PRISCILLA DE SOUZA SILVA

ROTULAÇÃO DE DADOS PARA A TAREFA DE RECONHECIMENTO DE ENTIDADES NOMEADAS NO DOMÍNIO DA BEBIDA CACHAÇA

LAVRAS – MG
2022

PRISCILLA DE SOUZA SILVA

ROTULAÇÃO DE DADOS PARA A TAREFA DE RECONHECIMENTO DE ENTIDADES NOMEADAS NO DOMÍNIO DA BEBIDA CACHAÇA

Projeto de mestrado apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Inteligência Artificial e Otimização, para a obtenção do título de Mestre.

Prof. Dr. Denilson Alves Pereira
Orientador

LAVRAS – MG
2022

Ficha Catalográfica preparada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Silva, Priscilla Souza
Rotulação de Dados para a Tarefa de Reconhecimento de Entidades Nomeadas no Domínio da Bebida Cachaça / Priscilla de Souza Silva. – Lavras : UFLA, 2022.
111 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2022.
Orientador: Prof. Dr. Denilson Alves Pereira.
Bibliografia.

1. Reconhecimento de Entidades Nomeadas. 2. Cachaça. 3. Aprendizagem de Máquina. I. Pereira, Denilson Alves. II. Título.

PRISCILLA DE SOUZA SILVA

ROTULAÇÃO DE DADOS PARA A TAREFA DE RECONHECIMENTO DE ENTIDADES NOMEADAS NO DOMÍNIO DA BEBIDA CACHAÇA
DATA LABELING FOR THE TASK OF NAMED ENTITY RECOGNITION IN THE DOMAIN OF CACHAÇA BEVERAGE

Projeto de mestrado apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Inteligência Artificial e Otimização, para a obtenção do título de Mestre.

APROVADA em 25 de Novembro de 2022.

Prof. Dr. Luiz Henrique de Campos Merschmann	UFLA
Prof. Dr. Mozar Jose de Brito	UFLA
Prof. Dr. Daniel Hasan Dalip	CEFET-MG

Prof. Dr. Denilson Alves Pereira
Orientador

LAVRAS – MG
2022

Dedico este trabalho primeiramente a Deus, pois sem Ele nada seria possível. Aos meus pais, minha irmã, meu namorado, meu orientador e aos amigos mais próximos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela dádiva da vida. Agradeço a Ele pelas inúmeras bençãos recebidas, dentre essas o diploma de Mestre em Ciência da Computação, em uma universidade tão prestigiada como a UFLA.

A minha família por todo amor gratuito, sem reservas e incondicional. Em especial aos meus pais, Juvenilda S. Silva e Jodimar G. Silva, que sempre batalharam com afinco para me proporcionar a oportunidade de crescer espiritualmente e profissionalmente. A minha irmã, Mikaelhy, pelo amor e companheirismo.

Agradeço ao meu namorado Kamil, que mesmo a 9.999 km de distância se fez extremante presente nesta etapa da minha vida, por meio de seu amor, companheirismo, fé e orações.

Agradeço especialmente ao professor Denilson Alves Pereira, por acreditar e confiar no meu trabalho, pela paciência, ensinamentos e por me entregar essa incrível pesquisa.

Aos alunos Arthur Franco e Thiago Salles Santos, que com presteza, afinco e paciência cooperaram com este trabalho.

Ao professor Mozar Jose de Brito, pela paciência, atenção e prestatividade durante o processo de elaboração deste trabalho.

A todos os meus colegas de turma com os quais partilhei bons momentos durante o curso.

Por fim e não menos importante, gostaria de agradecer a todos os professores do Instituto de Ciências Exatas e Tecnológicas da UFLA, com os quais tive a honra de trabalhar e aprender, bem como a todos os colaboradores que foram fundamentais a minha formação acadêmica.

A todos meu sincero obrigado!

RESUMO

O Reconhecimento de Entidade Nomeada (NER) é a tarefa de identificar *tokens* em textos livres e os classificar de acordo com um conjunto de categorias pré-definidas, tais como, nome de pessoa, organização e local. Conjuntos de dados rotulados para essa tarefa são essenciais para treinar modelos de aprendizagem de máquina supervisionados. Entretanto, apesar de existirem muitos conjuntos de dados rotulados com textos em inglês, para a língua portuguesa eles ainda são escassos. Portanto, este trabalho contribui com a criação e avaliação de um conjunto de dados rotulado manualmente para a tarefa de NER, com textos escritos em português brasileiro, no domínio específico da bebida destilada cachaça. Essa é uma bebida popular no Brasil e de grande importância econômica. O conjunto de dados proposto neste trabalho é o primeiro em português no domínio de bebidas e pode ser útil para outros tipos de bebidas com categorias de entidades semelhantes a cachaça, como o vinho e a cerveja. Neste trabalho é descrito o processo de coleta e extração de dados textuais, criação e rotulação do conjunto de dados NER e sua avaliação experimental. Como resultado obteve-se um *dataset* chamado de cachacaNER, o qual contém mais de 180.000 *tokens* rotulados em 17 categorias de entidades nomeadas específicas ao contexto da cachaça e categorias genéricas. De acordo a métrica Kappa de Fleiss a concordância (0,857) obtida entre os diferentes rotuladores foi quase perfeita, garantindo a confiabilidade do *dataset* em relação às rotulações feitas manualmente. O tamanho do conjunto de dados, bem como o resultado de sua avaliação experimental, são comparáveis a outros conjuntos de dados em língua portuguesa, embora o deste trabalho tenha um número maior de categorias de entidades nomeadas. Além da rotulação manual, também foi avaliada uma técnica de rotulação automática de entidades, com os dados do cachacaNER, a fim de propor uma rotulação mais rápida e com menos trabalho manual. Como resultado, identificou-se que o modelo de NER treinado com os dados rotulados automaticamente obteve um bom desempenho (F1 de 0,808), considerando o resultado do mesmo modelo treinado com os dados rotulados manualmente (F1 de 0,899).

Palavras-chave: Reconhecimento de Entidades Nomeadas. Cachaça. Aprendizagem de Máquina. Processamento de Linguagem Natural.

ABSTRACT

Named Entity Recognition (NER) is the task of identifying tokens in free text and classifying them according to a set of predefined categories such as person name, organization and location. Datasets labeled for this task are essential for training supervised machine learning models. However, although there are many datasets labeled with texts in English, for the Portuguese language they are still scarce. Therefore, this work contributes with the creation and evaluation of a manually labeled dataset for the NER task, with texts written in Brazilian Portuguese, in the specific domain of the distilled beverage cachaça. Essa é uma bebida popular no Brasil e de grande importância econômica. The dataset proposed in this work is the first in Portuguese in the field of beverages and may be useful for other types of beverages with categories of entities similar to cachaça, such as wine and beer. This work describes the process of textual data collection and extraction, creation and labeling of the NER data set and its experimental evaluation. As a result, a dataset called cachacaNER was obtained, which contains more than 180,000 tokens labeled in 17 categories of named entities specific to the cachaça context and generic categories. According to Fleiss' Kappa metric, the agreement (0.857) obtained between the different labelers was almost perfect, guaranteeing the reliability of the dataset in relation to manual labeling. The size of the dataset, as well as the result of its experimental evaluation, are comparable to other datasets in Portuguese, although the one in this work has a greater number of categories of named entities. In addition to manual labeling, an automatic entity labeling technique was also evaluated, with cachacaNER data, in order to propose faster labeling with less manual work. As a result, it was identified that the NER model trained with automatically labeled data performed well (F1 of 0.808), considering the result of the same model trained with manually labeled data (F1 of 0.899).

Keywords: Named Entity Recognition. Cachaça. Machine Learning. Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O Reconhecimento de Entidade Nomeada, do inglês *Named Entity Recognition* (NER), é uma tarefa computacional que permite o reconhecimento de entidades (ou conceitos) relevantes em textos, bem como a classificação dessas entidades de acordo com um conjunto de categorias semânticas pré-definidas, tais como, nome de pessoa, organização, local, data e hora. Uma entidade pode ser qualquer palavra ou sequência de palavras que se refiram ao mesmo conceito. Por exemplo, na sentença "Desde 2015 a Cachaça Princesa Isabel é produzida em Linhares pelo mestre Leandro Marelli, um mago dos destilados", tem-se as entidades: "2015" (tempo), "Princesa Isabel" (nome da bebida), "Linhares" (local) e "Leandro Marelli" (nome de pessoa).

A tarefa de NER surgiu em 1996 na 6ª *Message Understanding Conference* (MUC-6) (GRISHMAN; SUNDHEIM, 1996), um evento focado na extração automática de informações em mensagens militares. Nesse evento, identificou-se a importância do reconhecimento de determinadas entidades para a extração de informações relevantes em conjuntos de dados textuais, tais como, dados de páginas Web e artigos jornalísticos. Atualmente, essa tarefa também desempenha um papel fundamental em sistemas de recuperação de informação, resposta a perguntas, tradução automática, resumo automático de texto e monitoramento de eventos/produtos. Ela permite a desambiguação do contexto de um conteúdo textual. Por exemplo, reconhecer a entidade "São Paulo" como um local (nome de uma cidade brasileira) em uma frase, pode ser importante para detectar onde um determinado evento ocorreu e diferenciá-la de uma entidade da categoria referente a nome de santo no contexto religioso.

As pesquisas atuais relacionadas a NER são em sua grande maioria baseadas em técnicas de aprendizagem de máquina supervisionada (GOYAL; GUPTA; KUMAR, 2018). Esse tipo de aprendizagem requer dados anotados/rotulados, para que um modelo de NER seja treinado e então consiga identificar e categorizar entidades em textos livres. Entretanto, a rotulagem manual tem um alto custo, pois é um processo caro, tedioso e demorado. Embora existam vários conjuntos de dados textuais (*datasets*) rotulados manualmente na língua inglesa, em português eles são mais raros. Entre eles estão o primeiro e o segundo HAREM (MOTA; SANTOS, 2008; ALBUQUERQUE et al., 2022), o Paramopama (JUNIOR et al., 2015) e o LeNER-Br (JUNIOR et al., 2015). Os dois primeiros são formados por textos de diferentes contextos e o último, por textos na área específica do judiciário.

Considerando a problemática descrita anteriormente, este trabalho teve como um de seus objetivos a criação e avaliação de um novo conjunto de dados NER em português, anotado

manualmente, no domínio específico da bebida cachaça, o qual recebeu o nome de cachacaNER. Além disso, também foi avaliada uma técnica de rotulação automática de dados, a fim de propor uma rotulação mais rápida, que não depende tanto de esforço manual.

A cachaça é uma bebida destilada, tipicamente brasileira, cuja principal matéria-prima é a cana-de-açúcar. Consiste em uma das bebidas mais conhecidas dentro e fora do país e compõe o coquetel mundialmente conhecido como caipirinha. Essa bebida faz parte da história do Brasil, pois surgiu ainda no perídio colonial, quando os portugueses improvisaram uma bebida destilada a partir da fermentação e destilação de derivados do caldo da cana-de-açúcar, que produzia o mesmo efeito alcoólico da Bagaceira, um destilado de origem portuguesa feito de casca de uva (Brasil Travel New, 2021). Uma das principais características da cachaça é sua versatilidade, pois pode ser consumida de várias formas, pura, gelada ou misturada com outras bebidas.

De acordo com o Instituto Brasileiro da Cachaça (2021), a cachaça é a segunda bebida alcoólica mais consumida no Brasil, atrás apenas da cerveja. Representa 72% do mercado de destilados do país, além de ser um dos quatro mais consumidos no mundo (Brasil Travel New, 2021). Atualmente é exportada para 67 países, dentre eles, os maiores importadores são os EUA, a Alemanha e o Paraguai.

Segundo o IBRAC (2021), Instituto Brasileiro da Cachaça, o Brasil possui a capacidade de produzir 1,2 bilhão de litros de cachaça por ano e produz efetivamente aproximadamente 800 milhões de litros, o que gera mais de 600 mil empregos diretos e indiretos no país (SEBRAE, 2022). Apenas em 2020 foram exportados 7,22 milhões de litros de cachaça, o que resultou em um faturamento de 13,17 milhões de dólares. Esses números representam um crescimento de 38,39% em valor e de 29,52% em volume, em comparação a 2019 (IBRAC, 2021). Além disso, o Anuário da Cachaça 2021 registrou a existência de 1.131 produtores de cachaça e aguardente no Brasil em 2020, aumento de 4,14% em relação aos 1.086 de 2019 (MAPA, 2021).

Apesar da importância e contribuição da cachaça para o mercado e economia brasileira, após uma revisão da literatura e até onde vai nosso conhecimento, não existem *datasets* de NER anotados em português para o domínio de bebidas, tal como a cachaça. Essa bebida em particular compartilha características em comum com outros tipos de bebidas, tais como, vinho, cerveja e whisky. O vinho, por exemplo, assim como a cachaça, possui características sensoriais (cor, aroma, sabor e consistência), recipiente de armazenamento, local de origem e preço. Na literatura, é possível encontrar alguns trabalhos que buscam extrair informações relevantes de

textos em inglês sobre o vinho, tal como, Katumullage et al. (2022), Lefever et al. (2018), Palmer e Chen (2018) que aplicaram técnicas de mineração de texto em revisões de vinho, a fim de identificar características relevantes sobre esse tipo de bebida. Assim sendo, um conjunto de dados de NER no domínio da cachaça também pode ser útil para extrair informações relevantes sobre outros tipos de bebidas.

Dados rotulados com entidades nomeadas sobre cachaça podem ser úteis na construção de sistemas de recomendação de bebidas, os quais facilitem o processo de tomada de decisão ao mostrar e recomendar por meio de anúncios de empresas específicas uma seleção de cachaças que melhor se encaixem no que o cliente está procurando ou que gostaria de consumir. Além disso, também podem ser utilizados por máquinas de busca para recuperar informações relevantes sobre a bebida cachaça, dentre outras bebidas com mesmas características.

Para ilustrar os dados contidos no cachacaNER, *dataset* criado neste trabalho, segue o trecho de um texto extraído do mesmo contendo a análise de um cachacier (análogo a um *sommelier*[1] de vinho) sobre uma cachaça específica: “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.”.

1.1 Contribuições

Dentre as principais contribuições desta dissertação para a área da computação destacam-se: (i) identificação das principais categorias de entidades nomeadas descritas em textos sobre cachaça, (ii) coleta, extração e rotulagem (manual) de dados textuais em português para a tarefa de NER, (iii) avaliação experimental do *dataset* criado e (iv) avaliação de uma técnica de rotulação automática de entidades nomeadas. O *dataset* cachacaNER, encontra-se disponível publicamente em <https://github.com/LabRI-Information-Retrieval-Lab/CachacaNER>.

Os resultados deste trabalho também contribuem das seguintes maneiras para o mercado de cachaça: (v) construção de mercados e melhoria da comunicação entre produtores e consumidores de cachaça, (vi) avanço do conhecimento sobre o consumo de cachaça e outras bebidas, (vii) as categorias identificadas podem servir de referência para o desenvolvimento ou

[1] Profissional responsável por cuidar da carta de bebidas de restaurantes, bares, importadoras e lojas especializadas.

construção de marcas de cachaça, (vii) construção da identidade da cachaça por meio da escrita apresentada neste trabalho.

1.2 Objetivos Geral e Específicos

O objetivo geral desta pesquisa consiste na rotulação e avaliação de dados textuais para a tarefa de reconhecimento de entidades nomeadas na língua portuguesa, no domínio da bebida cachaça.

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

a) criar uma base de dados composta por documentos textuais sobre a bebida cachaça;
b) identificar as categorias de entidades nomeadas que caracterizem a cachaça;
c) criar e avaliar um *dataset* para a tarefa de NER rotulado manualmente;
d) avaliar uma técnica de rotulação automática de entidades nomeadas para o *dataset* criado.

1.3 Estrutura deste Documento

O restante deste documento está organizado da seguinte maneira. O Capítulo 2 apresentada os pressupostos teóricos que embasam esta pesquisa, tais como, Aprendizagem de Máquina, PLN, Mineração de Texto, Reconhecimento de Entidades Nomeadas e Cachaça. O Capítulo 3 apresenta os principais trabalhos na literatura sobre criação de conjuntos de dados, em português, para a tarefa de NER. O Capítulo 4 descreve em detalhes a criação, as características do *dataset* cachacaNER, bem como a avaliação experimental dos dados rotulados manualmente. O Capítulo 5 traz a descrição de como foi realizada a rotulação automática de entidades nomeada, bem como a sua avaliação experimental. O Capítulo 6 destaca as conclusões e trabalhos futuros. E por fim, é apresentado um anexo com informações sobre atributos sensoriais da cachaça e um manual de rotulação automática de categorias de entidades nomeadas, proposto nesta pesquisa.

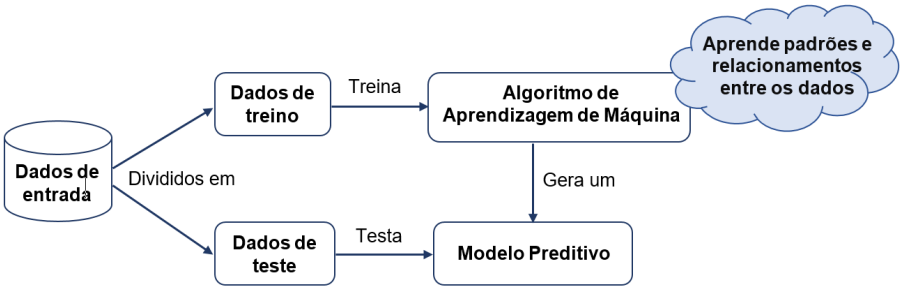
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, é descrito o contexto no qual a presente pesquisa se insere: Aprendizagem de Máquina, Processamentos de Linguagem Natural, Mineração de Texto, Reconhecimento de Entidades Nomeadas e Cachaça.

2.1 Aprendizagem de Máquina

Aprendizagem de Máquina (AM) é uma disciplina da inteligência artificial que fornece aos computadores a capacidade de aprender automaticamente com dados de exemplo, identificando padrões para fazer previsões com o mínimo de intervenção humana. Para isso, cria modelos que recebem dados de entrada como exemplo, e depois utiliza os próprios modelos treinados para realizar novas predições em dados não rotulados (SIMON et al., 2016). Essa tecnologia envolve basicamente dois tipos de abordagens: aprendizagem supervisionada e não supervisionada. Segundo Becker (2013), o objetivo principal de ambas abordagens é a descoberta automática de regras ou padrões gerais, implícitos em grandes conjuntos de dados.

Figura 2.1 – Processo de treinamento e teste de um modelo baseado em aprendizagem de máquina.



Fonte: Adaptado de Escovedo (2020).

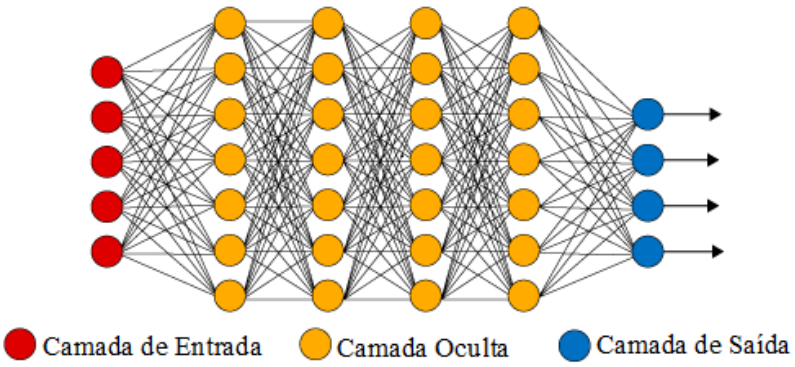
Na aprendizagem supervisionada, o modelo é treinado a partir de um conjunto de dados predefinido, isto é, ele conhece previamente as classes ou categorias que os dados pertencem, o que permite a capacidade de chegar a uma conclusão precisa ao receber novos dados de entrada. A Figura 2.1, ilustra o processo básico de treinamento e teste de um modelo baseado em aprendizagem de máquina supervisionada. Na aprendizagem não supervisionada, o algoritmo recebe os dados e deve encontrar informações e relações existentes neles, mas sem ter acesso às classes dos dados.

Além dos métodos tradicionais de aprendizagem de máquina, tem-se também uma técnica conhecida como Aprendizagem Profunda, do inglês *Deep Learning*, uma rede artificial

que simula a estrutura de nós de neurônios conectados entre si como uma teia, a fim de detectar objetos, reconhecer a fala humana, traduzir idiomas e tomar decisões. Os algoritmos baseados nessa técnica produzem representações hierárquicas com processamento não linear, por meio de camadas de processamento sequencial em uma rede neural artificial (RNA) (WANG et al., 2014).

Diferentemente dos métodos de aprendizagem tradicional, a aprendizagem profunda se baseia na utilização de redes neurais multicamadas, isto é, várias camadas de neurônios matemáticos (artificias) para processar os dados de entrada e extrair informações relevantes. Conforme ilustrado no exemplo da Figura 2.2, a arquitetura de uma rede neural de aprendizagem profunda se dá por meio de vários neurônios alinhados verticalmente constituindo uma camada, os quais se conectam aos neurônios da camada anterior e da seguinte. A primeira camada é denominada de camada de entrada, enquanto a última é chamada de camada de saída, e todas as camadas entre essas duas são conhecidas como camadas ocultas (NEU; LAHANN; FETTKE, 2021).

Figura 2.2 – Rede Neural de aprendizagem profunda.



Fonte: Adaptado de Deep Learning Book (2020).

Durante o processamento, as redes neurais multicamadas fazem com que os dados de entrada sejam passados por meio de cada camada, com a saída da camada anterior fornecendo entrada para a próxima camada. Cada neurônio processa uma transformação linear nos sinais recebidos da camada anterior. Uma função de ativação é aplicada ao resultado da transformação linear para atingir a não linearidade, e o sinal se propaga para a próxima camada de neurônios. Os pesos dos neurônios são os parâmetros do modelo, os quais são treinados iterativamente, propagando de volta o erro entre a previsão da rede e o resultado correto (NEU; LAHANN;

FETTKE, 2021). Dentre as diferentes arquiteturas de redes neurais de aprendizagem profunda existentes, tem-se a Rede Neural Convolucional (CNN) e a Rede Neural Recorrente (RNN).

Rede Neural Convolucional é um tipo de rede projetada originalmente para a análise de imagens, todavia, a sua aplicação em processamento de linguagem natural também tem trazido resultados interessantes para a área, dada sua capacidade de analisar dados sequenciais (ZHOU; RUECKERT; FICHTINGER, 2019). A CNN extrai toda e qualquer porção da imagem de entrada, conhecida como campo receptivo. Ela atribui pesos para cada neurônio, a fim de discriminar a importância deles entre si. Uma rede neural convolucional possui duas operações básicas, a saber, convolução e *pooling*. A operação de convolução utiliza vários filtros para extrair feições (mapa de feições) do conjunto de dados, por meio do qual suas informações espaciais correspondentes podem ser preservadas. A operação de *pooling*, no que lhe concerne, é usada para reduzir a dimensionalidade dos mapas de características da convolução. Na parte mais profunda das convoluções, espera-se que os dados num espaço dimensional reduzido contenham informação suficiente sobre os mapas de feições, para ser possível atribuir um valor semântico ao dado original (ZHU et al., 2018).

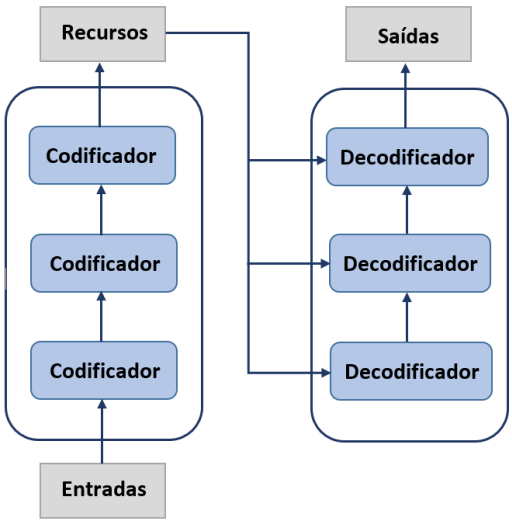
Rede Neural Recorrente processa dados de séries temporais ou dados sequenciais, e é comumente utilizada para tarefas de tradução de linguagem, reconhecimento de fala, geração de texto (legendagem de imagens, por exemplo) e processamento de linguagem natural. As RNNs são projetadas para reconhecer as características sequenciais dos dados e usar padrões para prever o próximo cenário provável, isto é, recebem informações de entradas anteriores para influenciar a entrada e a saída atuais. Essa estrutura de rede se distingue das demais, pois usa *loops* de *feedback* para processar uma sequência de dados que informa a saída final, que também pode ser uma sequência de dados. Esses *loops* de *feedback* permitem que as informações persistam, efeito conhecido como "memória" (MIKOLOV et al., 2010). Os casos de uso de RNN tendem a ser conectados a modelos de linguagem em que saber a próxima letra em uma palavra, ou a próxima palavra em uma sentença depende dos dados que vêm antes dela. Por exemplo, uma RNN pode ser treinada com as obras do autor José de Alencar para produzir automaticamente textos de prosas semelhantes as do referido autor.

Uma estrutura de rede neural que tem se destacado, atualmente, no processamento de linguagem natural é o *Transformer*, o qual consiste em um modelo de *deep learning* que trabalha com o mecanismo de autoatenção. Esse mecanismo possibilita a aprendizagem de contextos linguísticos, pois consegue processar e rastrear as relações/associações entre as palavras que

compõem um texto de entrada. (CRUZ, 2022; VASWANI et al., 2017). Tal como nas redes neurais tradicionais, os modelos *transformers* transformam uma sequência de entrada em outra sequência de saída, por meio da arquitetura codificador-decodificador. O codificador extrai recursos de uma sentença de entrada, e o decodificador utiliza esses recursos para produzir uma sentença de saída (ALAMMAR, 2018; KIKABEN, 2021). Pode ser aplicado, por exemplo, na tradução ou sumarização de textos.

Para processa um texto de entrada, o *transformer* utiliza vários blocos de codificadores e decodificadores, de maneira que a saída do último bloco codificador é a entrada dos blocos decodificadores, conforme ilustrado na Figura 2.3.

Figura 2.3 – Arquitetura codificador-decodificar do modelo *transformer*.



Fonte: Adaptado de Alammar (2018).

Diferentemente dos modelos tradicionais, o *transformer* não recebe uma entrada por vez, mas sim uma frase completa no formato de uma sequência de vetores incorporados. Esses vetores representam a semântica e a posição de cada palavra/*token* na frase. Para conseguir manter a ordem sequencial das palavras referente a sentença original, o modelo aplica a "codificação posicional" em cada vetor, ou seja, ele modifica os valores de cada vetor para representar sua localização na sentença. Após isso, a entrada é passada entre os blocos codificadores, os quais a processam por meio de uma "camada de autoatenção", a fim enriquecer cada vetor de incorporação com informações contextuais de toda a frase (KIKABEN, 2021; VASWANI et al., 2017).

A camada de autoatenção, visa capturar diferentes tipos de relações existentes entre as palavras da frase. Por meio da aplicação de múltiplos cálculos de atenção paralela, o modelo

pode examinar diferentes subespaços de incorporação. Por exemplo, na frase “A UFLA é uma universidade de destaque internacional, e ela merece o reconhecimento que tem.”, o modelo deve estabelecer associações como “UFLA” e “ela’, “UFLA” e “universidade” e“UFLA” e “destaque’. Em suma, a camada de atenção recebe vetores com os valores de palavras individuais e gera vetores que representam palavras individuais com as relações que possuem com as demais palavras (KIKABEN, 2021; VASWANI et al., 2017).

O decodificador utiliza a mesma tokenização, incorporação de palavras e mecanismo de autoatenção do codificador, a fim de traduzir (transformar) o vetor (saída) do codificador nos dados de saída esperados. Por exemplo, traduzir a frase em inglês “I love Brazil”, na frase em português “Eu amo o Brasil”.

2.2 Processamento de Linguagem Natural

Processamento de linguagem natural é uma área da computação que estuda as interações entre computadores e a linguagem humana, de modo que os sistemas computacionais consigam entender, interpretar e utilizar essa linguagem (POWERS; TURK, 1989). PLN também é responsável por grande parte das técnicas computacionais utilizadas na etapa de pré-processamento de dados, etapa que demanda considerável esforço no processo de descoberta de conhecimento (MORENO, 2015). Nessa etapa os dados textuais são geralmente transformados em representações numéricas, as quais podem ser processadas pelos algoritmos de aprendizagem de máquina tradicionais ou de *deep learning*.

Para processar a linguagem humana, o processamento de linguagem natural trabalha com vários níveis de conhecimentos linguísticos, tais como (FELDMAN, 1999; MORENO, 2015):

a) morfológico: lida com a estrutura, a forma e as inflexões das palavras;
b) léxico: trabalha com o acervo de palavras de uma determinada língua;
c) sintático: lida com a gramática e a estrutura das frases;
d) fonético: trabalha com a fala;
e) semântico: traduz o significado das palavras e frases;
f) discurso: analisa a estrutura de diferentes tipos de texto;
g) pragmático: investiga o uso da língua em diferentes contextos, e como esses contextos afetam o significado e interpretação do texto.

Os modelos de PLN propostos para interpretação da linguagem humana escrita não buscam aprender apenas a sequência do texto, mas também compreender como funciona a relação entre o texto e o contexto no qual ele se encontra inserido. O PLN dispõe de diferentes tarefas e técnicas que trabalham especificamente com textos, tais como (OLIVEIRA, 2020):

a) *lowercasing*: tarefa que transforma todas as letras do texto em minúsculas;
b) *word tokenization*: consiste no processo de dividir/quebrar um texto em unidades menores, chamadas de *tokens*. Um *tokem* pode ser uma única palavras, uma junção de palavras, parte de uma frase, ou até mesmo uma frase inteira (MUJTABA, 2020);
c) remoção de *stopwords*: equivale à remoção de palavras que possuem pouco ou nenhum significado dentro do texto, tais como, artigos, preposições e pronomes (K.; SAINI, 2016);
d) *stemming*: tarefa responsável pela redução de uma palavra ao seu radical, a partir da retirada de letras adicionais, isto é, sufixos e prefixos;
e) *lemmatization*: redução de uma determinada palavra ao seu lema, a partir do sentido da palavra;
f) normalização: tarefa que transforma palavras com desvios de escrita, por exemplo, erros ortográficos ou abreviações, em sua forma canônica;
g) remoção de ruídos: consiste na remoção de caracteres que atrapalham o entendimento do texto na totalidade;
h) *part-of-speech tagging*: tarefa de marcar uma palavra em um texto como correspondendo a uma classe gramatical específica. O conjunto mais comum de *tags* é formado por artigos, substantivos, verbos, adjetivos, preposições e nomes próprios;
i) *dependency parsing*: tarefa de reconhecer uma frase e atribuir uma estrutura sintática a ela. Uma das estruturas sintáticas mais utilizadas é a árvore de análise, que pode ser gerada usando alguns algoritmos de análise;
j) *named entity recognition*: visa extrair e classificar as entidades mencionadas em um texto de acordo com diferentes categorias, tais como, pessoas, lugares, preço, etc.

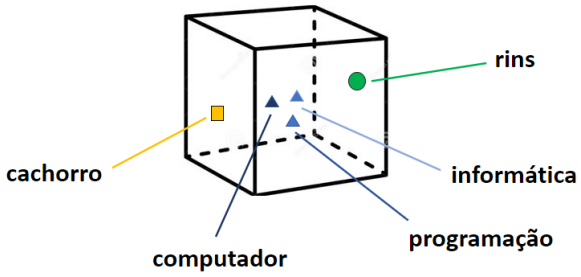
2.3 Representação de Textos

Para realizar operações sobre dados textuais, os algoritmos de aprendizagem de máquina utilizam representações de texto na forma de vetores numéricos, os quais guardam o significado das palavras do texto. A seguir são apresentados alguns métodos de representação de palavras.

O *Bag-of-words* é uma representação, simplificada, de um texto como uma bolsa composta por palavras, que formam o texto. Essa técnica desconsidera a gramática e até mesmo a ordem das palavras, mas mantém a multiplicidade. Por exemplo, para criar o *bag-of-words* da sentença "A cachaça Princesa Isabel foi criada no estado de Espírito Santo.", é necessário quebrar cada uma das palavras, 'A", "cachaça", "Princesa", "Isabel", "foi", "criada", "no", "estado", "do", "Espírito", "Santo". Após essa quebra, pode-se agrupar as palavras e listar a frequência de cada uma no texto. Esses valores podem ser utilizados pelos algoritmos de aprendizagem de máquina, como uma representação matemática das características dos textos de entrada. Posto que algoritmos de aprendizado de máquina necessitam de valores numéricos para realizar suas operações.

Os *Word embeddings*, no que lhe concerne, geram modelos de vetores de palavras que consideram o contexto no qual elas se encontram inseridas (OLIVEIRA, 2020). Nesse método, cada palavra do conjunto de textos de entrada vira um vetor em um espaço multidimensional chamado de *embedding space*. Nesse espaço, a proximidade entre os vetores representa a proximidade entre as palavras, isto é, palavras usadas no mesmo contexto ficam mais próximas umas das outras, enquanto palavras de contextos diferentes se distanciam. A Figura 2.4 ilustra a representação de um espaço vetorial tridimensional, no qual as palavras computador, informática e programação encontram-se agrupadas em uma mesma área, mostrando que fazem parte de um mesmo contexto. Entretanto, as palavras cachorro e rins estão em áreas totalmente diferentes, pois fazem parte de contextos completamente diferentes.

Figura 2.4 – *Embedding space* representado em três dimensões.



Fonte: Adaptado de
https://www.kaggle.com/sbongo/do-pretrained-embeddings-give-you-the-extra-edge.

Na literatura há diferentes modelos de *word embeddings* que podem ser utilizados como representação de palavras para aprendizagem de máquina, tais como: Word2Vec, GloVe, ELMo, FastText e BERT.

Word2Vec (MIKOLOV et al., 2013) é um modelo baseado em rede neural linear com apenas uma camada, capaz de aprender associações de palavras a partir de uma grande quantidade de dados textuais. Esse modelo não é único, pois inclui dois modelos de aprendizagem não supervisionada: *Continuous Bag-of-Words* (CBOW) e *Continuous Skip-gram* (Skip-gram). O CBOW tenta prever uma palavra-alvo com base nas palavras do contexto de origem (palavras ao redor), sendo o contexto interpretado como uma sentença. Todas as palavras que fazem parte de um mesmo contexto são combinadas para prever uma palavra-alvo. Já o Skip-gram, geralmente tenta alcançar o reverso do que o CBOW faz, pois visa prever as palavras do contexto de origem dada uma palavra alvo (SARKAR; BALI; GHOSH, 2018).

O GloVe (PENNINGTON; SOCHER; MANNING, 2014) trabalha de forma semelhante ao Word2Vec, mas ao invés de prever o contexto de uma palavra, ele aprende construindo uma matriz de coocorrência, isto é, palavras *versus* contexto. Essa matriz contabiliza com que frequência uma dada palavra aparece em um contexto.

O FastText (BOJANOWSKI et al., 2017) é uma extensão do modelo Word2vec, que ao invés de aprender vetores para palavras diretamente, usa caracteres de *n-gram* como a menor unidade necessária para se criar um vetor de palavras. Ele considera cada palavra como um saco de caracteres de tamanho *n-gram*, por exemplo, a palavra "polônia", pode ser dividida em unidades de vetores de palavras separadas com 3-gram: "pol", "olô", "lôn", "ôni", "nia". Essa segmentação permite ao modelo entender a estrutura morfológica das palavras como, por exemplo, o significado de palavras mais curtas e a identificação de sufixos e prefixos. No FastText o vetor de uma palavra é considerado a soma de todos os vetores de seus *char-ngrams* componentes.

ELMo (PETERS et al., 2018) é uma representação de palavras contextualizadas, que modela características complexas do uso da palavra (sintaxe e semântica, por exemplo) e como esses usos variam entre os contextos linguísticos, ou seja, para modelar a polissemia[2]. Os vetores de palavras criados com o ELMo são funções aprendidas dos estados internos de um modelo de linguagem bidirecional profundo (biLM), que é pré-treinado em um grande *corpus* de texto. As representações ELMo podem ser: contextual, onde a representação de cada palavra depende de todo o contexto em que é usada; profunda, onde as representações de palavras combinam todas as camadas de uma rede neural profunda pré-treinada; baseada em caracteres, na qual as representações são puramente baseadas em caracteres, permitindo que a rede neural

[2] Multiplicidade de sentidos de uma palavra ou locução.

use pistas morfológicas para formar representações robustas para *tokens* fora do vocabulário invisível no treinamento.

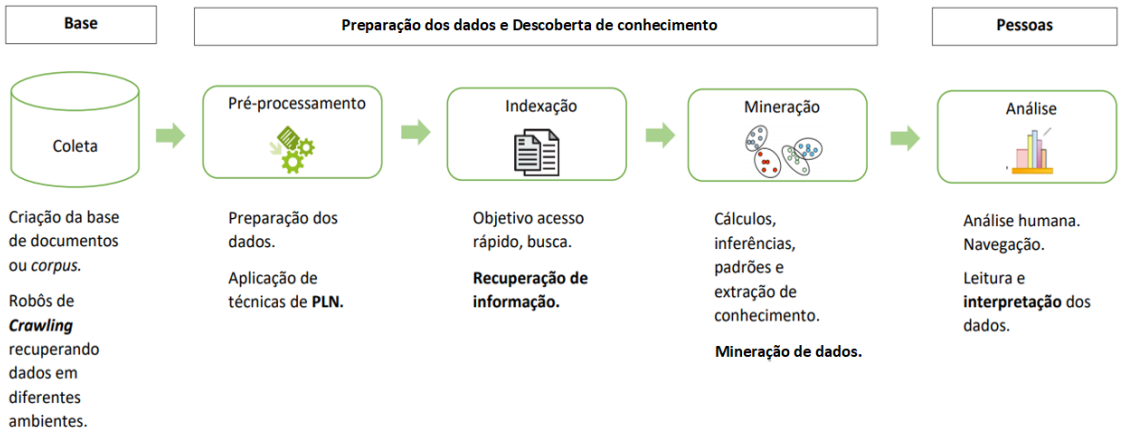
BERT (DEVLIN; CHANG, 2019) é um modelo de representação de palavras recente, desenvolvido por pesquisadores do Google AI Language. Esse modelo utiliza uma técnica inusitada chamada de Masked LM (MLM), a qual permite o treinamento bidirecional da sequência de texto. Esse tipo de treinamento garante ao modelo de representação de palavras um senso mais profundo de contexto e fluxo da linguagem humana se comparado aos modelos de direção única, isto é, que leem a entrada de texto sequencialmente (da esquerda para a direita ou da direita para a esquerda).

2.4 Mineração de Texto

Para tratar dados no formato de texto, compartilhados principalmente por meio da Web, surgiu um campo da computação denominado de Mineração de Texto (MT), do inglês *Text Mining*. Segundo Shelley, Jerry e J.Robert (2007), Mineração de Texto é um processo de descoberta de conhecimento usado para extrair padrões interessantes e não triviais de bases de dados textuais, por meio de princípios da linguística computacional e aprendizagem de máquina. Park e Ryu (2014) definem mineração de texto como uma variedade de técnicas computacionais utilizadas para classificar e organizar uma grande quantidade de texto, com o objetivo de extrair temas comuns, tendências lexicais, tendências frasais ou sentenciais dos dados.

O processo de descoberta de conhecimento na mineração de texto é composto basicamente por cinco etapas (ARANHA; PASSOS, 2009), conforme ilustrado na Figura 2.5.

Figura 2.5 – Processo de descoberta de conhecimento na mineração de texto.



Fonte: Adaptado de Aranha e Passos (2009).

As etapas são descritas a seguir:

a) **coleta** é a etapa inicial do processo, e tem por objetivo criar um conjunto/base de dados textuais, denominada na literatura de *corpus* ou *dataset*. Os dados podem ser extraídos a partir de fontes diversas, tais como, páginas Web, mídias sociais, fóruns, blogs, entre outros;

b) **pré-processamento** é uma das etapas cruciais do processo, pois impacta diretamente na qualidade no conjunto de dados criado. Consiste na identificação e eliminação de anomalias que capazes de comprometer a eficiência do processo, bem como na transformação dos dados textuais em representações numéricas ou vetoriais, que possam ser entendidas pelos algoritmos de mineração de dados. Para tratar e preparar os dados são aplicadas diferentes técnicas de PLN, tais como: remoção de palavras irrelevantes, normalização, *lemmatization*, *stemming*, criação de vetores de palavras (*word embedding*), entre outras;

c) **indexação** é aplicada quando se tem um grande volume de dados, pois permite a organização de todos os termos adquiridos a partir do conjunto de dados, acelerando assim o processamento dos textos;

d) **mineração de texto** é a fase relacionada á aplicação de tarefas de mineração texto e algoritmos de aprendizagem de máquina, para a extração de informações relevantes dos dados resultante das etapas anteriores. Por tanto, dependendo da necessidade de informação em questão, pode-se optar por tarefas como clusterização, classificação, sumarização ou extração de informação;

e) **análise da informação**, consiste na observação dos resultados gerados pela mineração de texto, para serem utilizados como suporte no processo de tomada de decisão. Isto é, consiste em analisar de maneira qualitativa ou quantitativa os resultados gerados pelos algoritmos utilizados na etapa anterior.

Tarefas típicas de mineração de texto incluem classificação, agrupamento, extração de informação, análise de sentimentos, reconhecimento de entidades nomeadas, sumarização, extração de palavras-chave, regressão e nuvem de palavras.

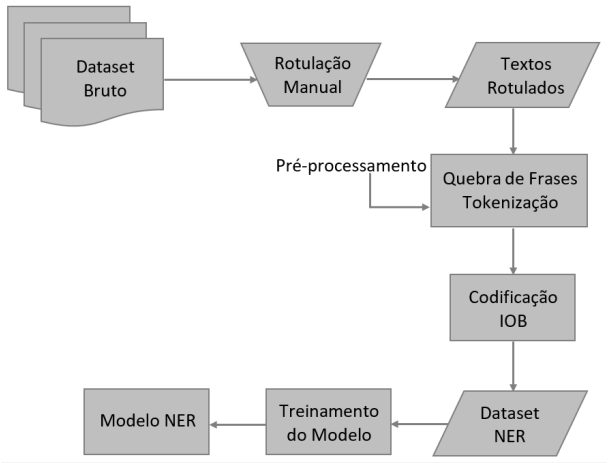
2.5 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas, uma tarefa tipicamente baseada em aprendizagem supervisionada, é responsável pela identificação de entidades nomeadas em docu-

mentos textuais, bem como a atribuição dessas entidades a categorias semânticas pré-definidas (LEITNER; REHM; MORENO-SCHNEIDER, 2019). Entende-se por entidades nomeadas os termos identificados em um texto que dão nomes às coisas reais, por exemplo, pessoa (Isabel), organização (UFLA), lugar (Brasil), tempo (10/02/2021), dentre outras categorias consideradas relevantes em domínios específicos.

A entrada de sistemas ou modelo NER são textos em sua forma livre, e a saída são conjuntos de textos rotulados. Para melhor exemplificar, suponha que a tarefa de reconhecimento de entidades foi aplicada ao texto "O professor Denilson trabalha na UFLA desde 2010, uma universidade localizada em Minas Gerais". O resultado do modelo seria o seguinte conjunto de anotações, formado por categorias *versus* entidades: "Pessoa: Denilson", "Organização: UFLA", "Tempo: 2010", "Lugar: Minas Gerais". A Figura 2.6, traz a ilustração do processo típico de treinamento de um modelo NER, para rotulação de entidades.

Figura 2.6 – Processo típico de treinamento de um modelo NER.



Fonte: Adaptado de Devopedia (2020).

De maneira geral, o processo de treinamento de um modelo NER inicia com a rotulação manual dos dados coletados, por um ou mais especialistas. Em seguida, os textos rotulados podem ser segmentados, isto é, quebrados em sentenças menores, para que cada sentença seja processada individualmente sem depender do contexto da anterior. Posteriormente, ainda como parte do pré-processamento, as sentenças são divididas em *tokens*, isto é, palavras ou conjuntos de palavras. O objetivo da tokenização é representar o texto como um vetor de palavras ou conjunto de palavras. Entidades nomeadas precisam ser marcadas de maneira adequada, para que os algoritmos de aprendizagem consigam identificar seu início e fim nas sentenças. O IOB[3]

[3] Abreviação em inglês de *Inside*, *Outside* e *Begin*.

é um dentre os diferentes esquemas de marcação disponíveis. Após todos os *tokens* serem marcados, obtém-se um *dataset*, o qual pode ser utilizado para a tarefa de NER. A etapa de treinamento do modelo acontece por meio de um algoritmo baseado em alguma abordagem de aprendizagem de máquina, tal como, aprendizagem tradicional ou *deep learning*, o qual irá aprender regras e padrões através dos dados rotulados. Por fim, é gerado um modelo de NER capaz de identificar e rotular entidades de acordo com as categorias utilizadas nos dados de treinamento.

2.5.1 Abordagens de NER

Os sistemas de NER podem ser divididos basicamente em quatro tipos de abordagens: a) baseada em dicionários, b) baseada em regras, c) baseada em aprendizagem de máquina tradicional e d) baseada em *deep learning*.

A abordagem baseada em dicionários utiliza léxicos construídos a partir de fontes externas de conhecimento, a fim de combinar partes do texto com os elementos (entidades) do dicionário. A combinação entre os elementos representados nos dicionários com as palavras do texto de entrada é feita por meio de algoritmos básicos de correspondência de *strings*, os quais verificam se as entidades estão ocorrendo no texto fornecido. Entretanto, essa abordagem lida com o problema da falta de recursos lexiais completos, dada a variedade de termos existentes em cada língua (ALBARED et al., 2019).

A abordagem baseada em regras utiliza conjuntos pré-definidos de regras criadas manualmente por especialistas linguísticos para identificar tipos específicos de entidades nomeadas. As regras são baseadas em conhecimento sintático, gramatical, ortográfico e de domínio (ANANDIKA; MISHRA, 2019). Nessa abordagem, os algoritmos realizam buscas nos textos, e retornam apenas as entidades que se encaixam nas regras estabelecidas. Existem basicamente dois tipos de regras, baseadas em padrões, que dependem do padrão morfológico das palavras usadas, e regras baseadas em contexto, que dependem do contexto da palavra usada no texto. Apesar de regras serem eficazes, em alguns domínios específicos, elas podem ser difíceis de serem criadas e mantidas, pois precisam ser atualizadas a cada nova mudança do sistema de NER, além de que não generalizam bem entre domínios ou idiomas diferentes (PALSHIKAR, 2012).

Abordagem baseada em aprendizagem de máquina tradicional utiliza modelos estatísticos para identificar padrões e relacionamentos nos textos, os quais possibilitam a detecção e

categorização de entidades. Esses modelos aprendem por meio das características existentes nos dados, e possuem melhor capacidade de generalização, pois não dependem de domínio (GARCIA, 2021). Apesar de serem eficazes, geralmente dependem de grandes quantidades de dados de treinamento, o que pode ser um desafio em termos de custo e tempo (ANANDIKA; MISHRA, 2019).

A abordagem de *deep learning* é composta por várias camadas de processamento, por meio das quais tenta aprender representações dos dados de entrada. O NER aproveita o recurso não linear e as funções de ativação do *deep learning* para descobrir relações e características complexas embutidas nos dados de entrada. Ele extrai automaticamente informações dos dados de entrada e aprende as representações das informações (GUO; WANG; WAN, 2020; BAIGANG; YI, 2022).

De acordo com Jing et al. (2018), alguns dos benéficos da aplicação de técnicas de *deep learning* ao NER são: i) permite o aprendizado de recursos complexos e intrínsecos de dados, por meio de funções de ativação não lineares, ii) economiza tempo e esforço na criação de recursos NER, algo necessário para outras abordagens tradicionais, pois são eficazes no aprendizado automático de representações úteis e fatores subjacentes a partir de dados brutos, iii) podem ser treinados em um paradigma de ponta a ponta, por gradiente descendente, permitindo assim a criação de sistemas NER possivelmente complexos. A desvantagem desse tipo de abordagem é que ela exige muito mais dados do que um algoritmo de aprendizagem tradicional, para conseguir identificar diferentes elementos dentre as camadas da rede neural. Entretanto, mesmo com essa desvantagem, segundo (BAIGANG; YI, 2022), nos últimos anos, os modelos baseados em *deep learning* tornaram-se dominantes na área de PLN, e vem alcançando resultados significativos.

2.5.2 Aplicações de NER

De acordo com (GOYAL; GUPTA; KUMAR, 2018), o reconhecimento de entidades nomeadas atua como base para diferentes aplicações complexas de PLN, tais como, extração de informação, perguntas e respostas, tradução automática de textos, sumarização automática de textos, recuperação de informação, mineração de opinião e anotação semântica.

Extração de informações é uma tarefa que, como o próprio nome diz, extrai informações relevantes de textos conforme solicitação realizada pelo usuário. Segundo Goyal, Gupta e Kumar (2018), a precisão de um sistema de extração de informação depende de nomes próprios,

isto é, entidades nomeadas, pois elas contém informações relevantes sobre o próprio texto. Portanto, empregar o reconhecimento de entidades nomeadas melhora significativamente a precisão da extração de informação.

Os sistemas de perguntas e respostas objetivam gerar resposta para perguntas realizadas por seres humanos em linguagem natural, a partir de documentos textuais (GOYAL; GUPTA; KUMAR, 2018). As respostas de perguntas baseadas em fatos são entidades nomeadas, de modo que a incorporação do sistema de entidades nomeadas melhora a velocidade e a precisão de obter respostas corretas (ARCHANA; MANISH; VISHAL, 2017).

A tradução automática consiste na conversão automática, sem interferências humana, de um texto ou fala de uma língua natural de origem para outro idioma de destino. Diferentes regras de tradução são aplicadas em entidades nomeadas e outras palavras, então a extração de entidades nomeadas torna a tarefa de tradução mais fácil (ARCHANA; MANISH; VISHAL, 2017; JING et al., 2018).

O objetivo da sumarização automática é extrair um resumo condensado e preciso do texto de entrada, de maneira que o texto de saída contenha todas as informações mais relevantes do documento original. Entidades nomeadas são informações importantes do texto, e aumentam o desempenho de identificação de segmentos de texto incluídos em dados resumidos (ARCHANA; MANISH; VISHAL, 2017).

Recuperação de informação lida com o armazenamento de documentos e a recuperação automática de informações relevantes associada a eles. As consultas consistem em uma coleção de *strings* que incluem palavras-chave ou entidades nomeadas. Essas palavras-chave ou entidades são combinadas com o conteúdo da informação armazenada em grandes bancos de dados para tornar o acesso à informação rápido. O NER, no que lhe concerne, desempenha um papel importante nesse tipo de aplicação, pois as pesquisas podem ser feitas a partir da perspectiva de entidades nomeadas (ARCHANA; MANISH; VISHAL, 2017).

A mineração de opinião, também conhecida como análise de sentimentos, coleta e classifica opiniões, emoções, avaliações e atitudes das pessoas em relação a diferentes entidades que podem aparecer em documentos textuais. Essas entidades podem se referir a indivíduos, organizações, produtos, serviços, eventos, entre outros. A tarefa de NER, neste contexto, identifica as entidades que serão classificadas com algum sentimento pelo sistema de mineração de opinião (JING et al., 2018).

A anotação semântica é o processo de anexar a um documento de texto metadados que o descrevam, por meio de referências a conceitos e entidades (ex.: pessoas, lugares, organizações, produtos ou tópicos) relevantes, mencionados no texto. Ao informar para um computador como os itens de dados do documento estão relacionados e como essas relações podem ser avaliadas automaticamente, torna-se possível processar operações complexas de filtro e pesquisa. Essa automação é implementada com técnicas de extração de informações, entre as quais o reconhecimento de entidades nomeadas é usado para identificar conceitos a serem anotados (JING et al., 2018).

2.6 A Cachaça

Cachaça é uma denominação típica da aguardente de cana fabricada no Brasil. Consiste em uma bebida fermento-destilada que apresenta graduação alcoólica de 38% a 48% (v/v), a 20 graus Celsius, obtida por meio da destilação e fermentação do caldo de cana-de-açúcar com características sensoriais peculiares, podendo ser adicionada de açúcares até seis gramas por litro, expressos em sacarose[4] (BRASIL, 2005).

De acordo com Meneghin e Barboza (2014), existem dois aspectos que diferenciam a cachaça da aguardente de cana, são eles, o teor alcoólico e a origem de produção. Enquanto a cachaça apresenta graduação alcoólica entre 38ºGL a 48ºGL, a aguardente pode possuir um teor entre 38ºGL a 54ºGL. A cachaça precisa ser produzida obrigatoriamente no Brasil, já a aguardente não.

2.6.1 Fabricação

A fabricação da cachaça é composta basicamente pelas etapas de colheita, moagem, fermentação, destilação e envelhecimento, as quais são descritas por Borragine (2009), INMETRO (2009), Oliveira (2010), Oliveira (2016), Lacerda (2018), Mapa da Cachaça (2021a), Cachaça Gestor (2022):

a) **colheita ou corte:** a colheita da cana-de-açúcar, matéria-prima da cachaça, pode ser realizada por meio de máquinas ou manualmente. A cana madura, fresca e limpa deve ser espremida no máximo vinte e quatro horas após o corte, pois quanto mais fresca ela for, melhor será o caldo (suco da cana);

[4] Sacarose é um carboidrato dissacarídeo natural encontrado em frutas, vegetais e grãos. É composta por uma molécula de glicose e outra de frutose.

b) **moagem:** depois de cortada, a cana é submetida ao processo de moagem, que consiste na extração do seu caldo por pressão mecânica nos rolos de uma moenda, máquina com cilindros giratórios que espremem o líquido, conforme apresentado na Figura 2.7. O bagaço em alguns casos é utilizado como combustível para aquecer os alambiques;

Figura 2.7 – Máquina de moagem da cana-de-açúcar.

Fonte: Meneghin e Barboza (2014).

c) **filtragem e fecantação**: o caldo obtido após a moagem é filtrado e decantado, isto é, limpo de impurezas como pedaços de bagaço, terra e areia, garantindo assim a separação entre líquido e sólido. Em seguida, o caldo limpo é preparado com a adição de nutrientes e levado às dornas de fermentação;

Figura 2.8 – Equipamento de decantação.

Fonte: Meneghin e Barboza (2014).

d) **fermentação:** nessa etapa, o açúcar é transformado em álcool. Para que isso ocorra, na fabricação artesanal, são acrescentados ao caldo da cana produtos naturais como fubá, farelo de trigo, arroz, soja ou milho, que estimulam a multiplicação de leveduras, isto é, fungos microscópios. Na fabricação industrial, são acrescentados produtos quimicos, como sulfato de amonia e antibioticos. São essas leveduras que transformam o açúcar em álcool. Após ser fermentado, o caldo passa a ser chamado de vinho, o qual é

composto por até 12% de álcool. A fermentação é normalmente feita em recipientes chamados de dorna, os quais são de madeira ou aço inox, conforme a Figura 2.9;

Figura 2.9 – Dorna de fermentação.

Fonte: Meneghin e Barboza (2014).

e) **destilação:** processo de separação de compostos voláteis do vinho, obtido na etapa de fermentação, a fim de purificar a bebida, de forma que ela possa ser consumida. A destilação da cachaça, em particular, consiste em aquecer o vinho de cana até a sua fervura, gerando vapores que ao serem condensados constituirão um destilado com teor alcoólico de cinco a seis vezes maior do que o líquido inicial. Para destilar, utiliza-se basicamente dois tipos de equipamentos, alambique, o mais comum, apresentado na Figura 2.10, ou coluna de aço inox, utilizado na produção de cachaça industrial;

Figura 2.10 – Alambique de destilação.

Fonte: Meneghin e Barboza (2014).

f) **armazenamento e envelhecimento:** após a destilação, a bebida deve ser armazenada em um recipiente de madeira ou outro material inerte que não influência negativamente

no aroma e sabor da cachaça, por um período inferior a 12 meses, conforme exemplo da Figura 2.11. Segundo a Portaria 276 de 2009 do Inmetro (INMETRO, 2009), o armazenamento da bebida deve ser feito em recipientes de madeira apropriada, aço inoxidável ou aço carbono revestido internamente com madeira, de maneira que as perdas por evaporação sejam reduzidas. Envelhecimento é o processo de armazenar a cachaça em recipiente de madeira, com capacidade máxima de setecentos litros, por um período mínimo 1 um ano. Entre as madeiras utilizadas para armazenamento ou envelhecimento pode-se citar: amburana, ipê, bálsamo, jequitibá-branco, jequitibá-rosa, ariribá, canela-sassafrás, tapinhoã, grápia, amendoim e freijó. O tempo de envelhecimento e as propriedades da madeira, desencadeiam alterações na composição química do destilado, refletindo em suas características sensórias de cor, aroma e sabor.

Figura 2.11 – Barril de armazenamento.

Fonte: Meneghin e Barboza (2014).

2.6.2 Tipos de Cachaça

De acordo com a Portaria 276 de 2009 do Inmetro, a cachaça pode ser classificada em diferentes tipos, são eles (INMETRO, 2009; Mapa da Cachaça, 2021b):

a) **cachaça adoçada:** produto que contém açúcares em quantidade superior a 6 gramas por litro e inferior a 30 gramas litro, expressos em sacarose;

b) **cachaça envelhecida:** bebida composta por, no mínimo, 50% de cachaça ou aguardente de cana envelhecidas em recipiente de madeira apropriado, com capacidade máxima de 700 litros, e que envelheceu por um período não inferior a 1 um ano;

c) **cachaça premium:** contém 100% de cachaça ou aguardente de cana envelhecidas em recipiente de madeira apropriado, com capacidade máxima de 700 litros, por um período mínimo de 1 ano. Podem ter cor ou ser incolor, dependendo do tipo de madeira em que

foi armazenada. As madeiras mais comuns para esse tipo de cachaça são: carvalho, amburana e bálsamo;

d) **cachaça extra premium:** contém 100% de cachaça ou aguardente de cana envelhecidas em recipiente de madeira apropriado, com capacidade máxima de 700 litros, por um período mínimo de 3 anos. Pode ser colorida ou incolor, sendo essa última bastante rara no mercado. As madeiras comumente utilizadas para esse tipo de bebida são: carvalho, amburana, jequitibá-rosa, bálsamo, sendo bastante comum a presença de *blends* (em português significa misturas) com diferentes madeiras;

e) **cachaça clássica, tradicional ou prata:** uma cachaça pode ser considerada clássica, tradicional ou prata quando não possui coloração, e foi ou não armazenada em recipiente de madeira. Algumas madeiras brasileiras não alteram a coloração da bebida, tais como, freijó, jequitibá-branco, amendoim, jequitibá-rosa. Algumas cachaças que passam por carvalho são filtradas posteriormente, para continuarem sem coloração;

f) **cachaça ouro:** para ser considerada ouro, pelo menos 50% da bebida deve ter sido armazenada ou envelhecida em recipiente de madeira, além de apresentar alteração substancial na sua coloração, isto é, uma coloração amarelada;

g) **cachaça reserva especial:** para ser considerada reserva especia, a cachaça deve possuir características sensoriais diferentes do padrão usual das outras cachaças elaboradas pelo estabelecimento produtor, e essas diferenças devem ser devidamente comprovadas para os fiscais do Ministério da Agricultura, Pecuária e Abastecimento.

3 TRABALHOS RELACIONADOS

Nesta seção, são apresentados alguns dos principais trabalhos correlatos a esta pesquisa, os quais discorrem sobre a criação e utilização de *datasets* NER para a língua portuguesa.

De acordo com Silva et al. (2021), as aplicações de NER para a língua portuguesa ainda possuem baixa acurácia em comparação aos resultados obtidos para o inglês. Isso se deve ao fato de que os conjuntos de dados rotulados existentes geralmente são pequenos e/ou insuficientes para treinar os modelos de reconhecimento de entidades nomeadas, haja vista que a rotulação é uma tarefa que demanda tempo, esforço e investimento financeiro. Uma alternativa é anotar os dados automaticamente. Mesmo alcançando desempenho um pouco menor que a rotulação manual, as anotações geradas automaticamente são importantes para o treinamento de sistemas NER, especialmente para linguagens com poucos recursos, pois possibilitam uma rápida aquisição de dados rotulados. Considerando essa problemática, alguns trabalhos encontrados na literatura, assim como esta dissertação, propuseram soluções relacionadas a rotulação manual e automática de *datasets* para treinamento e/ou teste de modelos NER. Alguns desses trabalhos são descritos a seguir, e estão organizados de acordo com o tipo de rotulação do *dataset* utilizado, isto é, rotulação manual ou automática.

Em Mota e Santos (2008), foi elaborado o HAREM, um dos primeiros conjuntos de dados rotulados manualmente com entidades nomeadas em português. Esse *dataset* contém dados de diferentes domínios e estilos, tais como textos jornalísticos, de *wikis*, *blogs* e avaliações de produtos. Ele possui 7.834 entidades rotuladas com as categorias Pessoa, Local, Tempo, Organização, Obra, Valor, Abstração, Coisa, Acontecimento e Outro. O HAREM foi testado a partir de diferentes ferramentas NER, as quais foram treinadas durante a competição de avaliação conjunta de entidades nomeadas em português, organizada pelo Linguateca[5]. Tanto este trabalho de dissertação quanto o de Mota e Santos (2008) realizam a análise e rotulação manual de todos os textos coletados. Entretanto, os dados do HAREM são de contextos diversos, enquanto nesta pesquisa são de domínio específico, isto é, cachaça.

Pires (2017) apresenta em sua dissertação diferentes configurações para as ferramentas spaCy, Stanford CoreNLP, OpenNLP e NLTK, a fim de identificar a melhor abordagem para a tarefa de NER em textos compartilhados no SIGARRA, um sistema de busca de notícias da Universidade do Porto, em Portugal. Para comparar as ferramentas, foram realizados testes

[5] Linguateca é um centro de recursos distribuído para o processamento computacional da língua portuguesa.

(baterias de treinamento) com dois conjuntos de dados diferentes, HAREM e SIGARRA *News Corpus*. O segundo *dataset* foi criado manualmente, especificamente para o objetivo da dissertação. Ele é composto por 905 textos em português extraídos do SIGARRA, os quais foram rotulados manualmente com entidades relacionadas a Hora, Evento, Organização, Curso, Pessoa, Localização, Data ou Unidade Orgânica. Segundo o autor, a Stanford CoreNLP obteve o melhor resultado de reconhecimento de entidades nomeadas, tanto para o HAREM (56% de *F1-measure*) quanto para o SIGARRA *News Corpus* (86,86%). Semelhantemente ao trabalho de Pires (2017), nesta pesquisa o *dataset* proposto também foi rotulado manualmente e avaliado por meio do treinamento de um modelo de NER.

Peres, Esteves e Maheshwari (2017) criaram um *dataset* de NER composto por 3.968 *tweets* de assuntos gerais escritos em PtBR, rotulados manualmente com entidades do tipo Pessoa, Local e Organização. Para testar o *dataset*, os autores realizaram experimentos utilizando uma variedade de modelos baseados em redes neurais do tipo *Long Short-Term Memory* (LSTM). O modelo com melhor resultado obteve F1 igual a 52,78. De acordo com os autores, para o contexto em que trabalharam, esse resultado foi maior que o alcançado pelo sistema *baseline* Stanford NER, que obteve F1 de 38,06%. Neste trabalho, assim com no de Peres, Esteves e Maheshwari (2017), a rotulação do *dataset* de NER foi realizada manualmente.

Araujo et al. (2018) criaram um *dataset* composto por 70 documentos para o reconhecimento de entidades nomeadas em textos jurídicos brasileiros, o qual denominaram de LeNER-Br. Esse conjunto de dados foi rotulado manualmente com entidades genéricas (Pessoa, Local, Tempo e Organização) e entidades específicas (Jurisprudência e Legislação). Para avaliar o *dataset* proposto treinaram um modelo de redes neurais LSTM-CRF com os dados do LeNER-Br e com os dados de outro *dataset* já existente chamado Paramopama, dado que ambos possuem as mesmas categorias genéricas. Assim como em Araujo et al. (2018), neste trabalho também foi utilizado o esquema de marcação IOB, os dados foram rotulados manualmente e utilizada a estratégia de modificar os valores dos parâmetros do algoritmo de NER para verificar se isso impactaria no desempenho do modelo.

Albuquerque et al. (2022) desenvolveram o UlyssesNER-Br, um *dataset* composto por 600 documentos legislativos brasileiros[6] para NER. Os documentos contém 10.226 sentenças e 216.182 *tokens*. As sentenças foram rotuladas manualmente com entidades nomeadas que representam organizações, pessoas, produtos jurídicos, localização, fundamentos jurídicos, even-

[6] Documentos referentes a projetos de lei e consultas legislativas da câmara dos deputados do Brasil.

tos e datas. A fim de validar confiabilidade da rotulação manual, os autores calcularam o coeficiente de concordância da Kappa de Cohen, a partir do qual identificaram 88% de concordância entre os rotuladores. Para avaliar o *dataset* UlyssesNER-Br os autores, treinaram o Modelo Oculto de Markov (HMM) e o modelo de aprendizado de máquina *Conditional Random Fields* (CRF). Os resultados mostraram que o CRF tem uma maior chance de sucesso em tarefas de NER, com uma pontuação média de *F1-measure* igual a 80,8%. Assim como em Albuquerque et al. (2022), neste trabalho também utilizou-se a estratégia de avaliar o *dataset* proposto por meio do treinamento e teste de um modelo NER com os dados do próprio cachacaNER.

Nothman et al. (2013) apresentaram o WikiNER, uma abordagem para criar automaticamente *datasets* de treinamento multilíngue para NER explorando a estrutura de *links* e os textos da Wikipedia [7]. Para alcançar tal objetivo, primeiramente os autores classificaram cada documento da Wikipédia em tipos de entidades nomeadas, através do treinamento e avaliação de diferentes artigos da Wikipédia rotulados manualmente em 9 idiomas: inglês, português, alemão, francês, polonês, italiano, espanhol, holandês e russo. Em seguida, os *links* da Wikipédia foram convertidos em rótulos, classificando os artigos de destino em Pessoa, Organização, Localização ou Entidades Diversas. Como resultado, os autores identificaram que para o idioma inglês as anotações realizadas automaticamente superaram o treinamento tradicional em uma coleção anotada manualmente de artigos da Wikipédia em 10–12% de *F-measure*. O modelo da Wikipédia pode ser melhor em alguns domínios do que os padrões de ouro existentes, mas também geralmente aplicável onde os dados de treinamento não estão disponíveis para corresponder a um alvo específico. Assim como na pesquisa de Nothman et al. (2013), neste trabalho também houve um esforço para rotular dados de treinamento automaticamente.

Junior et al. (2015) propuseram o Paramopama, um *dataset* de NER que consiste na extensão da versão PtBR[8] do *dataset* WikiNER. Para isso, primeiramente, realizaram a revisão manual de rótulos atribuídos incorretamente aos dados e em seguida rotularam automaticamente, por meio de um classificador treinado com os dados revisados, 2.500 novas sentenças extraídas de sites de notícias de domínios diversos. O resultado desse processo foi um conjunto de dados composto pelas sentenças revisadas do WikiNER e pelos textos sobre notícias, além das entidades rotuladas com as categorias Pessoa, Localização, Organização e Tempo. Para a tarefa de avaliação, um classificador NER foi treinado separadamente com o Paramopama e com outros dois *datasets*, HAREM e o WikiNER. O Paramopama obteve a melhor pontuação

[7] https://pt.wikipedia.org/

[8] Textos escritos em português do Brasil.

de *F1-measure* (81,58%). Diferentemente de Junior et al. (2015), nesta pesquisa, a ideia de rotular automaticamente o conjunto de treinamento se baseou em um algoritmo de comparação de *tokens*, ao invés de um classificador de NER.

No trabalho de Silva et al. (2021) foi realizada a criação de um *dataset* de treinamento NER padronizado para a língua portuguesa, isto é, um conjunto de dados formado por textos jornalísticos que possuem estruturas morfossintáticas e contextuais iguais. Os textos utilizados foram extraídos de um *corpus* já existente chamado CETENPublico[9], o qual não possui entidades rotuladas, mas sim classes gramaticais. O *dataset* resultante foi rotulado automaticamente, por meio de uma metodologia que utiliza as classes gramaticais para definir os *tokens* como Organização, Pessoa ou Localização. Para avaliar o *dataset* proposto, utilizaram a Bi-LSTM[10], uma variação da rede neural recorrente. Diferentemente do trabalho de Silva et al. (2021), nesta pesquisa não foi proposta uma nova metodologia de rotulação automática de entidades, mas sim a avaliação de uma técnica de rotulação automática de dados baseada em um algoritmo já existente na literatura.

Em Melo e Figueiredo (2021), foi proposta uma abordagem metodológica baseada em modelagem de tópicos, reconhecimento de entidades nomeadas e análise de sentimentos, a fim de identificar em textos jornalísticos (18.413) e *tweets* (1.597.934) escritos em português do Brasil os principais temas em discussão sobre COVID-19, e como os sentimentos sobre esses temas evoluíram ao longo do tempo. Dada a grande quantidade de dados a ser utilizada para o treinamento de um modelo NER, os autores propuseram um algoritmo que recebe como entrada o texto a ser rotulado e uma lista de exemplos composta por *tokens*/entidades rotulados manualmente com categorias de entidade nomeadas específicas. Então, o algoritmo rotula no texto apenas os *tokens* que já existem na lista de entidades rotuladas. Vale enfatizar que o trabalho de Melo e Figueiredo (2021) difere dos demais descritos neste capítulo, pois ele não gera um *dataset* para ser utilizado na tarefa de NER. A rotulação deles serve apenas para identificar os termos mais representativos nos conjuntos de dados avaliados. Nesta pesquisa, utilizou-se o algoritmo proposto por Melo e Figueiredo (2021) para simular a rotulação automaticamente dos dados do cachacaNER.

O diferencial deste trabalho em relação aos demais é o fato dele propor o reconhecimento de entidades nomeadas em textos compartilhados na Web sobre a bebida brasileira cachaça. Isso poderá auxiliar na identificação de características e aspectos específicos a esse tipo de

[9] CETENFolha é um conjunto de dados composto por 190,6 milhões de palavras em português do Brasil.
[10] Bi-LSTM significa *Bi-directional long short term memory.*

bebida, por parte de sistemas computacionais, tais como, análise de sentimentos, extração de informação e recuperação de informação. Além disso, também foi avaliada uma técnica de rotulação automática de entidade nomeadas.

4 CRIAÇÃO E AVALIAÇÃO DE UM *DATASET* PARA NER

Na tarefa de reconhecimento de entidades nomeadas, um *dataset* é uma coleção de dados textuais que podem ser tratados por algoritmos computacionais para alcançar fins analíticos e de previsão. Para que esses dados sejam compreensíveis e gerem informações relevantes, é necessário que sejam previamente tratados, limpos e rotulados com categorias de entidades nomeadas que os representem.

Nesse contexto, este capítulo compreende a descrição da metodologia, ferramentas, tarefas, experimentos e avaliações aplicadas durante o processo de construção de um *dataset* para NER, composto por textos sobre a bebida cachaça. As palavras que compõem os textos desse *dataset* foram rotuladas segundo a categoria de entidade nomeada que representam, sendo elas, específicas ou genéricas. As categorias específicas representam aspectos ou características relacionadas a cachaça, e genéricas se referem as entidades que podem ser encontradas em textos de diferentes domínios.

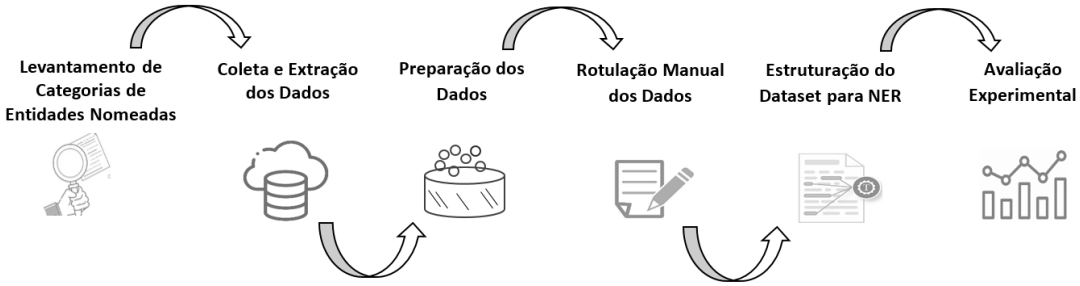
O *dataset* proposto recebeu o nome de "cachacaNER" e foi criado a partir de um conjunto de dados bruto maior denominado de "datasetBASE", o qual também foi criado por meio dos esforços empregados nesta pesquisa. O datasetBASE é composto por dados textuais extraídos de páginas Web sobre cachaça. Cada instância (documento textual) desse *dataset* corresponde a um produto do tipo cachaça, e é composto por atributos que representam informações relacionadas ao produto. Por exemplo, nome da bebida, preço e análise do cachacier, graduação alcoólica, local de produção, entre outros.

4.1 Metodologia

A construção do cachacaNER contou com uma sequência de tarefas que iniciou com a coleta e extração dos dados textuais em páginas Web e finalizou com a sua avaliação para a tarefa de NER. O fluxo da Figura 4.1 apresenta de maneira resumida cada uma dessas tarefas.

A tarefa inicial consistiu na leitura e análise de diferentes textos compartilhados na Web sobre cachaça, a fim de identificar características e aspectos relacionados a esse tipo de bebida. O resultado dessa análise foi a criação de categorias de entidades específicas à bebida cachaça e o levantamento de categorias genéricas.

Figura 4.1 – Etapas para criação e avaliação do *dataset* cachacaNER.



Fonte: Elaboração própria.

Após a construção do conhecimento sobre as categorias de entidades, realizou-se a coleta de páginas Web sobre o produto cachaça. Posteriormente, os textos e demais dados relevantes (ex.: graduação alcoólica e preço) foram extraídos de cada página coletada e utilizados para criar um conjunto de dados inicial, nomeado nesta etapa da pesquisa de "datasetBASE". Esse conjunto é uma base de dados relacional onde os registros representam as bebidas coletadas e os atributos as características relevantes dessas bebidas. A partir desses dados foram extraídas as sentenças utilizadas no conjunto apelidado de "conjuntoRotulacaoManual". E o *dataset* resultante da rotulação manual dessas sentenças recebeu o apelido de cachacaNER.

A preparação dos dados consistiu, basicamente, na remoção de itens não relevantes e ruidosos, preenchimento de dados sem contexto e divisão dos textos em sentenças. Sentença é uma unidade mínima de comunicação que produz efeitos de sentidos em um contexto específico. Uma sentença é formada por uma ou várias palavras, e encerrada pelos seguintes sinais de pontuação: ".", "!" e "?".

Na rotulação manual, três pessoas realizaram separadamente a leitura e rotulação das entidades nomeadas identificadas nas sentenças pré-processadas na etapa anterior. Para isso utilizaram as categorias de entidades nomeadas específicas e genéricas, relacionadas à cachaça.

Na tarefa de estruturação do *dataset* para NER, as sentenças rotuladas manualmente na etapa anterior foram divididas em *tokens*[11] de tamanho um, e seus rótulos foram convertidos para o formato de marcação IOB, criando-se assim, o cachacaNER, *dataset* proposto por esta pesquisa. Esse *dataset* também foi convertido para o formato processado pelo spaCy (HONNIBAL; MONTANI, 2017), para que os dados fossem reconhecidos pelo algoritmo de NER dessa ferramenta.

[11] Um *token* pode ser uma palavra, um fragmento de texto ou um elemento de pontuação (SILVA et al., 2021).

Por fim, foram realizados experimentos relacionados ao treinamento e teste de um modelo NER do spaCy, com o cachacaNER. O objetivo foi verificar a viabilidade dos dados desse *dataset* para a tarefa de reconhecimento de entidades nomeadas.

A seguir, são apresentados mais detalhes sobre a metodologia e o *dataset* criado.

4.2 Levantamento das Categorias de Entidades Nomeadas

O levantamento das categorias de entidades nomeadas teve como principal objetivo a identificação de aspectos que melhor descrevam a bebida cachaça, a partir de textos escritos em português do Brasil. Aspectos são características, atributos ou componentes de uma dada entidade/objeto. Para isso, realizou-se a leitura analítica de vários textos e comentários sobre cachaça, a fim de identificar informações intrínsecas a esse tipo de bebida, as quais pudessem ser transformadas em categorias semânticas de entidades nomeadas. Como resultado, foram criadas as seguintes categorias de entidades específicas, 11 no total:

a) **nome de bebida**: nome comercial do produto, tais como, Pinga Ni Mim, Prazer de Minas e 51;
b) **graduação alcoólica**: volume alcoólico que a bebida possui. Por exemplo, 42% e 27GL;
c) **classificação da bebida**: a cachaça pode ser classificada de diferentes maneiras: Clássica, Tradicional, Prata, Ouro, Premium, Extra Premium, Reserva Especial, Envelhecida e Adoçada. A classificação depende de algumas variáveis, tais como, tempo de envelhecimento, madeira utilizada, características sensoriais identificadas e misturas entre diferentes bebidas ou madeiras;
d) **equipamento de destilação**: tipo de equipamento/aparelho utilizado no processo de destilação da cachaça. De modo geral, existem dois tipos de equipamentos, alambique e coluna, os quais assumem diferentes tipos e tamanhos. Cada um desses equipamentos impacta diferentemente no processo produtivo e na qualidade sensorial percebida na bebida;
e) **tempo de armazenamento**: quantidade de tempo que a bebida fica armazenada ou envelhecendo antes de ser distribuída;
f) **recipiente de armazenamento**: recipiente no qual a bebida é acomodada antes de ser comercializada;
g) **tipo de madeira**: nome da madeira utilizada para a confecção do recipiente onde a bebida é armazenamento, envelhecidas ou fermentada;

h) **característica sensorial cor**: coloração que a bebida possui. Essa coloração pode ser alterada conforme o tipo da madeira ou pelo acréscimo de outras substâncias colorantes. Alguns exemplos de coloração são: amarela, clara, translucida, branca e brilhante;

i) **característica sensorial aroma**: aroma ou cheiro exalado pela bebida. Exemplos: canela, toque de especiarias e floral;

j) **característica sensorial sabor**: sabor sentido na boca quando a bebida é ingerida. Os sabores são divididos basicamente em doce, azedo, ácido, amargo, salgado e adstringente, mas também podem ser descritos de maneiras mais específicas, por exemplo, através de sabores de frutas e plantas;

k) **característica sensorial consistência**: consistência ou textura se refere a sensação percebida na boca em relação à bebida. Alguns dos termos utilizados para descrever essa sensação são: aveludado, macio, viscoso, cremoso, oleoso, licoroso, encorpado e pesado.

Além das categorias específicas, também foram levantadas outras seis categorias genéricas, as quais podem ser encontradas em textos que abordam diferentes assuntos, isto é, que não sejam especificamente sobre bebidas alcoólicas. A seguir, tem-se a descrição das categorias genéricas:

a) **nome de pessoa:** são considerados como pessoas as menções de nomes próprios e apelidos que correspondam a um ser humano, por exemplo, Pedro, Maria S. Silva e Denilson Alves Pereira;

b) **nome de local:** são menções que podem ser traduzidas como um local geográfico, tais como, Lavras, Minas Gerais e Brasil;

c) **nome de organização:** nome de entidade que possui vida própria, tendo uma administração própria e que não é caracterizada como Pessoa. Por exemplo, empresas privadas ou publicas, bancos, hospitais e fundações;

d) **tempo:** são as menções no texto que podem ser traduzidas como a representação do tempo. Exemplos: 10/10/202, agosto de 2021 e 15/02/20 às 10:33hr;

e) **preço:** valores monetários, por exemplo, R$120,00 e 20 reais;

f) **volume**: são medições usadas para verificar qual o volume que pode ser ocupado dentro de um objeto ou de um espaço. Por exemplo, a quantidade de cachaça em mililitros que uma garrafa suporta (ex.: 250ml), ou a quantidade de litros que um barril comporta (ex.: 200 litros).

4.3 Ferramentas Utilizadas

Na etapa de coleta e extração dos dados, foram utilizadas as ferramentas de web scraping Beautiful Soup[12] e Selenium[13]. As páginas recuperadas durante as requisições foram armazenadas em um banco de dados no MongoDB[14]. Para a etapa de rotulação manual dos dados, optou-se pelo Doccano[15]. Para a realização dos experimentos com o *dataset* rotulado, utilizou-se a biblioteca Spacy[16]. A linguagem de programação selecionada para a implementação das tarefas que necessitavam de desenvolvimento foi o Python.

Beautiful Soup é uma biblioteca Python, de código aberto, para extração de dados em páginas Web estáticas, ou seja, o conteúdo baixado pela requisição Web é o mesmo que é exibido online na página Web. Essa ferramenta organiza a estrutura HTML das páginas coletadas em objetos que podem ser inspecionados, pesquisados e modificados. O Selenium, no que lhe concerne, é um conjunto de ferramentas de código aberto multiplataforma, usado para automação de navegadores. Ele simula a navegação humana, o que possibilita a coleta de dados em páginas carregadas dinamicamente. Nesta pesquisa, ele foi usando apenas para realizar a requisição e *download* de algumas páginas dinâmicas, pois a maior parte da extração dos dados foi realizada com o Beautiful Soup.

MongoDB é um sistema gerenciador de banco de dados orientado a documentos, possui código aberto e é multiplataforma. O Doccano é uma ferramenta de código aberto para anotação de textos de maneira intuitiva, por meio do qual é possível criar dados rotulados para as tarefas de reconhecimento de entidades nomeadas, análise de sentimentos, resumo de textos, entre outras.

A spaCy é uma biblioteca gratuita de processamento de linguagem natural escrita em Python e de código aberto. Oferece vários recursos, tais como, reconhecimento de entidades nomeadas, análise de dependência, marcação de parte da fala (POS) e lematização, dentre outros. Suporta 64 línguas, dentre as quais se encontra o português. Possui 18 categorias de entidades nomeadas previamente treinadas, além de permitir a inserção de novas categorias no módulo NER, e a atualização do modelo com novos exemplos de dados rotulados. Para a maioria das tarefas, implementa modelos estatísticos baseados em redes neurais profundas convolucionais.

[12] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[13] https://www.selenium.dev/
[14] https://www.mongodb.com/
[15] https://github.com/doccano/doccano
[16] https://spacy.io

Especificamente para o reconhecimento de entidades nomeadas, usa uma abordagem baseada em transição apresentando por Lample et al. (2016).

4.4 Coleta e Extração dos Dados

A coleta e extração de dados textuais em páginas Web pode ser realizada por meio de *scripts* ou programas baseados em uma técnica conhecida como Web Scraping. Essa técnica automatiza a coleta de páginas, bem como a extração de seu conteúdo. Em vez de realizar manualmente o processo de entrar em várias páginas, verificar o conteúdo, copiar e salvar os dados, com web scraping é possível simplificar esse processo ao criar um *script* que itere automaticamente sobre as várias páginas e extraia os dados relevantes do HTML. Para isso utilizou-se as ferramentas Beautiful Soup e Selenium.

Antes de iniciar a coleta e extração propriamente dita, foi necessário primeiramente realizar o levantamento dos sites que possuíam textos sobre cachaça. Após uma minuciosa busca e análise manual de várias páginas Web, foram selecionados 24 sites de venda de produtos do tipo bebida alcoólica, os quais possuíam diversos tipos de cachaça.

Em seguida, foram criados e executados diferentes *scripts* baseados na técnica de web scraping para: (i) coletar dos 24 sites somente as páginas Web referentes aos produtos do tipo cachaça, e (ii) extrair dessas páginas os dados textuais relevantes ao contexto da cachaça, a fim de criar uma base de dados relacional composta por documentos textuais representado os diferentes tipos de cachaça coletados. No total, foram coletadas 3.381 páginas Web no formato HTML, das quais foram extraídos os seguintes tipos de dados:

a) nome comercial da bebida. Exemplos: Cachaça Três Corações 700ml, Cachaça Theodoro Jequitibá, Wiba Amburana;
b) graduação alcoólica presente na bebida. Exemplos: Grad. Alcool. 35%, 40 porcento de graduação alcoólica, 28^{o}GL;
c) tipo de classificação da bebida. Exemplos: Clássica, Tradicional ou Ouro;
d) equipamento onde a bebida é produzida. Por exemplo: alambique de cobre e coluna de aço;
e) tempo de armazenamento da bebida. Exemplos: 1 ano, oito meses, 5 anos;
f) recipiente onde a bebida é armazenada ou envelhecida. Exemplos: barril de carvalho, dorna de aço inoxidável;

g) análise sensorial. Exemplos: Cor dourada com reflexos verdes, intensidade média no nariz, aromas florais e frutados muito agradáveis com notas de anis, pimenta e frutas vermelhas, persistência elegante da madeira;
h) local de produção da bebida. Exemplo: Minas Gerais - MG;
i) Preço do produto. Exemplos: R$120,00, Setenta reais, R$20;
j) volume do recipiente onde a bebida é engarrafada. Exemplos: 700ml, volume 700, vol.500ml;
k) texto contendo informações descritivas sobre a bebida. Exemplo: "A Cachaça Wiba Amburana é armazenada por cerca de 1 ano em tonéis de jequitibá e em sequência mais 6 meses em barris de carvalho. Cristalina, límpida e de brilho intenso, tem uma tonalidade amarelo bem claro com nuances esverdeadas";
l) texto com a análise do cachacier. Exemplo: "Cachaça levemente amadeirada, mas com o sabor e predominância do carvalho francês, podendo ser degustada pura ou em *drinks* mais elaborados. Equilíbrio perfeito entre o amadeirado e o frutado.";
m) história da empresa ou da bebida. Exemplo: "A Cachaça Costa Rica é produzida em Guarani, Minas Gerais. O alambique foi projetado e concebido para produzir de forma sustentável, preservando as tradições seculares de fabricação da cachaça mineira de qualidade. Mesmo assim, não abre mão do que existe de mais moderno e eficiente em boas práticas de fabricação e segurança dos alimentos, resultando em um produto sem igual e sem comparação. No Alambique Guarani, o que faz a diferença na bebida é o carinho com que é produzida.";
n) nome do produtor. Exemplos: Glauri Ind Com. Ltda, Princesa Isabel, RXM Agropecuária;
o) premiações que a bebida recebeu. Exemplos: Medalha de Prata na Cacharitiba – 2019; Ouro - 2011 Concours Mondial de Bruxelas;
p) harmonização da bebida com alguns tipos de comidas. Exemplo: Harmoniza com carnes e chocolates, além de ser um excelente digestivo;
q) ingredientes que compõem a bebida. Exemplos: cana-de-açúcar, corantes, essências;
r) modo como a bebida foi produzida. Exemplos: artesanal ou orgânica;
s) informações técnicas. Exemplos: tamanho da garrafa e peso médio do produto.

O resultado da etapa de coleta e extração dos dados foi a criação do datasetBASE, o qual é composto por documentos que representam diferentes tipos de cachaças, por meio de textos

com informações relevantes sobre os produtos. Na Tabela 4.1, são apresentas as quantidades de páginas Web coletadas em cada site de venda de bebidas, onde os sites estão representados por números, mas podem ser identificados por nome na Tabela 4.2. O Quadro 4.1 mostra uma parte de como ficou a organização do *dataset* em questão. Destaca-se que nem todos os documentos possuem valores para todos os atributos.

Tabela 4.1 – Total de páginas coletadas por site.

Sites	**Total de Páginas**	**Sites**	**Total de Páginas**	**Sites**	**Total de Páginas**
1	576	2	111	3	61
4	31	5	13	6	164
7	90	8	299	9	1.080
10	4	11	183	12	139
13	98	14	15	15	10
16	5	17	69	18	6
19	227	20	170	21	4
22	8	23	12	24	6

Fonte: Elaboração própria.

Quadro 4.1 – Exemplo de alguns documentos e atributos do datasetBASE com dados brutos.

*	NOME_DA_CACHAÇA	PREÇO	ANÁLISE_DO_CACHACIER	...
0	Cachaça Costa Rica Carvalho 670ml	R$ 55,00	Cachaça levemente amadeirada, mas com o sabor e predominância do carvalho francês, podendo ser degustada pura ou em drinks mais elaborados.	...
1	Cachaça Famosinha de Minas 600ml	R$ 29,90	De cor amarelo-palha, possui uma mescla de aromas frutado e de cana-de-açúcar, com notasde amêndoas. No paladar, traz um gosto predominantemente doce.	...
...	...	...	...	...
3381	Cachaça Weber Haus Rota 48 Pura	R$ 70,00	Transparente, possui aroma de cana-de-açúcar. No paladar, traz um doce. Além disso, causa uma sensação aveludada e licorosa.	...

Fonte: Elaboração própria.

Vale ressaltar que a coleta e extração dos dados foram coordenados pela autora desta dissertação e executadas por dois alunos de iniciação científica ligados ao projeto. Portanto, a coleta das páginas e extração dos textos das estruturas HTML não são contribuições desta dissertação. Mas toda a estratégia relacionada ao que coletar, onde e como são contribuições, pois foram planejadas e lideradas pela autora deste trabalho.

4.5 Preparação dos Dados

Esta etapa consistiu basicamente em: (i) limpar os dados (ii), completar os dados textuais que possuíam pouca ou nenhuma contextualização, (iii) separar os textos em sentenças e (iv) criar um novo conjunto de dados textuais a partir do dataseBASE, para ser utilizado na etapa de rotulação manual.

Na limpeza, foram aplicadas as seguintes correções ao datasetBASE: remoção de atributos não relevantes, sendo eles, link da página coletada, data da coleta e ids gerados automaticamente pela biblioteca pandas durante a criação dos arquivos, remoção de espaçamento e quebra de linhas replicadas, substituição de caracteres unicodes gerados pela ferramenta Beautiful Soup ou que já faziam parte das páginas HTML.

Alguns dados coletados possuíam pouca ou nenhuma informação textual para lhes contextualizar, por exemplo, no atributo graduação alcoólica alguns dados consistiam apenas de um número seguido ou não do símbolo de percentual. No atributo local de produção, entre a maioria dos dados havia apenas o nome do local onde a bebida foi produzida. No atributo volume, normalmente aparecia apenas números seguidos da sigla de ml, da palavra volume ou de sua contração (vol.). Para resolver esse problema de dados sem contexto, optou-se por inserir automaticamente textos que descrevessem o que os dados representavam, por exemplo, nos dados relacionados a graduação alcoólica foi inserido o texto “graduação alcoólica”. Antes de inserir os textos, primeiramente, foram definidos manualmente quais atributos possuíam dados sem contexto e em seguida os textos específicos foram inseridos automaticamente para todos os dados de cada atributo escolhido. Esses ajustes foram realizados em uma pequena quantidade de dados em comparação a base de dados total, de maneira que não houvesse impacto na estrutura original do mesmo. O Quadro 4.2 mostra alguns dados antes e após a aplicação desse ajuste.

Quadro 4.2 – Exemplo de dados sem contextualização preenchidos.

Volume Antes	Volume Depois	Graduação Alcoólica Antes	Graduação Alcoólica Depois
670ml	VOLUME: 670ml	40%	GRADUAÇÃO ALCOÓLICA: 40%
600ml	VOLUME: 600ml	42%	GRADUAÇÃO ALCOÓLICA: 42%
1LT	VOLUME: 1LT	48GL	GRADUAÇÃO ALCOÓLICA: 48GL
500ml	VOLUME: 500ml	40	GRADUAÇÃO ALCOÓLICA: 40

Fonte: Elaboração própria.

Após a limpeza e ajuste, os dados textuais do datasetBase que estavam separados por atributo foram divididos em sentenças. A divisão por sentença é a forma tradicional de organização de *datasets* para NER. Para isso, os textos foram quebrados por ponto final (.), interrogação (?) ou exclamação(!), criando-se assim um novo *dataset* composto por sentenças. O Quadro 4.3 traz o exemplo de um texto antes e após a quebra por sentenças.

Quadro 4.3 – Exemplo de texto dividido em sentenças.

Texto sem divisão	Texto divido por sentenças
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.	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.

Fonte: Elaboração própria.

Por fim, foram escolhidos estrategicamente 1.000 documentos do datasetBASE para serem rotulados manualmente e compor o *dataset* de NER proposto neste trabalho. Selecionou-se essa quantidade de dados específica, ao invés de todos os dados coletados, porque o custo de rotulação manual é muito alto. Para garantir diversidade entre os textos que comporiam esse conjunto de sentenças, apelidado neste trabalho de conjuntoRotulacaoManual, optou-se por selecionar porções equilibradas de dados referentes a cada site de venda de bebidas, as quais podem ser observadas na Tabela 4.2. Vale enfatizar que, foram escolhidas todas as páginas dos sites com menos de 69 páginas, conforme números apresentados na Tabela 4.1. O conjuntoRotulacaoManual não é um novo *dataset*, mas sim um arquivo .txt composto por 13.628 sentenças provenientes dos documentos textuais escolhidos para compor o cachacaNER.

4.6 Rotulação Manual dos Dados

Nessa etapa, as sentenças do conjuntoRotulacaoManual foram analisadas e rotuladas manualmente, por três pessoas, com as categorias de entidades nomeadas levantadas previamente, específicas e genéricas. Para isso, utilizou-se a ferramenta Doccano, que permitiu a inserção manual de rótulos nos textos, conforme exemplo real da Figura 4.2.

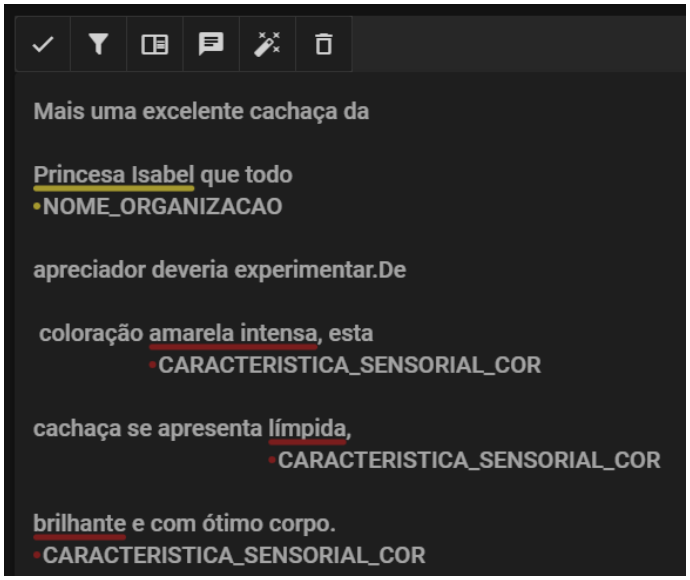
O processo de rotulação manual ocorreu basicamente da seguinte maneira:

Tabela 4.2 – Distribuição da quantidade de páginas por site.

Número	Site	Quantidade de Documentos
1	Amburana	69
2	Ararauna	68
3	Bebida Online	61
4	Blubeer	31
5	Cachaça Companheira	13
6	Cachaça e Pinga	69
7	Cachaça e Presente	68
8	Cachaçaria dos Amigos	69
9	Cachaçaria Nacional	69
10	Cachaça Sagatiba	4
11	Cachaças Brasileiras	69
12	Canela de Ema	69
13	Casa da Bebida	68
14	Ceia Clandestina	15
15	Cia Muller	10
16	Dom Tapparo	5
17	Ethylica	69
18	Cachaça Magnifica	6
19	Moça Bonita	69
20	Salinas	69
21	Sanhaçu	4
22	Sapucaia	8
23	Velho Barreiro	12
24	Cachaça Wiba	6

Fonte: Elaboração própria..

Figura 4.2 – Exemplo de sentença rotulada manualmente com o Doccano.



Fonte: Elaboração própria.

a) primeiramente, cada um dos três rotuladores analisou e rotulou individualmente todas as sentenças do conjuntoRotulacaoManual, por meio do Doccano. No final dessa primeira rotulação foram gerados separadamente três arquivos no formato jsonl, cada qual contendo as sentença e rotulações realizadas. A Figura 4.3 mostra exemplos de como são os dados gerados pelo Doccano, o quais são compostos por: a) um id, para enumerar cada item passado como entrada ao Doccano, isto é, as sentenças, b) o texto analisado e c) uma lista contendo a posição inicial e final da entidade rotulada e o rótulo atribuído à entidade;

b) em seguida, os resultados obtidos foram comparados automaticamente para identificar onde houve divergências entre os rotuladores e definir qual dentre as rotulações seria utilizada como a correta;

c) por fim, os três rotuladores conjuntamente reanalisaram as 403 divergências e definiram a qual categoria de entidade cada uma delas pertenceria. Vale ressaltar que, nos casos onde dois participantes rotularam as entidades igualmente e apenas um rotulou de maneira diferente, considerou-se como rotulação final aquela realizada pelos dois rotuladores. Nos casos onde os três rotularam igualmente, considerou-se a marcação como rotulação final. Estes dois últimos casos não entraram no processo de reanálise;

Figura 4.3 – Exemplo de dados gerados pelo Doccano.

Fonte: Elaboração própria.

4.6.1 Diretrizes para Rotulação Manual

Antes de iniciar a rotulação manual das sentenças, foi elaborado um documento de orientação para a rotulação. Em seguida foi realizada uma amostragem dos dados, 30 produtos no total, onde cada rotulador fez sua rotulação seguindo as orientações do documento. Posteriormente, realizou-se uma discussão sobre dúvidas e problemas encontrados. A partir disso, o documento de orientação foi ajustado, para que a rotulação completa do conjuntoRotulacaoManual fosse realizada. Esse documento é apresentado no Apêndice A.

Para orientar e auxiliar o processo de rotulação das características sensoriais, principalmente aroma e sabor, que dependem fortemente da percepção de quem avalia a bebida, optou-se por utilizar a Roda Sensorial da Cachaça proposta por Bortoletto (2016) e o levantamento de atributos sensoriais elaborado no trabalho de Pinheiro (2010), apresentado no Anexo A. Na Figura 4.4 é mostrada a roda sensorial da cachaça, a qual é composta por diferentes termos utilizados para descrever as características sensoriais percebidas ao se degustar uma cachaça.

Figura 4.4 – Roda sensorial da cachaça.

Fonte: Bortoletto (2016).

4.6.2 Concordância entre os Rotuladores

Para verificar e validar a consistência da rotulação manual realizada pelos rotuladores, utilizou-se o coeficiente de concordância Kappa de Fleiss (FLEISS, 1971). Essa métrica mede a concordância entre três ou mais avaliadores[17] ao atribuírem classificações categóricas a um mesmo conjunto de dados. Ela expressa o grau em que a proporção observada de concordância entre os avaliadores excede o que seria esperado se todos fizessem suas classificações de forma totalmente aleatória.

O coeficiente Kappa basicamente calcula uma medida de quão consistentes são as rotulações feitas, pois remove a concordância esperada devido ao acaso. Pode ser expresso da

[17] Nesta pesquisa, avaliador é o mesmo que rotulador, ou seja, quem atribuiu as categorias de entidades nomeadas as sentenças textuais.

seguinte maneira:

$$kappa(k) = \frac{\overline{P}_o - \overline{P}_e}{1 - \overline{P}_e} \tag{4.1}$$

$\overline{P}_o$ é o acordo observado, e $\overline{P}_e$ o acordo esperado. O fator $1 - \overline{P}_e$ equivale ao grau de concordância que é atingível acima do acaso, e $\overline{P}_o - \overline{P}_e$ o grau de concordância realmente alcançado acima do acaso. Se os avaliadores estiverem de acordo, então tem-se $k = 1$. Se não houver concordância entre eles, então $k \leq 1$. Quando k é igual a 0, então a concordância não é melhor do que a que seria obtida por acaso. Para k negativo, a concordância é menor do que a esperada ao acaso. Para k positivo, a concordância excede a concordância ao acaso.

Para auxiliar na interpretação dos valores de concordância do kappa, Landis e Koch (1977) propuseram uma tabela que contém intervalos de valores e, para cada intervalo, uma interpretação do que o valor significa. Por meio dessa tabela, apresentada pela Tabela 4.3, foi possível verificar a qualidade das rotulações de entidades nomeadas realizadas nesta pesquisa.

Tabela 4.3 – Interpretação dos valores da concordância Kappa.

Valor Kappa	**Interpretação**
Menor que zero	Concordância Ruim
Entre 0 e 0,20	Concordância Fraca
Entre 0,21 e 0,40	Concordância Razoável
Entre 0,41 e 0,60	Concordância Moderada
Entre 0,61 e 0,80	Concordância Forte
Entre 0,81 e 1,00	Concordância Quase perfeita

Fonte: Adaptado de Landis e Koch (1977).

Nesta pesquisa, calculou-se a estatística Kappa de Fleiss para todo o *dataset* rotulado manualmente (kappa total), e também para cada categoria de entidade nomeada individualmente. Para obtenção dessas estatísticas, foi necessário:

a) tokenizar todas as sentenças, sem perder a referência entre os *tokens* (palavras que compõem as entidades) e suas respectivas categorias de entidades nomeadas, atribuídas durante a etapa de rotulação manual. A Figura 4.5 traz um exemplo real de como os dados rotulados por cada analisador foram tokenizados, o rótulo ‘o’ significa que o *token* não foi rotulado, por não pertencer a nenhuma categoria;

b) criar um único *dataset*, composto pelos *tokens* e suas respectivas rotulações feitas por cada um dos analisadores, conforme exemplo da Figura 4.6. O proposito dessa etapa foi ter uma estrutura que possibilitasse a identificação do que cada analista rotulou para cada *token* do conjuntoRotulacaoManual;

c) criar uma matriz $M[i, j]$ para ser utilizada no cálculo Kappa. Essa matriz guarda a quantidade de avaliadores que atribuíram o 'i-ésimo' *token* à 'j-ésima' categoria de entidade nomeada. As colunas representam as categorias e as linhas os *tokens*. Assim, tem-se uma matriz que contém o somatório de quantos analistas rotularam cada *token* para cada categoria. A Figura 4.7 ilustra a matriz em questão, as categorias estão representadas por números de 1 a 17 e os itens destacados compreendem o somatório de rotuladores por categoria;

d) por fim, foi realizado por meio da ferramenta IBM SPSS*Statistics*[18], o cálculo da estatística Kappa de Fleiss total e por categoria. A entrada dessa ferramenta foi o *dataset* que representa a matriz de **tokens** *versus* categorias. Essa ferramenta oferece diferentes tipos de análises estatísticas avançadas, bibliotecas de algoritmos de *machine learning* e análise de texto.

Figura 4.5 – Exemplo de como as sentenças rotuladas manualmente foram tokenizadas.

ROTULADOR 1	ROTULADOR 2	ROTULADOR 3
['Cachaça', 'o'],	['Cachaça', 'o'],	['Cachaça', 'o'],
['Colombina', 'NOME_BEBIDA'],	['Colombina', 'NOME_BEBIDA'],	['Colombina', 'NOME_LOCAL'],
['Chita', 'NOME_BEBIDA'],	['Chita', 'NOME_BEBIDA'],	['Chita', 'NOME_LOCAL'],
['PREÇO', 'o'],	['PREÇO', 'o'],	['PREÇO', 'o'],
[':', 'o'],	[':', 'o'],	[':', 'o'],
['R$', 'PRECO'],	['R$', 'PRECO'],	['R$', 'PRECO'],
['100', 'PRECO'],	['100', 'PRECO'],	['100', 'PRECO'],
[',', 'PRECO'],	[',', 'PRECO'],	[',', 'PRECO'],
['00', 'PRECO'],	['00', 'PRECO'],	['00', 'PRECO'],
['DESCRIÇÃO', 'o'],	['DESCRIÇÃO', 'o'],	['DESCRIÇÃO', 'o'],
['DA', 'o'],	['DA', 'o'],	['DA', 'o'],
['CACHAÇA', 'o'],	['CACHAÇA', 'o'],	['CACHAÇA', 'o'],
[':', 'o'],	[':', 'o'],	[':', 'o'],
['Colombina', 'NOME_BEBIDA'],	['Colombina', 'NOME_BEBIDA'],	['Colombina', 'NOME_BEBIDA'],
['Chita', 'NOME_BEBIDA'],	['Chita', 'NOME_BEBIDA'],	['Chita', 'NOME_BEBIDA'],
['700ml', 'VOLUME'],	['700ml', 'VOLUME'],	['700ml', 'VOLUME'],
['Volume', 'o'],	['Volume', 'o'],	['Volume', 'o'],
[':', 'o'],	[':', 'o'],	[':', 'o'] 'VOLUME'],
['700ml', 'VOLUME'],	['700ml', 'VOLUME'],	['700ml', 'VOLUME'],

Fonte: Elaboração própria.

Figura 4.6 – Exemplo do *dataset* composto pelos *tokens* rotulados por cada analista.

	Tokens	Rotulador 1	Rotulador 2	Rotulador 3
0	Três	NOME_BEBIDA	NOME_BEBIDA	NOME_BEBIDA
1	Coronéis	NOME_BEBIDA	NOME_BEBIDA	NOME_BEBIDA
2	R$	PRECO	PRECO	PRECO
3	35	PRECO	PRECO	PRECO
4	,	PRECO	PRECO	PRECO
5	00	PRECO	PRECO	PRECO
6	Barril	RECIPIENTE_ARMAZENAMENTO	TIPO_MADEIRA	RECIPIENTE_ARMAZENAMENTO
7	Amburana	TIPO_MADEIRA	TIPO_MADEIRA	RECIPIENTE_ARMAZENAMENTO
8	Floral	CARACTERISTICA_SENSORIAL_AROMA	CARACTERISTICA_SENSORIAL_SABOR	CARACTERISTICA_SENSORIAL_SABOR
9	6	TEMPO_ARMAZENAMENTO	TEMPO_ARMAZENAMENTO	TEMPO_ARMAZENAMENTO
10	anos	TEMPO_ARMAZENAMENTO	TEMPO_ARMAZENAMENTO	TEMPO_ARMAZENAMENTO

Fonte: Elaboração própria.

[18] https://www.ibm.com/br-pt/analytics/spss-statistics-software

Figura 4.7 – Matriz de *token* por categoria de entidade nomeada.

Token	Categoria																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Três	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Coronéis	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	R$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0
	35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0
	,	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0
	00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0
	Barril	0.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Amburana	0.0	0.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	floral	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	6	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	anos	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fonte: Elaboração própria.

Os níveis de concordância Kappa identificados para as rotulações realizadas neste trabalho são apresentados na Tabela 4.5, onde é possível observar que a concordância geral foi quase perfeita (0,857). Dentre as categorias individuais, mesmo o menor resultado (0,723) pode ser interpretado como uma forte concordância. Dada a satisfatoriedade dos resultados obtidos, foi possível constatar a viabilidade de dar continuidade a esta pesquisa, através das seguintes tarefas: criação de um *dataset* no formato IOB contendo todos os *tokens* rotulados, e avaliação desse *dataset* por meio do treinamento e teste de um modelo NER.

A Tabela 4.4 apresenta quantas entidades cada rotulador, individualmente, rotulou diferente dos demais rotulados. Também mostra quantas entidades os três analistas rotularam diferente, 403 no total.

Tabela 4.4 – Total de divergências entre os rotuladores.

Rotuladores	**Total de Divergências**
Rotulador 1 *versus* Rotulador 2 e Rotulador 3	3.706
Rotulador 2 *versus* Rotulador 1 e Rotulador 3	3.020
Rotulador 3 *versus* Rotulador 1 e Rotulador 2	1.846
Rotulador 1 *versus* Rotulador 2 *versus* Rotulador 3	**403**

Fonte: Elaboração própria.

Tabela 4.5 – Resultados do coeficiente de concordância da Kappa de Fleiss.

Categorias	**Kappa**
NOME_BEBIDA	0,899
GRADUACAO_ALCOOLICA	0,962
CLASSIFICACAO_BEBIDA	0,878
EQUIPAMENTO_DESTILACAO	0,911
TEMPO_ARMAZENAMENTO	0,955
RECIPIENTE_ARMAZENAMENTO	0,905
TIPO_MADEIRA	0,964
CARACTERISTICA_SENSORIAL_COR	0,924
CARACTERISTICA_SENSORIAL_AROMA	0,804
CARACTERISTICA_SENSORIAL_SABOR	0,723
CARACTERISTICA_SENSORIAL_CONSISTENCIA	0,803
NOME_PESSOA	0,899
NOME_LOCAL	0,897
NOME_ORGANIZACAO	0,793
TEMPO	0,901
PRECO	0,967
VOLUME	0,964
Resultado da Kappa Geral	**0,857**

Fonte: Elaboração própria.

4.7 Estruturação do *Dataset* para NER

Datasets são conjuntos de dados que podem ser utilizados para o treinamento e teste de algoritmos de aprendizagem de máquina, os quais buscam prever automaticamente diferentes tipos de informações, tais como, categorias de entidades nomeadas. Os dados contidos em um *dataset* também podem servir para a extração de informações estatísticas, tais como, a quantidade de *tokens* rotulados por categoria e a média de *tokens* por sentença. São estruturas comumente organizadas no formato de tabela, as quais possuem linhas e colunas preenchidas com informações específicas acerca de sua finalidade. O formato pode variar entre txt, json, csv e xml.

Para que um *dataset* possa ser utilizado na tarefa de reconhecimento de entidades nomeadas, ele precisa primeiramente ser convertido para um formato adequado a ferramenta ou algoritmo escolhido. Considerando essa particularidade, realizou-se a conversão do cachaca-NER para o formato IOB e também para o formato aceito pelo spaCy.

4.7.1 *Dataset* no Formato IOB

O IOB é um formato de marcação proposto Ramshaw e Marcus (1995) para marcar *tokens* em tarefas de agrupamento linguístico computacional, tal como o reconhecimento de entidades nomeadas. As marcações podem denotar o interior, o exterior e o início de um pedaço[19]. O rótulo **O** indica que o *token* não pertence a nenhum pedaço, isto é, não faz parte de uma entidade. O prefixo **B-** antes do nome do rótulo indica que o *token* é o início de um pedaço que segue imediatamente outro pedaço sem rotulações **O** entre eles. O prefixo **I-** antes do nome do rótulo indica que o *token* está dentro de um pedaço, ou seja, faz parte da entidade.

Nesta pesquisa optou-se pelo formato IOB2 ao invés do IOB, pois é o mais amplamente utilizado na literatura e pelos códigos compartilhados na Web. Ambos são semelhantes, exceto pelo fato de que no IOB2 o rotulo B é usado no início de cada pedaço, ou seja, todas as entidades começam com B, enquanto no IOB a entidade inicia com I, e o rótulo B só é usado para separar duas entidades adjacentes do mesmo tipo. O Quadro 4.4 apresenta o exemplo de uma sentença *tokenizada* e marcada com o formato IOB2.

Quadro 4.4 – Exemplo de uma sentença do cachacaNER no formato IOB2.

Sentença	Tokens	Rotulação
A Cachaça Serra Limpa 355ml é uma bebida armazenada por seis meses em tonéis de inox.	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_ARAZENAMENTO
	em	O
	tonéis	B-RECIPIENTE_ARMAZENAMENTO
	de	I-RECIPIENTE_ARMAZENAMENTO
	inox	I-RECIPIENTE_ARMAZENAMENTO
	.	O

Fonte: Elaboração própria.

Além do *token* (Palavra) e rótulo, acrescentou-se também ao *dataset* os seguintes atributos: (i) “Sentença”, guarda a referência de qual sentença cada *token* pertence, (ii) “Início”,

[19] Pedaço se refere a um composto de *tokens*, que no caso desta pesquisa são as palavras que representam as entidades nomeadas.

indica a posição inicial do *token* em sua respectiva sentença, (iii), “Fim”, indica a posição final mais 1 do *token* em relação à sua sentença e (iv) “Documento”, contém números de 1 até 1000 para referenciar o documento ao qual o *token* faz parte. Esse *dataset* recebeu o nome de cachacaNER, o qual é mostrado na Figura 4.8.

Figura 4.8 – *Dataset* real no formato IOB.

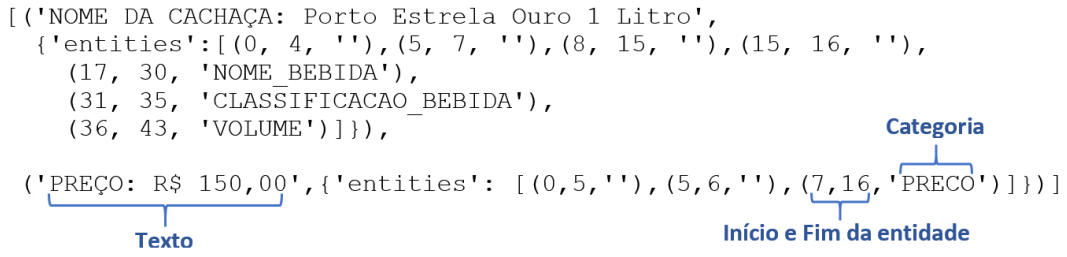
	Palavras	Rotulo	Sentenca	Inicio	Fim	Documento
0	NOME	O	1	0	4	1
1	DA	O	1	5	7	1
2	CACHAÇA	O	1	8	15	1
3	:	O	1	15	16	1
4	Cachaça	O	1	17	24	1
...	...	...	...	...	...	...
183014	Destilados	O	13628	54	64	1000
183015	do	O	13628	65	67	1000
183016	Brasil	B-NOME_LOCAL	13628	68	74	1000
183017	2019	B-TEMPO	13628	75	79	1000
183018	.	O	13628	79	80	1000

Fonte: Elaboração própria.

4.7.2 *Dataset* no formato aceito pelo spaCy

Para treinar um modelo de NER no spaCy, são necessários três dados: (i) uma *string* de texto, (ii) índices de início e fim de cada entidade no texto e (iii) a categoria que cada entidade nomeada representa. Esses dados devem ser passados para o algoritmo de treinamento do modelo como uma lista, a qual é composta por tuplas. Cada tupla contém o texto analisado e um dicionário. O dicionário é composto por uma lista, a qual guarda tuplas com a posição inicial e final das entidades identificadas no texto e também a categoria que essa entidade representa. As palavras que não representam entidades são rotuladas com aspas simples **‘’**, representando assim um intervalo de início e fim sem categoria. A Figura 4.9 traz alguns exemplos reais de como os dados do cachacaNER ficaram após a transformação para o formato do spaCy. O *dataset* nesse formato encontra-se disponível publicamente em <https://github.com/PriscillaIA/cachacaNER/blob/main/cachacaNER_Formato_spaCy.bin>.

Figura 4.9 – Exemplo de dados no formato processado pelo spaCy.



Fonte: Elaboração própria.

4.8 Divisão do *Dataset* em Treino e Teste

Para dividir proporcionalmente as entidades por tipo de categoria, efetuou-se as seguintes tarefas: particionamento dos documentos em conjuntos de dados de mesmo tamanho, com seleção aleatória dos documentos para compor cada partição, e divisão dos dados em conjuntos de treino e teste.

Os documentos foram divididos entre 10 partições de tamanho 100, cada uma. Para selecionar quais documentos comporiam as partições, realizou-se uma escolha aleatória excludente, isto é, cada documento foi alocado em apenas uma das dez partições. Para construir cada partição, foram selecionados 10% dos documentos coletados em cada site de venda de bebidas, conforme apresentado na Tabela 4.6.

Após o particionamento, os dados foram concatenados novamente, de maneira que às sete primeiras partições foram definidas como o conjunto de dados de treinamento e às três últimas, os dados de teste. Dessa maneira, foi verificado que aproximadamente 70% das entidades de cada categoria foram alocadas no *dataset* de treinamento, e 30% no de teste, conforme apresentando na Tabela 4.7. Além disso, o particionamento também permitiu que os dados de treinamento e teste possuíssem sentenças diversas, isto é, de diferentes sites de venda de bebidas.

Para facilitar a identificação dos dados de treinamento, de teste e as partições dentro do cachacaNER, acrescentou-se os atributos “Particao” e “Identificacao_Treino_Teste” ao *dataset*, conforme mostrado na Figura 4.10.

Nas Figuras 4.11, 4.12, 4.13, 4.14 e 4.15 são apresentados gráficos com as quantidades de entidades por categoria em cada partição do *dataset*. Por meio desses gráficos, é possível observar que as partições possuem quantidades semelhantes de entidades rotuladas com cada tipo de categoria.

Tabela 4.6 – Distribuição dos documentos em cada partição.

Sites	**Partições**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Amburana	7	7	7	6	6	7	8	7	7	7
Ararauna	6	7	6	7	7	7	7	7	7	7
Bebida Online	6	6	6	6	6	6	6	6	6	7
Blubeer	3	3	3	3	3	3	3	3	3	4
Cachaça Companheira	1	1	1	1	1	1	1	2	2	2
Cachaça e Pinga	6	6	7	7	7	7	8	8	6	7
Cachaça e Presente	7	7	7	7	6	7	7	7	8	5
Cachaçaria dos Amigos	7	7	6	7	7	7	7	7	7	7
Cachaçaria Nacional	7	6	7	7	7	7	7	7	7	7
Cachaça Sagatiba	1	1	1	1	0	0	0	0	0	0
Cachaças Brasileiras	7	7	7	6	7	7	7	7	7	7
Canela de Ema	7	6	7	7	7	7	7	7	7	7
Casa da Bebida	6	7	7	7	7	7	7	7	7	6
Ceia Clandestina	2	2	2	2	2	1	1	1	1	1
Cia Muller	1	1	1	1	1	1	1	1	1	1
Dom Tapparo	1	1	1	0	1	0	0	0	0	1
Ethylica	7	7	6	7	7	7	7	7	7	7
Cachaça Magnifica	1	0	1	1	1	1	0	0	0	1
Moça Bonita	7	7	7	6	7	7	7	7	7	7
Salinas	7	7	7	7	7	7	7	7	7	6
Sanhaçu	1	1	1	1	0	0	0	0	0	0
Sapucaia	0	1	1	1	1	1	1	1	1	0
Velho Barreiro	1	1	1	1	1	1	1	1	2	2
Wiba	1	1	0	1	1	1	0	0	0	1
Total de Documentos	100	100	100	100	100	100	100	100	100	100

Fonte: Elaboração própria.

Tabela 4.7 – Percentual de entidades por categoria para os conjuntos de treinamento e teste.

Categorias	**Total de Entidades do cachacaNER**	**Conjunto de Treino**	**Conjunto de Teste**
NOME_BEBIDA	3.171	70%	30%
GRADUACAO_ALCOOLICA	1.144	69%	31%
CLASSIFICACAO_BEBIDA	1.325	67%	33%
EQUIPAMENTO_DESTILACAO	292	71%	29%
TEMPO_ARMAZENAMENTO	1.210	74%	32%
RECIPIENTE_ARMAZENAMENTO	991	70%	30%
TIPO_MADEIRA	2.557	68%	32%
CARACTERISTICA_SENSORIAL_COR	562	73%	27%
CARACTERISTICA_SENSORIAL_AROMA	935	70%	30%
CARACTERISTICA_SENSORIAL_SABOR	906	69%	31%
CARACTERISTICA_SENSORIAL_CONSISTENCIA	278	69%	31%
NOME_PESSOA	743	70%	30%
NOME_LOCAL	4.232	70%	30%
NOME_ORGANIZACAO	974	67%	33%
TEMPO	1.302	69%	31%
PRECO	885	70%	30%
VOLUME	2.532	70%	30%
Total de entidades	**24.039**	**69%**	**31%**

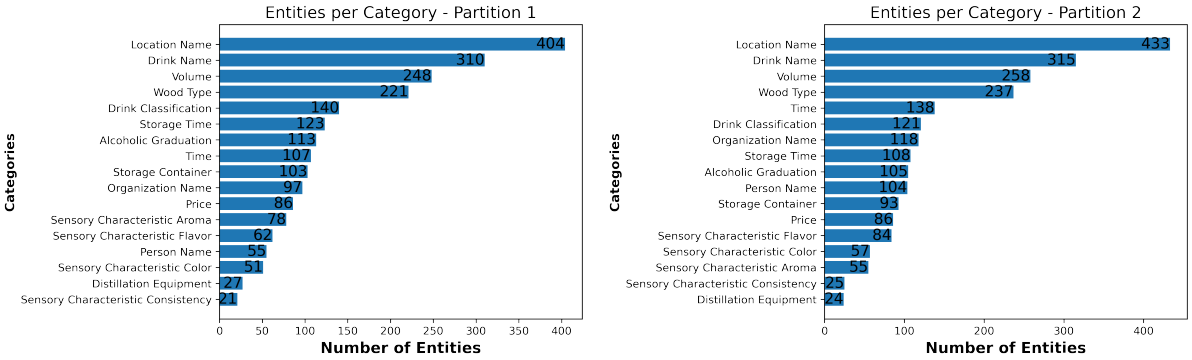
Fonte: Elaboração própria.

Figura 4.10 – *Dataset* cachacaNER versão final.

	Palavras	Rotulo	Sentenca	Inicio	Fim	Documento	Particao	Identificacao_Treino_Teste
0	NOME	O	130	0	4	12	1	treino
1	DA	O	130	5	7	12	1	treino
2	CACHAÇA	O	130	8	15	12	1	treino
3	:	O	130	15	16	12	1	treino
4	Porto	B-NOME_BEBIDA	130	17	22	12	1	treino
...	...	...	...	...	...	...	...	...
183014	Destilados	O	13618	53	63	999	10	teste
183015	do	O	13618	64	66	999	10	teste
183016	Brasil	B-NOME_LOCAL	13618	67	73	999	10	teste
183017	2019	B-TEMPO	13618	74	78	999	10	teste
183018	.	O	13618	78	79	999	10	teste

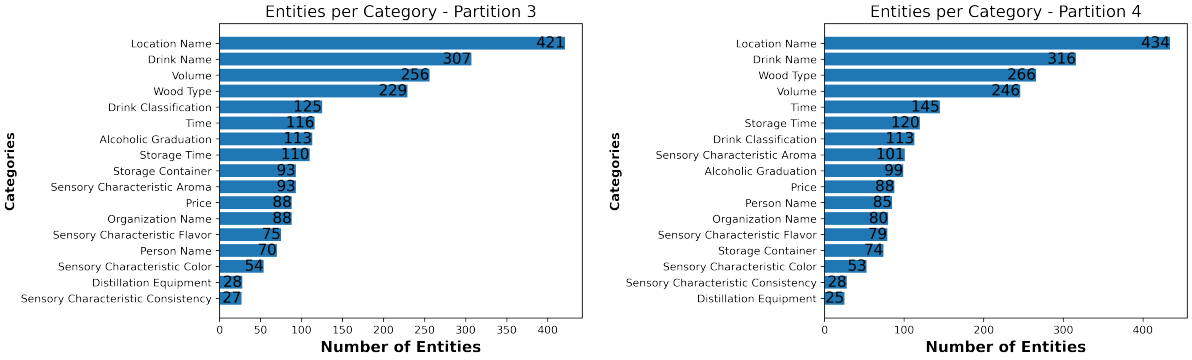
Fonte: Elaboração própria.

Figura 4.11 – Partições 1 e 2.



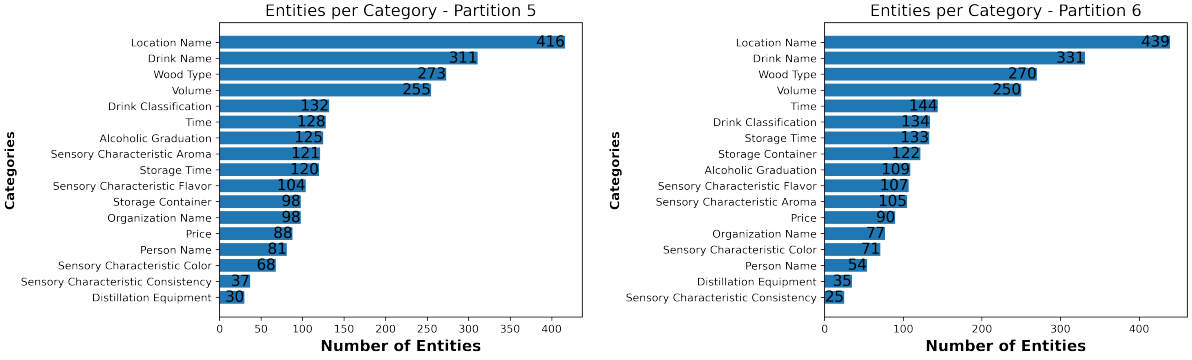
Fonte: Elaboração própria.

Figura 4.12 – Partições 3 e 4.



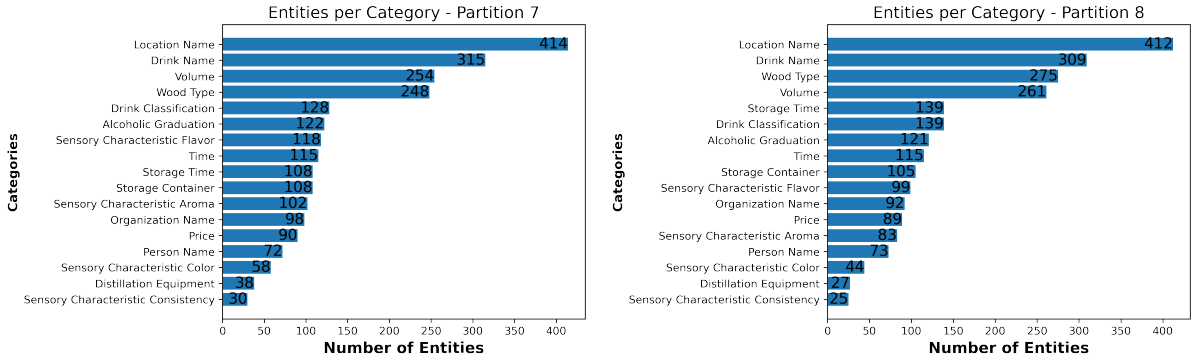
Fonte: Elaboração própria.

Figura 4.13 – Partições 5 e 6.



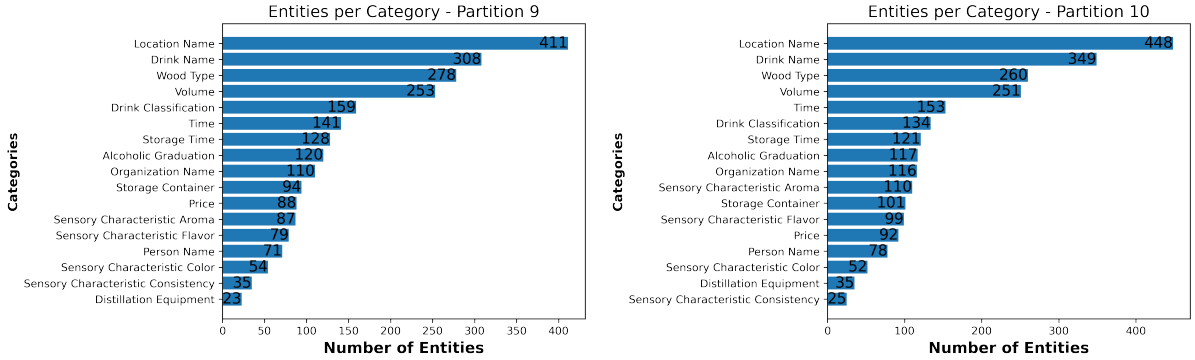
Fonte: Elaboração própria.

Figura 4.14 – Partições 7 e 8.



Fonte: Elaboração própria.

Figura 4.15 – Partições 9 e 10.



Fonte: Elaboração própria.

4.9 Estatísticas Extraídas do *Dataset*

Nesta seção, são apresentadas algumas informações e estatísticas descritivas extraídas do *dataset* cachacaNER.

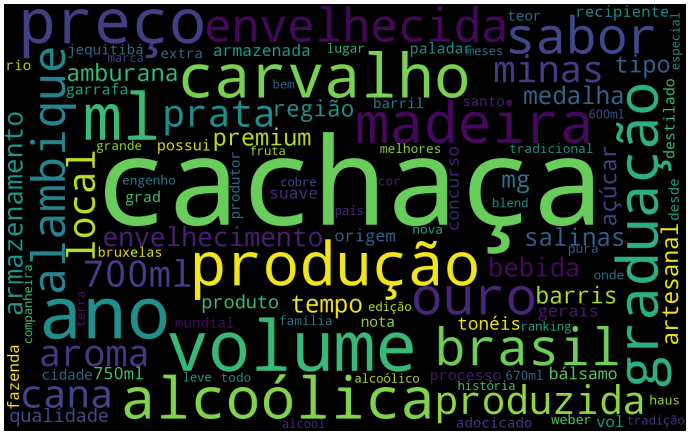
Para identificar as palavras que aparecem com maior frequência, implementou-se a técnica de nuvem de palavras, apresentada na Figura 4.16. Dentre as palavras mais comuns tem-se: cachaça, volume, graduação, alcoólica, carvalho, preço, madeira, envelhecida, entre outras.

O gráfico da Figura 4.17 mostra a distribuição da quantidade de entidades por categoria. As categorias com maior e menor número de entidades rotuladas foram NOME_LOCAL (4.232) e CARACTERISTICA_SENSORIAL_CONSISTENCIA (278), respectivamente.

O histograma da Figura 4.18 apresenta a frequência com que determinados intervalos de quantidade de *tokens* aparecem nas sentenças. Por meio desses dados, é possível observar que 7.359 sentenças possuem entre 1 e 10 *tokens*, o equivalente a 53,99% das sentenças do *dataset*, ou seja, a maioria das sentenças se enquadra neste intervalo, o que reflete um conjunto de dados

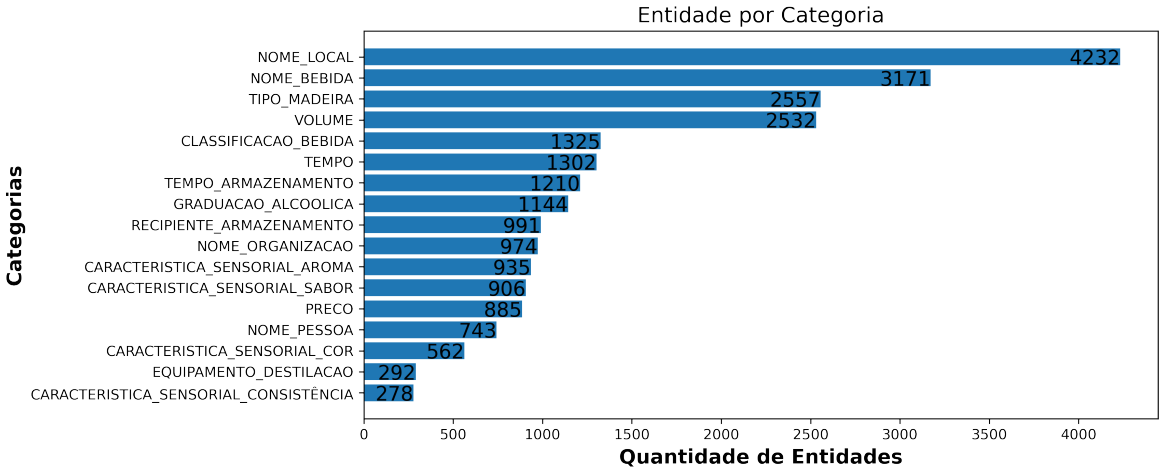
com sentenças não muito longas. Vale ressaltar que a maioria dos *tokens* de tamanho um, dois e três são menções a volume (ex.: 700ml), local (ex.: Brasil, Minas Gerais) e graduação alcoólica (ex.: 23%).

Figura 4.16 – Nuvem de Palavras do *dataset* cachacaNER.



Fonte: Elaboração própria.

Figura 4.17 – Quantidade de entidades por categoria.

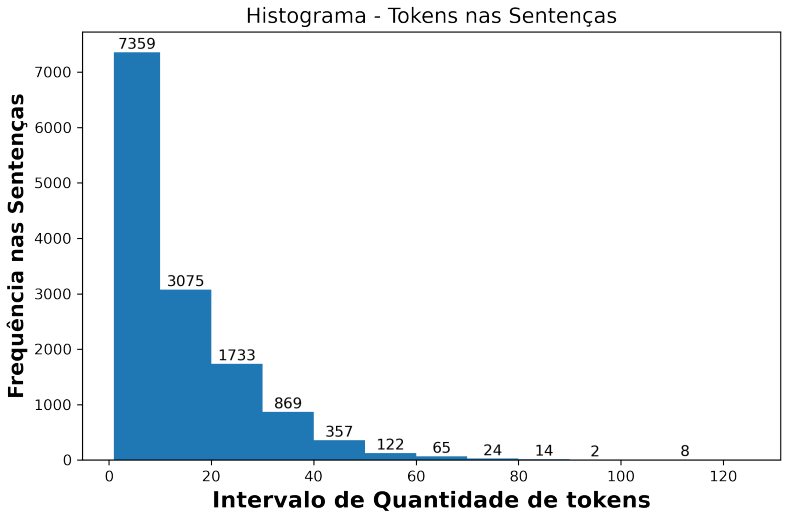


Fonte: Elaboração própria.

No histograma da Figura 4.19 tem-se a relação do intervalo de quantidade de *tokens* por documento. O intervalo mais frequente é entre 100 e 200 *tokens* por documento, equivalendo a 381 documentos.

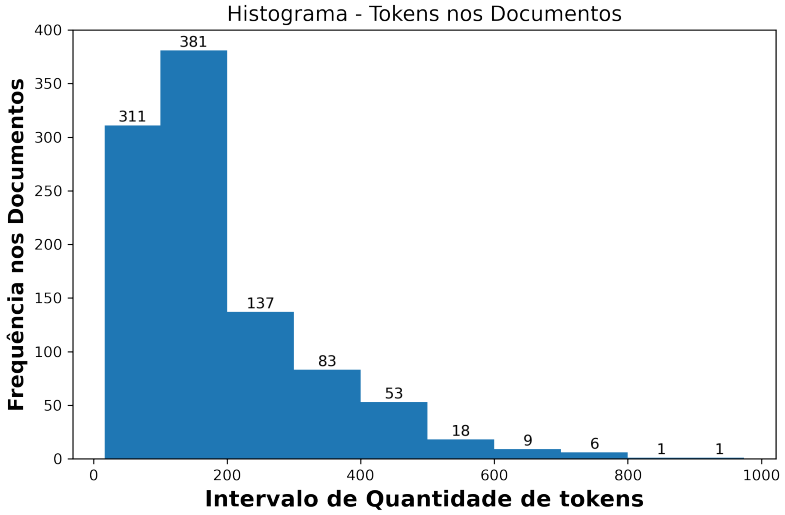
Na relação entre sentenças e documentos, observa-se por meio da Figura 4.20 que o intervalo entre 10 e 20 sentenças é o mais frequente nos documentos, 394 documentos se enquadram neste intervalo.

Figura 4.18 – Frequência de intervalo de quantidade de *tokens* nas sentenças.



Fonte: Elaboração própria.

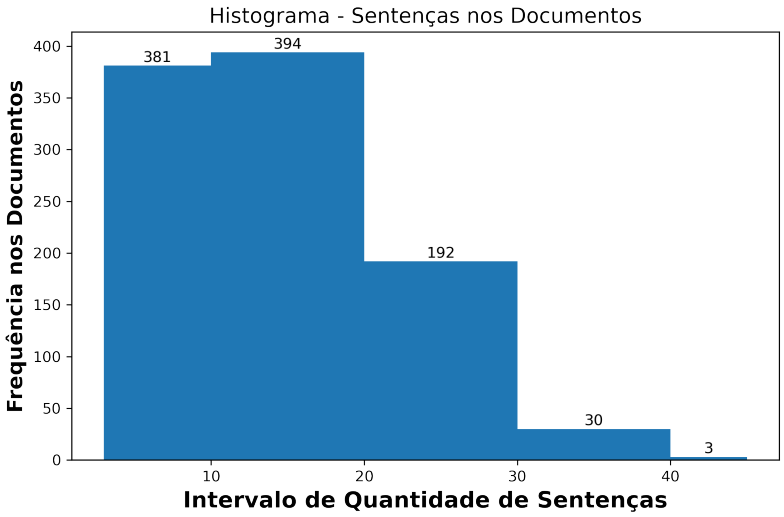
Figura 4.19 – Frequência de intervalo de quantidade de *tokens* nos documentos.



Fonte: Elaboração própria.

Na Tabela 4.8 são apresentadas algumas medidas de tendência central (média aritmética, média harmônica, moda e mediana), de dispersão (desvio padrão, variância populacional) e de forma (assimetria e curtose), extraídas do *dataset* cachacaNER. As medidas de tendência central mostram valores centrais significativos de um conjunto de dados. Medidas de dispersão descrevem o grau de variabilidade dos dados observado, ou o quão espalhados eles se encontram. As medidas de forma auxiliam na identificação de onde os dados estão concentrados. Essas medidas foram calculas para o nível de *token*, sentença e documento, sendo descritas a seguir:

Figura 4.20 – Frequência de intervalo de quantidade de sentenças nos documentos.



Fonte: Elaboração própria.

a) quantidade mínima e máxima: menor e maior quantidade de dados que podem compor uma sentença ou documento;
b) média aritmética: indica o valor onde estão concentrados os dados de um conjunto de valores;
c) média harmônica: tipo especial de média, aplicada quando a série de dados apresenta uma relação inversa entre os dados;
d) moda: valor que mais se repete dentro do conjunto de dados;
e) mediana: valor encontrado no meio do conjunto de dados. Esse valor divide o conjunto ao meio em uma metade superior e outra inferior;
f) variância populacional: mede o quão afastados os dados estão de seu valor médio, isto é, o quão dispersos são os dados;
g) desvio padrão: mede a dispersão dos dados a partir da variância populacional, pois é a raiz quadrada desse parâmetro;
h) assimetria: grau de desvio da simetria de uma distribuição de dados em torno da média. Uma distribuição será considerada assimétrica negativa quando o valor de desvio da média for negativo, e assimétrica positiva quando o valor de desvio for positivo;
i) curtose: é o grau de achatamento de uma distribuição de dados em relação à distribuição normal. Se o valor da curtose for igual a 3, então tem o mesmo achatamento que a distribuição normal, e recebe o nome de distribuição mesocúrtica. Se o valor for maior que 3, então é considerada uma distribuição mais alta, afunilada e mais concentrada que

a distribuição normal, e se chama leptocúrtica. Se o valor for menor que 3, então tem-se uma distribuição mais achatada, denominada de platicúrtica.

Tabela 4.8 – Estatísticas extraídas do *dataset* cachacaNER.

Estatísticas	**Tokens por Sentença**	**Tokens por Documento**	**Sentenças por Documento**
Quantidade Mínima	1	17	3
Quantidade Máxima	125	974	45
Média Aritmética	13,42	183,01	13,62
Média Harmônica	6,72	112,14	10,10
Moda	6	175	9
Mediana	9	140,5	12
Desvio Padrão	12,32	136,98	7,38
Variância Populacional	151,82	1.8763,69	54,56
Assimetria	2	1,72	0,91
Curtose	5,84	3,72	0,37

Fonte: Elaboração própria.

De acordo com os valores da Tabela 4.8, as sentenças possuem em média 13,42 *tokens* e os documentos 183,01 *tokens*, treze vezes mais do que uma sentença. Os documentos possuem uma média de 13,62 sentenças. A dispersão (variância populacional) da quantidade de *tokens* nas sentenças e nos documentos é alta, dado que o *dataset* é composto por tamanhos de textos variados, sendo assim algumas sentenças e documentos possuem muitos *tokens* e outras menos. Isso se explica, pois alguns sites possuem apenas a ficha técnica do produto, a qual é resumida em poucas informações, enquanto em outros sites há textos com a história da bebida, análise do cachacier, explicação de como degustar a cachaça, além das informações mais técnicas. Os dados estão distribuídos de maneira assimétrica a direita, e possuem um achatamento do tipo leptocúrtica. Na Tabela 4.9 são apresentados alguns dados gerais e por partição referentes ao *dataset* cachacaNER

Tabela 4.9 – Dados gerais e por partição referentes ao *dataset* cachacaNER.

*	**Conjunto de Treino**	**Conjunto de Teste**	**Total do Dataset**
Quantidade de documentos	700	300	1.000
Quantidade de sentenças	9.454	4.174	13.628
Quantidade de t*okens*	129.380	53.639	183.019
Quantidade de entidades	16.651	7.388	24.039

Fonte: Elaboração própria.

4.10 Avaliação Experimental

A avaliação experimental teve como objetivos avaliar o *dataset* cachacaNER para a tarefa de reconhecimento de entidades nomeadas e estabelecer valores *baseline* de desempenho para a implementação de outros modelos de NER. Para isso, utilizou-se o conjunto de dados de treinamento do cachacaNER (70% dos dados) para treinar o modelo NER do spaCy com diferentes configurações de parâmetros e o conjunto de teste, também do cachacaNER (30% dos dados), para testar o modelo nesses diferentes cenários.

A biblioteca spaCy requer o carregamento de *pipelines* treinados para conseguir definir anotações, por exemplo, determinar se uma palavra é sujeito ou verbo. Um *pipeline* treinado normalmente consiste em algumas tarefas (NER, POS-tagging, tokenizador, classificador, entre outros) que usam um modelo estatístico de CNN treinado com dados rotulados. Os *pipelines* previamente treinados para várias linguagens podem ser instalados como módulos de Python individuais. Neste trabalho, utilizou-se o pacote de *pipeline* "pt_core_web_sm" para português, o qual possui uma tabela de vetores de palavras reduzidas com 20 mil vetores exclusivos para aproximadamente 500 mil palavras. Esse *pipeline* apresenta uma das melhores relações entre dimensão/velocidade de execução e precisão (HONNIBAL; MONTANI, 2017).

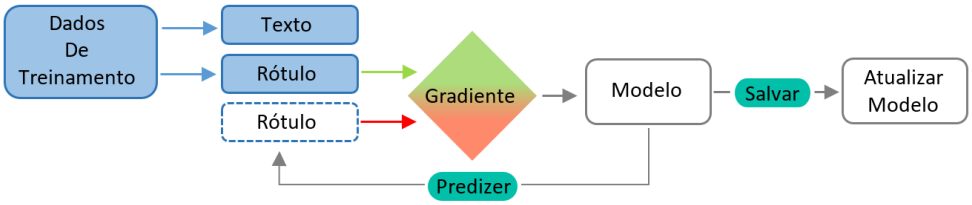
O treinamento de um modelo NER no spaCy é um processo iterativo, no qual as classificações do modelo são comparadas com as rotulações de referência para estimar o gradiente da perda[20]. O gradiente da perda é então usado para calcular o gradiente dos pesos através de retropropagação, uma abordagem que permite expressar as operações da rede neural como funções de ordem superior. Os gradientes indicam como os valores do peso devem ser alterados para que as classificações do modelo se tornem mais semelhantes aos rótulos de referência ao longo do tempo. Este fluxo iterativo é apresentado de maneira simplificada na Figura 4.21.

Um modelo spaCy não busca apenas memorizar os dados de referência, mas sim apresentar uma teoria que possa ser generalizada em dados ainda não vistos. O objetivo é que o modelo aprenda com base no contexto em que as entidades aparecem no texto, isto é, aprender que a instância "Princesa Isabel" provavelmente será um nome de bebida, quando ela aparecer em um contexto semelhante aos dos dados de treinamento. Por exemplo, aprender que no texto "A cachaça Prince Isabel possui 49% de graduação alcoólica.", o *token* **Prince Isabel** é

[20] Gradiente é a direção e a taxa de mudança de um valor numérico. Minimizar o gradiente dos pesos deve resultar em previsões mais próximas dos rótulos de referência dos dados de treinamento.

uma entidade refente a nome de bebida, mas no texto “A empresa Princesa Isabel tem ganhado destaque no mercado de cachaça nos últimos anos.”, **Prince Isabel** é nome de organização.

Figura 4.21 – Treinamento iterativo de modelos no spaCy.



Fonte: Adaptado de Honnibal e Montani (2017).

4.10.1 Configuração do Experimento

Os parâmetros do algoritmo de treinamento do modelo NER que sofreram alterações durante o experimento foram os seguintes:

a) *epoch* (época): quantas passagens o algoritmo de aprendizado faz pelo conjunto de treinamento;

b) *batch* (tamanho de lote): quantidade de amostras do conjunto de treino utilizada para treinar o modelo antes que seus pesos sejam atualizados;

c) *drop* (taxa de abandono): percentual de neurônios que são descartados aleatoriamente durante o treinamento do modelo, para evitar *overfitting*[21]. Essa técnica dificulta a memorização dos dados por parte do modelo.

Além dos parâmetros descritos acima, também utilizou-se os seguintes componentes para compor a arquitetura do modelo utilizado:

a) otimizador “SGD” com um valor fixo igual a “nlp.create_optimizer()”. Esse parâmetro atualiza os pesos do modelo após seu treinamento, com cada amostra do conjunto de treinamento. Quando nenhum valor é definido, o spaCy utiliza o otimizador “Adam” com configurações padrão, mas nesta pesquisa utilizou-se apenas o SGD, por ser o mais comum entre os códigos encontrados na literatura;

b) *word embbeding* chamado Tok2Vec, o qual se concentra na ordem das palavras nas frases;

[21] Cenário que ocorre quando um modelo está muito alinhada a um conjunto limitado de dados. Consequentemente, ele será útil apenas para seu conjunto de dados inicial e não a quaisquer outros conjuntos de dados.

c) random.shuffle (lista) para embaralhar aleatoriamente os dados da lista de exemplos rotulados para treinar o algoritmo.

4.10.2 Execução do Experimento

Para execução do experimento, as seguintes tarefas foram realizadas: (i) treinamento e teste do modelo e (ii) avaliação da capacidade de rotulação do modelo.

Na tarefa de treinamento, os parâmetros do algoritmo de implementação do modelo NER foram alterados para diferentes tamanhos de *epoch*, *batch* e *drop*, a fim de treinar e testar o modelo com diferentes configurações.

Por fim, os resultados obtidos em cada execução foram utilizados para calcular as métricas de precisão, revocação e micro-F1[22], as quais permitiram avaliar a capacidade de predição do modelo. Essas métricas foram calculadas para os resultados do modelo de forma global, isto é, predição de todas as categorias juntas, e também para os resultados individuais de cada categoria.

4.10.3 Métricas de Avaliação

A avaliação de modelos de reconhecimento de entidades nomeadas é geralmente baseada na comparação das saídas geradas automaticamente pelo modelo com os textos rotulados manualmente pelos analistas. As métricas, por sua vez, são maneiras de representar o desempenho preditivo dos modelos por meio de valores numéricos. Dentre as principais métricas utilizadas, encontram-se a precisão (em inglês *precision*), revocação (em inglês *recall*) e medida-F1 (em inglês *F1-measure*) (KUPERUS; VEENMAN; KEULEN, 2013), as quais são calculadas a partir dos valores extraídos de uma matriz de confusão, tal como a ilustrada no Quadro 4.5.

Quadro 4.5 – Matriz de confusão.

Rótulo Real	**Previsão do Rótulo**	
	Previsão (+)	**Previsão (-)**
Observação Positiva (+)	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Observação Negativa (-)	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Adaptado de Garcia (2021).

No caso do reconhecimento de entidades nomeadas VP, FP, FN e VN representam as seguintes circunstâncias (GARCIA, 2021):

[22] Cada instância em cada rótulo é ponderada igualmente.

a) verdadeiro positivo: ocorre quando uma entidade (*token*) é classificada como pertencendo a determinada categoria de entidade e realmente pertence a essa categoria assinalada;
b) falso positivo: se dá quando o modelo afirma que uma entidade pertence a determinada categoria, porém ela pertence a outra categoria;
c) falso negativo: acontece quando o modelo afirma que uma entidade não pertence a determina categoria, mas na realidade ela pertence a essa referida categoria;
d) verdadeiro negativo: ocorre quando o modelo afirma que a entidade não pertence a determinada categoria e ela realmente não pertence.

Para mensurar a capacidade de predição de um modelo NER, considera-se que uma entidade está correta apenas se for uma correspondência exata da entidade correspondente no *dataset* rotulado, isto é, *dataset* de comparação (SANG; MEULDER, 2003). Correspondência exata, significa que o intervalo de início e fim da *string* (*token*) identificada pelo modelo, bem como o tipo/categoria de entidade atribuído a essa *string*, são exatamente iguais ao intervalo e tipo encontrados no *dataset* de comparação. Esse é o formato de correspondência utilizado pelo spaCy.

As métricas são definidas da seguinte forma:

Precisão: é uma medida da qualidade da resposta do modelo, isto é, mede a proporção de entidades identificadas e classificadas corretamente dentre todas as repostas fornecidas pelo modelo (MOTA; SANTOS, 2008). Calcula-se da seguinte forma:

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos + Falsos Positivos}} \tag{4.2}$$

Revocação: mede a proporção de entidades corretamente identificadas e classificadas pelo modelo em relação às entidades efetivamente corretas. Calcula-se da seguinte forma:

$$\text{Revocação} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos + Falsos Negativos}} \tag{4.3}$$

Medida-F1: é a média harmônica que combina as medidas de precisão e revocação, conforme a seguinte fórmula:

$$F1 = 2 \times \frac{\text{Precisão } \times \text{ Revocação}}{\text{Precisão } + \text{ Revocação}} \tag{4.4}$$

As três métricas podem ser calculadas para cada categoria separadamente (avaliação em nível de categoria) e para todas as categorias conjuntamente (avaliação em nível global). Para melhor exemplificar, considere o seguinte texto “A cachaça Sanhaçu é envelhecida por 2 anos em Jatobá, 1 ano em Amburana, e descansa por seis meses em Carvalho”, e as seguintes previsões de categorias de entidades feitas por um modelo NER *versus* as categorias corretas, conforme ilustrado no Quadro 4.6.

Quadro 4.6 – Exemplo de predições do modelo NER *versus* categorias corretas.

Entidade	Predição do Modelo	Categoria Correta
Sanhaçu	NOME_BEBIDA	NOME_BEBIDA
2 anos	TEMPO_ARMAZENAMENTO	TEMPO_ARMAZENAMENTO
Jatobá	TIPO_MADEIRA	TIPO_MADEIRA
1 ano	TEMPO_ARMAZENAMENTO	TEMPO_ARMAZENAMENTO
Amburana	TIPO_MADEIRA	TIPO_MADEIRA
seis meses	TIPO_MADEIRA	TEMPO_ARMAZENAMENTO
Carvalho	TEMPO_ARMAZENAMENTO	TIPO_MADEIRA

Fonte: Elaboração própria.

De acordo com as previsões apresentadas no Quadro 4.6, o modelo teria a seguinte avaliação em nível de categoria, para a categoria TIPO_MADEIRA:

a) 2 Verdadeiros Positivos, porque “Jatobá” e “Amburana” foram corretamente previstos como TIPO_MADEIRA;
b) 1 Falso Positivo, pois “seis meses” foi incorretamente previsto como TIPO_MADEIRA, enquanto deveria ter sido previsto como TEMPO_ARMAZENAMENTO;
c) 1 Falso Negativo, dado que “Carvalho” foi incorretamente previsto como TEMPO_ARMAZENAMENTO, em vez de TIPO_MADEIRA.

Avaliação do modelo a nível de categoria para TEMPO_ARMAZENAMENTO:

a) 2 Verdadeiros Positivos: “2 anos” e “1 ano” foram preditos corretamente como TEMPO_ARMAZENAMENTO;
b) 1 Falso Positivo: “Carvalho” foi incorretamente previsto como TEMPO_ARMAZENAMENTO, enquanto deveria ter sido previsto como TIPO_MADEIRA;
c) 1 Falso Negativo: “seis meses” foi incorretamente previsto como TIPO_MADEIRA, em vez de TEMPO_ARMAZENAMENTO.

Avaliação do modelo a nível de categoria para NOME_BEBIDA:

a) 1 Verdadeiro Positivo: “Sanhaçu” foi predito corretamente como NOME_BEBIDA;
b) 0 Falso Positivo;

c) 0 Falso Negativo.

Na Tabela 4.10, são apresentados os cálculos das métricas para o nível de categoria.

Tabela 4.10 – Cálculo das métricas de avaliação para o nível de categoria.

Categoria	Precisão	Revocação	F1
TIPO_MADEIRA	$\frac{2}{2+1} = 0,66$	$\frac{2}{2+1} = 0,66$	$2 \times \frac{0,66 \times 0,66}{0,66+0,66} = 0,66$
TEMPO_ARMAZENAMENTO	$\frac{2}{2+1} = 0,66$	$\frac{2}{2+1} = 0,66$	$2 \times \frac{0,66 \times 0,66}{0,66+0,66} = 0,66$
NOME_BEBIDA	$\frac{1}{1+0} = 1$	$\frac{1}{1+0} = 1$	$2 \times \frac{1 \times 1}{1+1} = 1$

Fonte: Elaboração própria.

Para a avaliação global, teria-se os seguintes cenários e cálculos:

a) 5 Verdadeiros Positivos: “Sanhaçu”, “2 anos”, “Jatobá”, “1 ano” e “Amburana” foram corretamente previstos, então 5 é a soma dos verdadeiros positivos para todas as categorias;

b) 2 Falsos Positivos: “seis meses” e “Carvalho” foram incorretamente previstos;

c) 2 Falsos Negativos: “seis meses” e “Carvalho”, somando dos falsos negativos para todas as categorias.

Para o nível global as métricas seriam calculadas da seguinte maneira:

a) Precisão: $\frac{5}{5+2} = 0,71$;

b) Revocação: $\frac{5}{5+2} = 0,71$;

c) F1: $2 \times \frac{0,71 \times 0,71}{0,71+0,71} = 0,71$.

4.10.4 Resultados e Discussões

Nesta seção são discutidos os resultados do experimento de avaliação do *dataset* cachacaNER. Todos os resultados do modelo para as diferentes configurações de parâmetros foram avaliados a nível de categoria e global. O desempenho do modelo é calculado com base nos dados do teste.

Para iniciar a avaliação, executou-se primeiramente um teste piloto, por meio do treinamento e teste do modelo NER dez vezes com parâmetros diferentes. A quantidade de épocas foi aumentada em mais 10 unidades a cada novo ciclo de treinamento e teste, iniciando em 10 e terminando em 100. Os parâmetros *batch* e *drop* não sofreram alterações, permanecendo com os valores 32 e 0,25, respectivamente. Esses valores foram escolhidos por serem os mais comumente utilizados na literatura. O objetivo desse teste piloto foi verificar o desempenho do

modelo para diferentes tamanhos de época. Como resultado, identificou-se que 50 é a menor quantidade de épocas com que o modelo obteve maior resultado de desempenho.

A partir do conhecimento obtido com o teste piloto, foram realizados novos ciclos de treinamento e teste do modelo, com os seguintes parâmetros: (i) épocas de tamanho 50 e 100, (ii) *batchs* potências de 2, iniciando em 16 e terminando em 128, tamanhos comumente utilizados na literatura, (iii) *drops* de tamanho 0,10, 0,20 e 0,30, pois o tamanho padrão utilizado pelo algoritmo do spaCy é 0,20, então 0,10 seria o valor inferior e 0,30 o superior ao padrão. Os resultados de desempenho do modelo a nível global e por categoria, para os ciclos de treinamento e teste com as 24 configurações de parâmetros, são apresentados nas Tabelas 4.11 e 4.12.

Tabela 4.11 – Resultados de desempenho do modelo a nível global para diferentes configurações de parâmetros.

Época	**Batch**	**Drop**	**Precisão**	**Revocação**	**micro-F1**
50	16	0,10	0,893	0,871	0,882
100	16	0,10	0,895	0,871	0,883
50	16	0,20	0,882	0,874	0,878
100	16	0,20	0,893	0,871	0,882
50	16	0,30	0,881	0,878	0,879
100	16	0,30	0,901	0,867	0,883
50	32	0,10	0,892	0,870	0,881
100	32	0,10	0,883	0,866	0,874
50	32	0,20	0,891	0,877	0,884
100	32	0,20	0,891	0,875	0,883
50	32	0,30	0,889	0,880	0,884
100	32	0,30	0,888	0,876	0,882
50	64	0,10	0,886	0,874	0,880
100	64	0,10	0,900	0,865	0,882
50	64	0,20	0,895	0,874	0,884
100	64	0,20	0,886	0,878	0,882
50	64	0,30	0,897	0,880	**0,889**
100	64	0,30	0,900	0,876	0,888
50	128	0,10	0,874	0,878	0,876
100	128	0,10	0,899	0,873	0,885
50	128	0,20	0,895	0,878	0,886
100	128	0,20	0,896	0,875	0,885
50	128	0,30	0,888	0,881	0,884
100	128	0,30	0,885	0,877	0,881

Fonte: Elaboração própria.

Para avaliar se os resultados obtidos pelo modelo, no contexto da classificação global e por categoria, são significativamente diferentes, realizou-se um teste estatístico, com os valores da métrica F1, através da análise de variância (ANOVA, abreviação em inglês). ANOVA é um

teste paramétrico utilizado para determinar se há ou não diferença estatisticamente significativa entre as médias de três, ou mais grupos independentes (JAPKOWICZ; SHAH, 2011; BOBBITT, 2021). O resultado da fórmula ANOVA, a estatística F, permite a análise de vários grupos para determinar qual é a variabilidade entre as amostras e dentro das amostras.

Na avaliação global, organizou-se os dados em grupos, onde um grupo se refere ao conjunto de valores de F1 que possuem o mesmo tamanho de *batch*. No total tem-se 4 grupos, representando os *batchs* de tamanho 16, 32, 64 e 128.

O resultado obtido, com a aplicação da ANOVA Unidirecional[23], $F(3, 20) = 1.03819$, $p < 0,05$, mostra que o resultado da estatística do teste F é 1.03819, e o valor p correspondente é 0.397108. Como o valor p não é inferior a $0,05$ (nível alfa utilizado para teste de confiança), então deixa-se de rejeitar a hipótese nula. Não rejeitar a hipótese nula significa dizer que não existem evidências suficientes para dizer que há uma diferença estatisticamente significativa entre os valores de F1 referentes ao desempenho global do modelo. A mesma análise foi feita com um conjunto de valores agrupados conforme o tamanho do *Drop* (0,10, 0,20 e 0,30), e a hipótese nula também não pode ser rejeitada. Os resultados foram F igual a 2,4497 e p igual a 0,1105.

Conforme os resultados apresentados na Tabela 4.11 e os resultados da análise de variância, é possível extrair as seguintes informações:

a) nenhuma configuração de parâmetros obteve resultado estatisticamente melhor do que outra, segundo o resultado da ANOVA;
b) analisando as métricas de precisão e revocação separadamente, observa-se que para todos os testes a precisão é ligeiramente maior que a revocação;
c) o maior valor de precisão foi 0,901, onde o modelo foi configurado com 100 épocas, 16 *batchs* e *drop* igual a 0,30. A menor precisão foi de 0,874, com 50 épocas, 128 *batchs* e 0,10 de *drop*;
d) o maior valor de revocação foi 0,881, o qual se refere a 50 épocas, 128 *batchs* e *drop* igual a 0,30. O menor valore foi 0,865, com 100 épocas, 64 *batchs* e *drop* igual a 0,10.

A principal conclusão obtida, em relação ao nível global, é que a variação nos valores dos parâmetros não altera o resultado de desempenho do modelo.

[23] ANOVA Unidirecional é uma versão da análise de variância que identifica se existe ou não diferenças estatísticas entre as médias dos valores entre diferentes grupos, mas não consegue identificar quais grupos foram estatisticamente significativamente diferentes entre si.

Na Tabela 4.12, são apresentados os resultados das métricas de avaliação calculados individualmente para cada categoria de entidade nomeada. Esses resultados se referem à configuração de parâmetros com o maior resultado de micro-F1 (0,889), apresentado na Tabela 4.11.

Tabela 4.12 – Resultados de desempenho do modelo a nível de categoria.

Categorias	**Precisão**	**Revocação**	**F1**
NOME_BEBIDA	0,836	0,804	0,820
GRADUACAO_ALCOOLICA	0,980	0,983	**0,981**
CLASSIFICACAO_BEBIDA	0,824	0,856	0,839
EQUIPAMENTO_DESTILACAO	0,842	0,882	0,862
TEMPO_ARMAZENAMENTO	0,952	0,930	0,941
RECIPIENTE_ARMAZENAMENTO	0,930	0,930	0,930
TIPO_MADEIRA	0,948	0,926	0,937
CARACTERISTICA_ SENSORIAL_COR	0,872	0,913	0,892
CARACTERISTICA_ SENSORIAL_AROMA	0,737	0,679	0,707
CARACTERISTICA_ SENSORIAL_SABOR	0,711	0,624	0,665
CARACTERISTICA_ SENSORIAL_CONSISTENCIA	0,884	0,811	0,846
NOME_PESSOA	0,903	0,887	0,895
NOME_LOCAL	0,933	0,950	0,941
NOME_ORGANIZACAO	0,878	0,815	0,846
TEMPO	0,929	0,941	0,935
PRECO	0,858	0,903	0,880
VOLUME	0,957	0,887	0,921

Fonte: Elaboração própria..

Na avaliação por categorias considerou-se as próprias categorias para dividir os dados em grupos, isto é, 17 grupos no total. O resultado obtido com a aplicação da ANOVA Unidirecional é estatisticamente significativo, pois a estatística do teste F foi igual a $572,7055802404764$ e o valor de p igual a 8,173725409443941e-260.

Dentre todas as categorias, o modelo obteve maior desempenho (0,981 de F1) em rotular as entidades referentes a GRADUACAO_ALCOOLICA. Uma das possíveis razões para esse resultado é a regularidade e simplicidade estrutural das expressões quantitativas que representam essa categoria, tais como, “40%”, “45,8” e “40GL”, o que pode ter facilitado o trabalho do modelo. Após análise minuciosa das rotulações geradas pelo modelo, identificou-se que os poucos erros de rotulação relacionados a graduação alcoólica ocorreram porque o modelo rotulou apenas parte de algumas entidades, por exemplo, na sentença “’GRADUAÇÃO ALCOÓLICA:

38.0%", ao invés de rotular **38.0%** ele rotulou apenas "38", haja vista que ele também foi treinado com dados sem o símbolo de porcentagem.

Em contrapartida, CARACTERISTICA_SENSORIAL_SABOR foi a categoria que o modelo obteve menor desempenho (F1 igual 0,665), além de menor precisão (0,711) e revocação (0,624). Esse baixo desempenho se deve ao fato de que um mesmo *token* pode ser usado para representar tanto sabor quanto aroma (CARACTERISTICA_SENSORIAL_AROMA). Por exemplo, na sentença "A cachaça Cabaré possui um toque frutado de baunilha na boca", a palavra **baunilha** se refere a característica sabor, já na sentença "Dom Bré Ouro possui um aroma equilibrado entre baunilha e cravo, que exala no ambiente", baunilha representa aroma. Esse tipo de situação promove a ambiguidade entre as categorias, e consequentemente os erros de classificação das categorias.

Apesar das categorias TEMPO e TEMPO_ARMAZENAMENTO compartilharem *tokens* iguais, o modelo não cometeu muitos erros de ambiguidade, 6 total. Por exemplo, na sentença "Com o lucro obtido com a venda do café, Cyrineo produziu cachaça por apenas dois anos e optou por dedicar-se exclusivamente ao cafezal", o modelo rotulou **dois anos** como TEMPO_ARMAZENAMENTO, ao invés de TEMPO. Analisando os textos relacionados a essas duas categorias, identificou-se que grande parte dos textos que continham entidades referentes a tempo de armazenamento, possuíam próximo ao *token* de entidade palavras relacionadas ao nome da bebida, tipo de madeira ou recipiente de armazenamento, além de palavras como envelhecida ou armazenada. Isso nos leva a deduzir que possivelmente, mesmo com um cenário de ambiguidade, o algoritmo rotulou as entidades corretamente, porque nas sentenças haviam palavras vizinhas que cooperaram para o aprendizado do modelo.

O modelo NER também atribuiu incorretamente entidades do tipo NOME_BEBIDA a categoria NOME_ORGANIZACAO, e vice-versa. Isso se deve ao fato de que algumas bebidas possuem o mesmo nome comercial das empresas que lhe produzem, gerando assim ambiguidade também. Por exemplo, na sentença "A identidade visual do produto segue o conceito da marca WIBA!" a palavra **WIBA!** representa uma entidade do tipo NOME_ORGANIZACAO, já na sentença "Em um copo *long drink*, encha de gelo até a borda, esprema metade de um limão e adicione 60ml de WIBA!", a mesma palavra se refere a NOME_BEBIDA. O modelo também cometeu alguns erros relacionados a identificação das entidades, pois rotulou entidade maiores ou menores do que deveria. Por exemplo, na sentença "A Santo Grau Reserva Itirapuã é uma

cachaça tradicional do interior paulista", o modelo rotulou "Santo Grau Reserva Itirapuã" como nome de bebida, ao invés de **Santo Grau**.

Em relação a categoria CLASSIFICACAO_BEBIDA, o modelo também cometeu alguns erros relacionados a identificação de entidades, pois rotulou apenas parte das entidades. Por exemplo, na sentença "DESCRIÇÃO DA CACHAÇA: Cachaça Premissa Extra Premium 670ml", ele rotulou separadamente apalavras "Extra" e "Premium", como classificação da bebida, ao invés de rotular **Extra Premium** como uma única entidade.

Os erros relacionados a categoria EQUIPAMENTO_DESTILACAO, ocorreram porque o modelo rotulou como equipamento de destilação alguns nomes de organizações que continham a palavra alambique. Por exemplo, na sentença "O Alambique Caialua segue uma tradição familiar de cultivo e amor pela terra.", o modelo rotulou incorretamente "Alambique Caialua" como EQUIPAMENTO_DESTILACAO, ao invés de NOME_ORGANIZACAO.

Os pouco erros do modelo em relação à categoria VOLUME ocorreram porque em alguns casos ele rotulou apenas os números, que representavam as entidades, sem a partícula "ml". Por exemplo, na sentença "Cachaça Cabaré Extra Premium 15 Anos 700 ml", o modelo rotulou como graduação alcoólica "700", ao invés de **700 ml**. Talvez, esse equivoco ocorreu porque no conjunto de treinamento há dados com e sem o termo "ml", consequentemente o modelo pode ter aprendido o padrão com e sem a partícula "ml".

Em relação às categorias recipiente de armazenamento, tipo de madeira, característica sensorial de consistência, característica sensorial de cor, não foram identificados padrões que explicassem os erros do modelo relacionados a elas.

4.10.5 Resultados Obtidos por Outros *Datasets*

Para que esta pesquisa não se restringisse apenas aos relatos de autoavaliação do cachacaNER, realizou-se também uma análise dos resultados de desempenho obtidos por outros dois *datasets* de NER em português, são eles: HAREM[24] e LeNER-BR (ARAUJO et al., 2018). Os resultados não são diretamente comparáveis, pois foram obtidos por métodos diferentes, mas podem ser usados como referência para o desempenho dos conjuntos de dados. A Tabela 4.13 traz os resultados de precisão, revocação e F1-*measure* desses *datasets* e do cachacaNER.

O *dataset* HAREM é um dos principais conjuntos de dados anotados com diferentes categorias de entidades nomeadas para a língua portuguesa. Ele é composto por 129 documentos

[24] https://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip

extraídos de diferentes fontes, tais como, blogs, textos jornalísticos e Wikipédia, todos escritos em português do Brasil e de Portugal. Os textos correspondem a 2.274 parágrafos construídos por 147.991 palavras, e contém 7.272 entidades categorizadas em Pessoa (2.036), Local (1.311), Tempo (1.189), Organização (961), Obra (449), Valor (353), Coisa (308), Acontecimento (300), Abstração (286) e Outro (79) (MOTA; SANTOS, 2008). Na Tabela 4.14 são apresentados, por categorias, os maiores resultados de desempenho alcançados por sistemas de NER que utilizaram o HAREM. Vale ressaltar que no trabalho de Mota e Santos (2008) não foram realizados testes com a categoria Outro.

Tabela 4.13 – Resultados a nível global entre diferentes *datasets* para NER.

Datasets	**Precisão**	**Revocação**	**F1-measure**
cachacaNER	0,901	0,881	0,889
HAREM	0,834	0,535	0,590
LeNER-BR	0,9321	0,9191	0,9253

Fonte: Elaboração própria.

Tabela 4.14 – Resultados do HAREM a nível de categoria, conforme relatado em Mota e Santos (2008).

Entidades	**Precisão**	**Revocação**	**F1-measure**
Pessoa	0,777	0,723	0,638
Local	0,692	0,701	0,607
Tempo	0,748	0,733	0,709
Organização	0,685	0,558	0,418
Obra	0,629	0,407	0,324
Valor	0,413	0,717	0,524
Coisa	1,000	0,513	0,128
Acontecimento	0,783	0,503	0,356
Abstração	0,236	0,567	0,181

Fonte: Adaptado de Mota e Santos (2008).

O LeNER-BR é um *dataset* composto por 60 documentos legais extraídos de leis e decisões jurídicas escritas em português do Brasil. Possui 10.392 sentenças, 318.073 *tokens* e 44.513 entidades categorizadas em Pessoa (6.241), Casos legais (5.370), Tempo (3.146), Localização (1793), Legislação (18.317) e Organização (9.646). Na Tabela 4.15, são apresentados os resultados de desempenho de um modelo treinado e testado com o LeNER-BR (ARAUJO et al., 2018).

Tabela 4.15 – Resultados do LeNER-Br a nível de categoria, conforme relatado em Araujo et al. (2018).

Entidades	**Precisão**	**Revocação**	**F1-measure**
Pessoa	0,9444	0,9252	0,9347
Casos legais	0,8739	0,903	0,8882
Tempo	0,9115	0,9115	0,9115
Localização	0,6124	0,5985	0,6054
Legislação	0,9708	0,97	0,9704
Organização	0,9127	0,8566	0,8838

Fonte: Adaptado de Araujo et al. (2018).

Tomando como referência os resultados obtidos com os *datasets* HAREM e LeNER-Br, é possível observar que o *dataset* cachacaNER encontra-se dentro de um padrão de desempenho de classificação de categorias de entidades nomeadas próximo ao do LeNER-Br e do HAREM. Em relação ao tamanho dos *datasets*, o cachacaNER possui uma quantidade de documentos (1.000), sentenças (13.628), *tokens* (183.019) e entidades (24.039) maior que o HAREM e menor que o LeNER-Br. Em comparação com o LeNER-Br e o HAREM, o cachacaNER é composto por uma quantidade maior de categorias de entidades nomeadas (17 no total), o que o torna mais complexo.

5 ROTULAÇÃO AUTOMÁTICA

Um dos principais problemas relacionados ao treinamento de modelos de reconhecimento de entidades nomeadas é o esforço envolvido na criação e obtenção de dados rotulados (SILVA et al., 2021). Considerando essa problemática, neste trabalho, avaliou-se a rotulação automática de entidades do *dataset* cachacaNER, por meio do algoritmo proposto no trabalho de Melo e Figueiredo (2021). Para avaliar a qualidade da rotulagem feita pelo algoritmo, calculou-se a diferença entre o que o algoritmo deveria ter rotulado pela quantidade de acertos do algoritmo. Além disso, também realizou-se o experimento de avaliar o desempenho de um modelo NER ao ser treinado com conjuntos de dados que contém diferentes quantidades de entidades rotuladas automaticamente.

Desta maneira, este capítulo traz a descrição da metodologia adotada, o algoritmo utilizado para a rotulação automática das entidades, o experimento e a avaliação de um modelo NER treinado com diferentes quantidades de entidades rotuladas.

5.1 Metodologia

O objetivo da rotulação automática é permitir que entidades sejam rotuladas sem a necessidade de que um humano as rotule manualmente. Para alcançar esse objetivo, foram realizadas duas tarefas principais: (i) implementação de um algoritmo de rotulação automática pré-existente e (ii) simulação da rotulação manual de entidades de diferentes tipos de categorias, para compor a lista de entidades rotuladas utilizada pelo algoritmo.

Para rotular textos de entrada, o algoritmo utiliza uma lista composta por exemplos de entidades acompanhadas de suas respetivas categorias. Vale enfatizar que o algoritmo considerada apenas uma categoria por entidade. Ele verifica se o texto de entrada possui *tokens* idênticos aos da lista, se a resposta for verdadeira, ele marca os *tokens* do texto com as categorias associadas aos *tokens* da lista. Essa é uma rotulação às cegas, pois o algoritmo não interpreta o contexto no qual a entidade marcada/rotulada se encontra. Consequentemente, ele pode cometer alguns erros, tais como, rotular palavras que não são entidades de fato, ou que não representam as categorias em questão. Por exemplo, algumas palavras como baunilha e amadeirado podem representar tanto aroma quanto sabor, o que induziria o algoritmo ao erro.

Apesar do algoritmo realizar rotulação automática, para funcionar ele precisa de uma lista com exemplos de entidades rotuladas corretamente. A criação dessa lista demanda tempo

e esforço humano, entretanto, esse esforço é muito menor do que a rotulação manual de um *dataset*. Para otimizar o processo de criação dessa lista, adotou-se a estratégia de reaproveitar os dados (entidades e categorias) previamente rotulados do cachacaNER, mais especificamente os dados do conjunto de treinamento. A ideia principal por trás dessa estratégia foi utilizar os rótulos de categorias que já existiam no *dataset* cachacaNER para simular uma pessoa rotulando as entidades da lista de entidades e as fornecendo ao algoritmo de rotulação automática.

O bom funcionamento do algoritmo de rotulação automática também depende da quantidade de dados que a lista de entidades possui, pois ele só rotula o que está nela. Então, a fim de tentar identificar a quantidade ideal de entidades rotuladas para compor essa lista, propôs-se, nesta pesquisa, a estratégia de criar várias listas com diferentes quantidades de entidades rotuladas. Para determinar o tamanho das listas, cada uma recebeu uma porcentagem diferente do total de entidades contidas na lista base, a qual foi construída a partir de todas as entidades e categorias extraídas do conjunto de treinamento do cachacaNER. Essa porcentagem foi cumulativa, isto é, a cada nova lista acrescentava-se uma porcentagem maior de entidades rotuladas, até que a última lista possuísse o mesmo total de entidades que a lista base. Essas listas foram criadas com base na ideia de simular a rotulação humana, descrita no parágrafo anterior.

Para avaliar a variabilidade dos tamanhos das listas, optou-se por treinar um modelo de NER com textos rotulados automaticamente a partir desses diferentes tamanhos de listas. Para isso, foi necessário criar diferentes conjuntos de dados de treinamento, contendo os mesmos textos, mas com quantidades diferentes de entidades rotuladas. Esses conjuntos de treinamento, foram criados por meio do algoritmo de rotulação automática, o qual foi executado um número de vezes igual ao total de lista de entidades criadas. Em cada execução, o algoritmo recebeu como parâmetros os mesmos textos de entrada, extraídos do conjunto de treinamento do cachacaNER, e as listas com diferentes quantidades de exemplos de entidades rotuladas, também extraídas do conjunto de treinamento do cachacaNER.

De maneira mais detalhada, a estratégica metodológica adotada para realização da rotulação automática e avaliação do modelo NER treinado com diferentes quantidades de entidades rotuladas consistiu em:

a) criar uma lista com exemplos, sem repetições, de pares formados por entidades *versus* categorias, a qual é chamada neste trabalho de **"lista de entidades"**. Essa lista é composta por 2.844 entidades distintas rotuladas, as quais serviram de base para a criação das outras listas com diferentes porcentagens de entidades. Essa lista foi criada por

meio da extração de entidades e categorias do conjunto de treinamento do cachacaNER. Alguns exemplos reais dos dados que compõem essa lista são apresentados a seguir:

- (‘Porto Estrela’, ‘NOME_BEBIDA’);
- (‘Ouro’, ‘CLASSIFICACAO_BEBIDA’);
- (‘Jequitibá’, ‘TIPO_MADEIRA’);
- (‘Cachaçaria Companheira’, ‘NOME_ORGANIZACAO’);
- (‘18%’, ‘GRADUACAO_ALCOOLICA’);

b) criar uma lista com sentenças extraídas do conjunto de treinamento do cachacaNER. Essas sentenças, 9.454 no total, serviram de entrada para o algoritmo de rotulação automática, e após serem rotuladas, foram utilizadas no experimento de treinamento do modelo NER;

c) implementar o algoritmo de rotulação automática proposto por Melo e Figueiredo (2021). Nessa etapa, realizou-se a codificação do algoritmo, em Python, de maneira que ele recebesse como parâmetros a lista de entidades e a lista de sentenças a serem rotuladas, além de gerar como saída um arquivo no formato aceito pelo modelo NER do spaCy, apresentado anteriormente na Figura 4.9;

d) rotular as sentenças de entrada, a partir das listas de entidades de diferentes tamanhos. Para isso, adotou-se a estratégia de dividir a lista de entidades base (com 2.844 entidades) em 10 conjuntos de dados sem repetições, cada um contendo de 10% a 100% do total de entidades da lista base. Os dados que formariam cada lista de entidades rotuladas foram selecionados aleatoriamente, e continham 10% das entidades de cada tipo de categoria existente na lista base. Dessa maneira, o algoritmo de rotulação automática foi executado dez vezes, e a cada nova execução ele acrescentava aproximadamente mais 10% de entidades rotuladas ao conjunto anterior, até obter uma lista com 100%. Essa foi a estratégia utilizada para simular uma pessoa criando listas de entidades e as fornecendo ao algoritmo de rotulação automática. No final dos dez ciclos, foram criados 10 conjuntos de dados de treinamento contendo as mesmas sentenças, mas com quantidades (percentuais) diferentes de entidades rotuladas;

e) treinar o modelo NER com cada um dos 10 conjuntos de dados rotulados automaticamente na etapa anterior. Para cada um dos 10 conjuntos, contendo quantidades diferentes de entidades rotuladas, o modelo foi testado com o mesmo conjunto de teste do cachacaNER;

f) avaliar o desempenho do modelo para cada um dos 10 conjuntos de dados rotulados automaticamente. Para isso, utilizou-se as métricas de precisão, revocação e F1-*measure*, a nível global e por categoria.

5.2 Algoritmo de Rotulação Automática

A estratégia do algoritmo de rotulação automática consiste basicamente em utilizar uma lista semente de *tokens*-chave associados as suas respectivas categorias, a partir da qual as demais entidades existentes nos textos de entrada serão rotuladas, conforme descrito no Algoritmo 1.

Algorithm 1 - Rotulação automática de entidades

seja K um conjunto de *tokens*-chave;
seja E o conjunto categorias;
Entrada: Um conjunto $P = \{\langle k, e\rangle \mid k \in K \text{ e } e \in E\}$;
Entrada: Um conjunto S de sentenças não rotuladas;
Saída: Um conjunto de pares de treinamento $\{\langle s, L\rangle \mid s \in S \text{ e } L \text{ é uma lista de } p \in P\}$

```
1: listaEntidades ← {∅}
2: for s ∈ S do
3:     L ← {∅}
4:     for k, e ∈ P do
5:         if k ∈ {s} then
6:             L ← L ∪ {⟨k, e⟩}
7:         end if
8:     end for
9:     listaEntidades ← listaEntidades ∪ {⟨s, L⟩}
10: end for
11: return listaEntidades
```

O algoritmo recebe duas entradas. A primeira é um conjunto de pares $\langle k, e\rangle \in P$, onde k é um *token*-chave, isto é, uma entidade e e uma categoria de entidade nomeada, por exemplo, ⟨Pinga Ni Mim, NOME_BEBIDA⟩ e ⟨Amburana, TIPO_MADEIRA⟩. A segunda entrada é um conjunto S de sentenças não rotuladas. A saída gerada pelo algoritmo é uma lista composta de pares $\langle s, L\rangle$, onde s é uma sentença e L uma lista de pares $\langle k, e\rangle$. Cada um desses pares representa que a sentença s contém um ou mais *tokens*-chave $k \in K$ associado a alguma categoria $e \in E$.

O algoritmo itera através do conjunto de sentenças $s \in S$ (linhas 2-10), tentando combinar qualquer um dos *tokens* que compõem a sentença com algum *token* $k \in P$ (linha 5). Se houver uma ocorrência de k com qualquer *token* da sentença s, então o par $\langle k, e\rangle$ é adicionado à

lista L (Linha 6), é nesse momento que ocorre a associação entre os *tokens* (entidades) da sentença com os rótulos da lista de entidades, pois o k representa a própria entidade identificada na sentença, e o e seu respectivo rótulo. Para marcar a posição da entidade na sentença, considerou-se a posição inicial e final do *token* (entidade) na sentença. Após todos os pares pertencentes a P serem processados, um dado de treinamento $\langle s, L \rangle$ é adicionado à $listaEntidades$ (linha 9). Depois que todas às sentenças $s \in S$ são processadas, o algoritmo retorna $listaEntidades$ (linha 11), lista que contém todas as sentenças rotuladas.

5.3 Avaliação Experimental

O objetivo principal desta etapa consistiu em avaliar o desempenho de um modelo NER, ao ser treinado com conjuntos de dados que possuem diferentes quantidades de entidades rotuladas, isto é, com os dados rotulados pelo algoritmo de rotulação automaticamente ao considerar diferentes listas de entidades nomeadas, bem como, avaliar também o desempenho do próprio algoritmo de rotulação automática.

5.3.1 Configuração do Experimento

O experimento foi dividido em 10 execuções de treinamento e teste do modelo NER. Em cada execução, o modelo foi treinado com um dos dez conjuntos de dados gerados pelo algoritmo de rotulação automática.

Na segunda e terceira colunas da Tabela 5.1 são apresentadas as porcentagens e respectivas quantidades de entidades rotuladas, através das listas de entidades, pelo algoritmo para rotular automaticamente os conjuntos de dados de treinamento. A quarta coluna traz o total de entidades rotuladas pelo algoritmo, de acordo com cada percentual de entidades rotuladas.

Em relação aos parâmetros de configuração do modelo, utilizou-se a mesma configuração em todas as execuções (50 épocas, 64 *batchs* e *drop* de 0,30), isto é, a configuração associada ao maior valor de micro-F1 da rotulação manual, apresentado na Tabela 4.11. E no que diz respeito ao conjunto de teste, foi utilizado o mesmo conjunto nos dez ciclos, o qual corresponde ao conjunto de teste do *dataset* cachacaNER.

Tabela 5.1 – Porcentagens de entidades rotuladas utilizadas pelo algoritmo para rotular automaticamente os conjuntos de treinamento.

Execução	Porcentagem de entidades rotuladas	Quantidade de entidades referentes a porcentagem	Total de entidades rotuladas pelo algoritmo
1	10,27%	292	2.870
2	21,53%	584	5.147
3	30%	876	6.701
4	41,07%	1.168	9.920
5	51,34%	1.460	11.242
6	61,60%	1.752	13.142
7	71,87%	2.044	14.790
8	82,03%	2.333	17.082
9	92,09%	2.619	17.908
10	100%	2.844	19.171

Fonte: Elaboração própria.

5.3.2 Resultados e Discussões

Nesta seção, são apresentados os resultados da rotulação automática das sentenças com o algoritmo e os resultados do modelo NER a nível global e por categoria. O desempenho do modelo foi calculado com base nos dados do conjunto de teste do cachacaNER.

5.3.2.1 Rotulação Automática com o Algoritmo

Na Tabela 5.2, são apresentados por categoria o total de entidades rotuladas automaticamente pelo algoritmo, o total de entidades corretamente rotuladas, o total de erros cometidos pelo algoritmo e o percentual de acertos do algoritmo. Por meio destes resultados, é possível observar que o algoritmo conseguiu rotular corretamente 95,16% de todas as entidades do conjunto de treinamento, um percentual ótimo, dado os problemas relacionados a uma rotulação as cegas, como ambiguidade e rotulação de entidades equivocadas. O cálculo desse percentual consiste na subtração do total de acertos do algoritmo pelo total de erros.

TIPO_MADEIRA foi a categoria que o algoritmo mais acertou (99,94%), o que se deve ao fato de que ela é representada por um conjunto pequeno de palavras específicas, as quais não sofrem muitas mudanças em sua estrutura morfológica. Dessa maneira, o algoritmo conseguiu na maioria dos casos identificar corretamente esse conjunto de palavras dentro dos textos. Entretanto, ele cometeu alguns poucos erros ao rotular algumas dessas entidades em sentenças onde os referidos *tokens* não eram entidades de fato, ou representavam outras categorias. Por exemplo, na sentença “Harmonização: Indicado para ser apreciada antes e após

as refeições e acompanhar queijos, amendoim e azeitonas.", o algoritmo rotulou "amendoim" como TIPO_MADEIRA, mas de acordo com o contexto ele representa uma comida, não um tipo de madeira. O algoritmo também rotulou incorretamente palavras ambíguas, por exemplo, em "Sabor suave e doce com aroma exclusivo de canela.", ele rotulou "canela" como TIPO_MADEIRA, ao invés de CARACTERISTICA_SENSORIAL_AROMA, pois na rotulação manual "canela" também aparece como tipo de madeira.

Tabela 5.2 – Resultados da rotulação com o algoritmo, a nível de categoria, para o conjunto de treinamento rotulado com 100% dos dados da lista de entidades rotuladas.

Categorias	**Entidades rotuladas manualmente**	**Entidades rotuladas pelo algoritmo**	**Entidades rotuladas corretamente**	**Erros de rotulação**	**Percentual de acertos**
NOME_BEBIDA	2.205	2.466	2.044	422	92,70%
GRADUACAO_ ALCOOLICA	786	839	783	56	99,62%
CLASSIFICACAO_ BEBIDA	893	1.344	880	464	98,54%
EQUIPAMENTO_ DESTILACAO	207	325	205	120	99,03%
TEMPO_ ARMAZENAMENTO	822	881	795	86	96,72%
RECIPIENTE_ ARMAZENAMENTO	691	799	687	112	99,42%
TIPO_MADEIRA	1.744	1.804	1.743	61	99,94%
CARACTERISTICA_ SENSORIAL_COR	412	484	399	85	96,84%
CARACTERISTICA_ SENSORIAL_AROMA	655	1.075	532	543	81,22%
CARACTERISTICA_ SENSORIAL_SABOR	629	1.496	491	1.005	78,06%
CARACTERISTICA_ SENSORIAL_ CONSISTENCIA	193	229	192	37	99,48%
NOME_PESSOA	521	510	496	14	95,20%
NOME_LOCAL	2.961	2.971	2.889	82	97,57%
NOME_ ORGANIZACAO	656	658	511	147	77,90%
TEMPO	893	894	872	22	97,65%
PRECO	616	624	609	15	98,86%
VOLUME	1.767	1.772	1717	55	97,17%
Total	16.651	19.171	15.845	3.337	95,16%

Fonte: Elaboração própria.

A categoria com menor percentual de acertos foi NOME_ORGANIZACAO (77,90%). Isso se deve, primeiramente, ao fato de que o algoritmo rotulou incorretamente como nome de

organização vários *tokens* que não eram entidades de fato, dado o contexto em que se encontravam. O segundo maior motivo de erros está relacionado com a restrição do algoritmo de não conseguir associar a uma mesma entidade mais de uma categoria. Todavia, no conjunto de treinamento do cachacaNER, de onde as entidades rotuladas foram extraídas para criar as listas de entidade, há entidades associadas a mais de uma categoria, de maneira que foi necessário escolher qual das categorias entraria na lista de entidades. Para realizar essa escolha, optou-se por durante a criação da lista de entidades base, automaticamente, definir a primeira categoria que aparecesse associada a entidade ambígua como a categoria padrão, isto é, a que entraria na lista de entidades. Entretanto, com essa estratégia, em algumas sentenças o algoritmo continuou identificando as entidades corretamente, mas as rotulou com a categoria incorreta, por causa da troca de categorias. Vale enfatizar que mesmo com esses erros, o algoritmo conseguiu resultados próximos de 100%. No caso de nome de organização, o algoritmo identificou corretamente a posição inicial e final das entidades dentro dos textos, mas rotulou incorretamente com a categoria nome de bebida, dado que houve na lista de entidades essa mudança de categoria.

Em relação à categoria NOME_PESSOA, o algoritmo rotulou a menos 11 entidades, pois durante a criação da lista de entidades as categorias dessas onze entidades foram trocadas pelas categorias NOME_BEBIDA, NOME_ORGANIZACAO e NOME_LOCAL, assim como ocorreu com nome de organização em relação a nome de bebida. Por pertencerem a mais de uma categoria, ao identificar essas entidades nos textos, o algoritmo as rotulou com as categorias incorretas, por isso, na Tabela 5.2 ao invés de haver 25 erros, há apenas 14. Os demais erros relacionados a nome de pessoa, 14 no total, ocorreram devido a rotulações de entidades que não eram entidades de fato, dado o contexto em que se encontravam, e devido à rotulação de categorias incorretas, por exemplo, na sentença “A Cachaça Leandro Batista Envelhecida 750ml, foi premiada em diversas competições importantes.”, o algoritmo rotulou **Leandro Batista** como nome de pessoa, ao invés de nome de cachaça.

As categorias com o maior número de erros por troca de nome de categoria durante a criação da lista de entidades foram CARACTERISTICA_SENSORIAL_AROMA e CARACTERISTICA_SENSORIAL_SABOR, pois boa parte dos termos utilizados para descrevê-las são iguais, provocando assim ambiguidade, e consequentemente a troca de nomes de categorias.

De forma geral, a maioria dos erros relacionados às demais categorias consistiram na rotulação de entidades que não eram entidades de fato, dado o contexto em que se encontram.

5.3.2.2 Rotulação com o Modelo de NER

Em relação ao treinamento do modelo NER com conjuntos de dados contendo diferentes quantidades de entidades rotuladas, observa-se nos resultados apresentados na Tabela 5.3, que houve uma melhoria gradual de desempenho do modelo à medida que a porcentagem de entidades rotuladas aumentou, atingindo F1 máximo (0,808) com o conjunto de treinamento contendo 100% das entidades rotuladas.

Tabela 5.3 – Resultados de desempenho do modelo, a nível global, treinado com diferentes porcentagens de entidades rotuladas.

Porcentagem de entidades rotuladas	**Precisão**	**Revocação**	**micro-F1**
10,27%	0,602	0,099	0,171
21,53%	0,603	0,175	0,272
30%	0,726	0,273	0,396
41,07%	0,659	0,375	0,478
51,34%	0,698	0,448	0,546
61,60%	0,696	0,514	0,591
71,87%	0,707	0,603	0,651
82,03%	0,732	0,726	0,729
92,09%	0,746	0,785	0,765
100%	0,761	0,860	0,808

Fonte: Elaboração própria.

Comparando, a nível global, o resultado do modelo treinado com os dados rotulados manualmente (Tabela 4.11) *versus* o modelo treinado com os dados rotulados automaticamente com 100% das entidade rotuladas (Tabela 5.3), observa-se que a diferença de desempenho foi de apenas 0,081 (0,889 - 0,808). Quando o modelo foi treinado com 92,09% das entidades rotuladas, a diferença foi de 0,124 (0,889 - 0,765). Com 82,03% das entidades rotuladas, a diferença aumentou para 0,16 (0,889 - 0,729). Com 71,87% das entidades rotuladas, a diferença foi de 0,238 (0,889 - 0,651). E com 61,60% das entidades rotuladas, a diferença foi mais expressiva, isto é, 0,298 (0,889 - 0,591).

Apesar da redução de desempenho do modelo, a medida que o percentual de entidades rotuladas diminui dentro do conjunto de treinamento, pode ser viável ou até melhor optar pela rotulação automática, em detrimento da rotulação manual, pois fornecer exemplos de entidades rotuladas e implementar um algoritmo simples de verificação de *tokens*, é mais barato, mais simples e mais rápido do que rotular dados manualmente.

A nível de categoria, conforme apresentando na Tabela 5.6, o modelo obteve melhor desempenho (F1 de 0,951) para a categoria GRADUACAO_ALCOOLICA, o que pode ser atribuído à simplicidade estrutural desse tipo de entidade, que também vem associada ao símbolo curinga de porcentagem.

Os menores desempenho ficaram com as categorias CARACTERISTICA_SENSORIAL_SABOR (F1 igual a 0,329) e CARACTERISTICA_SENSORIAL_AROMA (F1 de 0,485), o que se deve a ambiguidade existente entre os termos que representam sabor e aroma. Por meio dos resultados apresentados nas Tabelas 5.4, 5.5 e 5.6, também é possível observar que a precisão, revocação e F1 de cada categoria aumentam de acordo com a porcentagem de entidades rotuladas dentro do conjunto de dados de treinamento. Os resultados, por categoria, de F1 para a rotulação automática são de forma geral menores que os da rotulação manual (Tabela 4.12), possivelmente devido à menor quantidade de dados rotulados para treinar o modelo.

Tabela 5.4 – Resultados de precisão, a nível de categoria, do modelo treinado com diferentes porcentagens de entidades rotuladas.

Categorias	**Precisão**									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
NOME_BEBIDA	0,473	0,630	0,698	0,632	0,688	0,681	0,719	0,735	0,679	0,700
GRADUACAO_ALCOOLICA	0,898	0,909	0,795	0,897	0,925	0,918	0,895	0,884	0,922	0,921
CLASSIFICACAO_BEBIDA	0,581	0,537	0,380	0,448	0,472	0,511	0,579	0,583	0,593	0,584
EQUIPAMENTO_DESTILACAO	0,428	0,435	0,512	0,562	0,555	0,552	0,515	0,510	0,510	0,530
TEMPO_ARMAZENAMENTO	0,647	0,824	0,877	0,848	0,865	0,810	0,752	0,769	0,900	0,901
RECIPIENTE_ARMAZENAMENTO	0,622	0,813	0,698	0,815	0,858	0,871	0,855	0,802	0,837	0,833
TIPO_MADEIRA	0,614	0,647	0,820	0,746	0,778	0,810	0,834	0,856	0,897	0,917
CARACTERISTICA_SENSORIAL_COR	0,6	0,4	0,307	0,493	0,424	0,427	0,423	0,513	0,602	0,710
CARACTERISTICA_SENSORIAL_AROMA	0,612	0,4	0,377	0,350	0,366	0,378	0,361	0,363	0,379	0,389
CARACTERISTICA_SENSORIAL_SABOR	0,210	0,215	0,253	0,281	0,296	0,219	0,200	0,208	0,226	0,245
CARACTERISTICA_SENSORIAL_CONSISTENCIA	0,411	0,857	0,785	0,575	0,585	0,510	0,674	0,707	0,762	0,783
NOME_PESSOA	0,625	0,514	0,687	0,392	0,681	0,700	0,767	0,865	0,864	0,840
NOME_LOCAL	0,883	0,918	0,906	0,921	0,915	0,914	0,919	0,929	0,932	0,936
NOME_ORGANIZACAO	0,653	0,687	0,565	0,558	0,586	0,617	0,6	0,677	0,669	0,701
TEMPO	0,695	0,727	0,896	0,938	0,942	0,948	0,953	0,950	0,932	0,944
PRECO	0,821	0,270	0,929	0,478	0,585	0,657	0,718	0,794	0,831	0,856
VOLUME	0,576	0,555	0,737	0,744	0,794	0,84	0,833	0,918	0,9	0,882

Fonte: Elaboração própria.

Tabela 5.5 – Resultados de revocação, a nível de categoria, do modelo treinado com diferentes porcentagens de entidades rotuladas.

Categorias	**Revocação**									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
NOME_BEBIDA	0,037	0,102	0,213	0,220	0,363	0,469	0,577	0,638	0,712	0,765
GRADUACAO_ALCOOLICA	0,173	0,223	0,108	0,561	0,723	0,751	0,762	0,793	0,863	0,983
CLASSIFICACAO_BEBIDA	0,231	0,266	0,099	0,437	0,476	0,553	0,643	0,877	0,879	0,877
EQUIPAMENTO_DESTILACAO	0,317	0,317	0,247	0,317	0,352	0,682	0,952	0,894	0,882	0,929
TEMPO_ARMAZENAMENTO	0,028	0,193	0,314	0,505	0,561	0,608	0,657	0,713	0,912	0,948
RECIPIENTE_ARMAZENAMENTO	0,11	0,45	0,516	0,486	0,526	0,543	0,61	0,893	0,943	0,95
TIPO_MADEIRA	0,174	0,210	0,218	0,549	0,666	0,704	0,793	0,841	0,889	0,931
CARACTERISTICA_SENSORIAL_COR	0,06	0,093	0,133	0,253	0,373	0,453	0,5	0,626	0,746	0,866
CARACTERISTICA_SENSORIAL_AROMA	0,202	0,270	0,245	0,359	0,391	0,441	0,444	0,480	0,519	0,644
CARACTERISTICA_SENSORIAL_SABOR	0,072	0,079	0,072	0,256	0,285	0,310	0,342	0,422	0,472	0,501
CARACTERISTICA_SENSORIAL_CONSISTENCIA	0,082	0,211	0,388	0,270	0,282	0,282	0,682	0,741	0,870	0,894
NOME_PESSOA	0,045	0,162	0,247	0,279	0,414	0,495	0,581	0,693	0,801	0,878
NOME_LOCAL	0,077	0,132	0,509	0,357	0,383	0,413	0,527	0,693	0,763	0,939
NOME_ORGANIZACAO	0,053	0,103	0,176	0,226	0,245	0,305	0,367	0,540	0,559	0,606
TEMPO	0,039	0,176	0,212	0,481	0,564	0,721	0,797	0,838	0,875	0,921
PRECO	0,085	0,156	0,245	0,334	0,434	0,527	0,617	0,802	0,862	0,907
VOLUME	0,088	0,150	0,260	0,324	0,358	0,43	0,550	0,794	0,8	0,873

Fonte: Elaboração própria.

Tabela 5.6 – Resultados de F1, a nível de categoria, do modelo treinado com diferentes porcentagens de entidades rotuladas.

Categorias	**F1-measure**									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
NOME_BEBIDA	0,069	0,176	0,326	0,326	0,475	0,556	0,640	0,683	0,695	0,731
GRADUACAO_ALCOOLICA	0,290	0,358	0,191	0,690	0,811	0,826	0,823	0,836	0,891	**0,951**
CLASSIFICACAO_BEBIDA	0,331	0,356	0,157	0,443	0,474	0,531	0,609	0,700	0,708	0,701
EQUIPAMENTO_DESTILACAO	0,364	0,367	0,333	0,406	0,431	0,610	0,669	0,649	0,646	0,675
TEMPO_ARMAZENAMENTO	0,054	0,313	0,462	0,633	0,681	0,695	0,701	0,740	0,906	0,924
RECIPIENTE_ARMAZENAMENTO	0,186	0,579	0,593	0,609	0,652	0,669	0,712	0,845	0,887	0,887
TIPO_MADEIRA	0,272	0,317	0,345	0,633	0,718	0,753	0,813	0,848	0,893	0,924
CARACTERISTICA_SENSORIAL_COR	0,109	0,151	0,186	0,334	0,397	0,440	0,458	0,564	0,666	0,780
CARACTERISTICA_SENSORIAL_AROMA	0,304	0,322	0,297	0,355	0,378	0,407	0,398	0,414	0,438	0,485
CARACTERISTICA_SENSORIAL_SABOR	0,107	0,116	0,112	0,268	0,290	0,257	0,253	0,278	0,306	0,329
CARACTERISTICA_SENSORIAL_CONSISTENCIA	0,137	0,339	0,519	0,368	0,380	0,363	0,678	0,724	0,813	0,835
NOME_PESSOA	0,084	0,246	0,364	0,326	0,515	0,580	0,661	0,769	0,831	0,859
NOME_LOCAL	0,143	0,231	0,652	0,515	0,540	0,569	0,670	0,794	0,839	0,937
NOME_ORGANIZACAO	0,098	0,180	0,268	0,322	0,345	0,408	0,456	0,601	0,609	0,650
TEMPO	0,074	0,283	0,343	0,636	0,706	0,819	0,868	0,890	0,902	0,933
PRECO	0,154	0,198	0,388	0,393	0,498	0,585	0,664	0,798	0,846	0,880
VOLUME	0,154	0,236	0,384	0,451	0,493	0,576	0,662	0,852	0,847	0,877

Fonte: Elaboração própria.

6 CONCLUSÃO

Neste trabalho, foi apresentado o cachacaNER, um *dataset* de textos escritos em português do Brasil sobre a bebida Cachaça, para a tarefa de Reconhecimento de Entidades Nomeadas. Os textos que compõem esse *dataset* foram coletados em sites de venda de bebidas alcoólicas. Por meio da análise desses textos, foram criadas onze categorias de entidades nomeadas específicas, as quais representam características e aspectos intrínsecos da cachaça, além do levantamento de outras seis categorias genéricas.

Para criação do cachacaNER, os textos foram coletados e extraídos por meio da técnica de Web Scraping, pré-processados, rotulados manualmente e anotados nos formatos IOB2 e do spaCy. Ele também foi estrategicamente divido em dados de treinamento (70% dos dados) e teste (30%), de maneira que é possível reproduzir os experimentos realizados neste trabalho ou realizar novos experimentos e aplicações. Para avaliar a qualidade da rotulação manual, calculou-se o coeficiente de concordância Kappa de Fleiss, por meio do qual se identificou uma concordância quase perfeita, de 0,857.

A partir das estatísticas extraídas do *dataset* proposto, foi possível identificar que os textos desse conjunto de dados possuem tamanhos diferentes, dado que alguns sites de venda de bebidas utilizam mais textos para apresentar o produto, por exemplo, textos com a descrição da bebida, história do alambique, análise do cachacier e ficha técnica, enquanto outros sites apresentam apenas a ficha técnica que possui informações simples e precisas. Essa particularidade influência na distribuição dos dados, por exemplo, na quantidade de *tokens* por sentença (13,42) e por documento (183,01), e de sentenças por documento (13,62).

Com a avaliação do cachacaNER por meio do treinamento de um modelo NER, identificou-se que a nível global o modelo obteve um bom desempenho de 0,889. Descobriu-se também que a variação dos parâmetros do modelo, referentes a quantidade de épocas, *batchs* e tamanho do *drop*, não alteram muito o resultado de desempenho do modelo. No nível de categoria, o modelo obteve maior desempenho em identificar e rotular entidades referentes a categoria GRADUACAO_ALCOOLICA, com F1 de 0,981. A parir de uma análise comparativa com outros *datasets* de NER, identificou-se que o cachacaNER é maior em quantidade de documentos, de sentenças, de *tokens* e de categorias semânticas, do que o pioneiro HAREM.

No experimento de rotulação automática de entidades, foi apresentada uma técnica de identificação do, possível, percentual mínimo de entidades rotuladas que são necessárias em um conjunto de dados de treinamento para treinar um modelo NER. Essa técnica pode auxi-

liar pesquisadores e empresas a terem uma base de quantos dados rotular antes de construir um *dataset* para NER. Além disso, também foi criada uma lista com 2.844 entidades rotuladas, a qual pode ser utilizada por empresas ou pesquisadores em tarefas que envolvam o reconhecimento de entidades nomeadas. Essa lista encontra-se publicamente disponível em <https://github.com/PriscillaIA/cachacaNER/blob/main/lista_de_entidades_rotuladas.txt>.

Este trabalho, portanto, é inovador, pois contribui para o meio acadêmico e empresarial com a criação e disponibilização de um *dataset* para tarefa de NER exclusivo. Exclusivo, porque até onde vai nosso conhecimento, não há na literatura um *dataset* de NER com textos em português sobre cachaça. Além disso, as categorias semânticas de entidades nomeadas também podem ser utilizadas para a rotulação de outros *datasets*, pois neste trabalho também é disponibilizado um manual de rotulação com tais categorias. Para além do *dataset* cachacaNER, das categorias específicas e do manual de rotulação, neste trabalho também foram apresentados valores *baseline*, que podem ser utilizados por outros trabalhos para comparação de desempenho de modelos NER e percentuais de entidades rotuladas que podem servir de referência para definir a porcentagem mínima de entidades rotuladas em um *dataset*.

A cachaça possui categorias de entidades comuns a outros tipos de bebidas. Assim, em trabalhos futuros, pretende-se avaliar o *dataset* cachacaNER na previsão de entidades para outras bebidas, como vinho e cerveja. Esse conjunto de dados também será avaliado para o uso de reconhecimento de entidades em um modelo de recuperação de informação para um motor de busca no domínio específico da cachaça.

REFERÊNCIAS

ALAMMAR, J. **The Illustrated Transformer**. 2018. Acessado em 03 de Novembro de 2022. Disponível em: <http://jalammar.github.io/illustrated-transformer/>.

ALBARED, M.; OCAÑA, M. G.; GHAREB, A.; AL-MOSLMI, T. Recent progress of named entity recognition over the most popular datasets. In: **2019 First International Conference of Intelligent Computing and Engineering (ICOICE)**. Hadhramout, Yemen: IEEE Yemen Subsection, 2019. p. 1–9.

ALBUQUERQUE, H. O.; COSTA, R.; SILVESTRE, G.; SOUZA, E.; SILVA, N. F. F. da; VITÓRIO, D.; MORIYAMA, G.; MARTINS, L.; SOEZIMA, L.; NUNES, A.; SIQUEIRA, F.; TARREGA, J. P.; BEINOTTI, J. V.; DIAS, M.; SILVA, M.; GARDINI, M.; SILVA, V.; CARVALHO, A. C. P. L. F. de; OLIVEIRA, A. L. I. Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In: **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2022. p. 3–14.

ANANDIKA, A.; MISHRA, S. A study on machine learning approaches for named entity recognition. In: **2019 International Conference on Applied Machine Learning (ICAML)**. Los Alamitos, CA, USA: IEEE Computer Society, 2019. p. 153–159. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/ICAML48257.2019.00037>.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v. 5, 01 2009. DOI 10.21529/RESI.2006.0502001.

ARAUJO, P. H. L.; CAMPOS, T. E.; OLIVERIA, R. R. R.; STAUFFER, M.; COUTO, S.; BERMEJO, P. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In: **International Conference on the Computational Processing of Portuguese (PROPOR)**. Canela, RS, Brazil: Springer, 2018. (Lecture Notes on Computer Science (LNCS)), p. 313–323. Disponível em: <https://cic.unb.br/~teodecampos/LeNER-Br/>.

ARCHANA, G.; MANISH, K.; VISHAL, G. Named entity recognition: Applications, approaches and challenges. v. 06, n. 10, 2017.

BAIGANG, M.; YI, F. A review: development of named entity recognition (ner) technology for aeronautical information intelligence. **Artificial Intelligence Review**, 2022. Disponível em: <https://doi.org/10.1007/s10462-022-10197-2>.

BECKER, K. **Introdução à Mineração de Opiniões**. 2013. 1-51 p. Acessado em 28 de agosto de 2020. Disponível em: <https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/>.

BOBBITT, Z. **One-Way vs Two-Way ANOVA: When to Use Each**. 2021. Acessado em 20 de Outubro de 2022. Disponível em: <https://www.statology.org/one-way-vs-two-way-anova/>.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. **Enriching Word Vectors with Subword Information**. 2017. cs.CL.

BORRAGINE, M. d. C. C. **ENVELHECIMENTO DA CACHAÇA COM CIRCULAÇÃO FORÇADA E AERAÇÃO**. 1-92 p. Dissertação (Mestrado) — Universidade Estadual Paulista, 2009. Acessado em 15 de Outubro de 2022. Disponível em: <https://www2.fcfar.unesp.br/Home/Pos-graduacao/AlimentoseNutricao/Michelle_CaiadoME.pdf>.

BORTOLETTO, A. M. **Influência da madeira na qualidade química e sensorial da aguardente de cana envelhecida**. Tese (Doutorado) — Escola Superior de Agricultura "Luis Queiroz", Piracicaba, 2016.

BRASIL. Instrução normativa nº 13, de 29 de junho de 2005. aprova o regulamento técnico para fixação do padrões de identidade e qualidade para aguardente de cana e para cachaça. **Diário Oficial da União, Ministério da Agricultura, Pecuária e Abastecimento**, Brasília, DF, 2005. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/inspecao/produtos-vegetal/legislacao-1/biblioteca-de-normas-vinhos-e-bebidas/instrucao-normativa-no-13-de-29-de-junho-de-2005.pdf>.

Brasil Travel New. **A importância da cachaça para a economia do Brasil**. 2021. Acessado em 24 de Agosto de 2022. Disponível em: <https://brasiltravelnews.com.br/noticias/a-importancia-da-cachaca-para-a-economia-do-brasil/>.

Cachaça Gestor. **A produção da cachaça em 8 passos**. 2022. Acessado em 15 de Outubro de 2022. Disponível em: <https://cachacagestor.com.br/blog/a-producao-da-cachaca-em-8-passos/>.

CRUZ, G. G. d. J. **Dos robôs alienígenas ao deep learning: o que é um modelo transformer?** 2022. Acessado em 03 de Novembro de 2022. Disponível em: <https://www.zup.com.br/blog/modelo-transformer>.

Deep Learning Book. **Deep Learning e a Tempestade Perfeita**. 2020. Acessado em 14 de maio de 2021. Disponível em: <https://www.deeplearningbook.com.br/deep-learning-a-tempestade-perfeita/>.

DEVLIN, J.; CHANG, M.-W. Open sourcing bert: State-of-the-art pre-training for natural language processing. Google IA Blog, 2019. Acessado em 17 de novembro de 2020.

Devopedia. **Named Entity Recognition**. 2020. Acessado em 18 de Outubro de 2022. Disponível em: <https://devopedia.org/named-entity-recognition>.

ESCOVEDO, T. **Machine Learning: Conceitos e Modelos — Parte I: Aprendizado Supervisionado**. 2020. Acessado em 02 de Novembro de 2022. Disponível em: <https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445>.

FELDMAN, S. Nlp meets the jabberwocky: Natural language processing in information retrieval: Search engine section. (online, ed.). Information Today, Incy, p. 273–297, 1999.

FLEISS, J. Measuring nominal scale agreement among many raters. **Psychological bulletin**, v. 76, n. 5, p. 378—382, November 1971. ISSN 0033-2909. Disponível em: <https://doi.org/10.1037/h0031619>.

GARCIA, G. C. **Reconhecimento de Entidades Nomeadas na base de notificações de eventos adversos e queixas técnicas de dispositivos médicos no Brasil**. 1-158 p. Dissertação (Mestrado) — Universidade Federal de Brasília, Brasília, 2021. Acessado em 12 de Outubro de 2022. Disponível em: <https://repositorio.unb.br/bitstream/10482/42718/1/2021_GustavoCunhaGarcia.pdf>.

GOYAL, A.; GUPTA, V.; KUMAR, M. Recent named entity recognition and classification techniques: A systematic review. **Computer Science Review**, v. 29, p. 21–43, 2018.

ISSN 1574-0137. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1574013717302782>.

GRISHMAN, R.; SUNDHEIM, B. Message understanding conference-6: A brief history. In: **Proceedings of the 16th Conference on Computational Linguistics - Volume 1**. USA: Association for Computational Linguistics, 1996. (COLING '96), p. 466–471. Disponível em: <https://doi.org/10.3115/992628.992709>.

GUO, Q.; WANG, S.; WAN, F. **Research on Named Entity Recognition for Information Extraction**. 2020. 121-124 p.

HONNIBAL, M.; MONTANI, I. spaCy 3: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

IBRAC. **MERCADO EXTERNO**. 2021. Acessado em 09 de Outubro de 2022. Disponível em: <https://ibrac.net/servicos/mercado-externo>.

INMETRO. Portaria nº 276 de 24 de setembro de 2009. aprova a revisão dos requisitos de avaliação da conformidade para cachaça. **Diário Oficial da União, Ministério do Desenvolvimento, Indústria e Comércio Exterior**, Rio de Janeiro, RJ, 2009. Disponível em: <www.inmetro.gov.br>.

Instituto Brasileiro da Cachaça. **A Cachaça no Brasil - Anuário da Cachaça 2021**. 2021. Acessado em 24 de Agosto de 2022. Disponível em: <https://ibrac.net/servicos/cartilhas>.

JAPKOWICZ, N.; SHAH, M. **Evaluating Learning Algorithms: A Classification Perspective**. Cambridge University Press, 2011. ISBN 9781139494144. Disponível em: <https://books.google.com.br/books?id=VoWIIOKVzR4C>.

JING, L.; AIXIN, S.; JIANGLEI, H.; CHENLIANG, L. A survey on deep learning for named entity recognition. **CoRR**, abs/1812.09449, 2018. Disponível em: <http://arxiv.org/abs/1812.09449>.

JUNIOR, C.; MACEDO, H.; BISPO, T.; OLIVEIRA, F.; SILVA, N.; BARBOSA, L. . Paramopama: a brazilian-portuguese corpus for named entity recognition. 12th National Meeting on Artificial and Computational Intelligence (ENIAC), p. 6, 2015. Disponível em: <https://github.com/davidsbatista/NER-datasets/blob/master/Portuguese/Paramopama/Paramopama.pdf>.

K., J.; SAINI, J. Stop-word removal algorithm and its implementation for sanskrit language. **International Journal of Computer Applications**, v. 150, p. 15–17, 09 2016.

KATUMULLAGE, D.; YANG, C.; BARTH, J.; CAO, J. Using neural network models for wine review classification. **Journal of Wine Economics**, Cambridge University Press, v. 17, n. 1, p. 27–41, 2022.

KIKABEN. **Transformer's Encoder-Decoder: Let's Understand The Model Architecture**. 2021. Acessado em 03 de Novembro de 2022. Disponível em: <https://kikaben.com/transformers-encoder-decoder/>.

KUPERUS, J.; VEENMAN, C. J.; KEULEN, M. van. Increasing ner recall with minimal precision loss. **2013 European Intelligence and Security Informatics Conference**, p. 106–111, 2013.

LACERDA, M. **Como é feita a cachaça? Conheça o processo de criação desta bebida histórica**. 2018. Acessado em 15 de Outubro de 2022. Disponível em: <https://super.abril.com.br/historia/como-e-feita-a-cachaca/>.

LAMPLE, G.; BALLESTEROS, M.; SUBRAMANIAN, S.; KAWAKAMI, K.; DYER, C. Neural architectures for named entity recognition. In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. San Diego, California: Association for Computational Linguistics, 2016. p. 260–270. Disponível em: <https://aclanthology.org/N16-1030>.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, International Biometric Society, v. 33, n. 1, 1977.

LEFEVER, E.; HENDRICKX, I.; CROIJMANS, I.; BOSCH, A. van den; MAJID, A. Discovering the language of wine reviews: A text mining account. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)**. European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1521>.

LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. Fine-grained named entity recognition in legal documents. In: ACOSTA, M.; CUDRÉ-MAUROUX, P.; MALESHKOVA, M.; PELLEGRINI, T.; SACK, H.; SURE-VETTER, Y. (Ed.). **Semantic Systems. The Power of AI and Knowledge Graphs**. Cham: Springer International Publishing, 2019. p. 272–287.

MAPA. A cachaça no brasil: dados de registro de cachaças e aguardentes ano 2021. Ministério da Agricultura, Pecuária e Abastecimento, 2021. Disponível em: <https://ibrac.net/public/uploads/cartilhas/162551157860e3569aed45f.pdf>.

Mapa da Cachaça. **As diferentes formas de se destilar cachaça**. 2021. Acessado em 16 de Outubro de 2022. Disponível em: <https://www.mapadacachaca.com.br/artigos/as-diferentes-formas-de-se-destilar-cachaca/>.

Mapa da Cachaça. **Os tipo de cachaça**. 2021. Acessado em 17 de Outubro de 2022. Disponível em: <https://www.mapadacachaca.com.br/os-tipos-de-cachaca/>.

MELO, T. de; FIGUEIREDO, C. M. S. Comparing news articles and tweets about covid-19 in brazil: Sentiment analysis and topic modeling approach. **JMIR Public Health Surveill**, v. 7, n. 2, p. e24585, Feb 2021. ISSN 2369-2960. Disponível em: <http://publichealth.jmir.org/2021/2/e24585/>.

MENEGHIN, M. C.; BARBOZA, R. A. B. **Requisitos Legais para a Produção da Cachaça**. São Paulo, 2014.

MIKOLOV, T.; CHEN, K.; CORRADO, G. S.; DEAN, J. Efficient estimation of word representations in vector space. **CoRR**, 2013.

MIKOLOV, T.; KARAFIáT, M.; BURGET, L.; CERNOCKY, J.; KHUDANPUR, S. Recurrent neural network based language model. ISCA, p. 1045–1048, 2010.

MORENO, A. C. **Análise de Sentimentos na Classificação de Comentários Online Aplicando Técnicas de Text Mining**. 1-72 p. Dissertação (Mestrado) — Instituto Universitário de Lisboa, 2015. Acessado em 01 de agosto de 2020. Disponível em: <https://repositorio.iscte-iul.pt/bitstream/10071/11504/1/ACMoreno_MSIAD_F.pdf>.

MOTA, C.; SANTOS, D. Book. **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. 1º. ed. Brasil: Linguateca, 2008. ISBN 978-989-20-1656-6. Disponível em: <https://www.linguateca.pt/HAREM/actas/Livro-MotaSantos2008.pdf>.

MUJTABA, H. **Tokenising into Words and Sentences | What is Tokenization and it's Definition?** 2020. Acessado em 28 de agosto de 2020. Disponível em: <https://www.mygreatlearning.com/blog/tokenization/>.

NEU, D. A.; LAHANN, J.; FETTKE, P. A systematic literature review on state-of-the-art deep learning methods for process prediction. **Artificial Intelligence Review**, 2021. https://doi.org/10.1007/s10462-021-09960-8.

NOTHMAN, J.; RINGLAND, N.; RADFORD, W.; MURPHY, T.; CURRAN, J. R. Learning multilingual named entity recognition from wikipedia. **Artificial Intelligence**, v. 194, p. 151–175, 2013. ISSN 0004-3702. Artificial Intelligence, Wikipedia and Semi-Structured Resources. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0004370212000276>.

OLIVEIRA, A. M. L. d. **O PROCESSO DE PRODUÇÃO DA CACHAÇA ARTESANAL E SUA IMPORTÂNCIA COMERCIAL**. 1-55 p. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, 2010. Acessado em 15 de Outubro de 2022. Disponível em: <https://repositorio.ufmg.br/bitstream/1843/BUOS-99VGVE/1/monografia_ana_marcia_2011_2.pdf>.

OLIVEIRA, A. M. L. d. **PATRIMÔNIO, HISTÓRIA E CULTURA DA CACHAÇA TABOA EM BONITO/MS: PERSPECTIVAS DE DESENVOLVIMENTO LOCAL**. 1-114 p. Dissertação (Mestrado) — Universidade Católica Dom Bosco, 2016. Acessado em 15 de Outubro de 2022. Disponível em: <https://site.ucdb.br/public/md-dissertacoes/22618-dissertacao-beatriz-carlini-garcia-de-oliveira.pdf>.

OLIVEIRA, L. **Um guia abrangente sobre NLP - Processamento de texto**. 2020. Acessado em 28 de agosto de 2020. Disponível em: <https://medium.com/@lucasoliveiras/um-guia-abrangente-sobre-nlp-processamento-de-texto-60b852125202>.

PALMER, J.; CHEN, B. Wineinformatics: Regression on the grade and price of wines through their sensory attributes. **Fermentation**, v. 4, n. 4, 2018. ISSN 2311-5637. Disponível em: <https://www.mdpi.com/2311-5637/4/4/84>.

PALSHIKAR, G. Techniques for named entity recognition: A survey. **Bioinformatics: Concepts, Methodologies, Tools, and Applications**, v. 1, p. 191–, 01 2012.

PARK, J.; RYU, Y. U. Online discourse on fibromyalgia: Text-mining to identify clinical distinction and patient concerns. **Medical Science Monitor**, v. 20, p. 1858 – 1864, 2014. ISSN 1643-3750. DOI 10.12659/MSM.890793.

PENNINGTON, J.; SOCHER, R.; MANNING, C. 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, 2014. p. 1532–1543. Disponível em: <https://www.aclweb.org/anthology/D14-1162>.

PERES, R.; ESTEVES, D.; MAHESHWARI, G. Bidirectional lstm with a context input window for named entity recognition in tweets. In: **Proceedings of the Knowledge Capture Conference**. Nova York, NY, EUA: Association for Computing Machinery, 2017. p. 4. ISBN 9781450355537. Disponível em: <https://doi.org/10.1145/3148011.3154478>.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. 2018.

PINHEIRO, S. H. d. M. **Avaliação sensorial das bebidas aguardente de cana industrial e cachaça de alambique**. Tese (Doutorado) — Universidade Federal de Viçosa, Viçosa, 2010.

PIRES, A. R. O. P. **Named entity extraction from Portuguese web text**. 1-104 p. Dissertação (Mestrado) — Faculdade de Engenharia da Universidade de Porto, 2017. Disponível em: <https://repositorio-aberto.up.pt/handle/10216/106094>.

POWERS, D.; TURK, C. **Machine Learning of Natural Language**. London: Springer-Verlag, 1989. 385 p. ISBN 978-1-4471-1697-4.

RAMSHAW, L. A.; MARCUS, M. P. Text chunking using transformation-based learning. arXiv, 1995. Disponível em: <https://arxiv.org/abs/cmp-lg/9505040>.

SANG, E. F. T. K.; MEULDER, F. D. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. p. 142–147, 2003. Disponível em: <https://aclanthology.org/W03-0419>.

SARKAR, D.; BALI, R.; GHOSH, T. **Hands-On Transfer Learning with Python: Implement Advanced Deep Learning and Neural Network Models Using TensorFlow and Keras**. [S.l.]: Packt Publishing, 2018. ISBN 1788831306.

SEBRAE. **Produção de cachaça no Brasil ainda tem muito potencial econômico**. 2022. Acessado em 09 de Outubro de 2022. Disponível em: <https://www.sebrae.com.br/sites/PortalSebrae/artigos/producao-de-cachaca-no-brasil-ainda-tem-muito-potencial-economico,578ed967936ef710VgnVCM100000d701210aRCRD#:~:text=Em%202021%20foram%20exportados%207,52%25%20em%20volume%20em%202021.>

SHELLEY, K.; JERRY, W.; J.ROBERT, B. **Natural language processing and text mining**. London: Springer-Verlag London, 2007. 1 - 272 p. ISBN 978-1-84628-175-4.

SILVA, R. A.; SILVA, L.; DUTRA, M. L.; ARAUJO, G. M. An improved ner methodology to the portuguese language. **The Journal of Supercomputing**, v. 26, p. 319–325, 2021.

SIMON, A.; DEO, M.; SELVAM, V.; BABU, R. An overview of machine learning and its applications. **International Journal of Electrical Sciences Engineering**, Volume, p. 22–24, 01 2016.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. **Attention Is All You Need**. arXiv, 2017. Disponível em: <https://arxiv.org/abs/1706.03762>.

WANG, H.; JIANG, M.; QI, J.; ZHANG, X.; WANG, Q.; ZHOU, Y.; BAI, M.; LIU, L.; PEI, Z. Application of deep learning in text mining. In: **International Conference on Mechatronics, Control and Electronic Engineering (MCE 2014)**. Shenyang, China: Atlantis Press, 2014.

ZHOU, S.; RUECKERT, D.; FICHTINGER, G. **Handbook of Medical Image Computing and Computer Assisted Intervention**. Elsevier Science, 2019. (The MICCAI Society book Series). ISBN 9780128165867. Disponível em: <https://books.google.com.br/books?id=aKO3DwAAQBAJ>.

ZHU, W.; MA, Y.; ZHOU, Y.; BENTON, M.; ROMAGNOLI, J. Deep learning based soft sensor and its application on a pyrolysis reactor for compositions predictions of gas phase components. **Computer Aided Chemical Engineering**, p. 2245–2250, 01 2018.

APÊNDICE A – Documento com Diretrizes para Rotulação Manual

As palavras sublinhadas se referem as entidades que devem ser rotuladas como pertencentes à categoria de entidade em questão, já as palavras em negrito são as que não devem ser rotuladas como parte da categoria.

i. **NOME_BEBIDA:** esta categoria se refere ao nome comercial da bebida.

- Exemplo 1: **Cachaça** Princesa Isabel **Blend Bálsamo** e **Jaqueira**.
  A palavra cachaça não foi rotulada porque se refere ao tipo da bebida não ao nome comercial. Blend, Bálsamo e Jaqueira fazem parte das categorias Classificação da Bebida e Tipo de Madeira, respectivamente, por isso também não foram rotuladas.
- Exemplo 2: **Cachaça** Vanderley Azevedo **600 ml 12 anos**.
  600 ml e 12 anos representam às categorias Volume e Tempo de Armazenamento.
- Exemplo 3: **Cachaça** de **Alambique** Carvalheira **750ml**.
  Alambique se refere ao equipamento ou processo de destilação em que a bebida é preparada.
- Exemplo 4: **Cachaça** Rapadura com Banana **de Minas 500ml**. Minas se refere ao local onde a bebida foi produzida, desta maneira pertence à categoria Local.

ii. **GRADUACAO_ALCOOLICA:** corresponde a quantidade de álcool que a bebida possui.

- Exemplo 1: A Linha Top 45 foi desenvolvida para paladares mais exigentes. Com grau alcoólico de 45%, veio para satisfazer os amantes de uma cachaça mais forte.
- Exemplo 2: A Cachaça Pinga ni mim é armazenada por 3 anos em barris de carvalho europeu, e possui 38 por cento de graduação alcoólica.
- Exemplo 3: Ficha Técnica da cachaça Ypióca: graduação alcoólica - 17% vol.; garrafa - 500ml; local de produção - Ouro Preto/MG.

iii. **CLASSIFICACAO_BEBIDA:** A cachaça pode ser classificada de acordo com os seguintes nomes: Clássica, Tradicional, Prata, Branca, Ouro, Amarela, Premium, Extra Premium, Super Premium e Reserva Especial, os quais foram identificados nos textos coletados.

- Exemplo 1: Cachaça Dom Bré Clássica Jequitibá 700ml.
- Exemplo 2: A Sagatiba Pura é uma cachaça premium com sabor suave, 38% vol., aroma agradável e aspecto cristalino ideal para a preparação de *drinks*.

iv. **EQUIPAMENTO_DESTILACAO:** equivale ao equipamento/aparelho utilizado no processo de destilação da cachaça.

- Exemplo 1: A destilação da Sanhaçu ocorre em alambiques de cobre, o que lhe propicia características especiais de aroma e sabor.
- Exemplo 2: A Cachaça Carvalheira Branca é um *blend* fantástico de cachaça produzida em coluna de aço.
- Exemplo 3: Produzida no **Alambique** Taverna de Minas, a cachaça Legítima de Minas é um reflexo da essência mineira deixada pelos tropeiros na Estrada.

  No exemplo 3, alambique não foi rotulado porque se refere a parte do nome da empresa.

v. **TEMPO_ARMAZENAMENTO:** tempo em que a bebida fica armazenada ou envelhecendo, antes de ser distribuída ou consumida.

- Exemplo 1: A Cachaça Cabaré é armazenada por 15 anos em tonéis de carvalho europeu e envasada em garrafa de *design* francês.
- Exemplo 2: A Cachaça Moreninha é armazenada durante nove meses em dornas de aço.

vi. **RECIPIENTE_ARMAZENAMENTO:** recipiente no qual a bebida é depositada para ser armazenada ou envelhecida.

- Exemplo 1: Envelhecida em seculares paróis de Jatobá, conserva o puro sabor da cana e uma sutil coloração da madeira.
- Exemplo 2: A cachaça Saracura Prata representa a cachaça *In natura*, e é descansada um ano em tonéis de bálsamo.
- Exemplo 3: A Cachaça Vale do Cedro 700ml é armazenada por no mínimo 5 anos em barris de madeira com capacidade para 200 litros.

vii. **TIPO_MADEIRA:** nome da madeira utilizada para a confecção do recipiente de armazenamento.

- Exemplo 1: As bebidas armazenadas por 3 anos em tonéis de bálsamo adquirem tonalidade e aroma característico deste tipo de madeira.
- Exemplo 2: A Cachaça Sanhaçu Amburana é proveniente de Chã Grande em Pernambuco. Ela é produzida em alambique e armazenada em tonel de amburana.

Nota explicativa: as características sensoriais levantadas nesta pesquisa permitem a identificação de menções relacionadas ao que se pode ver e sentir durante a degustação da bebida.

Observação 1: Palavras como coloração, cor, aroma, sabor, textura e consistência não deverão ser rotuladas, pois não são aspectos ou características sensoriais que descrevem a bebida de fato.

Observação 2: Termos que qualificam ou atribuem intensidade à bebida deverão ser rotulados apenas quando acompanharem uma entidade, pois o objetivo é identificar os termos que caracterizam os aspectos sensoriais da bebida, não apenas as palavras que lhe atribuem qualificação. Alguns desses termos são: rico, elegante, fina, agradável, intenso, equilibrado, suave e leve.

viii. **CARACTERISTICA_SENSORIAL_COR:** coloração que a bebida apresenta.

- Exemplo 1: Uma bela cachaça, com **coloração** amarelo-esverdeado, límpida, brilhante a muito encorpada.
- Exemplo 2: Um *blend* de três tostas diferentes de barris de carvalho americano (3 AOB – American Oak Barrel) que lhe confere a **cor** dourada alaranjada.

ix. **CARACTERISTICA_SENSORIAL_AROMA:** aroma exalado ou sentido em relação à bebida.

- Exemplo 1: A cachaça Três Coronéis é produzida em alambique e armazenada em amburana, o que lhe proporciona notas de canela com especiarias.
- Exemplo 2: Feita em alambique de cobre e envelhecida em tonéis novos de bálsamo, a Cachaça Porto Estrela Ouro ganha um toque especial em sua finalização ao ser armazenada em dornas de Jaqueira. Assim, temos um *blend* de aromas ricos e equilibrados que além das características herbais habituais e do aroma de anis marcante é enriquecido com o perfume floral e de frutas amarelas, proveniente da Jaqueira.

x. **CARACTERISTICA_SENSORIAL_SABOR:** sabor sentido na boca durante a ingestão da bebida.

- Exemplo 1: A Cachaça Formosinha possui baixa acidez, é levemente picante e adocicada, lembrando um pouco o sabor do gengibre. O retrogosto é agradável e inusitado de cravo.
- Exemplo 2: Essa cachaça possui uma das menores graduações alcoólicas entre as bebidas envelhecidas em bálsamo, além de possuir **sabor equilibrado**.

  Sabor e equilibrado não são rotulados neste caso porque não representam um sabor de fato, por exemplo, baunilha ou cravo.

xi. **CARACTERISTICA_SENSORIAL_CONSISTÊNCIA:** consistência ou textura percebida em relação à bebida.

- Exemplo 1: Essa é uma bela e deliciosa cachaça que possui uma coloração amarelo-esverdeado, e uma textura muito macia.
- Exemplo 2: A Cachaça Companheira Castanheira é proveniente do estado do Paraná, possui uma textura aveludada, e é levemente adocicada.

xii. **NOME_PESSOA:** são considerados como pessoas as menções de nomes próprios e apelidos que correspondam a um ser humano.

  – Exemplo: A Cachaçaria Nacional é a maior loja de cachaças *on-line* do mundo sendo fundada em 25 de janeiro de 2010. Idealizada por Rafael Araújo e Marcos Paolinelli, tem o objetivo de difundir e democratizar o consumo da Cachaça, a bebida genuinamente brasileira.

xiii. **NOME_LOCAL:** menções nos textos que podem ser traduzidas como um local geográfico.

Observação 1: Quando os nomes de lugares aparecerem juntos no texto, eles deverão ser rotulados separadamente, por exemplo: nome de município e estado (Iterava-MG); estado e país (Minas Gerais/BR); município, estado e país (Lavras-MG/Brasil).

Observação 2: Abreviações também deverão ser rotuladas como nome de local, por exemplo: BR, MG e PA.

- Exemplo: A Cachaçaria Nacional oferece mais de 2000 rótulos de cachaças artesanais de alambiques das principais regiões produtoras do Brasil. Tam-

bém comercializa dornas, barris e queijos produzidos por pequenos produtores de Iterava - MG.

xiv. **NOME_ORGANIZAÇÃO:** entidade que possui vida própria, tendo uma administração própria e que não é caracterizada como pessoa.

- Exemplo 1: A empresa Weber Haus é administrada por Evandro Weber, descendente de alemães de Ivoti, encosta da Serra Gaúcha, interior do Grande do Sul.
- Exemplo 2: A Cachaça Porto do Vianna Premium é produzida pelo grupo Gouveia Brasil.

xv. **TEMPO:** são menções que podem ser traduzidas como a representação do tempo. **Observação:** Quando as entidades que representam tempo aparecerem juntas deverão ser rotuladas separadamente.

- Exemplo 1: A empresa Cachaça Batista, engarrafou sua primeira cachaça em 01/05/1992 às 10:32hr.
- Exemplo 2: A Cachaça Weber Haus Rota 48 é uma homenagem da destilaria Weber Haus à Rota 48, um trajeto muito utilizado por tropeiros no século XIX.

xvi. **PREÇO:** se refe a valores monetários.

- Exemplo 1: A Cachaça Reserva do Gerente, pode ser encontrada no mercado com o valor médio de R$ 350,00.
- Exemplo 2: Obviamente essa é uma pinga gostosa e que vale a pena pagar 62 reais.

xvii. **VOLUME:** menções no texto que se referem a capacidade em volume de algo.

  - Exemplo 1: A Prosa Mineira Ouro 500ml é uma cachaça envelhecida por dois anos em barris de 200 litros.

ANEXO A – Atributos Sensoriais da Cachaça.

Tabela 1 – Atributos sensoriais da cachaça.

ATRIBUTOS	DESCRIÇÃO
APARÊNCIA	
Coloração	Cor característica de corante caramelo.
SENSAÇÃO NASAL	
Irritante	Impacto irritante e agressivo na mucosa nasal.
Pungente	Sensação de ardor e queimação na cavidade nasal.
AROMA	
Alcoólico	Aroma característico do etanol.
Adocicado	Aroma característico de soluções adocicadas.
Floral	Aroma que lembra flores.
Amadeirado	Aroma característico da madeira do carvalho utilizado no tonel para envelhecimento das aguardentes.
Caldo de cana	Aroma de caldo de cana que se evapora ao ser aquecido, diluído em álcool.
Baunilha	Aroma exalado por uma solução alcoólica de baunilha.
Cítrico	Aroma característico de frutas cítricas.
SENSAÇÃO NA BOCA	
Ardência	Sensação ardente percebida na língua e na garganta.
Agressividade	É o impacto agressivo de sabor inicial.
Adstringência	Sensação de secura na mucosa oral, semelhante àquela causada de forma intensa por certas frutas verdes, como banana.
Pungente	Sensação de ardor/queimação na cavidade oral
SABOR	
Doce	Gosto doce percebido pelas papilas gustativas.
Alcoólico	Sabor característico de soluções alcoólicas.
Amadeirado	Sabor característico promovido pela madeira do carvalho utilizado no tonel para envelhecimento das aguardentes.
Ácido	Gosto ácido característica de frutas cítricas.
Amargo	Gosto amargo , característico de cafeína.
Caldo de cana	Sabor proveniente do caldo de cana fervido presente em base alcoólica.
Frutal	Sabor que lembra frutas tropicais.
Floral	Sabor que lembra flores.
Sulfuroso	Sabor característico de ovo cozido.
Caramelo	Sabor característico de açúcar caramelizado.
Toffe	Sabor que lembra bala toffe.
Cítrico	Sabor associado a frutas cítricas.
SABOR RESIDUAL	
Doce	Gosto doce que permanece por um certo período na boca após a ingestão de uma determinada substância.
Alcoólico	Sabor de álcool que permanece por um certo período na boca após a ingestão de uma determinada substância.
Amadeirado	Sabor característico da madeira do carvalho que permanece por um período de tempo após a ingestão de uma determinada substância.

Fonte: Adaptado de Pinheiro (2010).