: sentence number, document number, starting and ending position of the token within the sentence, partition number for cross-validation experiments, and whether the item is part of the training or test set.

The dataset is divided into training and test partitions, in order to facilitate the execution of experiments and compare the performance of NER algorithms.

It contains 10 partitions, which allows for cross-validation experiments. Furthermore, the first 7 partitions are marked as part of the training set, and the last 3 as part of the test set, in order to establish a baseline pattern for experiments with a single training and test run.

Each of the 10 partitions has 100 documents. To create a proportional division and ensure diversity among the data, for each partition, we randomly selected approximately 10% of documents from each site. Proceeding in this way, we verify that about 10% of the entities of each category were allocated to each partition, and that about 70% of them are in the training set and 30% in the set of test. The graphics in Figure 2 show the distribution of the number of entities per category for each partition.

3.6 Dataset Statistics

In this section, we present some statistics extracted from the CachacaNER dataset. Figure 3 shows the word cloud for the dataset. Among the most common words are: cachaça, volume, graduação (graduation), alcoólica (alcoholic), carvalho (oak), madeira (wood), envelhecida (aged) and alambique (alembic).

The graphic in Figure 4 shows the distribution of the number of entities per category. The categories with the highest and lowest number of labeled entities are Location Name (4,232) and Sensory Characteristic - Consistency (278), respectively.

Table 3 presents some measures of central tendency (arithmetic mean and median) and dispersion (standard deviation) extracted from the dataset for tokens, sentences and documents.

Table 3: Statistics for tokens, sentences and documents

Measure	Tokens per sentence	Tokens per doc.	Sentences per doc.
Minimum	1	17	3
Maximum	125	974	45
Mean	13.42	183.01	13.62
Median	9	140.5	12
Std dev.	12.32	136.98	7.38

Table 4 presents the number of documents, sentences, tokens and entities for the training and test partitions, and the total for the dataset.

Table 4: Number of documents, sentences, tokens and entities per partition and total

Partition	# docs	# sentences	# tokens	# entities
Training set	700	9,454	129,380	16,651
Test set	300	4,174	53,639	7,388
Total	1,000	13,628	183,019	24,039

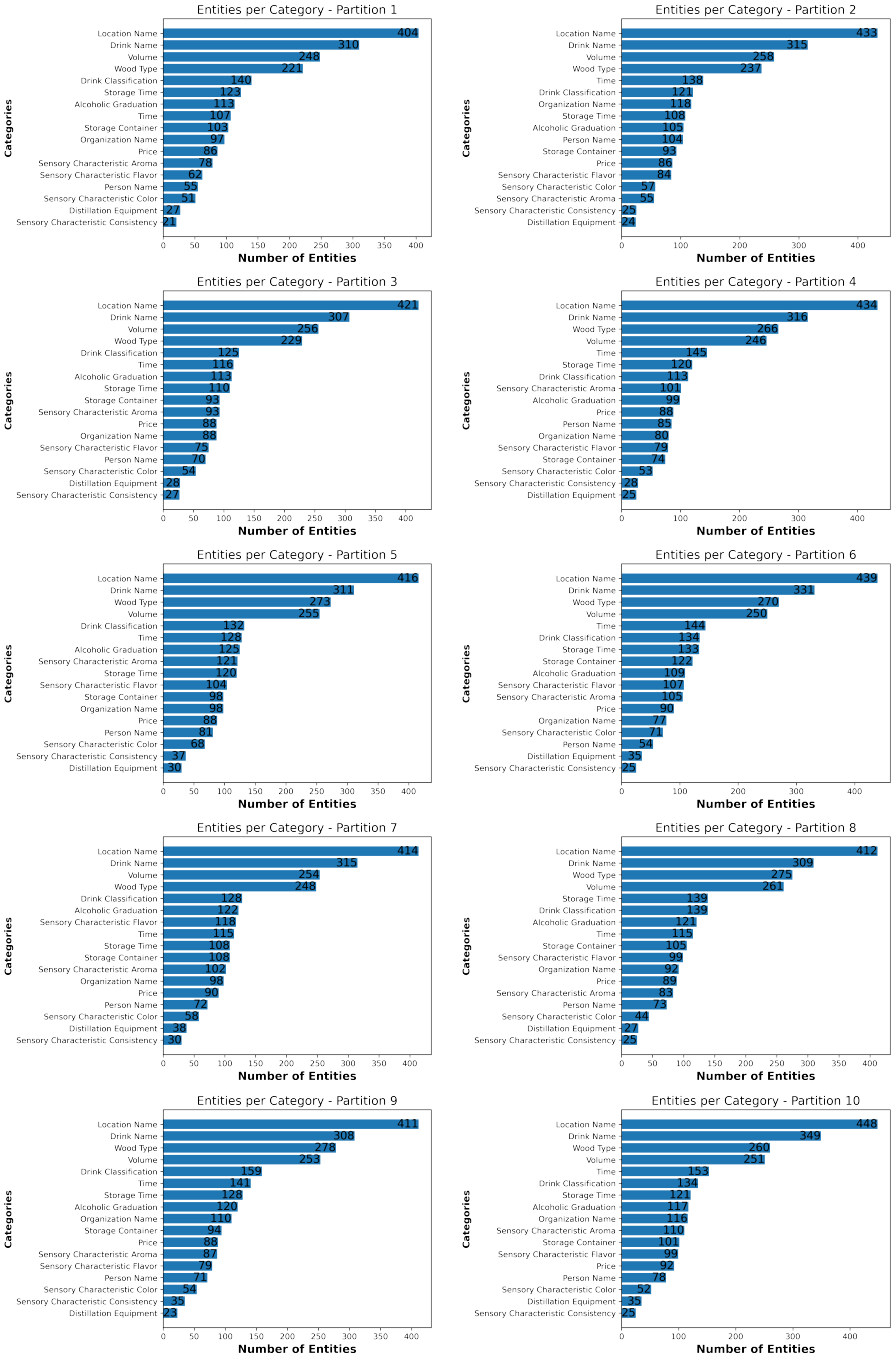

Fig. 2: Number of entities per category for each partition.

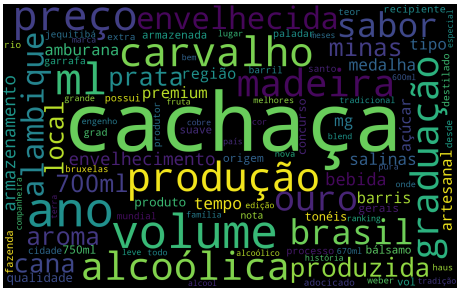

Fig. 3: Word cloud for the CachacaNER dataset

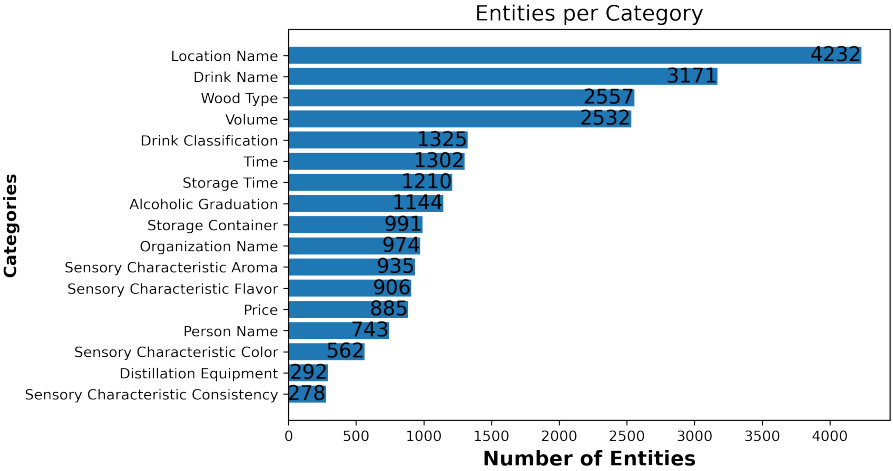

Fig. 4: Number of entities per category

4 Experimental Evaluation

The experimental evaluation aimed to evaluate the CachacaNER dataset for the named entity recognition task and establish a baseline value for the evaluation of NER models on this dataset.

4.1 Experimental Setup

We used the spaCy[5] tool to perform the experimental evaluation. spaCy offers pre-trained deep learning NER models, which we fine-tuned using the CachacaNER training partition. The reported results were obtained by evaluating the test partition of the dataset.

We ran the experiment with the SGD optimizer and with various combinations of values for the number of epochs, batch size and dropout rate hyperparameters. The results were very little affected by variations in these

[5] https://spacy.io/api/entityrecognizer

hyperparameters. The results reported in Section 4.3 were obtained with epochs = 50, batch = 64 and dropout = 0.30.

4.2 Evaluation Metrics

The prediction generated by the application of a NER model results in a sequence of labels of the categories corresponding to the classification of the tokens of each sentence. To measure the quality of extraction, we used the following criteria: (i) if all tokens of an entity are labeled correctly, this counts as a true positive (TP) and (ii) if tokens of an entity are labeled partially or if a token that is not an entity is labeled an entity, it is counted as a false positive (FP).

We used the metrics Precision, calculated as $P = \frac{TP}{TP+FP}$, Recall, calculated as $R = \frac{TP}{T_A}$, where T_A is the number of entities annotated, and F1-Measure, calculated as $F1 = 2 * \frac{P*R}{P+R}$.

4.3 Results and Discussions

Table 5 presents the NER model performance results for each entity category and overall performance for all categories.

Table 5: NER model performance results for each entity category and overall performance for all categories

Category	Precison	Recall	F1
Drink Name	0.836	0.804	0.820
Alcoholic Graduation	0.980	0.983	0.981
Drink Classification	0.824	0.856	0.839
Distillation Equipment	0.842	0.882	0.862
Storage Time	0.952	0.930	0.941
Storage Container	0.930	0.930	0.930
Wood Type	0.948	0.926	0.937
Sensory Characteristic - Color	0.872	0.913	0.892
Sensory Characteristic - Aroma	0.737	0.679	0.707
Sensory Characteristic - Flavor	0.711	0.624	0.665
Sensory Characteristic - Consistency	0.884	0.811	0.846
Person Name	0.903	0.887	0.895
Organization Name	0.878	0.815	0.846
Location Name	0.933	0.950	0.941
Volume	0.957	0.887	0.921
Time	0.929	0.941	0.935
Price	0.858	0.903	0.880
All categories	0.897	0.880	0.889

Among all the categories, Alcoholic Graduation reached F1 = 0.981, the highest value. One of the reasons is due to the regularity and structural simplicity of the sentences that represent this category, such as “40%”, “45.8%” and “40GL”. The few errors for this category occurred because the classifier labeled only part of the entity. For example, in the sentence “GRADUAÇÃO

ALCOÓLICA: 38.0%", instead of labeling "38.0%", it just labeled "38" because it was also trained with data without the percent symbol. The Volume and Price categories, for having numerical values and symbols, also had errors of this type.

On the other hand, the Sensory Characteristic - Flavor and Aroma categories were the ones with the lowest F1 values. This is due to the fact that sometimes the same token can be used to represent both flavor and aroma. For example, in the sentence "A cachaça Cabaré possui um toque frutado de baunilha na boca" (Cachaça Cabaré has a fruity touch of vanilla in the mouth), the word "baunilha" (vanilla) refers to the characteristic flavor, while in the sentence "Dom Bré Ouro possui um aroma equilibrado entre baunilha e cravo"(Dom Bré Ouro has a balanced aroma between vanilla and cloves), "baunilha" represents a type of aroma, which produces an ambiguity between these categories, and consequently, classification errors.

The classifier incorrectly assigned entities from the Drink Name category to the Organization Name category, and vice versa, due to the fact that some drinks have the same trade name as the companies that produce them. However, most of the mistakes related to the Drink Name category were by labeling entities composed by the drink name with other tokens. For example, in the sentence "Cachaça Mineiriana Clássica 500 ml" (Classic Mineiriana Cachaça 500 ml), the classifier labeled "Mineiriana Clássica" (Classic Mineiriana) as the name of the drink, instead of simply "Mineiriana".

In the Drink Classification category, the main error was due to the partial identification of some entities. For example, in the sentence "Cachaça Premissa Extra Premium 670ml", the classifier separately labeled the words "Extra" and "Premium", while the correct one would be "Extra Premium". This may be due to the fact that in most training data, the tokens representing this category are single words.

For the Distillation Equipment category, some errors were due to the classification of tokens relating to the drink name or organization name appearing together with the word "alambique" (alembic). For example, in the sentence "Cachaça de Alambique Doçura de Minas Ouro Quinta das Castanheiras 750 ml.", the tokens "Alambique Doçura de Minas" were incorrectly classified as being in the Distillation Equipment category.

Although the Time and Storage Time categories contain similar tokens, there were few errors due to the classification swapping between these categories. For the other categories, we did not identify a pattern for the errors made by the classified.

The overall performance of our dataset can be contrasted with that of other datasets in the Portuguese language. Table 6 presents the performance of two of the main manually labeled NER datasets in Portuguese, HAREM (Freitas et al., 2010; Santos et al., 2006) and LeNER-Br (de Araujo et al., 2018). The results reported for HAREM were obtained from Santos, Dutra, Parreiras, and Brandão (2021).

Table 6: Overall performance results for the CachacaNER, HAREM and LeNER-Br datasets

Dataset	Precison	Recall	F1
CachacaNER	0.897	0.880	0.889
HAREM	0.880	0.876	0.878
LeNER-Br	0.932	0.919	0.925

The results for the three datasets are not directly comparable, as they were obtained by different methods, however they can be used as a reference for the performance of the datasets. It can be noted that CachacaNER performs close to the other two datasets. It is also important to note that CachacaNER has a greater number of named entity categories than the other two datasets, which can make it more complex.

5 Conclusion

In this work, we developed and evaluated a NER dataset in Portuguese, manually annotated, in the specific domain of the cachaça drink. The dataset is composed of seventeen categories of entities, being eleven categories specific to the beverage domain and six generic categories. We identify the main categories of entities described in texts about cachaça, collect and label the data for the NER task and experimentally evaluate the created dataset. We also review and highlight the main NER datasets available for the Portuguese language.

CachacaNER is the first dataset labeled for the NER task in the beverage domain. It has 183,019 tokens, containing 24,039 entities, which is comparable to other manually labeled datasets for the Portuguese language. Labeling was carried out by three labelers, obtaining a high coefficient of agreement in the Kappa metric.

To facilitate experimental evaluations, the dataset is available in partitions for training and test. In our evaluation, we obtained the value of 88.9% of F1-measure in the test set. The Sensory Characteristic - Flavor and Aroma categories were the ones with the lowest F1 value, as they are more difficult to disambiguate.

Cachaça has entity categories that are common to other types of drinks. Thus, in future works, we intend to evaluate our dataset in the prediction of entities for other beverages such as wine and beer. We will also evaluate our dataset for the use of entity recognition in an information retrieval model for a search engine in the cachaça domain.

Acknowledgements

This work was partially supported by the Brazilian National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq) and the Minas Gerais Research Support

Foundation (Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG).

References

Adak, C., Chaudhuri, B.B., Blumenstein, M. (2016). Named entity recognition from unstructured handwritten document images. *Proceedings of the 12th IAPR workshop on document analysis systems* (p. 375-380). 10 .1109/DAS.2016.15

Albuquerque, H.O., Costa, R., Silvestre, G., Souza, E., da Silva, N.F.F., Vitório, D., ... Oliveira, A.L.I. (2022). UlyssesNER-Br: A corpus of brazilian legislative documents for named entity recognition. *Proceedings of the 15th international conference on computational processing of the portuguese language (propor)* (p. 3–14). Berlin, Heidelberg: Springer-Verlag. Retrieved from https://doi.org/10.1007/978-3-030-98305-5_1 10.1007/978-3-030-98305-5_1

Bortoletto, A.M. (2016). *Influência da madeira na qualidade química e sensorial da aguardente de cana envelhecida* (Unpublished doctoral dissertation). Escola Superior de Agricultura "Luis Queiroz", Piracicaba.

de Araujo, P.H.L., de Campos, T., Oliveira, R., Stauffer, M., Couto, S., de Souza Bermejo, P. (2018, September). LeNER-Br: A dataset for named entity recognition in brazilian legal text. *Proceedings of the 13th international conference on computational processing of the portuguese language (propor)* (p. 313-323). 10.1007/978-3-319-99722-3_32

ExpoCachaça (2022). *Números da cachaça: A importância do mercado da cachaça no brasil e no mundo.* Retrieved from https://www.expocachaca.com.br/numeros-da-cachaca (Accessed in September, 2022)

Fleiss, J.L. (1971, November). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378–382. Retrieved from https://doi.org/10.1037/h0031619

10.1037/h0031619

Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P. (2010, May). Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. *Proceedings of the seventh international conference on language resources and evaluation.* European Language Resources Association. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/412_Paper.pdf

Goyal, A., Gupta, V., Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, *29*, 21-43. Retrieved from https://www.sciencedirect.com/science/article/pii/S1574013717302782

https://doi.org/10.1016/j.cosrev.2018.06.001

Instituto Brasileiro da Cachaça (2022). *IBRAC.* Retrieved from https://ibrac.net/ (Accessed in September, 2022)

Katumullage, D., Yang, C., Barth, J., Cao, J. (2022). Using neural network models for wine review classification. *Journal of Wine Economics*, *17*(1), 27–41.

10.1017/jwe.2022.2

Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(suppl_1), i180-i182. Retrieved from https://doi.org/10.1093/bioinformatics/btg1023

10.1093/bioinformatics/btg1023

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lefever, E., Hendrickx, I., Croijmans, I., van den Bosch, A., Majid, A. (2018, May). Discovering the language of wine reviews: A text mining account. *Proceedings of the eleventh international conference on language resources and evaluation (LREC).* European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1521

Li, J., Sun, A., Han, J., Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50-70.

10.1109/TKDE.2020.2981314

Malarkodi, C., Lex, E., Devi, S.L. (2016). Named entity recognition for the agricultural domain. *Research in Computing Science*, *117*(1), 121–132.

Mendonça Jr., C.A.E., Barbosa, L.A., Macedo, H.T., São Cristóvão, S. (2015). Paramopama: a Brazilian-Portuguese corpus for named entity recognition. *XII encontro nacional de inteligência artificial e computacional*

(ENIAC). SBC.

Menezes, D.S., Savarese, P., Milidiú, R.L. (2019). Building a massive corpus for named entity recognition using free open data sources. *arXiv preprint arXiv:1908.05758v1*. Retrieved from https://arxiv.org/abs/1908.05758

10.48550/ARXIV.1908.05758

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, *30*(1), 3–26. Retrieved from https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad

https://doi.org/10.1075/li.30.1.03nad

Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, *194*, 151-175. Retrieved from https://www.sciencedirect.com/science/article/pii/S0004370212000276

https://doi.org/10.1016/j.artint.2012.03.006

Palmer, J., & Chen, B. (2018). Wineinformatics: Regression on the grade and price of wines through their sensory attributes. *Fermentation*, *4*(4). Retrieved from https://www.mdpi.com/2311-5637/4/4/84

10.3390/fermentation4040084

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., Ji, H. (2017, July). Cross-lingual name tagging and linking for 282 languages. *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 1946–1958). Association for Computational Linguistics. Retrieved from https://aclanthology.org/P17-1178 10.18653/v1/P17-1178

Peres, R., Esteves, D., Maheshwari, G. (2017). Bidirectional LSTM with a context input window for named entity recognition in tweets. *Proceedings of the knowledge capture conference.* Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3148011.3154478 10.1145/3148011.3154478

Pinheiro, S.H.d.M. (2010). *Avaliação sensorial das bebidas aguardente de cana industrial e cachaça de alambique* (Unpublished doctoral dissertation). Universidade Federal de Viçosa, Viçosa.

Ramshaw, L., & Marcus, M. (1995). Text chunking using transformation-based learning. *Proceedings of the third workshop on very large corpora.*

Retrieved from https://aclanthology.org/W95-0107

Sang, E.F.T.K. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. *Proceedings of the 6th conference on natural language learning* (pp. 155–158). Association for Computational Linguistics. Retrieved from https://aclanthology.org/W02-2024

Sang, E.F.T.K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the 7th conference on natural language learning* (pp. 142–147). Association for Computational Linguistics. Retrieved from https://aclanthology.org/W03-0419

Santos, D., & Cardoso, N. (2006). A golden resource for named entity recognition in portuguese. *International workshop on computational processing of the portuguese language* (pp. 69–79). Springer.

Santos, D., Dutra, F., Parreiras, F., Brandão, W. (2021). Assessing the effectiveness of multilingual transformer-based text embeddings for named entity recognition in portuguese. *Proceedings of the 23rd international conference on enterprise information systems* (p. 473-483). SciTePress. 10.5220/0010443204730483

Santos, D., Seco, N., Cardoso, N., Vilela, R. (2006, May). HAREM: An advanced NER evaluation contest for Portuguese. *Proceedings of the fifth international conference on language resources and evaluation.* European Language Resources Association. Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf

Silva, R.d.A., Silva, L.d., Dutra, M.L., Araujo, G.M.d. (2021). An improved ner methodology to the portuguese language. *Mobile Networks and Applications*, *26*, 319–325. Retrieved from https://doi.org/10.1007/s11036-020-01644-x

10.1007/s11036-020-01644-x

Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470v1*. Retrieved from https://arxiv.org/abs/1910.11470

10.48550/ARXIV.1910.11470

Yosef, M.A., Bauer, S., Hoffart, J., Spaniol, M., Weikum, G. (2012, December). HYENA: Hierarchical type classification for entity names. *Proceedings of the international conference on computational linguistics* (pp. 1361–1370). Retrieved from https://aclanthology.org/C12-2133