UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

BÁRBARA STÉPHANIE NEVES OLIVEIRA

APRENDIZADO PROFUNDO PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM NARRATIVAS DE ROUBOS

QUIXADÁ

2020

BÁRBARA STÉPHANIE NEVES OLIVEIRA

APRENDIZADO PROFUNDO PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM NARRATIVAS DE ROUBOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. Regis Pires Magalhães

Coorientadora: Profª. Drª. Ticiana Linhares Coelho da Silva

QUIXADÁ

2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O45a Oliveira, Bárbara Stéphanie Neves.
Aprendizado Profundo para Reconhecimento de Entidades Nomeadas em Narrativas de Roubos / Bárbara Stéphanie Neves Oliveira. – 2020.
99 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Ciência da Computação, Quixadá, 2020.
Orientação: Prof. Dr. Regis Pires Magalhães.
Coorientação: Profa. Dra. Ticiana Linhares Coelho da Silva.

1. Criminologia. 2. Reconhecimento de Entidade Nomeada. 3. Aprendizagem Profunda. 4. Mineração de Dados. 5. Semântica distribucional. I. Título.

CDD 004

BÁRBARA STÉPHANIE NEVES OLIVEIRA

APRENDIZADO PROFUNDO PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM NARRATIVAS DE ROUBOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: __/__/____.

BANCA EXAMINADORA

Prof. Dr. Regis Pires Magalhães (Orientador)
Universidade Federal do Ceará (UFC)

Profª. Drª. Ticiana Linhares Coelho da Silva (Coorientadora)
Universidade Federal do Ceará (UFC)

Profª. Drª. Maria Viviane de Menezes
Universidade Federal do Ceará (UFC)

Para todas e todos que amo, e para os que acreditam que a ciência possa transformar em uma sociedade melhor esta que recebemos.

AGRADECIMENTOS

Agradeço, inicialmente, aos meus pais, Marlene e Benedito, por tudo. Não sou nada sem o amor e incentivo de vocês.

Meus sinceros agradecimentos aos meus excelentes orientadores, Prof. Dr. Regis Pires Magalhães e Profª. Drª. Ticiana Linhares Coelho da Silva, pela oportunidade, conselhos, zelo, ensinamentos e, sobretudo, por se preocuparem com o meu bem-estar. Regis, agradeço por ter atuado ativamente para melhorar o trabalho e transmitir conhecimento. Ticiana, você é uma pessoa extraordinária; agradeço de coração pela amizade e por acreditar no meu potencial.

Agradeço à Profª. Maria Viviane de Menezes pela disponibilidade em participar da banca, pelas grandes contribuições e sugestões no trabalho, pela amizade, e por todo o apoio que me forneceu durante a graduação.

Agradeço aos meus demais professores e aos servidores da UFC – Quixadá, em especial ao Prof. Paulo de Tarso Guerra Oliveira, pela grande ajuda, cuidado e aprendizado.

Sou muito grata às queridas Bárbara Sheyla, Waleska Farias, Débora Reche, Isabel Lima, Brianna Neves e Bruna Neves, pela nossa amizade eterna, e por estarem comigo sempre.

Aos amigos que conquistei nesta jornada acadêmica, sou grata ao Lucas Benjamim, Michel de Melo, Dieinison Jack e à Darliene Ferreira, pela amizade e pelos momentos (estressantes e engraçados) que passamos juntos. Vocês foram de extrema importância durante os lapsos de incerteza.

Também agradeço aos meus colegas do *Insight Data Science Lab*, Andreza Fernandes, Ulisses Silva e Victor Vieira, por todo o suporte necessário neste semestre e ano difíceis.

Por fim, agradeço ao meu namorado Ícaro Farias, pelo apoio incondicional, (muita) paciência, e por sempre me fazer sentir corajosa o suficiente para que eu acredite em mim mesma.

"A fala humana é como uma chaleira rachada em que batemos ritmos grosseiros para os ursos dançarem, enquanto ansiamos por produzir uma música que derreta as estrelas."

(Gustave Flaubert, *Madame Bovary* – 1857)

RESUMO

Diferentes modalidades de violência têm revelado espiral acentuada de crescimento no mundo inteiro. Desde a primeira década do século XXI, o governo federal brasileiro e alguns governos estaduais vêm promovendo novas políticas públicas no combate à criminalidade violenta. Com o objetivo de acompanhar a evolução dos crimes e da violência, bem como promover o acesso da população às informações da segurança pública no estado do Ceará, a Secretaria da Segurança Pública e Defesa Social apresenta mensalmente suas estatísticas. Os dados utilizados para construção desses levantamentos são separados em diferentes indicadores de desempenho para análise criminal. Um dos indicadores-chave, chamado de Crimes Violentos Contra o Patrimônio (CVP), engloba todos os crimes classificados como roubo, exceto o roubo seguido de morte. As taxas de ocorrências para esse indicador são bastante altas e não vem diminuindo ao longo do tempo. Encontrar e divulgar informações relevantes e oportunas em narrativas CVP é crucial para a população, e pode desempenhar um papel central nas capacidades de combate aos tipos de roubos. No entanto, analisar esse grande volume de dados requer um trabalho manual e extenso. Uma forma de automatizar esse processo é utilizar a tarefa de Reconhecimento de Entidade Nomeada (NER) para extração da informação. Nos últimos anos, o Aprendizado Profundo, potencializado por representações vetoriais, tem sido amplamente empregado em sistemas NER. Este trabalho propõe a criação de um modelo NER com técnicas de Aprendizado Profundo capaz de reconhecer entidades nomeadas em textos de roubos, seguindo três questões de pesquisa principais. As questões de pesquisa definem um protocolo experimental abrangente, e exploram aspectos problemáticos presentes no domínio, como o desbalanceamento de dados, quantidade de entidades nomeadas, e representação vetorial eficiente do vocabulário (*CVP2Vec*). Os resultados demonstram alto desempenho, com a métrica Acurácia Balanceada alcançando $88,1\%$ e o F_1-*score* com $82,9\%$ para os melhores modelos, todos superando o *baseline*.

Palavras-chave: Domínio criminal. Processamento de Linguagem Natural. Reconhecimento de Entidade Nomeada. Aprendizado Profundo. Desbalanceamento de Dados. Word embeddings.

ABSTRACT

Different forms of violence have shown a marked growth spiral worldwide. Since the first decade of the 21st century, the Brazilian federal government and some state governments have been promoting new policies to combat violent crime. To monitor the evolution of crimes and violence, as well as promote the population's access to public security information in the state of Ceará (Brazil), the Secretariat of Public Security and Social Defense presents its statistics monthly. The data used to collect these statistics have different performance indicators for criminal analysis. One of the key-indicators, called Violent Crimes Against Patrimony (CVP in the Portuguese acronym), encompasses all crimes classified as theft, except robbery-homicide. The occurrence rates for this indicator are high and have not been decreasing over time. Finding and disseminating relevant and timely information in CVP Police reports is crucial for the population and play a central role in the capabilities to combat types of theft. However, analyzing this large volume of data requires manual and extensive work. One way to automate this process is to use the Named Entity Recognition (NER) task for information extraction. In recent years, Deep Learning, empowered by continuous real-valued vector representations, has been widely employed in NER systems. This work proposes the creation of a Deep Learning based NER model capable of recognizing named entities in CVP Police reports, following three main research questions. The research questions define a broad experimental protocol and explore problematic aspects present in the domain, such as data imbalance, number of named entities, and embedding representation (*CVP2Vec*). The results show high performance, with the Balanced Accuracy metric reaching 88.1% and the F_1-score with 82.9% for the best models, outperforming the baseline.

Keywords: Police reports. Natural Language Processing. Entity Named Recognition. Deep Learning. Data Imbalance. Word embeddings.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

SSPDS	Secretaria da Segurança Pública e Defesa Social
GEESP/SUPESP	Gerência de Estatística e Geoprocessamento
SIP/SIP3W	Sistema de Informações Policiais
CVP	Crimes Violentos Contra o Patrimônio
HNERD	*Human Named Entity Recognition with Deep Learning*
CVLI	Crimes Violentos Letais e Intencionais
AIS	Área Integrada de Segurança
NER	*Named Entity Recognition*
PLN	Processamento de Linguagem Natural
EN	Entidade Nomeada
CB	*Class-Balanced*
CCE	*Categorical Cross-Entropy*
DL	*Dice Loss*
RNA	Rede Neural Artificial
POS *tagging*	*part-of-speech tagging*
GloVe	*Global Vectors for Word Representation*
CBOW	*Continuous Bag of Words*
OOV	*out-of-vocabulary*
C*Window*	*Continuous Window*
NILC	Núcleo Interinstitucional de Linguística Computacional
RNN	*Recurrent Neural Networks*
LSTM	*Long Short-Term Memory*
BLSTM	*Bidirectional Long Short-Term Memory*
CNN	Redes Neurais Convolucionais, tradução de *Convolutional Neural Networks*
MLP	Rede *Multilayer Perceptron*
CRF	*Conditional Random Fields*
HITL	*Human in the Loop*
AB	Acurácia Balanceada
SVM	*Support Vector Machine*
NB	*Naïve Bayes*

LER	*Legal Entity Recognition*
CLM	Compreensão de Leitura de Máquina
BO	boletim de ocorrência
TCO	termo circunstanciado de ocorrência
BERT	*Bidirectional Encoder Representations from Transformers*
ELMo	*Embeddings from Language Models*

SUMÁRIO

1 INTRODUÇÃO

O debate sobre criminalidade violenta vem mobilizando uma série de estudos e pesquisas que visam entender este fenômeno social. Segundo Ayres (1998), as cidades estão cada vez mais inovando com programas de prevenção ao crime e à violência. Entretanto, em muitas partes do mundo, existem poucos incentivos e capacidade limitada para que os governos locais desempenhem um papel mais ativo na melhoria da segurança da população.

No Brasil, as ciências sociais se dedicaram ao estudo do crescimento do crime e da violência apenas durante a transição para a democracia ao longo dos anos 1980 (CEARÁ, 2019). Apesar dos reduzidos estudos e dados criminais sobre o período do regime militar, o primeiro registro do aumento das taxas criminais foi iniciado ainda durante esta época. Mais especificamente, de acordo com Adorno e Pasinato (2010), desde o último quartel do século XX e início do século XXI, é que vem crescendo o crime contra o patrimônio e contra a pessoa, associados ou não às formas organizadas de criminalidade, a par de graves violações de direitos humanos.

A evolução dos crimes e da violência estimula a difusão de sentimentos coletivos de medo e insegurança diante da falta de proteção de direitos fundamentais, como o direito à vida, à livre circulação das pessoas nos espaços públicos, e à posse privada de bens patrimoniais (ADORNO; PASINATO, 2010). A fim de acompanhar a evolução desse fenômeno, bem como promover o acesso público e irrestrito às informações referentes à segurança pública no estado do Ceará, a Secretaria da Segurança Pública e Defesa Social (SSPDS), por intermédio da Gerência de Estatística e Geoprocessamento (GEESP/SUPESP), apresenta mensalmente suas estatísticas em seu *site* oficial[1]. Os dados utilizados para construção desses levantamentos são oriundos do Sistema de Informações Policiais (SIP/SIP3W), e separados em diferentes indicadores de desempenho para análise criminal (SSPDS, 2020).

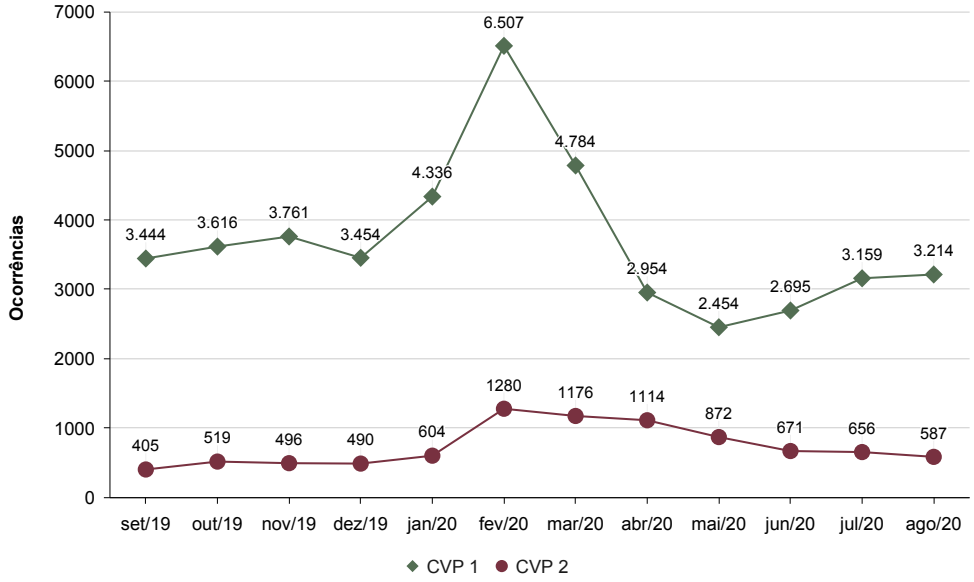
Segundo a SSPDS (2020), um dos indicadores-chave de desempenho criminal, conhecido como Crimes Violentos Contra o Patrimônio (CVP), engloba todos os crimes classificados como roubo, exceto o roubo seguido de morte (latrocínio). O indicador CVP é o único que possui dois tipos: (i) CVP 1, que compreende o roubo à pessoa, roubo de documentos e outros roubos que não estão incluídos no CVP 2; (ii) CVP 2, que abarca o roubo à residência, roubo com restrição de liberdade da vítima, e roubo de carga e veículos.

As estatísticas divulgadas mensalmente são computadas a partir da soma das ocorrên-

[1] https://www.sspds.ce.gov.br/estatisticas-2/

cias de todos os tipos de roubos (CVP 1 e 2) praticados no estado, que ocorrem entre o primeiro e o último dia do mês. A problemática social surge quando é apurada a quantidade de incidências desse tipo de crime. A Figura 1 apresenta um gráfico de série mensal com as tendências mais recentes das categorias de CVP no Ceará, de Setembro de 2019 a Agosto de 2020.

Figura 1 – Série mensal dos tipos de CVP no Ceará de Setembro de 2019 a Agosto de 2020



Fonte: Elaborado pela autora com os dados estatísticos divulgados pela SSPDS (2020).

Ao longo da série mostrada na Figura 1, observam-se variações acentuadas nos registros dos crimes violentos contra o patrimônio. Apenas o mês de Fevereiro contabiliza sozinho quase oito mil casos se somados os dois tipos de CVP. No total, a contagem de ocorrências dos 12 meses dispostos na figura chega a 53.248 roubos. Para os registros de CVPs cometidos em anos anteriores, em 2018 foram computados 64.513 roubos, enquanto que em 2019 cerca de 50 mil incidentes foram registrados. Além do gráfico de série mensal, os dados reportados pela SSPDS exibem outros dois gráficos com o percentual de CVP por dia da semana e por turno, como também uma tabela com o número total de ocorrências do ano.

Outros principais indicadores criminais divulgados oficialmente pela SSPDS, como Crimes Violentos Letais e Intencionais (CVLI), Apreensão de Entorpecentes, Apreensão de Armas, Furto, e Crimes Sexuais, também seguem o mesmo padrão que o de CVP na mostra de suas estatísticas. Salvo as estatísticas mensais, são publicados registros diários do índice de CVLI, relatórios com as principais ocorrências atendidas pelos órgãos públicos vinculados

à SSPDS, e a quantidade de crimes de cada indicador para as Aréas Integradas de Segurança (AISs)[2]. É notória a falta de uma análise e divulgação mais precisa dos crimes patrimoniais, dado que a distribuição dos CVPs é maior que a de qualquer outro indicador, e o seu volume de registros de ocorrências é o único que não vem diminuindo ao longo do tempo.

Encontrar e divulgar informações relevantes e oportunas em documentos CVP é crucial para a população e pode desempenhar um papel central na melhoria das capacidades de combate aos tipos de roubos, aumentando a segurança pública e reduzindo crimes futuros. Uma forma de processamento mais inteligente e rápido capaz de captar sentidos nos diversos textos de ocorrências CVP, de onde é possível extrair informações importantes que hoje não são disponibilizadas pela SSPDS, dado o trabalho extenso e manual que é analisar esse grande volume de dados, é utilizar a técnica de *Named Entity Recognition* (NER), ou Reconhecimento de Entidade Nomeada, oriunda do Processamento de Linguagem Natural (PLN).

Essa técnica extrai e classifica Entidades Nomeadas (ENs): termos que apresentam um ou mais designadores rígidos, em determinado texto (NADEAU; SEKINE, 2007). Geralmente, as ENs são divididas em duas categorias (LI *et al.*, 2020): ENs genéricas (identificam pessoas, lugares e organizações), e ENs específicas do domínio (para o domínio criminal, poderiam ser tipos de armas e veículos, tipo de propriedade privada, dentre outras). Nos últimos anos, o Aprendizado Profundo fortalecido por representações distribuídas, mais especificamente por *word embeddings*[3], e por demandar pouca engenharia de atributos e de recursos específicos de linguagens, tem sido amplamente empregado em modelos NER, gerando resultados que refletem o estado da arte (LI *et al.*, 2020).

Contudo, sistemas NER supervisionados, principalmente aqueles relacionados ao Aprendizado Profundo, requerem uma grande quantidade de dados anotados para treinamento, o que pode ser um problema para muitas linguagens e domínios específicos, pois são necessários especialistas do domínio para realização das tarefas de anotação. Além disso, devido à ambiguidade natural presente nas linguagens humanas, é indispensável que esses dados rotulados tenham qualidade e consistência (LI *et al.*, 2020). Tudo isso, somado ao fato de que é necessário obter uma representação textual eficiente e garantir que sejam extraídas informações diretamente do contexto, faz com que a adaptação de um modelo NER ainda seja um desafio para diferentes domínios e idiomas (LI *et al.*, 2020; COELHO DA SILVA *et al.*, 2019a).

[2] Áreas territoriais de atuação dos trabalhos policiais para potencializar a integração e a atuação conjunta na resolução dos problemas relacionados com a criminalidade (SSPDS, 2020).

[3] O termo *word embeddings* é amplamente utilizado na literatura, desta forma, o mesmo não foi traduzido.

Os trabalhos existentes na área que lidam com documentos de teor criminal limitam-se ao desenvolvimento de ferramentas e métodos de aprendizado supervisionado que reconhecem apenas alguns tipos específicos de entidades (CHEN *et al.*, 2004b; SHABAT *et al.*, 2014; COELHO DA SILVA *et al.*, 2019a; COELHO DA SILVA *et al.*, 2019b). Nenhum desses trabalhos fazem uso de textos sobre roubos, e nem tratam do problema de desbalanceamento de dados presente em muitas tarefas do PLN (LI *et al.*, 2019). Apenas alguns deles tentaram aperfeiçoar o reconhecimento de entidades explorando representações distribuídas para as palavras.

O presente trabalho visa automatizar o processo de obtenção de informações relevantes contidas em narrativas CVP, com o intuito de promover uma compreensão dos fenômenos relativos à segurança pública. Para esse propósito, este trabalho propõe a criação de um modelo NER com técnicas do Aprendizado Profundo capaz de reconhecer entidades nomeadas em textos de ocorrências de roubos, seguindo três questões de pesquisa:

1ª As representações distribuídas a nível de palavras (*word embeddings*) são suficientes para capturar as propriedades semânticas e sintáticas, que não aparecem explicitamente no texto? Qual abordagem é mais eficiente: *word embeddings* específicas do domínio, ou integração e/ou ajuste de *word embeddings* pré-treinadas?

2ª O uso de soluções aplicadas à arquitetura para lidar com o problema de desbalanceamento de classes, garante um aumento significativo na qualidade e no desempenho?

3ª A remoção da anotação de tipos de entidades nomeadas raras, mesmo que muito específicas, para o treinamento, melhora a capacidade do modelo em reconhecer as que possuem uma maior dependência de contexto?

Os modelos experimentados com técnicas para o desbalanceamento, e com *word embeddings* do *CVP2Vec*, que representam as relações sintáticas e semânticas dos termos usados em narrativas de roubos, superaram o *baseline* (HONNIBAL; MONTANI, 2017) em todas as métricas de avaliação. Assim sendo, as principais contribuições deste trabalho são: i) sistematizar a obtenção de informações referentes aos crimes violentos contra o patrimônio; ii) oferecer uma representação distribuída específica do domínio (*CVP2Vec*); e iii) prover um conjunto de modelos eficazes para o reconhecimento de entidade nomeada em textos CVP, tendo como base uma ferramenta NER bastante utilizada pela comunidade científica e pelo mercado, e uma arquitetura de rede neural para o problema de desbalanceamento de classes.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo deste trabalho é criar um modelo NER que trate o problema de desbalanceamento de classes e que seja capaz de reconhecer entidades nomeadas em narrativas CVP de acordo com rótulos predefinidos.

1.1.2 Objetivos Específicos

- Construir um conjunto de dados rotulado de textos CVP;
- Utilizar e avaliar diferentes vetores de *word embeddings* para a representação textual das narrativas de roubos;
- Aplicar abordagens de nível algorítmico (arquitetura do modelo) e dos dados para tratamento do problema de desbalanceamento de classes;
- Treinar o modelo NER do Aprendizado Profundo utilizando *word embeddings* e técnicas para o desbalanceamento de dados;
- Aperfeiçoar e validar o modelo NER proposto, para que possua um desempenho satisfatório no reconhecimento de entidades nomeadas em textos CVP.

Os próximos capítulos estão organizados da seguinte maneira: o Capítulo 2 apresenta a fundamentação teórica e os conceitos que embasam as abordagens propostas nesta monografia; o Capítulo 3 trata dos trabalhos relacionados, com descrição e comparação de projetos e pesquisas que possuem aspectos similares aos especificados neste trabalho; o Capítulo 4 apresenta os procedimentos metodológicos com a descrição das atividades relacionadas ao desenvolvimento e validação do modelo NER proposto; o Capítulo 5 aborda os experimentos e resultados obtidos, como também descreve mais detalhadamente as questões de pesquisa, e as estratégias e cenários de experimentação desenvolvidos para respondê-las; e, por fim, o Capítulo 6 conclui este trabalho, resumindo as contribuições e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os conceitos necessários para o desenvolvimento teórico deste trabalho. Inicialmente, é definido o conceito de Processamento de Linguagem Natural, focando na definição de Reconhecimento de Entidade Nomeada e no problema de Desbalanceamento de Classes. Em seguida, são apresentadas as técnicas de Aprendizado Profundo que embasam os métodos propostos nesta monografia. Na sequência, estão os conceitos de Humano no Ciclo, com a intenção de compreender a definição deste tipo de modelo de aprendizagem que requer interação humana. Por fim, são descritas as métricas adotadas para fins de avaliação e comparação de modelos NER.

2.1 Processamento de Linguagem Natural

Segundo Liddy (2001), Processamento de Linguagem Natural (PLN), tradução de *Natural Language Processing*, refere-se às teorias motivadas por uma série de técnicas computacionais relacionadas à geração e à compreensão de linguagens humanas naturais, geralmente na forma de textos, para uma variedade de tarefas e aplicações.

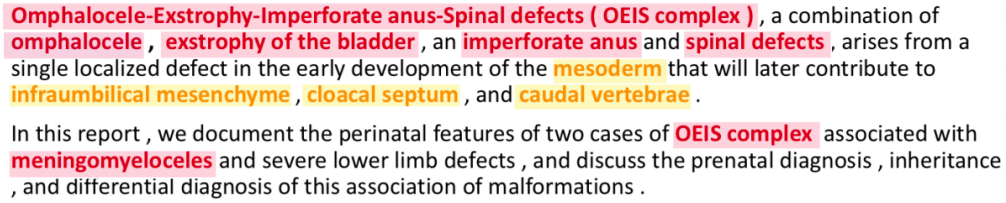
Como idiomas são o reflexo de uma cultura e de um contexto específico, não é uma tarefa fácil ensinar máquinas a entender a comunicação humana, o que faz com que este campo seja bastante pesquisado (MANNING; SCHÜTZE, 1999). Assim, existem diversas aplicações que envolvem PLN em variadas áreas, como tradução automática, interfaces de usuário, recuperação de informações em vários idiomas, e reconhecimento de fala e de entidades em textos (JACKSON; MOULINIER, 2007).

2.1.1 Reconhecimento de Entidade Nomeada

Os recentes avanços no PLN permitiram o desenvolvimento de ferramentas capazes de extrair informações de textos em grande escala, como o Reconhecimento de Entidade Nomeada (NER, sigla do termo em inglês *Named Entity Recognition*), um dos mais utilizados atualmente para a compreensão de dados textuais (WESTON *et al.*, 2019). Historicamente, o NER foi desenvolvido como uma técnica para a extração de tópicos gerais, como nomes de pessoas e localizações geográficas, a partir de textos não estruturados, como artigos de jornais (WESTON *et al.*, 2019; NADEAU; SEKINE, 2007).

Os sistemas que utilizam essa tarefa permitem monitorar tendências no enorme conjunto de documentos escritos em linguagem natural, produzido todos os dias por organizações, governos e indivíduos (NADEAU; SEKINE, 2007). A Figura 2 apresenta um exemplo do NER aplicado na Medicina e Biomedicina, com um texto em inglês que contém sete doenças e quatro estruturas anatômicas, destacadas em vermelho e em amarelo, respectivamente. A identificação dessas doenças e anatomias permitem que pessoas consigam coletar informações de várias partes distintas do texto, para a construção de explicações precisas e completas de domínio médico (SACHAN *et al.*, 2017).

Figura 2 – Texto médico com termos marcados que representam doenças (em vermelho) e estruturas anatômicas (em amarelo)



Fonte: Sachan *et al.* (2017).

À vista disso, o objetivo do NER é identificar menções de designadores rígidos em textos e categorizá-los em tipos semânticos predefinidos. Basicamente, duas tarefas principais são abordadas (SPECK; NGOMO, 2014): identificação de *tokens*[1] em um determinado texto não rotulado (processo chamado de tokenização), e classificação destes *tokens* em tipos de entidades nomeadas (ENs)[2], isto é, palavras ou frases que identificam termos de um conjunto no qual todos os itens possuem atributos semelhantes (LI *et al.*, 2020).

Formalmente, essas duas tarefas são reescritas da seguinte maneira: dada uma sequência de *tokens* $\mathbf{s} = \{w_1, w_2, \cdots, w_N\}$, um modelo NER deve gerar uma lista de tuplas $< I_{início}, I_{fim}, t >$, em que cada uma representa uma entidade nomeada em $\mathbf{s}$, onde $I_{início} \in [1,N]$ e $I_{fim} \in [1,N]$ são os índices que indicam o início e fim da EN, e t o tipo da entidade presente em um conjunto predefinido de categorias.

Em um exemplo prático, considere a Figura 3 que contém o primeiro parágrafo do *Capítulo 1: Uma Festa Muito Esperada* do livro "O Senhor dos Anéis: A Sociedade do Anel", de J.R.R. Tolkien (2017). Para analisar o texto segundo o NER, deve-se aplicar as tarefas descritas anteriormente.

[1] Também conhecido como palavras, nomes ou instâncias.
[2] Também conhecido como classes ou rótulos (*labels, tags*).

Figura 3 – Trecho do capítulo 1 do livro "O Senhor dos Anéis: A Sociedade do Anel"

Quando o Sr. Bilbo Bolseiro de Bolsão anunciou que em breve celebraria seu onzentésimo primeiro aniversário com uma festa de especial grandeza, houve muito comentário e agitação na Vila dos Hobbits.

Fonte: J.R.R. Tolkien (2017).

Na tokenização, o texto é colocado em uma lista, onde cada item se trata de um conjunto de caracteres significativo ao idioma, ou seja, uma palavra. Assim, é gerada uma lista **s** que contém uma sequência de *tokens* numerados, no seguinte formato:

$$\mathbf{s} = \{\text{“}Quando\text{”}_{w_1}, \text{“}o\text{”}_{w_2}, \text{“}Sr.\text{”}_{w_3}, \text{“}Bilbo\text{”}_{w_4}, \text{“}Bolseiro\text{”}_{w_5}, \text{“}de\text{”}_{w_6}, \text{“}Bolsão\text{”}_{w_7},$$
$$\vdots$$
$$\text{“}e\text{”}_{w_{27}}, \text{“}agitação\text{”}_{w_{28}}, \text{“}na\text{”}_{w_{29}}, \text{“}Vila\text{”}_{w_{30}}, \text{“}dos\text{”}_{w_{31}}, \text{“}Hobbits\text{”}_{w_{32}}, \text{“.”}_{w_{33}}\}$$

O próximo passo é identificar e classificar as entidades segundo rótulos predefinidos para que possuam sentido semântico. Geralmente, entidades nomeadas são dividas em duas categorias: genéricas (como indivíduos, lugares e organizações), e específicas do domínio (LI *et al.*, 2020). Tomando como base as ENs genéricas, como exemplificado a seguir em forma de tuplas, *"Sr. Bilbo Bolseiro"* é classificado como `PESSOA`, enquanto que *"Bolsão"* e *"Vila dos Hobbits"* são classificados como `LOCAL`. A Figura 4 mostra visualmente as tuplas com as ENs aplicadas ao texto.

$$< w_3, w_5, \texttt{PESSOA} > \quad \text{“}Sr.\ Bilbo\ Bolseiro\text{”}$$
$$< w_7, w_7, \texttt{LOCAL} > \quad \text{“}Bolsão\text{”}$$
$$< w_{30}, w_{32}, \texttt{LOCAL} > \quad \text{“}Vila\ dos\ Hobbits\text{”}$$

Figura 4 – Trecho da Figura 3 com suas entidades nomeadas e rótulos destacados

Quando o Sr. Bilbo Bolseiro PESSOA de Bolsão LOCAL anunciou que em breve celebraria seu onzentésimo primeiro aniversário com uma festa de especial grandeza, houve muito comentário e agitação na Vila dos Hobbits LOCAL .

Fonte: Elaborado pela autora.

O NER possui técnicas tradicionais que realizam esse processo de localização e classificação de entidades. Elas são separadas em três abordagens principais (LI *et al.*, 2020): 1) abordagens baseadas em regras, que não dependem de dados anotados, e sim de um conjunto de regras relacionais; 2) abordagens de aprendizagem não supervisionada, que utilizam algoritmos que também não precisam de dados rotulados à mão; 3) abordagens de aprendizado supervisionado, que contam com algoritmos que lidam com a engenharia de *features* (atributos ou características) para fornecerem informações que diferenciam da melhor maneira possível os padrões existentes nos dados.

No aprendizado supervisionado, o NER é escalado para uma classificação multiclasse ou uma tarefa de rotulagem de sequência. Como a engenharia de *features* é crítica em sistemas NER supervisionados, os modelos do Aprendizado Profundo tornaram-se dominantes e alcançaram resultados competitivos em relação ao estado da arte (LI *et al.*, 2020). Resumidamente, seus métodos descobrem automaticamente as representações necessárias para a classificação e/ou detecção de dados brutos, demandando pouca engenharia de *features* e de recursos específicos de linguagens. Esse conceito é apresentado e explicado com mais detalhes na Seção 2.2.

Tal abordagem, no entanto, possui um alto custo de desenvolvimento, especialmente para novas linguagens e novos domínios (COELHO DA SILVA *et al.*, 2019a). O primeiro desafio enfrentado está na alta ambiguidade presente nas escritas das linguagens humanas (NGUYEN *et al.*, 2020). O modelo pode se confundir ao tentar identificar ENs que possuem a mesma ortografia, sem contar que uma mesma entidade pode ser anotada diferentemente em conjuntos de dados distintos, mas de mesmo domínio. O termo *Bolsão*, por exemplo, classificado como `LOCAL` na Figura 4, é conhecido na obra de J.R.R. Tolkien como o lar de Bilbo Bolseiro, mas poderia facilmente ter sido anexado ao nome do personagem, então classificado como `PESSOA`.

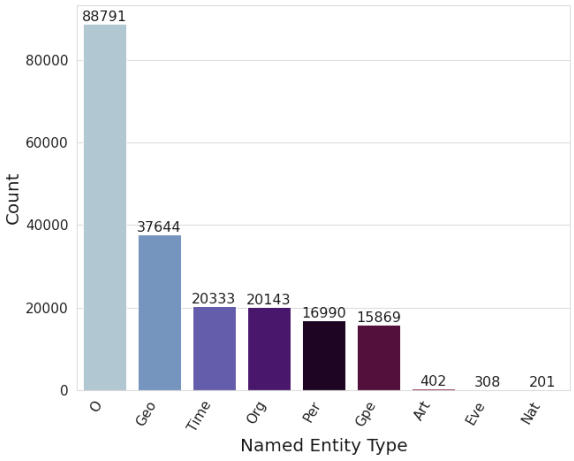
Em razão dessa inconsistência nos dados anotados, um modelo treinado com um conjunto de dados pode não funcionar bem em outro, mesmo que sejam de domínios iguais (LI *et al.*, 2020). Além disso, a complexidade aumenta à medida que são adicionadas mais classes, pois ocasiona em mais erros de incerteza e de predição (NGUYEN *et al.*, 2020). Desta forma, com o intuito de contornar esses problemas de ambiguidade e de desempenho, o presente trabalho propõe a criação de um modelo do Aprendizado Profundo para o reconhecimento de entidades em textos de ocorrências CVP, cujo processo de anotação utiliza uma ferramenta interativa para tarefas de classificação NER.

2.1.1.1 Desbalanceamento de Classes

Anotações de alta qualidade são críticas para o aprendizado e a avaliação de um modelo NER. Contudo, a distribuição natural dos dados textuais é frequentemente desbalanceada, devido à presença de classes que constituem uma minoria muito pequena dos dados (LIU *et al.*, 2009; CHAWLA *et al.*, 2004), o que faz com que o desbalanceamento de classes seja um problema comum para uma variedade de tarefas do PLN (LI *et al.*, 2019).

A Figura 5 apresenta um exemplo de distribuição desbalanceada de classes de um domínio padrão do NER. O gráfico exibido no lado esquerdo da figura contém nove tipos de entidades, cada uma representada com uma cor diferente. Observando o gráfico juntamente com a tabela exibida ao lado direito, pode-se concluir que o total de amostras do grupo que constitui as classes frequentes, formados pelos rótulos *O* (indica os termos que não são entidades nomeadas), *Geo*, *Time*, *Org*, *Per* e *Gpe*, é aproximadamente 50 vezes maior que o total das classes raras, que são: *Art*, *Eve* e *Nat*.

Figura 5 – Distribuição desbalanceada de dados de um domínio padrão do NER



Frequencies

Class	Count	Proportion
O	88.791	44.24
Geo	37.644	18.75
Time	20.333	10.13
Org	20.143	10.03
Per	16.990	8.46
Gpe	15.869	7.90
Art	402	0.20
Eve	308	0.15
Nat	201	0.10

Fonte: Adaptada de Nguyen *et al.* (2020).

O problema de desbalanceamento de classes representa uma dificuldade considerável para quase todos os algoritmos de classificação, uma vez que o grupo de classes minoritárias são mais difíceis de prever porque, por definição, possuem poucas amostras. Isso significa que é mais desafiador para um modelo aprender as características das classes raras já que tende a ser sobrecarregado pelo grupo das majoritárias (LIU *et al.*, 2009; CUI *et al.*, 2019). Na literatura, uma solução típica para esse problema é através da remoção completa dos rótulos raros (SECHIDIS *et al.*, 2011). No entanto, isso implica dizer que esse grupo não é importante

para o desempenho e a generalização dos modelos, o que é um argumento inválido, dado que esse tipo de problema deve considerar os erros de classificação desses rótulos (SECHIDIS *et al.*, 2011; CHEN *et al.*, 2004a).

Mesmo que o desbalanceamento de classes ainda possa variar dependendo do domínio, é necessário o uso de técnicas especializadas. Existem duas abordagens comumente aplicadas para resolver esse tipo de problema: uma baseada em aprendizado sensível a custos (nível algorítmico), e outra voltada para técnicas de amostragem (nível dos dados) (CHEN *et al.*, 2004a). No nível algorítmico, normalmente é feito o ajuste dos custos das classes a fim de combater o desbalanceamento e minimizar o custo geral, por meio da atribuição de custos relativamente mais altos aos rótulos raros. No nível dos dados, as soluções incluem muitas formas diferentes de reamostragem das classes, como amostragem aleatória, amostragem reduzida dos rótulos mais frequentes, sobreamostragem dos rótulos raros, dentre outras.

Muitos trabalhos utilizam a abordagem de nível dos dados (CHEN *et al.*, 2004a) e, para tarefas de classificação, a técnica mais recomendada é a amostragem estratificada, por evitar erros de avaliação em dados desbalanceados (SECHIDIS *et al.*, 2011). Esse método faz com que os dados sejam reorganizados de modo que a proporção de cada classe em cada subconjunto desejado seja aproximadamente igual à de todo o conjunto de dados. Verificando novamente a tabela presente na Figura 5, a estratificação pode ser exemplificada através da coluna *Proportion*, que lista as proporções de cada classe em relação a todo o conjunto de dados, com valores variando de zero (0) a um (1). Aplicando-se esse tipo de amostragem probabilística, as proporções das amostras das classes em cada subconjunto criado seriam as mesmas que as da coluna.

Em relação à abordagem de nível algorítmico, a heurística amplamente utilizada atribui pesos às classes ou organiza os dados de forma inversamente proporcional à frequência de cada classe. Entretanto, trabalhos recentes revelam baixo desempenho ao utilizar esse tipo de função de custo (CUI *et al.*, 2019). Uma estratégia eficaz que pode ser utilizada em uma ampla gama de modelos e funções de perda é a *Class-Balanced* (CB) *Loss*, um esquema de reponderação que é inversamente proporcional ao número efetivo de amostras por classe, denotada pela Equação 2.1:

$$\mathrm{CB}(\mathbf{p}, y) = \frac{1}{E_{n_y}} \mathscr{L}(\mathbf{p}, y) = \frac{1-\beta}{1-\beta^{n_y}} \mathscr{L}(\mathbf{p}, y). \tag{2.1}$$

Na Equação 2.1, n_y é o número de amostras de uma classe $y \in \{1, 2, \cdots, C\}$, dado que C é o número total de classes. E_{n_y} é a proposta para o número efetivo de amostras da classe y, o que indica que a expressão $(1-\beta)/(1-\beta^{n_y})$ é o fator de ponderação para uma função de perda genérica $\mathscr{L}$, com $\beta \in [0,1)$. Por fim, $\mathbf{p} = [p_1, p_2, \cdots, p_C]^\top$ indica as probabilidades das classe estimadas, em que $p_c \in [0,1] \, \forall \, c$.

Observe que a CB *Loss* considera apenas a atribuição de um rótulo para toda uma sentença, definição que se enquadra em um típico problema de classificação multi-classe. Neste trabalho, o NER é tratado como uma tarefa de rotulagem de sequência, onde cada amostra pode ser rotulada com y_c classes possíveis. Esse cenário é conhecido na literatura e comunidade científica por classificação multi-rótulo, uma generalização da multi-classe (PEDREGOSA *et al.*, 2011). Portanto, uma pequena adaptação é feita, primeiro considerando que cada *token* de uma sentença deve ser atribuído a uma das classes y_c mutuamente exclusivas, e trocando a função genérica $\mathscr{L}$ pela *Categorical Cross-Entropy* (CCE) *Loss*, definida abaixo, com N representando o tamanho de uma sentença:

$$\text{CCE} = -\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(p_{n,c}). \tag{2.2}$$

A função de perda CCE é a opção mais recomendada para problemas multi-classe e de aprendizagem de sequência, onde minimiza a distância entre as distribuições de probabilidade emitidas pela rede neural e a distribuição real dos alvos (CHOLLET, 2017). Assim, tem-se a função de perda *Class-Balanced* com *Categorical Cross-Entropy* (CB-CCE, Equação 2.3):

$$\begin{aligned} \text{CB} - \text{CCE} &= -\sum_{n=1}^{N}\sum_{c=1}^{C}\left(\frac{1}{E_{n,c}}\right) y_{n,c}\log(p_{n,c}) \\ &= -\sum_{n=1}^{N}\sum_{c=1}^{C}\left(\frac{1-\beta}{1-\beta^{n,c}}\right) y_{n,c}\log(p_{n,c}). \end{aligned} \tag{2.3}$$

Outra função de perda, a *Dice Loss* (DL), voltada especificamente para tarefas do PLN com dados desbalanceados, possui um esquema de ajuste de peso dinâmico que lida com a influência dominante das classes mais frequentes (LI *et al.*, 2019). Tomando como base a definição dos termos y e $\mathbf{p}$ da CB *Loss*, a função DL pode ser escrita da seguinte forma:

$$\text{DL} = 1 - \sum_{c=1}^{C}\frac{2\sum_i p_{i,c}y_{i,c} + \gamma}{\sum_i p_{i,c}^2 + \sum_i y_{i,c}^2 + \gamma}, \tag{2.4}$$

onde, dado um conjunto de instâncias X, cada instância $x_i \in X$ está associada a um rótulo $y_{i,c}$ e a uma probabilidade $p_{i,c}$. O fator γ é adicionado tanto ao numerador quanto ao denominador para fins de suavização.

Neste trabalho, o processo de estratificação é aplicado ao conjunto de dados CVP multi-rótulo para obtenção dos conjuntos oficiais de treinamento, desenvolvimento e teste. Para fins comparativos, as duas funções de perda apresentadas acima são adicionadas às arquiteturas dos modelos experimentados, com o propósito de alcançar um aumento significativo no desempenho.

2.2 Aprendizado Profundo

Do ponto de vista filosófico, o cérebro humano é uma máquina altamente complexa, capaz de processar uma grande quantidade de dados brutos, extrair informações relevantes destes dados, combiná-las com a memória de eventos passados para, enfim, gerar uma resposta motora em segundos. Biologicamente, cerca de 86 bilhões de neurônios são responsáveis por esse processamento poderoso (AZEVEDO *et al.*, 2009).

Diversos pesquisadores tentaram simular o funcionamento do cérebro humano, principalmente o processo de aprendizagem por experiência, a fim de criar sistemas inteligentes capazes de realizar tarefas consideradas fáceis para seres humanos, como classificação, reconhecimento de padrões, processamento de imagens, e até criação de obras de arte. Como resultado dessas pesquisas, surge então o modelo do neurônio artificial e, posteriormente, um sistema com vários neurônios interconectados (TALON, 2019).

Esse sistema em que neurônios são organizados em camadas, com conexões entre neurônios de outras camadas consecutivas, foi chamado de Rede Neural Artificial (RNA). Uma RNA é um modelo do Aprendizado Profundo (do inglês *Deep Learning*) que, por sua vez, é uma subárea do Aprendizado de Máquina, voltada para o estudo do aprendizado com múltiplas camadas de representações dos dados e vários níveis de abstração. Apesar de que o termo rede neural seja uma referência à neurobiologia, e embora alguns dos conceitos centrais do Aprendizado Profundo tenham sido desenvolvidos em parte por inspirar-se na compreensão do cérebro humano, RNAs não são modelos do cérebro (CHOLLET, 2017).

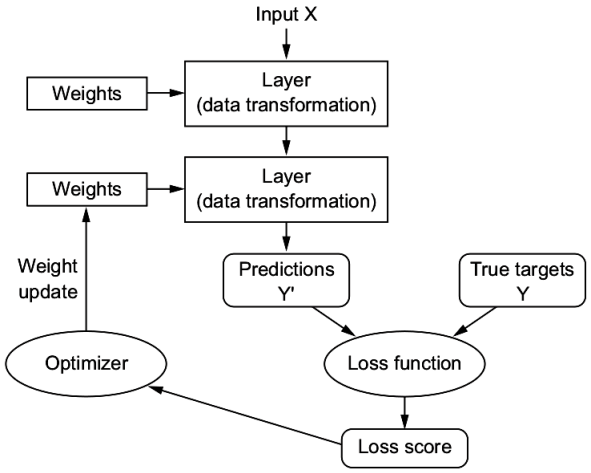
Desde a concepção do neurônio artificial e da formalização do termo Aprendizado Profundo, várias arquiteturas e modelos, com múltiplas combinações de técnicas matemáticas e estatísticas, surgiram e propiciaram a criação de arquiteturas avançadas de redes neurais

profundas, capazes de aprender atributos complexos e intrincados dos dados, por meio de funções de ativação não lineares, e de automatizar o processo da engenharia de *features* (CHOLLET, 2017). Nos últimos anos, os modelos de Aprendizado Profundo para o Reconhecimento de Entidade Nomeada tornaram-se dominantes, produzindo resultados superiores ao estado da arte (LI *et al.*, 2020).

Alguns passos são necessários para o desenvolvimento de RNAs, como coletar os dados e separá-los em subconjuntos (treino, teste e validação/desenvolvimento), para realização dos procedimentos de treino e avaliação de um modelo. Porém, a etapa crucial está na definição da arquitetura e da configuração da rede neural: determinar as suas camadas, seus parâmetros e funções de ativação, a função de perda (define o *feedback* usado para a aprendizagem), o otimizador (determina como a rede será atualizada com base na função de perda), etc.

É possível visualizar a interação desses elementos na Figura 6, na qual a rede, composta de camadas encadeadas, mapeia os dados de entrada para as suas predições. A função de perda, então, compara essas predições com os valores reais, produzindo um valor de perda: uma medida do quão bem as predições da rede correspondem ao que era esperado. Depois, o otimizador usa esse valor para atualizar os pesos da rede.

Figura 6 – Anatomia de uma rede neural: relação entre suas camadas, função de perda e otimizador

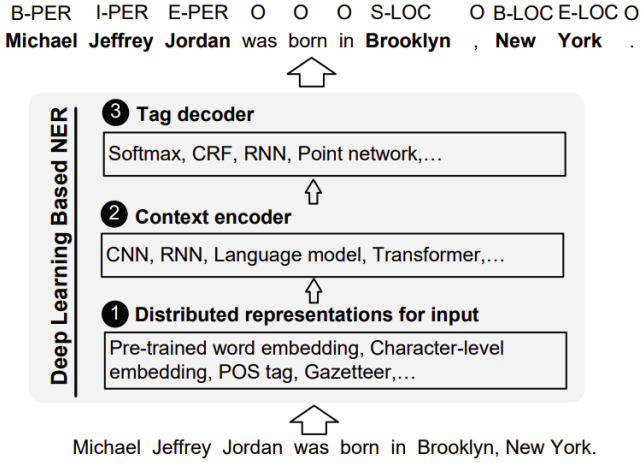


Fonte: Chollet (2017).

Baseada na anatomia de uma rede neural, a arquitetura geral de um modelo NER do Aprendizado Profundo é apresentada na Figura 7, com três etapas distintas: (i) representações distribuídas para o texto de entrada; (ii) codificador de contexto; e (iii) decodificador de rótulo. As representações distribuídas caracterizam as palavras do texto de entrada como vetores densos

de baixa dimensão, em que cada dimensão representa uma propriedade referente à estrutura, formação, classificação e/ou significado desses termos. Aprendidas a partir do texto, as representações distribuídas consideram *embeddings* a nível de palavras e/ou caracteres, bem como a incorporação de atributos adicionais como *gazetteer* (dicionário de entidades), e *part-of-speech tagging* (POS *tagging*, processo em que palavras são atribuídas com rótulos gramaticais, de acordo com o contexto).

Figura 7 – Taxonomia de um modelo NER neural profundo



Fonte: Li *et al.* (2020).

Na segunda etapa, o codificador de contexto captura as dependências de contexto usando redes neurais profundas e, por último, o decodificador de rótulos prevê rótulos para as entidades nomeadas do texto de entrada. Observe que na Figura 7, cada *token* é previsto como um rótulo pertencente a um esquema de rotulação, indicado por B-(*begin* ou início), I-(*inside* ou dentro), E-(*end* ou fim), S-(*singleton* ou único), seguido pelo tipo da entidade nomeada, ou O-(*outside* ou fora), caso não seja uma entidade nomeada.

A seguir, é descrito o funcionamento dos métodos do Aprendizado Profundo implementados neste trabalho, seguindo a taxonomia do modelo NER da Figura 7. A Seção 2.2.1 apresenta alguns tipos de representações textuais, com foco nas *word embeddings*. As Seções 2.2.2 e 2.2.3 descrevem o funcionamento das redes neurais profundas *Bidirectional Long Short-Term Memory* e Convolucionais, arquiteturas amplamente utilizadas como codificadores de contexto. A última seção (Seção 2.2.4), descreve as arquiteturas de decodificadores de rótulos *Conditional Random Fields*, e *Perceptron* Multicamadas com a função de ativação *Softmax*. Todos os métodos citados se tratam de componentes cruciais para muitas abordagens do NER, em que seu uso combinado permite um melhor treinamento e generalização (LI *et al.*, 2020).

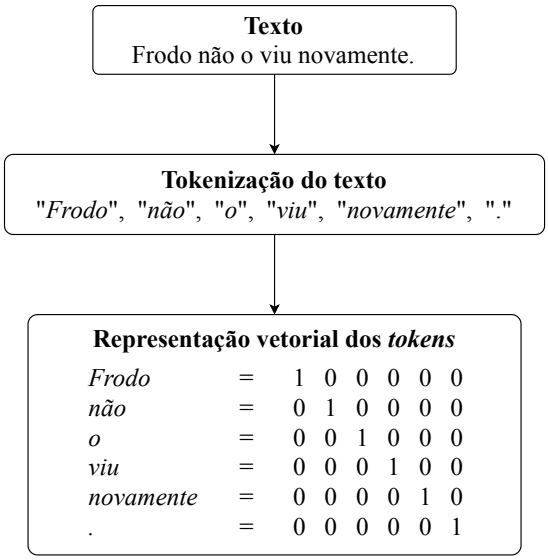
Especificamente, as redes neurais Convolucionais e *Perceptron* Multicamadas compõem a arquitetura do modelo de reconhecimento de entidade nomeada da biblioteca *spaCy*, utilizado como *baseline* por possuir integração com a ferramenta interativa HNERD (COELHO DA SILVA *et al.*, 2019b). Logo, a rede *Bidirectional Long Short-Term Memory* e o modelo *Conditional Random Fields* fazem parte da arquitetura proposta neste trabalho. A junção desses dois métodos é comumente conhecida na literatura como BLSTM-CRF, um modelo extensamente utilizado para tarefas NER (MA; HOVY, 2016; LAMPLE *et al.*, 2016).

2.2.1 Representação Textual

Os modelos de Aprendizado Profundo do Processamento de Linguagem Natural não recebem como entrada um texto composto por vários componentes e frases, mas sim a representação vetorial e numérica deste texto. Essa representação pode ser obtida mediante um procedimento chamado de *text-vectorization* (vetorização do texto), feito de várias maneiras: segmentando o texto em palavras ou caracteres, ou extraindo *n-grams*[3], e transformando cada componente em um vetor (CHOLLET, 2017).

Dividir um texto em unidades distintas (palavras, caracteres ou *n-grams*) nada mais é do que tokenizar este texto. Assim sendo, os processos de vetorização de textos envolvem duas tarefas principais, exemplificadas abaixo no fluxograma da Figura 8.

Figura 8 – Fluxograma com os passos da vetorização de um texto



Fonte: Elaborado pela autora.

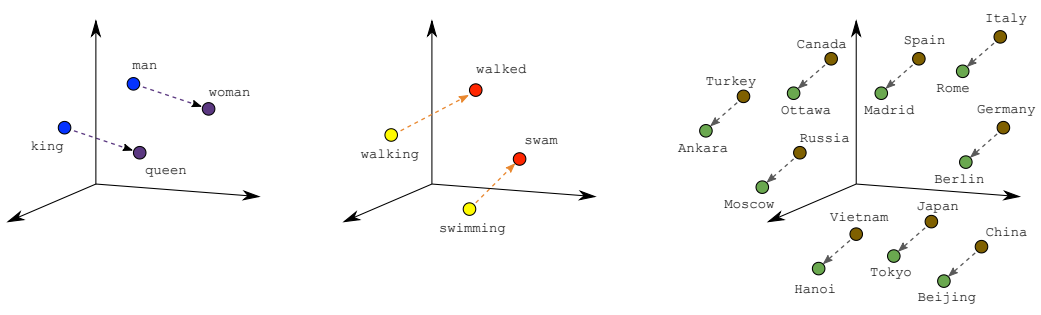
[3] Segundo Chollet (2017), *n-grams* (n-gramas) são grupos formados por *n* palavras ou caracteres consecutivos de uma determinada amostra de texto.

A primeira forma geométrica da Figura 8 contém o texto que será vetorizado. As duas últimas formas indicam as etapas da vetorização de um texto: primeiro, a aplicação de um esquema de tokenização, ilustrado na segunda forma, onde estão listados os *tokens* do texto; e, por fim, a transformação dos *tokens* em vetores numéricos, representados como vetores binários na última forma geométrica.

Existem duas maneiras principais de associar um vetor a uma palavra: codificação *one-hot* e *word embeddings*. A matriz apresentada na terceira forma geométrica da Figura 8 é um exemplo simples de aplicação da codificação *one-hot*, a forma mais comum e básica de transformar *tokens* em vetores, que consiste na conversão de um texto em uma matriz binária. Na figura, cada palavra está associada a um índice i de valor inteiro e é representada por um vetor de tamanho N (o tamanho do vocabulário, ou seja, a quantidade de palavras únicas existentes no texto) preenchido por zeros, exceto a i-ésima posição, ditada pelo índice desta palavra no vocabulário, que possui o valor um (1).

A segunda maneira de representar as palavras vetorialmente é fazendo uso de *word embeddings*. Em sua forma mais básica, *word embeddings* são técnicas capazes de identificar semelhanças entre palavras de um vocabulário, mapeadas em um espaço vetorial (CHOLLET, 2017). Nesse espaço, a proximidade entre os vetores representa as relações sintáticas, semânticas ou morfológicas entre as palavras. Isto é, palavras que possuem significados diferentes são colocadas em pontos distantes umas das outras, enquanto que palavras usadas em um mesmo contexto ficam em pontos próximos. Uma aplicação real das *word embeddings* é exposta na Figura 9, onde vetores representam relacionamentos semânticos entre gêneros, tempos verbais, e entre países e suas capitais.

Figura 9 – Representação vetorial entre *word embeddings*

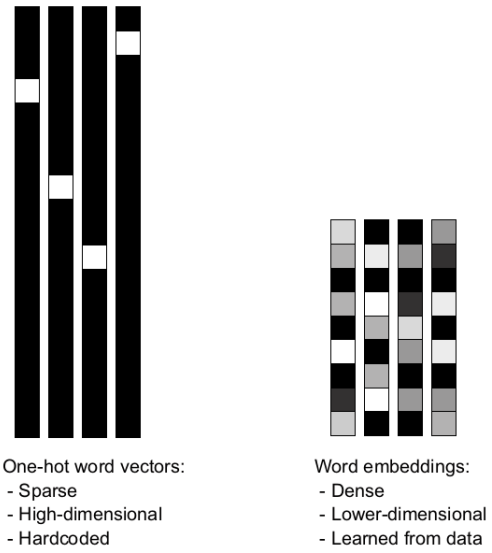


Male-Female Verb Tense Country-Capital

Fonte: Google Developers (2020).

Como ilustrado adiante na Figura 10, enquanto que os vetores obtidos com a codificação *one-hot* são binários, esparsos (em sua maioria compostos por zeros) e de alta dimensão (possuem a mesma dimensionalidade que o número de palavras do vocabulário, que pode ser vasto), a representação obtida com as *word embeddings* são de vetores densos, de baixa dimensionalidade e aprendidos através dos dados (CHOLLET, 2017). Por fornecerem representações textuais eficientes e por garantirem aumento na capacidade de generalização de um modelo, *word embeddings* são hoje a abordagem dominante para a vetorização de dados no PLN (HARTMANN *et al.*, 2017).

Figura 10 – Diferenças entre a codificação *one-hot* e *word embeddings*



Fonte: Chollet (2017).

Há duas maneiras principais de obter *word embeddings* (CHOLLET, 2017), conforme sumarizado a seguir:

1) **Aprender *word embeddings* junto com a rede neural principal.**

   Nesta configuração, é necessário adicionar à arquitetura da rede neural principal a camada responsável pela conversão das palavras em *word embeddings*. Essa camada, chamada de camada de *embedding* (*embedding layer*), mapeia índices inteiros que representam palavras específicas para vetores densos, inicializados aleatoriamente. Os índices são procurados no vocabulário passado como referência para, em seguida, serem retornados os vetores associados. Durante o treinamento, esses vetores de palavras são ajustados gradualmente, estruturando o espaço vetorial em algo que o modelo possa explorar. Esta abordagem fornece *word embeddings* específicas do domínio.

2) Carregar *word embeddings* pré-treinadas.

Invés de aprender *word embeddings* junto com o problema que se deseja resolver, *word embeddings* pré-treinadas possuem vetores treinados previamente, e podem ser usadas de duas maneiras: estática ou dinâmica. Na forma dinâmica, a camada de *embedding* é inicializada com os pesos das *word embeddings*, e a atualização é propagada pela rede neural durante o treino, processo esse conhecido como *fine-tuning*. Na forma estática, a camada de *embedding* é congelada, evitando que os pesos sejam atualizados.

Word embeddings pré-treinadas podem ser obtidas por meio de uma variedade de estratégias, algumas envolvendo ou não redes neurais. Também podem ser encontrados repositórios abertos, como *Facebook fastText*[4], CMU *Multilingual Embeddings*[5], *Stanford* NLP GloVe[6] e *Google Code Archive Word2Vec*[7], que fornecem vetores treinados em um grande volume de dados, para que possam ser aplicados em outros conjuntos de dados. Alguns desses repositórios geraram seus vetores de *word embeddings* utilizando algoritmos projetados especificamente para esta tarefa. Abaixo estão descritos os dois grupos principais aos quais esses algoritmos fazem parte, juntamente com a definição dos métodos mais populares.

- **Grupo 1:** Métodos que trabalham com uma matriz de palavras de coocorrência.
  - O método ***Global Vectors for Word Representation* (GloVe)** foi proposto por Pennington *et al.* (2014), e obteve resultados que refletem o estado da arte para tarefas que identificam analogias sintáticas e semânticas (PENNINGTON *et al.*, 2014). Este método consiste em uma matriz de coocorrência M, construída observando-se o contexto das palavras. Cada elemento M_{ij} na matriz representa a probabilidade da palavra i está próxima da palavra j. Na matriz M, as linhas (vetores) são geradas aleatoriamente e treinadas obedecendo à Equação 2.5, definida abaixo:

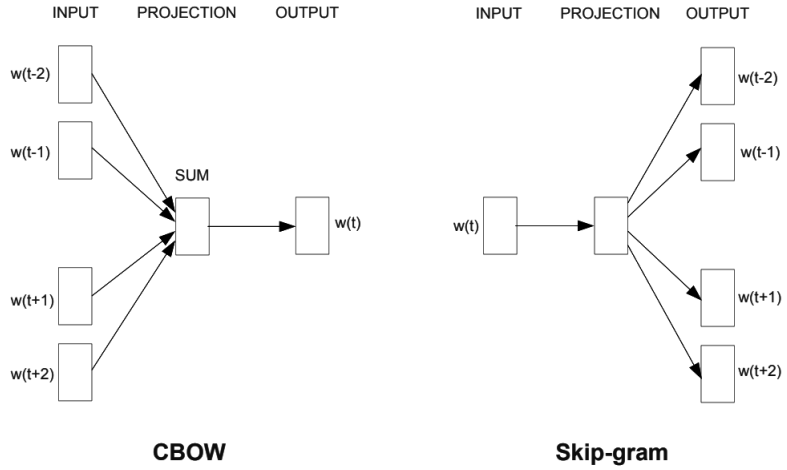
$$P(w_i, w_j) = \log(M_{ij}) = w_i w_j + b_i + b_j, \tag{2.5}$$

    onde w_i e w_j são vetores de palavras, e b_i e b_j são *bias*.
- **Grupo 2:** Métodos preditivos que tentam prever palavras vizinhas, dada uma ou mais palavras de contexto.

[4] https://fasttext.cc/
[5] http://www.cs.cmu.edu/~afm/projects/multilingual_embeddings.html
[6] https://nlp.stanford.edu/projects/glove/
[7] https://code.google.com/archive/p/word2vec/

– O ***Word2Vec*** é um método frequentemente usado no PLN para fornecer *word embeddings* (HARTMANN *et al.*, 2017), e que possui duas estratégias de treinamento, ilustradas na Figura 11: (i) *Continuous Bag of Words* (CBOW), em que o modelo recebe uma sequência de palavras com exceção da palavra alvo, e tenta prever essa palavra a partir do contexto; (ii) *Skip-Gram*, no qual o modelo recebe uma palavra alvo e tenta prever o contexto onde ela está inserida. Nos dois casos, o modelo resulta em um treinamento rápido, capaz de capturar informações semânticas (MIKOLOV *et al.*, 2013).

Figura 11 – Arquiteturas dos modelos CBOW e *Skip-Gram*



Fonte: Mikolov *et al.* (2013).

– O algoritmo ***Wang2Vec*** é uma modificação do *Word2Vec* e foi criado para capturar o comportamento sintático das palavras (LING *et al.*, 2015). A principal diferença em relação ao *Word2Vec* está nas arquiteturas usadas para gerar *word embeddings*. Na arquitetura da rede *Continuous Window* (C*Window*), uma variação do CBOW, ela recebe como entrada a concatenação dos vetores das palavras de contexto na ordem em que ocorrem. Na rede *Structured Skip-Ngram* (variação do *Skip-Gram*) é usado um conjunto diferente de parâmetros para prever cada palavra de contexto, dependendo da sua posição em relação à palavra alvo.
– O ***FastText*** se trata do método mais recente, desenvolvido por Bojanowski *et al.* (2017) para capturar informações morfológicas. Neste método, os vetores são associados a caracteres *n-grams*, e as palavras são representadas como o somatório destas representações. Assim, a representação de uma palavra é feita pela soma dos vetores de caracteres *n-grams* com os vetores das palavras circundantes.

Neste trabalho, a representação textual do conteúdo dos textos CVP é feita mediante o uso de *word embeddings*. As duas formas apresentadas anteriormente de como obter esses vetores são abordadas na experimentação: treinar *word embeddings* junto com a rede neural principal, e carregar *word embeddings* pré-treinadas. Em relação a essa última, este trabalho utiliza os vetores treinados pelos algoritmos *FastText*, GloVe, *Wang2Vec* e *Word2Vec*, disponibilizados pelo Núcleo Interinstitucional de Linguística Computacional (NILC)[8] da Universidade de São Paulo, Brasil.

O repositório do NILC[9] contém vetores gerados a partir de dezessete corpora linguísticos do português do Brasil e do português europeu, de fontes e gêneros variados, totalizando mais de um trilhão de *tokens*. Também são disponibilizados, para cada modelo e estratégia de treinamento (se tiver), vetores de palavras com variadas dimensões (de 50 a 1000 dimensões).

Dado que os modelos *FastText*, *Wang2Vec* e *Word2Vec* possuem as variações CBOW e *Skip-Gram*, este trabalho utiliza os vetores treinados com a arquitetura *Skip-Gram*, visto que representa bem palavras raras, identifica padrões, e compreende variados contextos eficientemente (MIKOLOV *et al.*, 2013). Essas características são essenciais para este trabalho, uma vez que o modelo NER deve ser capaz de: i) classificar corretamente ENs que podem ter rótulos diferentes, dependendo do contexto; e ii) generalizar palavras pouco frequentes.

Como não é possível obter todo o vocabulário de um idioma, em bases de dados de domínio específico podem aparecer problemas relacionados às palavras que estão fora do vocabulário (OOV, do inglês *out-of-vocabulary*): termos do conjunto de dados que não fazem parte do vocabulário usado para gerar *word embeddings* (PINTER *et al.*, 2017). No que diz respeito aos problemas de OOV, é feito um tratamento utilizando a métrica *Edit Distance* (Distância de Edição ou Distância Levenshtein), que mede o quão semelhantes são duas palavras, com base no número de edições (inserções, exclusões e substituições) necessárias para transformar uma sequência de caracteres em outra (JURAFSKY; MARTIN, 2017).

Cada uma das operações recebe um custo de valor um (1) como fator de ponderação mais simples para calcular a Distância de Levenshtein entre duas palavras (LEVENSHTEIN, 1966). Além de ser muito utilizada para a normalização e tratamento de erros dentro de um conjunto de dados textual, *Edit Distance* também está presente em uma variedade de pesquisas analíticas, tais como verificação/correção ortográfica, reconhecimento de fala, detecção de plágio, e resolução de correferência (RAJA *et al.*, 2019; JURAFSKY; MARTIN, 2017).

[8] http://www.nilc.icmc.usp.br/nilc/

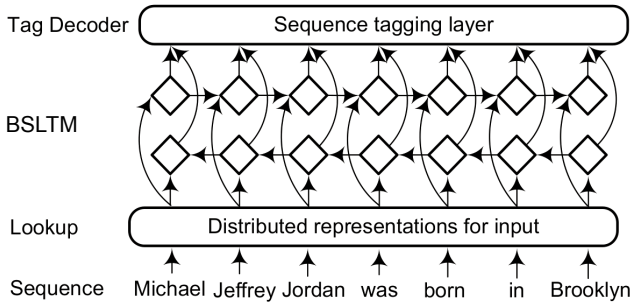
[9] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

2.2.2 Redes Neurais Bidirectional Long Short-Term Memory

As Redes Neurais Recorrentes ou *Recurrent Neural Networks* (RNN), junto com suas variantes, como a *Long Short-Term Memory* (LSTM), demonstraram resultados promissores na modelagem de dados sequenciais (LI *et al.*, 2020). Uma RNN mantém uma memória com base em um histórico e, em particular, as redes *Long Short-Term Memory* são iguais às RNNs, exceto que as atualizações das suas camadas ocultas são substituídas por células de memória específicas. Como resultado, elas são melhores para a localização e exploração de dependências de longo alcance nos dados (HUANG *et al.*, 2015).

Para muitas tarefas de rotulagem de sequência, como o Reconhecimento de Entidade Nomeada, é necessário ter acesso às informações de contexto passadas (via estados para frente) e futuras (via estados para trás) por um certo período de tempo, o que não pode ser obtido por uma típica LSTM, visto que obtém apenas informações do passado (MA; HOVY, 2016). Uma solução elegante que resolve esse problema é a rede neural profunda *Bidirectional Long Short-Term Memory* (BLSTM), presente na Figura 12. A ideia básica na qual é formada a arquitetura da figura está no uso de duas camadas LSTM regulares, cada uma das quais processa a sequência de entrada em uma direção (cronológica e anti-cronológica), para capturar informações passadas e futuras. Em seguida, as duas camadas LSTM são concatenadas para formar a saída final.

Figura 12 – Arquitetura de uma rede LSTM bidirecional para o NER



Fonte: Adaptada de Li *et al.* (2020).

Ao tratar dados sequenciais de duas maneiras distintas, uma BLSTM captura padrões que podem ser perdidos por uma LSTM tradicional. Essa característica a transformou na arquitetura padrão para representações de texto dependentes do contexto, por levar em consideração uma quantidade infinita de informações para toda uma sentença (LI *et al.*, 2020).

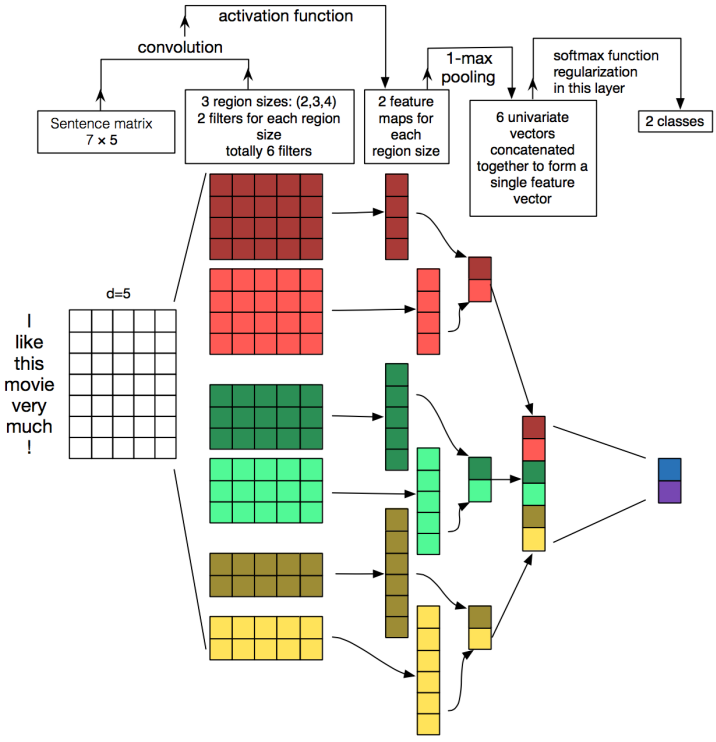
2.2.3 Redes Neurais Convolucionais

Redes Neurais Convolucionais, tradução de *Convolutional Neural Networks* (CNN), inspiram-se no funcionamento do córtex visual, parte do cérebro humano responsável por processar a informação que chega da retina, guardando-a em uma espécie de memória visual (BEZERRA, 2016). Na sua forma mais comum, uma CNN possui seis camadas: entrada, convolução, subamostragem, normalização, contraste e conexão.

Como as CNNs são bastante eficazes no reconhecimento e classificação de imagens, é comum achar que são especializadas apenas neste tipo de tarefa. Entretanto, mais recentemente, elas vêm sendo consideravelmente aplicadas em problemas do PLN, como análise e classificação de textos (TALON, 2019). Inclusive, foi demonstrado que esse tipo de arquitetura de rede alcança resultados rápidos para essas tarefas de classificação, e que captura dependências de contexto de forma eficiente (ZHANG; WALLACE, 2015; LI *et al.*, 2020).

Nesse cenário, em vez de *pixels* da imagem, uma CNN recebe como entrada um tipo de representação textual distribuída. Considere a Figura 13 como exemplo ilustrativo, em que a camada de entrada recebe a frase em inglês "*I like this movie very much!*", e incorpora cada palavra da sequência textual a um vetor de cinco posições.

Figura 13 – Exemplo de arquitetura da CNN para classificação de texto



Fonte: Zhang e Wallace (2015).

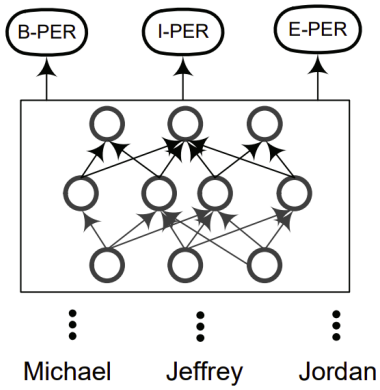
Após a camada de entrada, uma camada de convolução com seis filtros é usada para produzir características em torno de cada palavra: os dois primeiros filtros passam mais de quatro palavras por vez; os dois do meio passam mais de três palavras; e, os últimos dois passam mais de duas por vez. Os filtros da primeira camada capturam atributos de linguagem semelhantes e geram dois mapas de *features* com comprimento variado. Em seguida, o maior número de cada mapa é registrado. No final, todas as seis *features* registradas são concatenadas para formar um vetor na penúltima camada. A camada final recebe esse vetor como entrada e o usa para classificar a sentença de acordo com duas classes.

2.2.4 Decodificação de Rótulos

Para tarefas de rotulagem de sequência, é benéfico considerar as correlações entre os rótulos e decodificar em conjunto a melhor sequência para uma determinada sentença (MA; HOVY, 2016). Esse procedimento é chamado de decodificação de rótulos, realizado no estágio final de um modelo NER, no qual são utilizados métodos que recebem como entrada representações dependentes do contexto, para produzir uma sequência de rótulos correspondentes aos *tokens* da sentença de entrada.

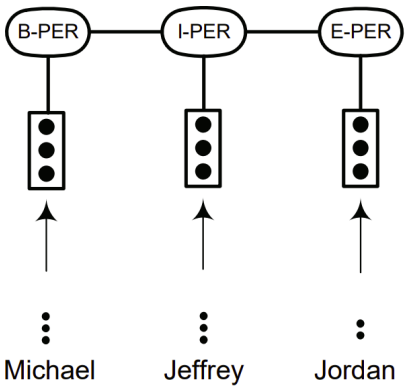
As Figuras 14 e 15 sumarizam duas arquiteturas de decodificadores de rótulos: Rede *Multilayer Perceptron* (MLP) com a função de ativação *Softmax* (MLP+*Softmax*) e *Conditional Random Fields* (CRF). Uma variedade de modelos NER usam essas arquiteturas como decodificadores de rótulos, sendo a CRF a escolha mais comum (LI *et al.*, 2020).

Figura 14 – MLP+*Softmax*



Fonte: Li *et al.* (2020).

Figura 15 – CRF



Fonte: Li *et al.* (2020).

Tipicamente, uma MLP consiste em uma camada de entrada, várias camadas ocultas, e uma camada de saída. Um ponto interessante da sua arquitetura é que representa uma gene-

ralização do *perceptron*: um único neurônio com pesos sinápticos e *bias*, tradicionalmente um algoritmo simples e uniforme contendo um tipo de função de ativação (HAYKIN, 2007). Esse modelo também utiliza uma técnica de aprendizado supervisionado chamada *back-propagation*, baseada na regra de aprendizagem por correção de erro, que consiste em dois passos através das diferentes camadas da rede: a propagação (um passo para frente), e a retro-propagação (um passo para trás).

Com a adição de uma camada com a função de ativação *Softmax* à arquitetura da rede MLP, é gerada uma distribuição de probabilidade consistente sobre as *N* classes de saída (CHOLLET, 2017). Como decodificador de rótulo, o modelo MLP+*Softmax* trata a tarefa de rotulagem de sequência como um problema de classificação multi-classe, em que as classes são preditas independentemente para cada palavra com base no contexto (LI *et al.*, 2020).

Diferentemente da arquitetura de rede neural MLP+*Softmax*, CRF é um modelo estatístico linear, grandemente utilizado em abordagens de aprendizagem supervisionada baseada em *features* (LI *et al.*, 2020). Devido ao fato de que campos aleatórios são condicionados globalmente dada uma sequência de observação, essa arquitetura concentra-se na sentença como um todo, em vez de cada *token* individualmente, no qual apenas as interações entre dois ou mais rótulos sucessivos que obtiveram uma pontuação de predição mais alta são consideradas.

O modelo de sequência CRF ainda oferece uma combinação única de propriedades para a segmentação e rotulagem de sequência: combinação de características de observação arbitrárias, sobrepostas, e aglomerativas do passado e do futuro; treinamento e decodificação eficientes baseados em programação dinâmica; e, estimativa de parâmetros garantida para encontrar o campo global ótimo (LAFFERTY *et al.*, 2001).

2.3 Humano no Ciclo

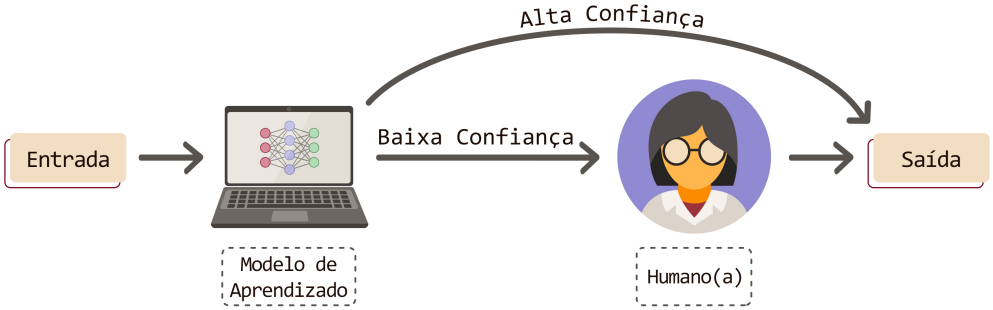
Song e Xiao (2015) afirmam que embora exista um progresso notável na melhoria de modelos do Aprendizado Profundo e no desenvolvimento de sistemas de treinamento de alto desempenho, faltam avanços na construção dos conjuntos de dados. Enquanto os modelos ficam mais “profundos” e o poder computacional aumenta, a quantidade de dados para treinamento e avaliação não está aumentando, mas tornando-se cada vez menor e desatualizada.

O que é um problema, já que os modelos do Aprendizado Profundo, incluindo os do Reconhecimento de Entidade Nomeada, requerem um grande volume de dados anotados para o treinamento. Além disso, o processo de anotação continua sendo caro e demorado, e para muitas

linguagens e domínios específicos, são necessários especialistas do domínio para realizar esta tarefa, em que a qualidade e consistência da anotação são preocupantes, por conta da ocorrência natural da ambiguidade nas línguas naturais (LI *et al.*, 2020).

Com o intuito de resolver esse problema, pesquisadores desenvolveram uma técnica capaz de ampliar o tamanho dos conjuntos de dados, e de garantir estabilidade e generalização aos modelos (SONG; XIAO, 2015). Essa técnica, chamada de *Human in the Loop* (HITL) ou Humano no Ciclo, é definida como um modelo de aprendizado híbrido, no qual um humano pode intervir e anular as decisões feitas pelo algoritmo quando são menos prováveis de serem corretas. A Figura 16 inclui os passos de um exemplo de aplicação do HITL.

Figura 16 – *Pipeline* de uma aplicação com HITL



Fonte: Elaborado pela autora.

Via HITL, dados adicionais são incorporados no processo de tomada de decisão e adicionados aos modelos de aprendizagem. Sistemas que utilizam HITL trazem benefícios para diversas áreas de pesquisa e, conforme analisado por Park *et al.* (2012), também é bastante importante na criminologia, contribuindo significativamente para a Teoria Criminológica e prevenção de crimes. Como o foco deste trabalho está em propor uma melhoria à divulgação de informações referentes à segurança pública, o HITL será usado, pois o fluxo de aprendizado da ferramenta HNERD (seleção e/ou correção dos dados anotados e treino do modelo) baseia-se nas ações de humanos presentes no ciclo.

2.4 Métricas de Avaliação

A fim de estimar de forma confiável o desempenho de modelos NER neurais profundos, nesta pesquisa, os modelos são avaliados com quatro métricas empíricas amplamente utilizadas para a estimativa de desempenho de sistemas do PLN: *Precision*, *Recall*, F_1-*score* e Acurácia Balanceada. Essas métricas estão definidas abaixo, mas, para entender melhor cada

uma, primeiro é necessário compreender alguns conceitos.

A Tabela 1 apresenta uma matriz de confusão multi-classe 3×3 definida como C^3, onde PER (PESSOA), LOC (LOCALIZAÇÃO) e ORG (ORGANIZAÇÃO) representam três ENs genéricas de um modelo NER tradicional qualquer. Esse modelo fictício nos diz que, dada a descrição de uma entidade, ele identifica se é uma pessoa, um local ou uma organização. A matriz de confusão C^3 possui 25 entidades classificadas, com as células marcadas em verde indicando os termos preditos corretamente, e as células em rosa indicando as entidades confundidas pelo modelo.

Tabela 1 – Matriz de confusão multi-classe 3×3 denotada por C^3

		Valores Reais			
		PER	LOC	ORG	Total
Valores Preditos	PER	4	6	3	13
	LOC	1	2	0	3
	ORG	1	2	6	9
	Total	6	10	9	25

Fonte: Elaborado pela autora.

Sabe-se que o NER envolve essencialmente duas sub-tarefas: detecção de limites de uma entidade e identificação do seu tipo. Na avaliação de correspondência exata, uma instância corretamente predita requer um meio de avaliação para identificar corretamente seu limite e tipo, simultaneamente (LI *et al.*, 2020). Através de uma matriz de confusão, é possível indicar os erros e acertos de um modelo, mais especificamente, os números de falsos positivos, falsos negativos e verdadeiros positivos.

- **Falso Positivo (FP):** entidade predita por um modelo, mas que não faz parte dos valores reais. Exemplo: na matriz de confusão C^3, todos os números pertencentes à linha da classe PER dos Valores Preditos ($4+6+3=13$) são falsos positivos.
- **Falso Negativo (FN):** entidade que não é predita por um modelo, mas que faz parte dos valores reais. Exemplo: na matriz de confusão C^3, todos os números pertencentes à coluna da classe PER dos Valores Reais ($4+1+1=6$) são falsos negativos.
- **Verdadeiro Positivo (VP):** entidade que é predita por um modelo e que faz parte dos valores reais. Exemplo: os valores marcados em verde (4, 2 e 6) na matriz de confusão C^3 são verdadeiros positivos.

A definição dos indicadores descritos acima é o que determina as métricas *Precision*,

Recall e F_1-*score*, listadas a seguir.

- ***Precision* (P, Precisão):** razão entre verdadeiros positivos e a soma de verdadeiros positivos com falsos positivos, definida pela Equação 2.6, indica a porcentagem dos resultados do modelo que foram preditos corretamente.

$$P = \frac{VP}{VP + FP} \tag{2.6}$$

- ***Recall* (R, Revocação):** razão entre verdadeiros positivos e a soma de verdadeiros positivos com falsos negativos, definida pela Equação 2.7, indica a porcentagem do total de entidades preditas corretamente pelo modelo.

$$R = \frac{VP}{VP + FN} \tag{2.7}$$

- **F_1-*score* (F_1, Medida F_1):** média harmônica das métricas *Precision* e *Recall*, definida pela Equação 2.8, é uma forma de observar somente uma métrica ao invés de duas, mostrando o balanço entre P e R.

$$F_1 = 2 \cdot \frac{P \times R}{P + R} \tag{2.8}$$

As métricas P, R e F_1 são usadas para computar cada rótulo individualmente. Uma forma de considerar o desempenho para múltiplos tipos de entidades é calcular a média ponderada dos valores das métricas para cada rótulo, utilizando como peso a quantidade de exemplos. Considere uma métrica M escolhida para medir a performance de cada EN, e a quantidade de ocorrências dos rótulos $\#Obs$ de tamanho N. A média ponderada de M para o conjunto de rótulos L pode ser definida como:

$$\sum_{l=1}^{|L|} M_l \frac{\#Obs_l}{N}. \tag{2.9}$$

Em relação a métrica Acurácia Balanceada (AB), um meio de avaliação geral intuitivamente simples, no qual a qualidade preditiva é medida para cada classe de forma independente e agregada, este tipo de acurácia evita uma estimativa de desempenho melhor do que realmente é, dado um conjunto de dados desbalanceado (MOSLEY, 2013; PEDREGOSA *et al.*, 2011). Seu cálculo é feito obtendo o *Recall* independentemente para diferentes tipos de entidades, e

dividindo a soma dos valores pela quantidade de rótulos. A AB é definida pela Equação 2.10 abaixo:

$$\mathrm{AB} = \frac{1}{|L|} \sum_{l=1}^{|L|} \mathrm{R}_l, \tag{2.10}$$

onde $|L|$ é a quantidade total de rótulos do conjunto L, e R_l o valor de *Recall* da classe l.

No próximo capítulo, trabalhos que utilizam alguns dos conceitos explicados até aqui são apresentados e discutidos.

3 TRABALHOS RELACIONADOS

Os trabalhos apresentados neste capítulo indicam o contexto no qual esta monografia está inserida. Inicialmente, apresenta-se os trabalhos relacionados que exploram dados de domínio criminal para o reconhecimento e extração de entidades nomeadas. Em seguida, são identificados os trabalhos que utilizam técnicas para o tratamento de dados desbalanceados, de domínios variados e com foco em outras tarefas do PLN, além do Reconhecimento de Entidade Nomeada. Na parte final deste capítulo, uma comparação é feita entre os trabalhos relacionados e este trabalho, com base em um conjunto de dez (10) características específicas.

3.1 Reconhecimento e Extração de Entidades para o Domínio Criminal

Dos trabalhos que aplicam técnicas automatizadas para analisar diferentes tipos de crimes ou outras informações importantes de domínio criminal, quatro foram selecionados por abordarem principalmente, dentre outras similaridades, o reconhecimento e extração de entidades nomeadas de relatórios de narrativas policiais e de outros tipos de textos criminais.

Chen *et al.* (2004b) e Shabat *et al.* (2014) possuem a mesma motivação social que o trabalho proposto: a de realizar uma análise precisa e eficiente do extenso volume de dados criminais existentes, por se tratar de um grande problema enfrentado pela população, por jornalistas, analistas, e autoridades policias. Apesar de Coelho da Silva *et al.* (2019a) e Coelho da Silva *et al.* (2019b) possuírem motivações diferentes, seus conjuntos de dados fazem parte do domínio criminal, devido à dificuldade existente no reconhecimento de certos tipos de entidades em textos deste domínio.

Resumidamente, todos esses trabalhos baseiam-se em um conjunto de dados de domínio criminal para obter resultados de qualidade, variando nas classes exploradas e soluções propostas. Seus contextos e principais características estão descritos a seguir.

3.1.1 Crime Data Mining: A General Framework and Some Examples

O estudo de Chen *et al.* (2004b) apresenta um *framework* que aplica distintas operações sobre dados criminais. A estrutura foi baseada na experiência adquirida com o projeto *Coplink*[1], onde foi desenvolvida uma aplicação voltada para a resolução de problemas enfrentados pela comunidade policial do estado do Arizona, Estados Unidos da América. O

[1] https://www.ncjrs.gov/pdffiles1/nij/grants/190988.pdf

projeto foi criado pelo Laboratório de Inteligência Artificial da Universidade do Arizona, com apoio dos Departamentos de Polícia de Tucson e Phoenix.

A ferramenta mostra visualmente as relações entre quatro categorias principais para mineração de dados criminais (Extração de Entidades, Associação, Predição e Visualização de Padrões), e os tipos de crimes (violações de trânsito, crime sexual, roubo, fraude, incêndio criminoso, gangues/drogas, crimes violentos e cibernéticos), selecionados por um especialista do domínio, um detetive local, com mais de 30 anos de experiência.

Cada categoria representa um conjunto de técnicas para uso em certos tipos de análise de crimes. Por exemplo, os investigadores podem usar uma rede neural MLP para o reconhecimento de entidades nomeadas em relatórios policiais relacionados a narcóticos. Para essa tarefa em específico, foram usados 36 relatórios não rotulados escritos em inglês do Departamento de Polícia de Phoenix. Antes de treinar o modelo, o especialista do domínio identificou manualmente todas as entidades que pertenciam às cinco classes de interesse (nomes de pessoas, endereços, veículos, narcóticos e características físicas), através de uma adaptação do sistema *AI Entity Extractor*.

3.1.2 Improving Named Entity Recognition using Deep Learning with Human in the Loop

Em Coelho da Silva *et al.* (2019b) é proposta uma estrutura interativa denominada *Human Named Entity Recognition with Deep Learning* (HNERD), que auxilia o usuário nas tarefas de classificação NER. A ferramenta incorpora modelos NER do *framework spaCy*[2] e baseados na biblioteca *Keras*[3], e considera dois humanos no ciclo: o revisor, um especialista do domínio, que pode corrigir possíveis erros de classificação; e, o cientista de dados, o responsável pelo ajuste dos modelos de Aprendizado Profundo.

Como forma de avaliar a qualidade desse *framework*, foi utilizado um conjunto de dados com textos não rotulados de documentos policiais em português sobre homicídios da cidade de Fortaleza, Ceará, Brasil. A validação do processo de anotação envolveu um modelo e um humano no ciclo: o modelo pré-treinado *Wiki*NER do *spaCy*, formado pelas redes neurais CNN e MLP+*Softmax*, classificou os textos de acordo com mais de 20 classes (como pessoa, local, organização, arma de fogo, arma branca, dentre outras), e revisores especialistas adicionaram novas anotações, como também editaram as feitas pelo modelo conforme necessário.

[2] https://spacy.io/usage/training#ner
[3] https://keras.io/why-use-keras

A partir dessas análises, o HNERD criou um novo modelo NER neural profundo para o domínio criminal capaz de reconhecer ENs em textos de homicídios.

3.1.3 Named Entity Recognition in Crime Using Machine Learning Approach

Shabat *et al.* (2014) propõem um modelo NER que obtém informações relevantes de textos criminais da *Internet*, encontrados em artigos de notícias, *blogs* e redes sociais. O modelo foi desenvolvido com o intuito de processar notícias e tópicos sobre crimes, bem como detectar novos textos sobre um tipo específico de delito, a partir do reconhecimento de tipos de armas, nacionalidades de vítimas e suspeitos, e localização dos crimes. Para a realização dessa pesquisa, um total de 500 documentos em língua inglesa foram coletados da Agência Nacional de Notícias da Malásia[4] e anotados manualmente, com cada arquivo relatando um ou mais crimes.

Na experimentação, a etapa de pré-processamento envolveu o processo de tokenização, remoção das *stopwords* (conjunto de palavras comumente usadas em um determinado idioma), e POS *tagging*. Também foi feita a extração de um conjunto de *features* e a criação de dois modelos NER com os algoritmos de Aprendizado de Máquina *Support Vector Machine* (SVM) e *Naïve Bayes* (NB). O classificador SVM atingiu o melhor desempenho na avaliação das três classes: armas, nacionalidade e localização, obtendo valores iguais a $91,08\%$, $96,25\%$ e $89,28\%$, respectivamente, para a métrica F_1-*score*.

3.1.4 Novel Approach for Label Disambiguation via Deep Learning

Tal como na pesquisa de Coelho da Silva *et al.* (2019b), também fez-se uso de relatórios policiais sobre homicídios da cidade de Fortaleza para a validação do trabalho de Coelho da Silva *et al.* (2019a), que propõe uma arquitetura de rede neural profunda capaz de resolver o problema de Desambiguação de Rótulos. Esse problema ocorre pois modelos NER tradicionais podem classificar duas entidades diferentes com um mesmo rótulo, por conta da mesma ortografia e da ambiguidade presente nas linguagens naturais.

O conjunto de dados possui 1.749 textos datados de 2014 até 2018, anotado manualmente com o auxílio de especialistas do domínio, que identificaram as classes VÍTIMA, ASSASSINO, e OUTROS. A arquitetura de rede neural proposta, chamada de *Char*-BLSTM-CRF, se trata da combinação do codificador de contexto BLSTM e do decodificador de rótulo CRF. O termo "*Char*" indica a concatenação feita entre uma camada de *word embeddings* e uma camada

[4] https://www.bernama.com/en/

LSTM para *character embeddings*. Os experimentos mostram que a rede *Char*-BLSTM-CRF supera os dois *baselines* (BLSTM-CRF, e RNN bidirecional com CRF ou BRNN-CRF) em termos de qualidade.

3.2 Aplicação de Técnicas para Dados Desbalanceados

Os trabalhos relacionados da Seção 3.1 não abordam técnicas para dados desbalanceados, apesar da existência de classes raras na distribuição dos conjuntos de dados de alguns trabalhos. Desta forma, para exemplificar casos de sucesso no tratamento do desbalanceamento de classes, são apresentados abaixo dois trabalhos similares, um sendo de domínio jurídico, e o outro de domínio genérico, com textos de fontes e assuntos variados.

3.2.1 Fine-Grained Named Entity Recognition in Legal Documents

A pesquisa de Leitner *et al.* (2019) descreve abordagens do NER para textos em alemão de domínio jurídico. Foi desenvolvida uma base de dados chamada de *Legal Entity Recognition* (LER), com 750 documentos de decisões judiciais publicadas no portal *Rechtsprechung im Internet*[5]. O processo de anotação foi feito manualmente por um estudante de Linguística Computacional usando o *WebAnno*[6], uma ferramenta *Web* para anotação de estruturas sintáticas e semânticas.

A base de dados LER inclui duas variantes de anotação: uma feita com 7 *coarse-grained* classes (ENs mais genéricas, como pessoa e localização), e outra feita com 19 *fine-grained* classes (ENs mais específicas. Por exemplo, a classe genérica pessoa é divida nas classes pessoa, juiz e advogado). Devido à presença de classes desbalanceadas, foi feita uma divisão dos dados em 10 subconjuntos proporcionais, com a aplicação do método *k-fold* da validação cruzada (10-*fold*), juntamente com a estratificação. Dez iterações foram feitas, de forma que cada parte do conjunto de dados fosse usada nove vezes para treinamento e uma vez para validação.

Para a tarefa de classificação NER, seis arquiteturas diferentes foram desenvolvidas: três com o modelo CRF e três com a rede BLSTM, cada uma das quais foi treinada com as anotações *coarse-* e *fine-grained*. De todos os métodos, o modelo BLSTM-CRF+, similar ao de Coelho da Silva *et al.* (2019a), mas com uma camada de *character embeddings* BLSTM, foi o que obteve a melhor performance, com F_1-*score* geral de $95,95\%$ para as classes *coarse-grained*

[5] http://www.rechtsprechung-im-internet.de/jportal/portal/page/bsjrsprod.psml
[6] https://webanno.github.io/webanno/

e de $95,46\%$ para as *fine-grained*.

3.2.2 Dice Loss for Data-imbalanced NLP Tasks

Li *et al.* (2019) propõem a função de custo *Dice Loss* (DL) em substituição da função *Categorical Cross-Entropy* (CCE) para tarefas do PLN com dados desbalanceados. Essa função é baseada no Coeficiente *Sørensen-Dice* ou Índice *Tversky*, que atribui um esquema de ajuste de peso dinâmico às classes, e foi anteriormente apresentada no Capítulo 2: Fundamentação Teórica, visto que é utilizada na experimentação deste trabalho.

O método proposto foi aplicado e avaliado para quatro tarefas diferentes do PLN: POS *tagging*, NER, Compreensão de Leitura de Máquina (CLM), e identificação de paráfrases. Analisando apenas a tarefa de NER, o modelo proposto BERT-MRC+DSC foi treinado com quatro base de dados diferentes com textos em inglês e chinês, e comparado com seis *baselines*, atingindo os melhores resultados. Desses quatro conjunto de dados pré-anotados, o modelo performou melhor (F_1-*score* geral de $96,72\%$) para o *Chinese* MSRA, que possui três tipos de ENs genéricas (PER, LOC e ORG), e mais de 45 mil textos.

3.3 Análise Comparativa e Considerações Finais

Esta seção apresenta uma análise comparativa dos trabalhos descritos neste capítulo. Para realizar essa comparação, foi definido um conjunto de 10 características principais. O Quadro 1 sumariza as principais características e os diferenciais de todos os trabalhos explanados, com os seguintes símbolos denotando que: (—), o trabalho não menciona a característica; (**?**), o trabalho não especifica como realiza a característica; (✕), a característica não é aplicável ao trabalho.

O '*domínio*' dos trabalhos é a primeira característica abordada. Dos seis trabalhos listados, os de Leitner *et al.* (2019), e Li *et al.* (2019) não possuem domínio criminal, uma vez que o primeiro possui domínio jurídico, e o segundo um domínio genérico. Além do '*domínio*', os desafios para a construção de um corpus rotulado sempre depende do idioma, por isso a escolha da segunda e terceira características: '*conjunto de dados*' e '*idioma*'. Todos os trabalhos dispõem de conjuntos de dados de documentos com linguagens diferentes, mas nenhum contendo narrativas CVP em português. Como o trabalho de Li *et al.* (2019) considera quatro conjuntos de dados diferentes (dois em chinês: *OntoNotes*4.05 e MSRA, e dois em inglês: *OntoNotes*5.0 e

*Co*NLL2003), é destacado a base de dados que obteve os melhores resultados para o NER.

A quarta característica indica o '*processo de anotação*', não realizado por Li *et al.* (2019), dado que seus diferentes conjuntos de dados são pré-anotados. Os trabalhos de Coelho da Silva *et al.* (2019a) e Shabat *et al.* (2014) o fazem manualmente, sem expor muitos detalhes sobre a realização do procedimento. Coelho da Silva *et al.* (2019b) utilizam um humano no ciclo e o modelo *Wiki*NER do *framework spaCy* presente na ferramenta HNERD, à medida que Chen *et al.* (2004b) e Leitner *et al.* (2019) dispõem de sistemas voltados para realizar essa técnica manualmente.

Apenas Coelho da Silva *et al.* (2019b), Chen *et al.* (2004b) e Leitner *et al.* (2019) abordaram a quinta característica, que diz respeito ao '*feedback do usuário*' durante o '*processo de anotação*'. Coelho da Silva *et al.* (2019b) fizeram correções sobre a predição do modelo *Wiki*NER e, em contrapartida, Leitner *et al.* (2019) e Chen *et al.* (2004b) selecionaram as entidades úteis com o auxílio de ferramentas. Este trabalho também seleciona as entidades úteis com o auxílio do HNERD, corrigindo as predições feitas pelo próprio modelo CVP criado na ferramenta.

O maior diferencial deste trabalho está no uso de 36 classes distintas para o Reconhecimento de Entidade Nomeada. No Capítulo 5: Experimentos e Resultados, é abordado com mais detalhes quais são elas e porque são tantas. Pode-se notar através da característica '*quantidade de classes*' que apenas Coelho da Silva *et al.* (2019a) e Leitner *et al.* (2019) possuem quantidades aproximadas ao do trabalho proposto.

Como justificado na Seção 3.2, dois trabalhos relacionados que não são do domínio criminal aplicam técnicas para o problema de desbalanceamento de classes, dispostas na sétima característica '*técnicas para desbalanceamento*'. Este trabalho realiza estratificação dos dados e adiciona funções de perda diferentes (*Dice Loss* e a CB-CCE, uma variação da CB *Loss*) às arquiteturas dos modelos experimentados.

Para a '*representação textual*', três trabalhos mencionam os métodos usados. Leitner *et al.* (2019) aplicam *word embeddings* pré-treinadas de língua alemã disponíveis *online*[7], criadas com o modelo *Word2Vec* (sem especificação da estratégia de treinamento: CBOW ou *Skip-Gram*). Em Coelho da Silva *et al.* (2019a) são geradas *word embeddinds* específicas do domínio, com os vetores da camada de *embedding* treinados a partir de um vocabulário formado com relatórios policiais da cidade de Fortaleza, do ano de 2010 a 2018. No trabalho de Shabat *et al.* (2014),

[7] http://www.dialog-21.ru/en/germeval2014/

os vetores para representação textual são formados agrupando-se três diferentes conjuntos de *features*: com base em POS *tagging*, nos afixos das palavras, e no contexto.

A nona característica '*tratamento de OOV*' só é aplicável aos trabalhos que utilizam *word embeddings*. Coelho da Silva *et al.* (2019a) e Leitner *et al.* (2019) implementam uma camada de *character embeddings*, e apesar de não explicitarem o propósito para seu uso, sabe-se que a representação a nível de caractere lida naturalmente com problemas de OOV (LI *et al.*, 2020). Como este trabalho utiliza *word embeddings* específicas do domínio e pré-treinadas nos experimentos, o tratamento de palavras OOV é feito utilizando a métrica *Edit Distance* para substituição destes termos. Mais informações sobre esse tratamento são descritas no Capítulo 5.

A última e décima característica presente no Quadro 1 indica o '*modelo NER*' utilizado. Com exceção da pesquisa de Shabat *et al.* (2014) que aborda modelos do Aprendizado de Máquina, todos os outros trabalhos relacionados empregam modelos do Aprendizado Profundo, com destaque para Coelho da Silva *et al.* (2019a) e Leitner *et al.* (2019), que fazem uso da arquitetura BLSTM-CRF, assim como o trabalho proposto.

Em suma, os esforços deste trabalho concentraram-se na criação de uma base de dados rotulada de narrativas CVP, na exploração de estratégias que forneçam contexto para este tipo de texto, e no problema de desbalanceamento de classes. O último ponto torna-se crucial quando são analisados os trabalhos relacionados que tratam desse problema, pois possuem tipos de entidades muito específicas, de forma que são facilmente reconhecidas pelos modelos, sem precisar atentar para o fato de que palavras ou sentenças iguais podem ter rótulos diferentes, dependendo do contexto.

Quadro 1 – Comparativo entre os trabalhos relacionados e o trabalho proposto

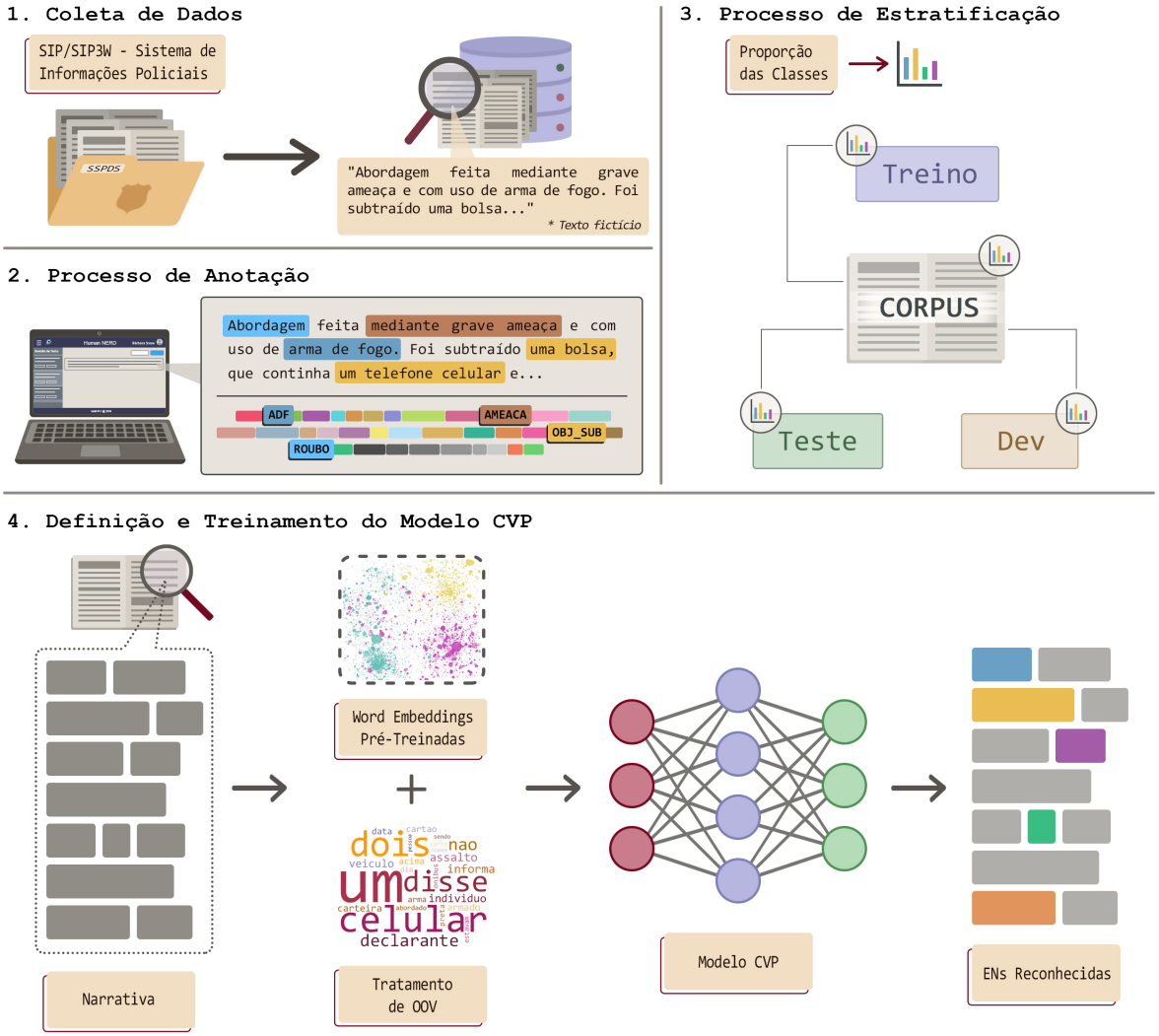
		Chen *et al.* (2004b)	**Coelho da Silva *et al.* (2019b)**	**Shabat *et al.* (2014)**	**Coelho da Silva *et al.* (2019a)**	**Leitner *et al.* (2019)**	**Li *et al.* (2019)**	**Este trabalho**
1	**Domínio**	Criminal	Criminal	Criminal	Criminal	Jurídico	Genérico	Criminal
2	**Conjunto de dados**	Relatórios policiais sobre narcóticos	Relatórios policiais sobre homicídios	Textos de artigos de notícias, *blogs* e redes sociais	Relatórios policiais sobre homicídios	Documentos de decisões judiciais	MSRA	Relatórios policiais sobre roubos
3	**Idioma**	Inglês	Português	Inglês	Português	Alemão	Chinês	Português
4	**Processo de anotação**	Manual, com o auxílio do AI *Entity Extractor*	Manual e com o modelo *Wiki*NER presente no HNERD	Manual	Manual	Manual, com o auxílio do *WebAnno*	×	Manual, com o auxílio do HNERD
5	***Feedback* do usuário**	Seleção de entidades	Correções sobre as predições do modelo	**?**	**?**	Seleção de entidades	×	Seleção de entidades e correções
6	**Quantidade de classes**	5 classes	+20 classes	3 classes	3 classes	7 genéricas & 19 específicas	3 classes	36 classes
7	**Técnica(s) para desbalanceamento**	—	—	—	—	Nível dos dados (*k-fold* e estratificação)	Nível algorítmico (DSC *Loss*)	Nível dos dados e algorítmico
8	**Representação textual**	—	—	Extração de um conjunto de *features*	*Word* e *character embeddings* específicas do domínio	*Word embeddings* pré-treinadas e *character embeddings*	—	*Word embeddings* específicas do domínio e pré-treinadas
9	**Tratamento de OOV***	—	—	×	Representação *character-level*	Representação *character-level*	—	Subs. com Levenshtein *Edit Distance*
10	**Modelo(s) NER**	MLP	CNN com MLP+*Softmax*	SVM	*Char*-BLSTM-CRF	BLSTM-CRF+	BERT-MRC	BLSTM-CRF

Fonte: Elaborado pela autora.
Significado dos símbolos: (—), a característica não é mencionada pelo trabalho; (**?**), a característica não é explanada pelo trabalho; (×), a característica não é aplicável ao trabalho; (*), esta característica é aplicável apenas para os trabalhos que utilizam *word embeddings*.

4 PROCEDIMENTOS METODOLÓGICOS

A metodologia deste trabalho consiste de cinco etapas: coleta de dados, processo de anotação, divisão estratificada dos dados, definição e treino do modelo CVP, e interpretação e análise dos resultados. A Figura 17 apresenta de forma ilustrativa o fluxo completo das quatro primeiras etapas.

Figura 17 – Fluxo de execução dos quatro primeiros procedimentos metodológicos



Fonte: Elaborado pela autora.
Descrição: (1) os documentos são coletados para criação do corpus CVP; (2) os dados são anotados de acordo com rótulos pré-definidos; (3) é feita a divisão estratificada dos dados em treino, teste e desenvolvimento; (4) alguns procedimentos importantes, como o tratamento de OOV, são realizados para geração do modelo CVP capaz de reconhecer ENs de narrativas de roubos.

Informações detalhadas sobre cada etapa da Figura 17 são descritas nas próximas seções deste capítulo, junto com os métodos selecionados para análise dos resultados dos experimentos.

4.1 Coleta de Dados

Tal como representado no passo um (1) da Figura 17, o primeiro procedimento metodológico consiste em coletar textos de naturezas diferentes, como boletins de ocorrência (BOs), termos circunstanciado de ocorrência (TCOs), inquéritos policiais, atos infracionais, e depoimentos, da base de dados que contém as duas categorias de CVP do Sistema de Informações Policiais (SIP/SIP3W) da Secretaria da Segurança Pública e Defesa Social (SSPDS).

4.2 Processo de Anotação

O processo de anotação é o ponto mais crítico na criação de bases de dados. Primeiramente, é definido o esquema de anotação para o corpus CVP, onde faz-se um levantamento dos rótulos que serão usados para a tarefa de anotação. A escolha desses rótulos é bastante importante, visto que eles são usados para identificar quais as entidades nomeadas mais relevantes para o domínio.

A ferramenta HNERD de Coelho da Silva *et al.* (2019b) é utilizada para anotação manual das narrativas CVP. Apesar de ser uma estrutura completa para auxiliar o usuário nas tarefas de classificação NER, esse *framework* interativo é utilizado apenas para o processo de anotação. Como o HNERD incorpora o modelo de aprendizado profundo do *spaCy* para o reconhecimento de entidades nomeadas, os dados são anotados com o formato de treinamento do *spaCy*, que considera os textos, os deslocamentos de caracteres e os rótulos de cada entidade.

4.3 Processo de Estratificação

Muitas tarefas do PLN enfrentam o grave problema de desbalanceamento de classes. No conjunto de dados CVP, a distribuição das entidades nomeadas é inerentemente desbalanceada por conta do domínio, pois a ocorrência natural de algumas classes facilmente dominam outras. Além disso, muitos dos termos presentes nas narrativas não são entidades nomeadas. A fim de compensar a existência de classes majoritárias e minoritárias, o processo de estratificação é aplicado para a obtenção dos conjuntos oficiais de treinamento, desenvolvimento e teste.

Porém, este trabalho emprega o NER como uma tarefa de rotulagem de sequência, em que, dada uma amostra de dados anotados, grupos podem ser formados com base nas diferentes combinações de rótulos. Nesse cenário de dados multi-rótulo, com rótulos sendo atribuídos a cada *token* de uma sentença, a capacidade de traçar uma distribuição semelhante para todo o

corpus ainda é restritiva (AGUILAR *et al.*, 2020). Como meio de fornecer divisões proporcionais considerando esses critérios, é utilizado o processo de estratificação iterativo proposto por Sechidis *et al.* (2011), voltado especificamente para a estratificação de dados multi-rótulo, disponível na biblioteca *Scikit-multilearn*[1] para a linguagem de programação *Python*.

4.4 Definição e Treinamento do Modelo CVP

A arquitetura de rede neural para o modelo CVP é baseada nas de Lample *et al.* (2016) e de Coelho da Silva *et al.* (2019a), que utilizam um método empregado em muitos modelos NER neurais profundos: uma camada CRF no topo de uma camada BLSTM, ou, apenas, modelo BLSTM-CRF. Como a integração de *word embeddings* (pré-treinadas ou não) é importante e necessária para modelos NER do Aprendizado Profundo, a camada CRF realiza a captura de forma bastante eficaz das dependências de transição de rótulos, até mesmo para os vetores treinados com modelos estatísticos, como *Word2Vec* e GloVe (LI *et al.*, 2020).

No modelo CVP, as palavras de um documento não são divididas em uma série de tarefas de classificação distintas e separadas (uma por entidade de destino), mas tratadas diretamente no nível da sentença. Um esquema caracterizando as camadas do modelo de rotulagem de sequência proposto é apresentado na Figura 18, que consiste, de baixo para cima, em:

Entrada: um conjunto de palavras de um texto $S = \{x_1, x_2, \cdots, x_N\}$, onde i representa a posição de uma palavra $x_i \in V$ na sentença, sendo V o vocabulário CVP. No final, cada símbolo x_i deve ser marcado com um rótulo $y_j \in L = \{y_1, y_2, \cdots, y_{N'}\}$.

Camada 1: uma camada de *word embeddings* que converte cada palavra x_i em um vetor D-dimensional e_{x_i}, via matriz de *embedding* $W \in \mathbb{R}^{dx|V|}$.

Camada 2: uma camada BLSTM que recebe os vetores de *word embeddings* para combinação das representações a nível de palavras. A primeira camada LSTM representa a palavra x_i e seu contexto à esquerda, e a segunda camada LSTM representa a palavra x_i e seu contexto à direita.

Camada 3: os vetores da camada BLSTM são alimentados em uma camada CRF para decodificar em conjunto a melhor sequência de rótulos y_j referente às classes em L.

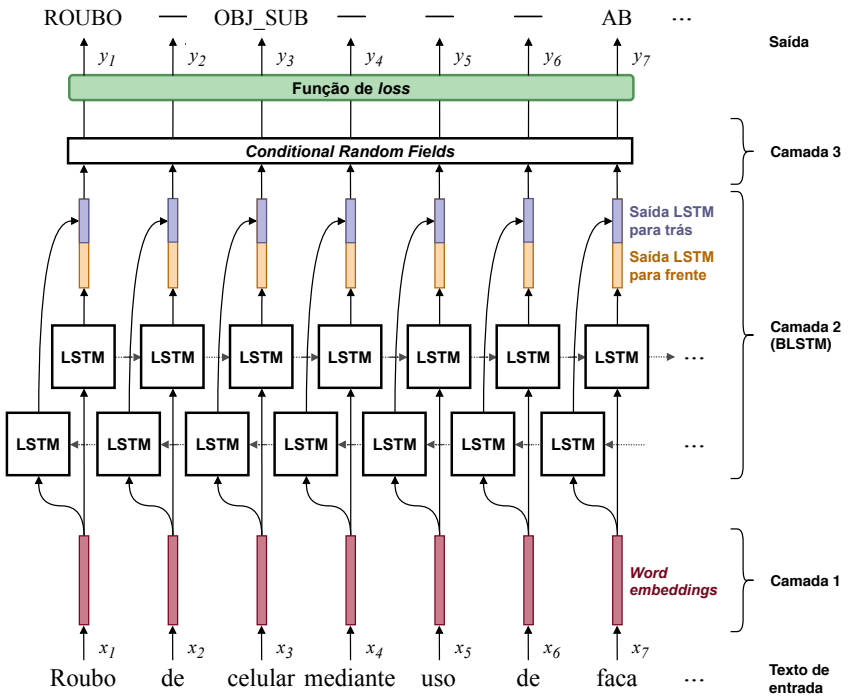
Função de *loss* (perda): à cada iteração do treinamento, a função de *loss* é usada para medir a

[1] http://scikit.ml/stratification.html

diferença entre o rótulo predito e o rótulo verdadeiro. A diferença é propagada de volta para as camadas ocultas para o ajuste dos pesos.

Saída: a melhor sequência de rótulos y_j para as palavras da sentença de entrada S.

Figura 18 – Arquitetura geral do modelo CVP: recebe como entrada uma sequência de palavras e retorna uma sequência de rótulos de entidade



Fonte: Elaborado pela autora.

Na exemplo prático da Figura 18, o modelo reconheceu para a sentença "*[Roubo] de [celular] mediante uso de [faca]...*" que: a entidade nomeada "*Roubo*" corresponde ao rótulo ROUBO, que indica o tipo do crime; "*celular*" foi o objeto subtraído, por isso marcado com o rótulo OBJ_SUB; e, "*faca*" foi o meio empregado na abordagem, tratando-se de uma AB (arma branca). Todos esses rótulos fazem parte do esquema de anotação do domínio CVP, definido e apresentado no próximo capítulo.

O objetivo geral da etapa de definição e treino do modelo CVP é fazer com que o modelo seja capaz de aprender sobre os rótulos presentes nos textos anotados do conjunto de treinamento, para prever os tipos de ENs de um novo texto CVP (conjuntos de desenvolvimento e teste), garantindo, assim, um processamento inteligente e um melhor aproveitamento do conteúdo dessas narrativas. O modelo CVP é treinado com a biblioteca *Keras*[2] da linguagem *Python* com

[2] https://github.com/keras-team/keras

TensorFlow no *back-end*.

4.4.1 Tratamento de OOV

Devido à ausência de formas adequadas de lidar com palavras OOV, trabalhos costumam simplesmente atribuir *embeddings* aleatórios às palavras desconhecidas, ou realizam um mapeamento para um tipo de *embedding* "desconhecido", esperando que seus modelos sejam bem generalizados (GARNEAU *et al.*, 2018).

Este trabalho propõe a realização de um procedimento capaz de contornar problemas de OOV, que consiste no uso da métrica Distância Levenshtein, posto que não se tem a representação vetorial destas palavras (palavras específicas do domínio, raras ou ruído), e é preciso garantir que a posição de termos similares no espaço de alta dimensão das *word embeddings* pré-treinadas possa permanecer a mesma ou melhorar durante o treino, a fim de alcançar uma representação de domínio específico consistente.

O procedimento consiste na comparação entre os termos OOV e o vocabulário das *word embeddings*, e a saída resultante contém as palavras do vocabulário que obtiveram o menor valor de *edit distance*. Primeiramente, um conjunto de palavras OOV $S' = \{s_1, s_2, \cdots, s_k, \cdots, s_K\}$ é identificado no corpus CVP. Em seguida, é realizada uma busca no vocabulário genérico das *word embeddings* pré-treinadas, onde é feita a substituição de um termo OOV s_k de comprimento n por uma palavra do vocabulário t_z de comprimento m, dado que possua o menor valor de *edit distance* entre $s[1..k]$ e $t[1..z]$, definido por $D[k, z]$. A distância de edição entre s_k e t_z é, portanto, $D[n, m]$. O conjunto de saída com as palavras similares é usado como entrada para o modelo.

Adicionando o tratamento de OOV à definição da arquitetura do modelo CVP:

Entrada: um conjunto de palavras de um texto $S = \{x_1, x_2, \cdots, x_N\}$, onde i representa a posição de uma palavra $x_i \in V$ na sentença, sendo V o vocabulário CVP. **Cada OOV x_i no texto é substituído por outra palavra com o valor mínimo de *edit distance* para x_i no vocabulário das *word embeddings* pré-treinadas.** No final, cada símbolo x_i deve ser marcado com um rótulo $y_j \in L = \{y_1, y_2, \cdots, y_{N'}\}$.

A Figura 17 ilustra no passo quatro o fluxo completo do tratamento de OOV e do treino do modelo.

4.5 Interpretação e Análise dos Resultados

Após a execução do modelo CVP, os resultados são coletados e comparados através das métricas de avaliação empírica: *Precision*, *Recall* e F_1-*score*, definidas na Seção 2.4 do Capítulo 2. Em especial, essas métricas são aplicadas para avaliação do reconhecimento de cada rótulo independentemente. Para a avaliação geral dos vários tipos de entidades, utiliza-se a média ponderada dos valores das métricas para cada rótulo, e a Acurácia Balanceada (AB). Visto que esses indicadores vêm desempenhando um papel central na estimativa do desempenho de sistemas do Processamento de Linguagem Natural, estas métricas foram escolhidas como forma de avaliar o desempenho do modelo.

No próximo capítulo são apresentados os resultados obtidos a partir da execução de todos os procedimentos metodológicos.

5 EXPERIMENTOS E RESULTADOS

Neste capítulo são apresentados os resultados obtidos a partir da execução dos procedimentos metodológicos. Inicialmente, são descritos os processos de obtenção, anotação e estratificação dos dados. Por fim, é descrita a metodologia e os resultados dos experimentos para a geração do modelo CVP, seguindo três questões de pesquisa principais.

Em relação às ferramentas utilizadas, a linguagem de programação *Python* foi empregada para o desenvolvimento de todo o trabalho. Também foram utilizadas várias bibliotecas complementares, das quais destacam-se *Scikit-multilearn* (estratificação), *spaCy* (pré-processamento e modelo *baseline*), *Gensim*[1] (criação do vocabulário e leitura das *word embeddings* pré-treinadas), e *Keras* (modelo CVP).

5.1 Coleta de Dados

Durante o período de um mês, foram coletados 3.981 textos da base de dados de CVPs do sistema SIP/SIP3W da SSPDS, de Janeiro de 2015 a Junho de 2019. Foram realizadas seis consultas no banco de dados, preenchidas pelo conteúdo de uma planilha com quase 300 mil linhas. Cada consulta retornou textos de naturezas diferentes, como BOs, TCOs, inquéritos policiais, atos infracionais, e depoimentos de testemunhas e vítimas. Como se tratam de informações confidenciais, não é possível especificar mais detalhes do procedimento, como o nome das tabelas e as relações feitas.

Após uma análise exploratória da base de dados, observou-se a presença de textos que poderiam comprometer a qualidade dos dados, como ruído, textos vazios (sem descrição), duplicados, e fora do domínio. A remoção desses textos foi feita durante o processo de anotação dos documentos, descrito na próxima seção, através da ferramenta HNERD.

5.2 Anotação dos Dados

O processo de geração do conjunto de dados anotado consiste de duas etapas principais: definição do esquema de rotulação e anotação do corpus. Os procedimentos e objetivos específicos de cada etapa são detalhados a seguir.

[1] https://radimrehurek.com/gensim/

5.2.1 Esquema de Anotação

Com a ajuda de especialistas do domínio, foi feito um levantamento de quais tipos de entidades que poderiam ser utilizadas para a tarefa de anotação dos textos CVP. Primeiramente, o esquema de anotação contou com 11 rótulos iniciais, para identificar alguns tipos de armas e de veículos, drogas, e características dos assaltantes. Depois de uma segunda análise das narrativas, os rótulos droga e papelote foram retirados, e mais 27 novos tipos de entidades foram adicionadas.

O esquema de anotação conta hoje com 36 rótulos no total, e são separados em quatro categorias distintas: rótulos que indicam locais, meios de transporte, meios empregados (abordagens), e outras informações, como comparsas e objetos subtraídos. Os Quadros 2 e 3 contém o nome e a descrição geral tipos de locais, meios empregados e meios de transporte. O Quadro 4 possui a descrição dos rótulos que não possuem categoria específica.

Quadro 2 – Esquema de anotação dos locais de A–Z

Tipo de local (A–G)	**Descrição**
AMB_VIRTUAL	Ambientes virtuais (Internet)
EMPRESA	Empresas ou indústrias
ESTAB_COMERCIAL	Estabelecimentos comerciais
ESTAB_ESPORTIVO	Estabelecimentos esportivos
ESTAB_MÉDICO	Estabelecimentos médicos
ESTAB_PÚBLICO	Estabelecimentos públicos
ESTAB_RELIGIOSO	Estabelecimentos religiosos
ESTAC	Estacionamentos
EVENTO	Nomes ou tipos de eventos
FAVELA	Favelas ou comunidades

Tipo de local (H–Z)	**Descrição**
HOSPEDARIA	Hotéis, motéis, pousadas, albergues, pensões e *hostels*
INST_FINANCEIRA	Instituições financeiras
LOC_DE_EMB_DESEMB	Terminais, paradas de ônibus, rodoviárias, portos e aeroportos
LOC_DE_ENSINO	Locais de ensino
LOC_DE_LAZER	Clubes, praças, praias, teatros, cinemas, etc.
LOC_DE_RES	Locais de residência, como casa, condomínios, etc.
OUTROS_LOC	Locais que não sejam qualquer um dos que já foram definidos
TERRENO	Terrenos baldios, matagais, etc.
VIA	Rodovias e vias férreas, públicas, urbanas e rurais

Fonte: Elaborado pela autora.

Quadro 3 – Esquema de anotação dos meios empregados e meios de transporte

Tipo de meio empregado	**Descrição**
AB	Armas brancas
ADF	Armas de fogo
AMEAÇA	Abordagem feita sob (grave) ameaça
SIMULACRO	Simulacro, arma de brinquedo
USO_FORÇA_FÍSICA	Abordagem feita sob uso de força física
ROUBO	Ação que indica roubo/assalto
FURTO	Ação que indica furto

Tipo de meio de transporte	**Descrição**
BICICLETA	Bicicletas
CARRO	Carros, automóveis
MOTO	Motos, motocicletas, ciclomotores, mobiletes e motociclos
OUTROS_MEIOS_TRANSP	Meios de transporte que não sejam qualquer um dos que já foram definidos
TRANSP_ALT	Táxis, moto-táxis, micro-ônibus, vans, topics e transportes por aplicativo, como *Uber*, 99Pop e *inDriver*
TRANSEUNTE	Transeuntes, que caminham a pé
VEIC_GRANDE	Veículos grandes, como ônibus, metrô, caminhão e carreta

Fonte: Elaborado pela autora.

Quadro 4 – Esquema de anotação das informações extras

Tipo de entidade	**Descrição**
COMPARSA	Cúmplices, comparsas, ou parceiros. Também é considerado a quantidade (se for maior que um) e características
MDI	Adolescentes, crianças e indivíduos com idade inferior a 18 anos
OBJ_SUB	Objetos subtraídos na abordagem

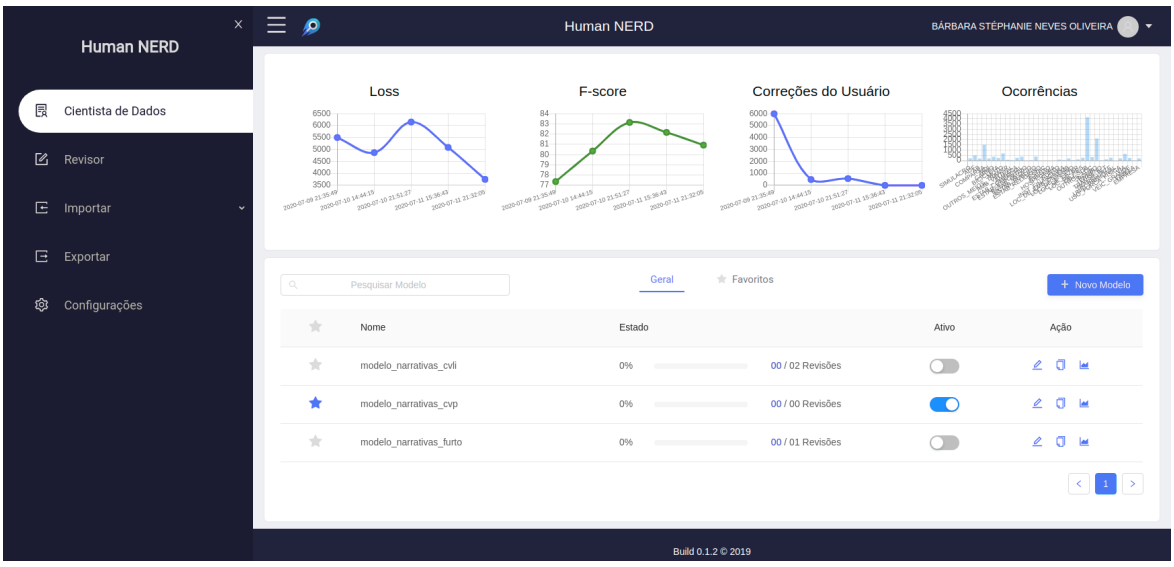
Fonte: Elaborado pela autora.

Para uma melhor compreensão e divisão aproximada, o esquema de anotação das 36 entidades é divido em dois grupos: Categoria 1, que identifica os meios empregados, meios de transporte e outras informações; e, Categoria 2, que reconhece apenas os tipos de locais. Além de conter uma quantidade considerável de classes e de todas serem importantes para o domínio, o esquema de anotação CVP é desafiador para a tarefa NER, dado que as narrativas concentram os detalhes do incidente em diferentes elementos. A próxima seção descreve o processo de anotação e apresenta essa complicação através da análise da distribuição dos rótulos.

5.2.2 Processo de Anotação

O *framework* interativo HNERD foi utilizado para anotar manualmente as narrativas do conjunto de dados. Na Figura 19 é possível notar a existência de alguns modelos, dentre eles, o que foi utilizado para o processo de anotação: `modelo_narrativas_cvp`.

Figura 19 – Página inicial do *Human* NERD na visão do Cientista de Dados

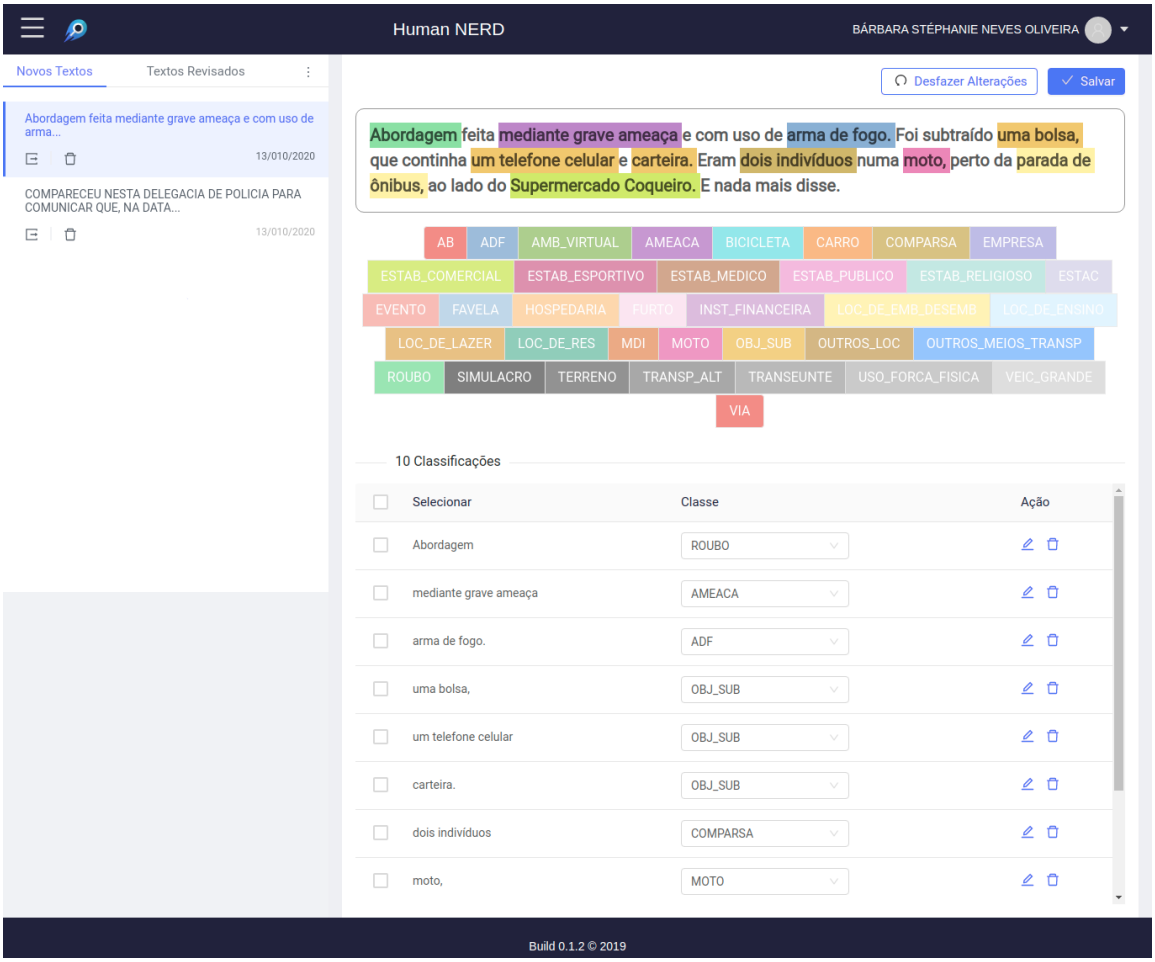


Fonte: Elaborado pela autora.

A Figura 19 apresenta a tela inicial da ferramenta, em que a parte superior contém alguns gráficos estatísticos sobre o modelo que está ativo. Todos os modelos existentes estão listados na parte inferior, onde podem ser feitas algumas operações sobre eles, como favoritar, editar, duplicar e verificar suas estatísticas. Além de também ser possível criar um novo modelo (inicial ou carregar um preexistente), todas as funcionalidades do HNERD podem ser acessadas a partir dessa tela, através do menu lateral esquerdo.

A Figura 20 exibe a página do Revisor, onde foi realizado o processo de anotação no decorrer de três meses. Por motivos de confidencialidade, a narrativa de roubo apresentada na figura é fictícia, visto que não é permitido exibir o conteúdo dos documentos CVP. No geral, a figura mostra o texto com as anotações feitas, todas as classes do esquema de anotação, e uma lista com as anotações. Essa tela também possibilita a realização de operações sobre os textos e anotações, e a visualização dos textos novos e revisados.

Figura 20 – Página do Revisor: processo de anotação de um texto CVP ilustrativo

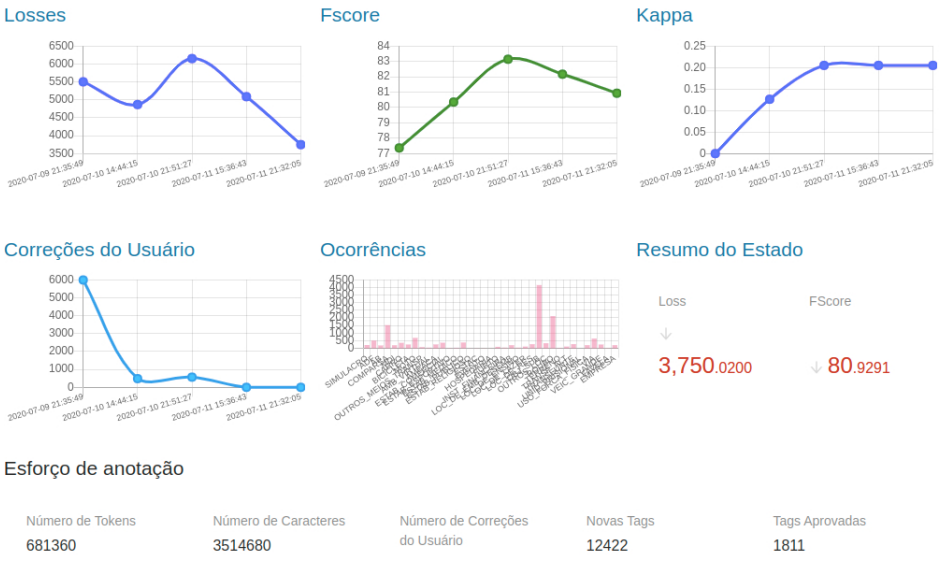


Fonte: Elaborado pela autora.

Para cada novo modelo inicial criado, o HNERD faz a integração de modelos NER pré-treinados da biblioteca *spaCy*. Para o idioma Português, é utilizado o *Wiki*NER, que reconhece quatro entidades genéricas: PER (PESSOA), LOC (LOCALIZAÇÃO), ORG (ORGANIZAÇÃO), e MISC (DIVERSAS). Adicionado o conjunto de dados CVP à ferramenta, o modelo classificou automaticamente as narrativas de acordo com as ENs genéricas. No entanto, como nenhuma delas foi apontada pelos especialistas do domínio, todas foram excluídas para serem adicionados os 36 rótulos do esquema de anotação.

A página das estatísticas na visão do Cientista de Dados é crucial para avaliação de um modelo. Na Figura 21 é possível visualizar as informações estatísticas sobre o `modelo_narrativas_cvp`, por meio de gráficos que retratam a sua performance. O modelo foi treinado cinco vezes por 75 épocas cada, com uma parcela do conjunto dados sendo adicionada após a finalização de cada treino. A intenção por trás dessa técnica era de automatizar ainda mais o processo de anotação, para correção apenas do que foi predito incorretamente pelo modelo.

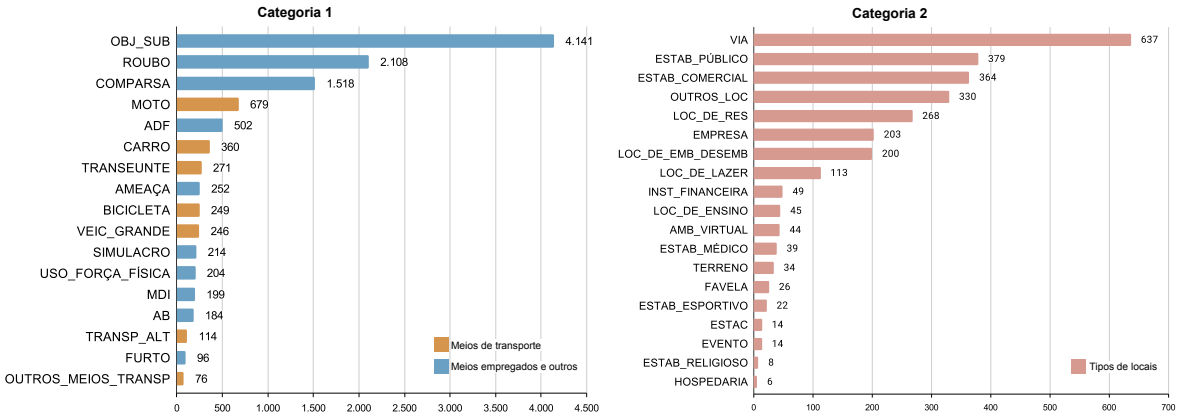
Figura 21 – Estatísticas do `modelo_narrativas_cvp` usado no processo de anotação



Fonte: Elaborado pela autora.

Finalizado o processo manual de anotação, foi então gerado o corpus anotado, com 1.510 textos dos 3.981 coletados do banco de dados, e cerca de 14.208 entidades nomeadas identificadas. Os gráficos da Figura 22 possuem a quantidade de ocorrências de cada rótulo, separados por categoria. O rótulo OBJ_SUB da Categoria 1 é o que possui a maior quantidade de menções (1.141 no total), e ESTAB_RELIGIOSO e HOSPEDARIA da Categoria 2 aparecem menos de 10 vezes.

Figura 22 – Distribuição dos dados para as Categorias 1 e 2



Fonte: Elaborado pela autora.

Pode-se notar pelos gráficos da Figura 22 que a distribuição dos rótulos é altamente desbalanceada. Para separar igualmente os conjuntos dos rótulos majoritários e minoritários, foi estabelecido um limite de 200 menções por rótulo: os que possuem um número maior ou igual a 200 ocorrências fazem parte do conjunto de tipos de entidades mais frequentes, enquanto que os que possuem menos de 200, fazem parte do conjunto de rótulos raros. O Quadro 5 contém a relação dos rótulos majoritários e minoritários, considerando o limite estabelecido.

Quadro 5 – Relação dos tipos de entidades majoritárias e minoritárias

Rótulos	**Categoria 1**	**Categoria 2**	**Total**
Majoritários	ADF, AMEAÇA, BICICLETA, CARRO, COMPARSA, MOTO, OBJ_SUB, ROUBO, SIMULACRO, TRANSEUNTE, USO_FORÇA_FÍSICA e VEIC_GRANDE	EMPRESA, ESTAB_COMERCIAL, ESTAB_PÚBLICO, LOC_DE_EMB_DESEMB, LOC_DE_RES, OUTROS_LOC e VIA	19
Minoritários	AB, FURTO, MDI, OUTROS_MEIOS_TRANSP e TRANSP_ALT	AMB_VIRTUAL, ESTAB_ESPORTIVO, ESTAB_MÉDICO, ESTAB_RELIGIOSO, ESTAC, EVENTO, FAVELA, HOSPEDARIA, INST_FINANCEIRA, LOC_DE_ENSINO, LOC_DE_LAZER e TERRENO	17

Fonte: Elaborado pela autora.

A partir da divisão presente no Quadro 5, nota-se uma quantidade aproximada entre os majoritários (19) e minoritários (17). A Categoria 1 é a que possui a maior quantidade de rótulos frequentes (12), ao passo que a Categoria 2 é a que contém os mais raros (12). Observe que os tipos de entidades raras possuem definições muito específicas, o que pode fazer com que o modelo CVP as reconheça facilmente. Contudo, a quantidade de exemplos é muito baixa e interferem no aprendizado das que possuem uma maior dependência de contexto. Detalhes sobre essas particularidades são fornecidos na Seção 5.4.

5.3 Estratificação dos Dados

Com o conjunto de dados anotado, são então realizadas a vetorização e divisão estratificada em subconjuntos para treino e validação dos modelos. Essas duas etapas são detalhadas nesta seção.

5.3.1 Vetorização do Conjunto de Dados

Como o HNERD incorpora o modelo *Wiki*NER do *spaCy*, os dados são anotados com o formato de treinamento aceito pela biblioteca, que considera uma lista de tuplas, em que cada tupla contém um texto e um dicionário que referencia as entidades pelos deslocamentos dos caracteres (índices que indicam o começo e o final da EN) e seus rótulos (y_j), da seguinte maneira:

$$< \texttt{TEXTO}, \{\text{``}\texttt{entidades}\text{''}: [< I_{início}, I_{fim}, y_j >]\} >$$

O tipo de anotação aceito pelo *spaCy* atribui um rótulo a cada entidade nomeada como se esta fosse um único *token*. Porém, este trabalho leva em consideração as anotações dos tipos de entidades para cada *token* presente no corpus, e o modelo CVP recebe como entrada as representações vetoriais e numéricas de tudo o que é texto. Como forma de garantir essas especificações, foi feita a conversão do conjunto de dados anotado, tanto para os textos quanto para os rótulos.

Para os textos, um pré-processamento foi efetivado para obtenção das palavras mais importantes, através da aplicação do processo de tokenização e de limpeza dos dados (conversão para letras minúsculas, e remoção de caracteres especiais e de *stopwords*). Não foi feita a remoção de dígitos e nem de números por extenso, pois são significativos para o domínio (indicam o número de comparsas, de objetos subtraídos, etc.). Em seguida, a partir da extração do conjunto de *tokens* (palavras sem repetição) do corpus, é que foi construído o vocabulário CVP, com 8.101 *tokens* no total, utilizado para o mapeamento vetorial das palavras de cada texto, transformando-as em seu índice correspondente.

A codificação *one-hot* foi aplicada no pré-processamento dos rótulos, posto que são dados categóricos. Desde que vários *tokens* não possuem nenhum tipo de entidade atribuído a eles, o rótulo OUTROS foi adicionado para indicar estes termos que não são ENs. Entretanto,

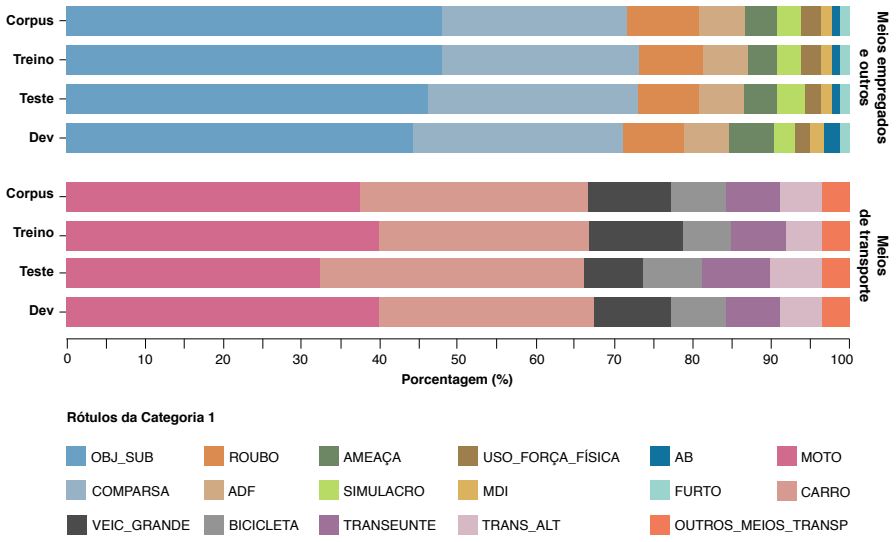
detectar esse rótulo não faz parte do problema, o que evita uma discussão extra sobre OUTROS na apresentação dos resultados. Com a codificação *one-hot*, cada rótulo é representado como um vetor binário, com o valor um (1) indicando seu índice (o índice 1 ficou para OUTROS, o 2 para AB, o 3 para ADF, e assim sucessivamente, em ordem alfabética, até o índice 37 para o rótulo VIA).

5.3.2 Processo de Estratificação

Feita a vetorização do corpus anotado, o método de estratificação iterativa de Sechidis *et al.* (2011) foi aplicado para obtenção dos conjuntos de treino (70%), teste (20%) e desenvolvimento (10%). Uma nova representação vetorial dos rótulos foi feita apenas para servir de entrada para o método iterativo: invés de uma matriz binária, um vetor de 37 posições (incluindo OUTROS) foi adicionado para cada texto. Cada posição representa um rótulo e contém a quantidade de vezes que o mesmo foi usado para marcar uma EN na sentença.

Com a estratificação, o conjunto de treino ficou com 1.062 textos, o de teste com 319, e o de desenvolvimento com 129 textos. Os gráficos das Figuras 23 e 24 confirmam que esse procedimento funciona na prática, mostrando que as proporções dos rótulos (com a exceção de OUTROS) de cada um dos subconjuntos possuem baixa divergência entre si, quando comparados com as distribuições do conjunto de dados completo. Analisando os gráficos da Figura 23, as proporções estão bem similares e apontam os rótulos da Categoria 1, justamente a que possui a maior quantidade de tipos de entidades majoritários.

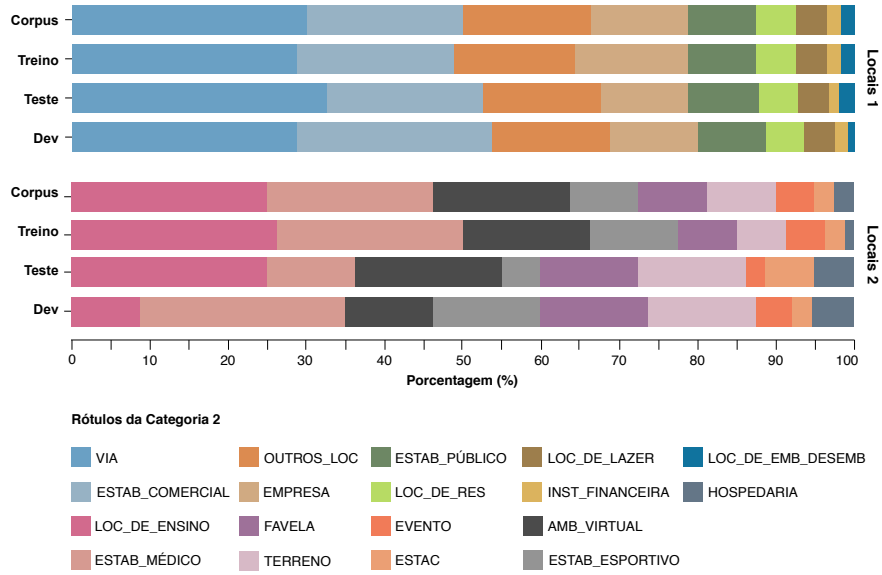
Figura 23 – Distribuição estratificada dos rótulos da Categoria 1



Fonte: Elaborado pela autora.

Para os gráficos da Figura 24, as proporções dos subconjuntos de teste e desenvolvimento diferem um pouco para os tipos de locais que possuem menos de 50 ocorrências no conjunto de dados (Figura 22). O algoritmo de estratificação iterativa possui a seguinte motivação: se os rótulos raros não forem tratados como prioridade, eles podem ser distribuídos de forma indesejada, o que não pode ser reparado posteriormente. Com rótulos majoritários, a chance de modificar a distribuição atual é bem maior, devido à disponibilidade de mais exemplos. Isso significa que o método iterativo examina inicialmente os rótulos com o menor número de exemplos a cada iteração. Assim, as divisões dos subconjuntos da Figura 24 estão corretas, pois partiram, antes de tudo, da verificação dos rótulos mais raros.

Figura 24 – Distribuição estratificada dos rótulos da Categoria 2



Fonte: Elaborado pela autora.

5.4 Modelo CVP

Para construir um modelo NER do Aprendizado Profundo capaz de reconhecer entidades nomeadas em narrativas de roubos, e que trate do problema de desbalanceamento de classes, este trabalho segue três questões de pesquisa principais. A presente seção apresenta detalhadamente cada questão de pesquisa, juntamente com a descrição dos procedimentos necessários para respondê-las, e os resultados obtidos.

5.4.1 Questões de Pesquisa

As seguintes questões de pesquisa são abordadas para a elaboração e avaliação do modelo CVP:

1ª **As representações distribuídas a nível de palavras (*word embeddings*) são suficientes para capturar as propriedades semânticas e sintáticas, que não aparecem explicitamente no texto? Qual abordagem é mais eficiente: *word embeddings* específicas do domínio, ou integração e/ou ajuste de *word embeddings* pré-treinadas?**
Esta questão de pesquisa concentra-se na camada de *embedding*. O desempenho do modelo CVP é avaliado para diferentes vetores pré-treinados: *FastText*, GloVe, *Wang2Vec* e *Word2Vec*, com seus pesos fixados e/ou ajustados durante o treino. Os resultados também são comparamos com *embeddings* específicos do domínio, mapeados por meio do vocabulário CVP e gerados a partir dos dados de treino.

2ª **O uso de soluções aplicadas à arquitetura para lidar com o problema de desbalanceamento de classes, garante um aumento significativo na qualidade e no desempenho?**
Para esta questão de pesquisa, é preciso comparar o desempenho dos modelos com diferentes *word embeddings*, para três funções de *loss*: CRF, CB-CCE e DL. A CRF *Loss* faz parte da arquitetura do modelo estatístico CRF, e é frequentemente utilizada quando tem-se uma camada deste decodificador de rótulo. Por meio desta questão, é possível ter uma melhor noção da importância de cada uma das camadas da arquitetura.

3ª **A remoção da anotação de tipos de entidades nomeadas raras, mesmo que muito específicas, para o treinamento, melhora a capacidade do modelo em reconhecer as que possuem uma maior dependência de contexto?**
A partir do melhor modelo obtido respondendo às duas primeiras questões, um novo treinamento deve ser feito sem a anotação dos rótulos raros. O objetivo desta questão de pesquisa é verificar se os rótulos minoritários influenciam na captura efetiva de dependências contextuais.

5.4.2 Estratégias de Treinamento

Duas estratégias de treinamento foram desenvolvidas para responder as questões de pesquisa. As estratégias elencadas a seguir descrevem as variações na arquitetura do modelo CVP com a adição das *word embeddings* (representadas por WE) e das funções de *loss*, indicadas

na 1ª e 2ª questões de pesquisa:

- **WE+BLSTM-CRF+*Loss*:** modelo BLSTM-CRF que usa *word embeddings* (pré-treinadas ou não) e um tipo específico de função de perda. Os vetores de *embeddings* pré-treinados podem ser fixados ou ajustados durante o treino. Com base na quantidade dos tipos de *word embeddings* e funções de perda, foram treinadas 27 variações desta arquitetura;
- **ConcatWE+BLSTM-CRF+*Loss*:** modelo BLSTM-CRF que possui a mesma arquitetura geral que o WE+BLSTM-CRF+*Loss*, mas com uma diferença: o termo "Concat" representa a concatenação feita entre uma camada de *embedding* com vetores pré-treinados e uma com vetores específicos do domínio. Para a camada com os pesos das *word embeddings* pré-treinadas, ela também pode ser fixada ou ajustada durante o treino. 24 variações desta arquitetura foram treinadas.

No total, considerando as diferentes estratégias[2] e os modelos para a 3ª questão de pesquisa, 53 redes neurais são treinadas e avaliadas. Todas utilizam o algoritmo de *back-propagation* durante o treino, e *word embeddings* com vetores de 50 dimensões. Observe que para os *embeddings* pré-treinados ajustados durante o treino, e para os modelos que possuem uma camada de concatenação, é realizado o processo de *fine-tuning*. Com o *fine-tuning*, as representações mais abstratas do método que está sendo reutilizado (*embeddings* pré-treinados ou camada de concatenação) tornam-se mais relevantes para o problema (CHOLLET, 2017).

5.4.3 Cenários de Experimentação

Definidas as estratégias de treinamento, os cenários de experimentação foram criados para o treinamento e avaliação dos modelos. Os objetivos específicos de cada experimento são detalhados abaixo.

- ***Experimento #1*:** avaliar o desempenho das duas estratégias de treinamento com os pesos das *word embeddings* pré-treinadas sendo fixados durante a execução;
- ***Experimento #2*:** avaliar o desempenho das duas estratégias de treinamento com os pesos das *word embeddings* pré-treinadas sendo atualizados durante a execução;
- ***Experimento #3*:** analisar a eficácia de dois (ou mais) modelos, os quais obtiveram bons resultados nos Experimentos #1 e #2, respectivamente, comparando-os com o *baseline*, para enfim responder a 3ª questão de pesquisa.

[2] A nomenclatura dos modelos que utilizam a CRF *Loss* não incluem o nome ou a sigla da função de perda no final, já que esta faz parte do decodificador de rótulo. Exemplo: o modelo GloVe+BLSTM-CRF usa a CRF *Loss* e o GloVe+BLSTM-CRF+DL usa a *Dice Loss*.

As performances das redes neurais são comparadas com o modelo *spaCy* NER, escolhido como *baseline* por fazer parte da ferramenta HNERD, e por ser popular na comunidade científica e no mercado devido à sua sofisticada arquitetura que alcança resultados do estado da arte (LI *et al.*, 2020). Na versão mais recente, os modelos de Aprendizado Profundo do *spaCy* v2.0 são reportados como "10 vezes menores, 20% mais precisos e práticos de rodar do que os das gerações anteriores" (HONNIBAL; MONTANI, 2017).

Como o *spaCy* oferece suporte ao aprendizado online (uma espécie de *fine-tuning*), este trabalho atualiza o modelo NER pré-treinado do Português (*Wiki*NER, o mesmo que é utilizado pelo HNERD) com os exemplos do conjunto de treino. O *spaCy* NER recebe como entrada o corpus anotado antes do processo de vetorização. Para a avaliação final, a vetorização é aplicada para obtenção dos resultados por *token*, e não por entidade, como normalmente é realizado pela biblioteca.

Os modelos do PLN do *spaCy*, especialmente o de NER, seguem quatro etapas simples: vetorização, codificação, atenção e predição. Primeiro, o texto é transformado para ser inserido na forma de valores numéricos. No estágio de *embedding*, atributos como o prefixo, o sufixo, a forma, e as letras minúsculas das palavras são usados para a extração de características que refletem suas semelhanças. Para codificar os *embeddings* independentes de contexto em uma matriz dependente de contexto, os valores passam por uma rede CNN para serem codificados. Antes da predição, a camada de atenção da CNN recebe a matriz para conversão desta em um único vetor. Então, uma camada MLP+*Softmax* é usada como decodificador de rótulo para a predição das classes.

Os 53 modelos BLSTM-CRF e o *baseline* foram treinados com o mesmo conjunto estratificado de treino por 75 épocas cada, e avaliados com os conjuntos estratificados de desenvolvimento e teste. No que diz respeito aos problemas de OOV, comparando os mais de oito mil *tokens* do vocabulário CVP com as 929.606 palavras do vocabulário das diferentes *word embeddings* pré-treinadas, constatou-se a presença de 679 termos OOV.

O tratamento de OOV foi realizado separadamente para os vocabulários de cada modelo com a variação *Skip-Gram* (com exceção do GloVe). Além disso, o mesmo pré-processamento aplicado nos textos do conjunto de dados descrito na Seção 5.3.1 foi efetuado para a construção de cada matriz de *embedding*, para a remoção dos vetores de palavras menos importantes. As quatro matrizes ficaram com uma dimensão final de 929.212×50.

5.4.3.1 Experimento #1

O objetivo deste experimento é avaliar nove diferentes abordagens para a camada de *embedding*, seguindo três funções de *loss*. A etapa principal envolveu o treino de 15 modelos WE+BLSTM-CRF+*Loss* e de 12 modelos ConcatWE+BLSTM-CRF+*Loss*, com os pesos das *word embeddings* pré-treinadas fixados durante a execução. Neste experimento, apenas para propósitos comparativos, o modelo que gera *word embeddings* específicas do domínio é considerado como *baseline*, visto que é o único que não utiliza vetores pré-treinados.

Os resultados detalhados por rótulo para os modelos avaliados neste experimento estão disponíveis no Apêndice A. As interpretações e conclusões feitas dos rótulos majoritários e minoritários para os tipos de *embeddings* agrupados por função de *loss*, estão elencadas a seguir:

- **Rótulos majoritários**
  - **CB-CCE *Loss*:** os modelos performaram melhor para os que possuem a camada de concatenação, com destaque para o *embedding* ConcatGloVe, que conseguiu resultados superiores para nove rótulos (ADF, AMEAÇA, BICICLETA, COMPARSA, OBJ_SUB, OUTROS_LOC, ROUBO, USO_FORÇA_FÍSICA e VEIC_GRANDE);
  - **CRF *Loss*:** os melhores resultados por rótulo estão bem distribuídos para os diferentes tipos de *embeddings*, com exceção do *FastText* que obteve os maiores valores para BICICLETA, CARRO, ESTAB_PÚBLICO, MOTO e OBJ_SUB;
  - ***Dice Loss*:** o *embedding* Concat*FastText* teve bons resultados para oito (AMEAÇA, BICICLETA, EMPRESA, ESTAB_COMERCIAL, LOC_DE_EMB_DESEMB, OUTROS_LOC, ROUBO e SIMULACRO) dos 19 rótulos majoritários.
- **Rótulos minoritários**
  - **CB-CCE *Loss*:** os modelos também performaram melhor com a camada de concatenação, empatando entre Concat*Wang2Vec* e Concat*Word2Vec*. Ambos obtiveram bons resultados para quatro rótulos distintos;
  - **CRF *Loss*:** os melhores resultados também estão bem distribuídos para esta função de *loss*, com o Concat*FastText* atingindo os maiores valores para ESTAB_ESPORTIVO, ESTAB_RELIGIOSO, INST_FINANCEIRA, LOC_DE_ENSINO e TRANSP_ALT;
  - ***Dice Loss*:** esta função de *loss* teve um desempenho ruim, não reconhecendo nenhum rótulo raro para quase todos os *embeddings*.

A análise individual de cada rótulo é significativa, mas não para este cenário de desbalanceamento em que os melhores resultados encontram-se espalhados entre os *embeddings*.

Para a avaliação geral dos vários tipos de entidades, a Tabela 2 apresenta a média ponderada dos resultados para as métricas *Precision*, *Recall* e F_1-*score*, e os valores para a Acurácia Balanceada. Cada linha na tabela representa um tipo de *embedding* ligado à uma função de *loss* (coluna).

Tabela 2 – Resultados do Experimento #1 para as variações do modelo BLSTM-CRF

Tipos de	**CB-CCE *Loss***				**CRF *Loss***				***Dice Loss***			
embeddings	**P**	**R**	**F_1**	**AB**	**P**	**R**	**F_1**	**AB**	**P**	**R**	**F_1**	**AB**
WE do domínio	67,484	81,507	73,556	*68,659*	73,117	78,981	75,464	*60,498*	67,622	67,339	67,050	*33,633*
FastText	60,785	82,428	68,524	***73,246***	84,483	74,481	78,602	*55,339*	70,049	69,486	68,518	*30,895*
GloVe	62,800	83,363	70,887	*71,905*	82,226	76,658	78,684	*55,913*	67,775	68,010	67,703	*29,861*
Wang2Vec	62,077	84,195	70,514	*73,197*	84,023	75,182	**78,932**	*56,480*	69,166	72,904	70,893	*33,344*
Word2Vec	55,815	79,930	64,289	*68,300*	81,494	72,860	76,210	*50,389*	66,450	66,418	66,160	*27,717*
Concat*FastText*	72,930	81,551	76,518	*72,824*	77,554	77,710	77,405	***66,013***	75,576	73,853	**74,248**	***41,414***
ConcatGloVe	72,792	82,106	**76,979**	*70,250*	77,035	78,089	77,221	*65,078*	74,125	73,678	73,713	*40,832*
Concat*Wang2Vec*	71,799	81,902	75,982	*73,140*	76,642	77,125	76,660	*65,183*	71,450	74,993	72,969	*40,426*
Concat*Word2Vec*	71,631	82,223	76,297	*71,482*	77,634	76,979	76,963	*62,675*	68,935	70,567	65,537	*36,236*

Fonte: Elaborado pela autora.

Os resultados destacados em negrito na Tabela 2 indicam os maiores valores apenas para as métricas F_1 e AB, dado que a Acurácia Balanceada é responsável por calcular a média da acurácia de cada rótulo (pelo *Recall*), na mesma proporção que o F_1-*score* é uma métrica semelhante, por medir a média harmônica do *Precision* e *Recall*. Portanto, como as distribuições dos rótulos são desbalanceadas, as avaliações gerais dos modelos não só deste cenário de experimentação, mas como dos próximos, leva em consideração a média ponderada do F_1-*score*, e a Acurácia Balanceada.

Analisando as conclusões por rótulo e os resultados da Tabela 2, *Wang2Vec*+BLSTM-CRF supera na métrica F_1 os melhores modelos para a CB-CCE (ConcatGloVe+BLSTM-CRF) e para a DL (Concat*FastText*+BLSTM-CRF), obtendo um F_1-*score* de $78,832\%$, com ganhos variando entre 2% e $4,7\%$, nesta ordem. Para a Acurácia Balanceada, o modelo *FastText*+BLSTM-CRF+CB-CCE obteve um resultado de $73,246\%$, com ganho superior quanto aos modelos com a CRF *Loss* ($+7,2\%$ para o Concat*FastText*) e a DL ($+31,8\%$ também para o Concat*FastText*). Apesar das duas métricas não indicarem o mesmo modelo, são considerados os valores diferentes de saída para ambas, como forma de avaliar o grau de discernimento de cada uma.

Conclui-se que as melhorias produzidas pelos *embeddings* pré-treinados do *FastText* e do *Wang2Vec* com a CB-CCE e CRF *Loss*, respectivamente, superam os resultados para os modelos que utilizam *word embeddings* específicas do domínio. Os valores confirmam que os vetores pré-treinados conseguem capturar de maneira mais eficiente os aspectos da linguagem dos textos CVP, principalmente porque a quantidade de dados de treinamento é pequena, o que não reproduz um tipo de estrutura especializada para o problema com *embeddings* específicos do

domínio. Também conclui-se que a *Dice Loss* não é eficaz para o reconhecimento dos rótulos raros.

5.4.3.2 Experimento #2

Este segundo experimento segue as mesmas especificações do Experimento #1. No entanto, são avaliadas as performances de 12 modelos WE+BLSTM-CRF+*Loss* e de 12 ConcatWE+BLSTM-CRF+*Loss*, em que os pesos das *word embeddings* pré-treinadas foram atualizados durante o treino. Como abordado no primeiro experimento, o modelo com *word embeddings* específicas do domínio é considerado como *baseline* para comparação com os *embeddings* pré-treinados.

Os resultados por rótulo dos modelos podem ser consultados no Apêndice B. As conclusões obtidas estão listadas a seguir:

- **Rótulos majoritários**
  - **CB-CCE *Loss*:** os modelos sem a camada de concatenação possuem valores bem mais distribuídos. Por outro lado, os que utilizam esta camada estão divididos entre dois *embeddings*, com o ConcatGloVe possuindo os melhores resultados para ADF, AMEAÇA, CARRO, OBJ_SUB e ROUBO;
  - **CRF *Loss*:** diferente do primeiro experimento, os melhores resultados para esta função de *loss* não estão espalhados entre os *embeddings*, mas concentram-se nos modelos sem a camada de concatenação, tendo o *Wang2Vec* os maiores valores para oito rótulos (ADF, BICICLETA, COMPARSA, LOC_DE_RES, OBJ_SUB, SIMULACRO, USO_FORÇA_FÍSICA e VIA);
  - ***Dice Loss*:** os melhores resultados estão distribuídos entre os *embeddings*, com destaque para o *Wang2Vec*, que identificou bem BICICLETA, COMPARSA, LOC_DE_RES, OBJ_SUB e VIA.
- **Rótulos minoritários**
  - **CB-CCE *Loss*:** assim como para os rótulos majoritários, o modelo ConcatGloVe obteve os maiores resultados para a maior quantidade de rótulos raros (ESTAB_MÉDICO, ESTAC, EVENTO, HOSPEDARIA, MDI e TRANSP_ALT);
  - **CRF *Loss*:** esta função de *loss* concentra seus melhores resultados no *embedding* *Wang2Vec*, que reconhece bem ESTAB_ESPORTIVO, ESTAB_RELIGIOSO, HOSPEDARIA, INST_FINANCEIRA, LOC_DE_ENSINO, LOC_DE_LAZER e TERRENO;

– ***Dice Loss*:** assim como reportado no Experimento #1, esta função de *loss* também obteve um desempenho ruim, não reconhecendo nenhum rótulo raro para quase todos os *embeddings*.

A Tabela 3 apresenta a média ponderada dos resultados das métricas *Precision*, *Recall* e F_1-*score*, como também os valores para a Acurácia Balanceada, de todos os modelos avaliados neste experimento.

Tabela 3 – Resultados do Experimento #2 para as variações do modelo BLSTM-CRF

Tipos de	**CB-CCE *Loss***				**CRF *Loss***				***Dice Loss***			
embeddings	**P**	**R**	**F_1**	**AB**	**P**	**R**	**F_1**	**AB**	**P**	**R**	**F_1**	**AB**
WE do domínio	67,484	81,507	73,556	*68,659*	73,117	78,981	75,464	*60,498*	67,622	67,339	67,050	*33,633*
*FastText**	70,244	83,976	76,054	*71,520*	81,891	79,316	80,268	*65,509*	74,859	78,922	76,675	*44,841*
GloVe*	74,846	83,625	**78,786**	*71,549*	79,797	80,398	79,749	*64,917*	75,645	76,001	75,613	*43,780*
*Wang2Vec**	72,966	83,450	77,699	*71,843*	82,505	79,959	**80,984**	*68,297*	73,161	77,768	75,324	*44,810*
*Word2Vec**	72,002	83,918	77,235	*71,337*	79,859	78,834	79,011	*64,305*	74,211	76,964	75,408	*42,579*
Concat*FastText**	73,045	82,194	77,061	*72,493*	77,211	78,995	77,965	***68,425***	74,834	74,832	**74,609**	*42,995*
ConcatGloVe*	75,343	80,719	77,760	*70,878*	77,990	78,863	78,190	*67,436*	75,222	74,029	74,413	*42,993*
Concat*Wang2Vec**	74,108	81,945	77,525	***72,494***	77,300	77,257	77,053	*65,934*	75,315	76,117	75,484	***45,352***
Concat*Word2Vec**	72,383	81,304	76,275	*70,604*	75,666	78,615	76,836	*66,360*	73,546	74,438	73,681	*42,866*

Fonte: Elaborado pela autora.
(*): Este símbolo indica que a camada das *word embeddings* pré-treinadas foi atualizada durante o treino.

Semelhante à Tabela 2, os valores em negrito da Tabela 3 indicam os maiores resultados apenas para as métricas F_1 e AB. A métrica F_1 é superior para o modelo *Wang2Vec*+BLSTM-CRF, que obteve um resultado de $80,984\%$, com ganho de $+2,2\%$ e $+6,3\%$ quando comparado com os melhores modelos da CB-CCE (GloVe+BLSTM-CRF) e DL (Concat*FastText*+BLSTM-CRF). No que se refere à Acurácia Balanceada, Concat*Wang2Vec*+BLSTM-CRF+CB-CCE atingiu um valor de $72,494\%$, com melhorias que variam de $4,1\%$ e $27,1\%$ para os modelos Concat*FastText*+BLSTM-CRF e Concat*Wang2Vec*+BLSTM-CRF+DL.

Wang2Vec+BLSTM-CRF, que nem no Experimento #1, foi o melhor modelo apontado pelo F_1-*score*, com um aumento de 2% para o que atualiza as *word embeddings* pré-treinadas. Em contrapartida, o melhor modelo *FastText*+BLSTM-CRF+CB-CCE para a Acurácia Balanceada no primeiro experimento, não obteve melhorias com o *fine-tuning*. Neste experimento, Concat*Wang2Vec*+BLSTM-CRF+CB-CCE foi indicado pela AB, com $-0,8\%$ se comparado com o *FastText*+BLSTM-CRF+CB-CCE do primeiro experimento. Como constatado anteriormente, a *Dice Loss* não é eficaz no reconhecimento dos rótulos raros.

Em resumo, os modelos com os *embeddings* pré-treinados do *Wang2Vec* combinados com as respectivas funções de perda CB-CCE e CRF, superam os resultados para os modelos que utilizam *word embeddings* específicas do domínio. Há um aumento significativo das métricas em relação a todos os modelos (com exceção da Acurácia Balanceada para a CB-CCE *Loss*) do

primeiro cenário de experimentação, que não ajusta os pesos das *word embeddings* pré-treinadas.

Na literatura, é recomendado manter a camada com os pesos congelada, sobretudo para poucos dados, pois quando partes de um modelo são pré-treinadas e outras são inicializadas aleatoriamente, as partes pré-treinadas não devem ser atualizadas durante o treinamento para evitar o esquecimento do que já sabem (CHOLLET, 2017).

Apesar disso, conclui-se que permitir que os vetores pré-treinados se adaptem ao conjunto de treino mesmo que com uma quantidade reduzida de exemplos, proporciona uma melhoria significativa, uma vez que esta abordagem consegue capturar ainda mais as especificidades do domínio. O termo *CVP2Vec* designa esse novo espaço de alta dimensão criado, mapeado pelas representações vetoriais pré-treinadas integradas ao contexto fornecido pelo vocabulário CVP.

5.4.3.3 Experimento #3

Os experimentos anteriores demonstram que o uso de *word embeddings* pré-treinadas, ajustadas ou não durante o treino, conseguem representar adequadamente o vocabulário dos textos CVP, com destaque para a construção do *CVP2Vec*. A substituição dos quase $8,3\%$ de termos OOV por palavras similares contribuiu para obtenção de uma representação consistente. No tocante às abordagens a nível de arquitetura dos modelos para lidar com dados desbalanceados, a CB-CCE *Loss* melhora o reconhecimento dos tipos de entidades raras.

Este terceiro experimento avalia a eficácia dos melhores modelos indicados nos experimentos anteriores pelas métricas F_1-*score* e Acurácia Balanceada, comparando-os com o *spaCy* NER, e seguindo as mesmas especificações dos dois primeiros cenários de experimentação. Em seguida, os modelos com o melhor desempenho são adaptados para responder à 3$^{\mathbf{a}}$ questão de pesquisa. Por fim, são feitas as avaliações por rótulo e gerais desses modelos.

Os resultados detalhados por rótulo do *baseline*, do *FastText*+BLSTM-CRF+CB-CCE e *Wang2Vec*+BLSTM-CRF com os pesos fixados, e do Concat*Wang2Vec*+BLSTM-CRF+CB-CCE e *Wang2Vec*+BLSTM-CRF com *fine-tuning*, podem ser consultados no Apêndice C. Basicamente, os modelos BLSTM-CRF superam o *baseline* para todos os rótulos: com a CB-CCE *Loss*, eles reconhecem de forma eficaz os rótulos raros; enquanto que os com a CRF *Loss*, possuem bons resultados para os mais frequentes.

Concat*Wang2Vec*+BLSTM-CRF+CB-CCE com *fine-tuning* classificou eficientemente os rótulos raros AB, AMB_VIRTUAL, ESTAB_ESPORTIVO, ESTAB_MÉDICO, ESTAB_PÚBLICO, ESTAC,

FAVELA e TERRENO. O modelo *Wang2Vec*+BLSTM-CRF também com *fine-tuning* atingiu os melhores resultados para 13 dos 19 rótulos majoritários. A Tabela 4 possui a avaliação geral para os modelos avaliados.

Tabela 4 – Resultados do Experimento #3: comparação geral dos melhores modelos BLSTM-CRF com o *baseline*

Modelos	**Métricas**			
	P	**R**	**F_1**	**AB**
spaCy NER	85,617	75,533	79,802	*63,272*
FastText+BLSTM-CRF+CB-CCE	60,785	82,428	68,524	***73,246***
Wang2Vec+BLSTM-CRF	84,023	75,182	78,932	*56,480*
Concat*Wang2Vec*+BLSTM-CRF+CB-CCE*	74,108	81,945	77,525	*72,494*
Wang2Vec+BLSTM-CRF*	82,505	80,984	**80,984**	*68,297*

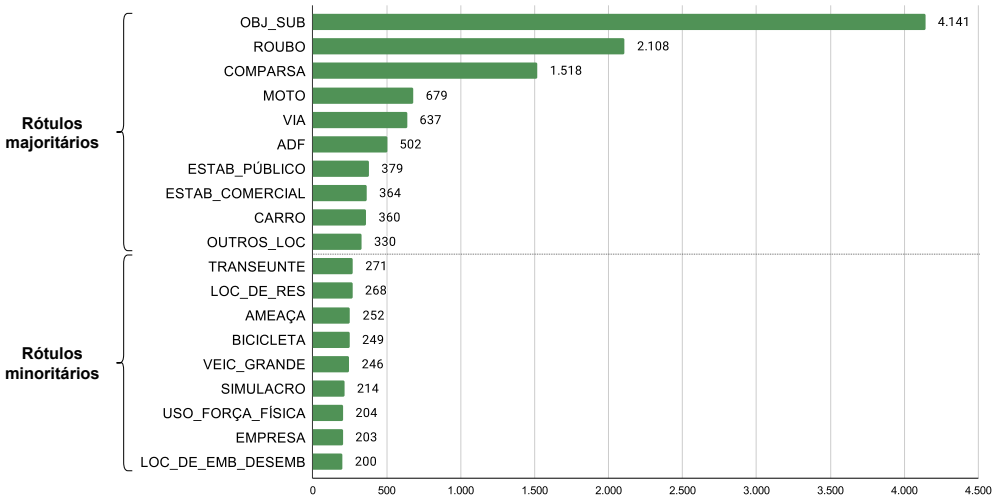
Fonte: Elaborado pela autora.
(*): Este símbolo indica que os modelos tiveram a camada das *word embeddings* pré-treinadas atualizada durante o treino.

Os resultados gerais presentes na Tabela 4 confirmam que os modelos BLSTM-CRF superam o *baseline* no reconhecimento de entidades nomeadas, principalmente para a métrica Acurácia Balanceada. Para a métrica F_1, *Wang2Vec*+BLSTM-CRF possui um ganho de $1,1\%$ quando comparado ao *baseline*. O *spaCy* NER possui um valor de F_1 superior ($+11,3\%$) para o melhor modelo indicado pela Acurácia Balanceada, mas com uma diferença de 10% para esta métrica. O segundo melhor modelo dado pela AB, o Concat*Wang2Vec*+BLSTM-CRF+CB-CCE, com $+9,2\%$ para esta métrica em comparação ao *baseline*, fica mais perto do *spaCy* NER com a métrica F_1 (diferença de $2,3\%$).

Como o *FastText*+BLSTM-CRF+CB-CCE e o *Wang2Vec*+BLSTM-CRF são os melhores modelos de acordo com as métricas Acurácia Balanceada e F_1-*score*, as duas arquiteturas são utilizadas para análise da 3ª questão de pesquisa. O modelo Concat*Wang2Vec*+BLSTM-CRF+CB-CCE também é considerado, por seus resultados próximos para as métricas AB e F_1. As conclusões para as 1ª e 2ª questões são dadas na Seção 5.5.

A 3ª questão de pesquisa foca apenas no treino e avaliação dos 19 rótulos majoritários. Para isso, as anotações dos rótulos raros precisam ser removidas do corpus CVP anotado, e os passos metodológicos de vetorização do conjunto de dados e processo de estratificação devem ser refeitos. Sem a anotação dos tipos de entidades raros, o desbalanceamento concentra-se em outros rótulos, como mostrado na Figura 25, já com o limite estabelecido para os rótulos frequentes e raros. Esse "novo" conjunto de dados possui 13.125 ENs anotadas, das mais de 14 mil do conjunto original.

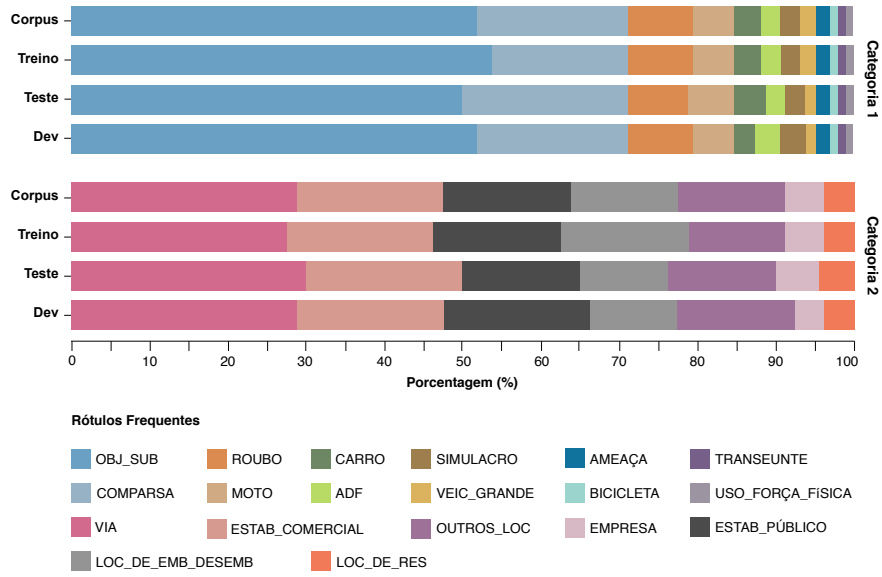
Figura 25 – Distribuição dos dados para o Experimento #3



Fonte: Elaborado pela autora.

Os *tokens* antes marcados com algum rótulo raro agora fazem parte de OUTROS, por não serem mais entidades nomeadas. Em virtude disso, uma nova vetorização e estratificação dos dados são feitas. Feita a vetorização, a Figura 26 contém as proporções dos novos subconjuntos estratificados de treino (70%), teste (20%) e desenvolvimento (10%), separados pelas duas categorias dos tipos de entidades. Com a nova estratificação, o conjunto de treino ficou com 1.069 textos, o de teste com 307, e o de desenvolvimento com 126 textos.

Figura 26 – Distribuição estratificada dos rótulos majoritários para as Categorias 1 e 2



Fonte: Elaborado pela autora.

A Tabela 5 detalha os resultados de todas as métricas. As linhas em itálico representam os tipos de ENs consideradas raras neste experimento. O modelo *Wang2Vec*+BLSTM-CRF obteve os melhores resultados da métrica F_1 para boa parte dos rótulos, com grandes diferenças quanto ao *FastText*+BLSTM-CRF+CB-CCE, por conta dos seus valores baixos no *Precision*.

Tabela 5 – Resultados do Experimento #3: análise dos rótulos majoritários para os melhores modelos BLSTM-CRF

Rótulos	**Modelos BLSTM-CRF**								
	***FastText*+CB-CCE**			**Concat*Wang2Vec*+CB-CCE***			***Wang2Vec****		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
ADF	55,333	87,368	67,755	79,245	88,421	83,582	88,462	84,737	**86,559**
AMEAÇA	*14,780*	*88,800*	*25,342*	*51,337*	*76,800*	*61,538*	*82,524*	*68,000*	***74,561***
BICICLETA	*54,264*	*97,222*	*69,652*	*78,947*	*83,333*	*81,081*	*82,192*	*83,333*	***82,759***
CARRO	50,360	92,920	65,319	64,688	91,593	75,824	90,286	69,912	**78,803**
COMPARSA	47,115	83,591	60,264	61,334	80,666	69,684	84,004	67,831	**75,056**
EMPRESA	*39,375*	*81,818*	*53,165*	*54,867*	*80,519*	*65,263*	*76,119*	*66,234*	***70,833***
ESTAB_COMERCIAL	35,072	81,757	49,087	66,292	79,730	72,393	79,231	69,595	**74,101**
ESTAB_PÚBLICO	35,714	95,000	51,913	70,229	92,000	79,654	84,259	91,000	**87,500**
LOC_DE_EMB_DESEMB	*73,000*	*82,955*	*77,660*	*86,517*	*87,500*	***87,006***	*80,612*	*89,773*	*84,946*
LOC_DE_RES	*54,839*	*75,000*	*63,354*	*96,552*	*82,353*	***88,889***	*91,071*	*75,000*	*82,258*
MOTO	81,995	94,398	87,760	88,064	92,997	**90,463**	92,941	88,515	90,674
OBJ_SUB	74,202	84,739	79,121	77,731	85,672	81,508	84,134	83,849	**83,992**
OUTROS_LOC	29,091	88,889	43,836	60,000	77,778	67,742	73,109	80,556	**76,652**
ROUBO	80,903	94,940	87,361	94,366	96,867	**95,600**	95,433	95,663	95,548
SIMULACRO	*40,278*	*95,868*	*56,724*	*68,212*	*85,124*	*75,735*	*80,469*	*85,124*	***82,731***
TRANSEUNTE	*40,136*	*96,721*	*56,731*	*72,840*	*96,721*	*83,099*	*96,667*	*95,082*	***95,868***
USO_FORÇA_FÍSICA	*21,809*	*70,690*	*33,333*	*46,753*	*62,069*	*53,333*	*74,419*	*55,172*	***63,366***
VEIC_GRANDE	*50,407*	*91,176*	*64,921*	*59,000*	*86,765*	*70,238*	*91,667*	*80,882*	***85,938***
VIA	69,249	91,615	78,877	85,965	91,304	88,554	92,926	89,752	**91,311**
Média Ponderada	61,824	86,987	70,833	75,865	84,743	79,859	85,795	80,546	**82,884**
Acurácia Balanceada		**88,183**			85,482			80,005	

Fonte: Elaborado pela autora.
(*): Este símbolo indica que a camada das *word embeddings* pré-treinadas foi atualizada durante o treino.

Verificando os resultados gerais desses modelos com os treinados com todos os rótulos, todos os modelos conquistaram um aumento considerável: o modelo *FastText*+BLSTM-CRF+CB-CCE teve um ganho de 15% para a métrica AB, e de $2,3\%$ para o F_1-*score*; o Concat*Wang2Vec*+BLSTM-CRF+CB-CCE com *fine-tuning* obteve $+13\%$ para a métrica AB, e $+2,3\%$ para o F_1; e, o *Wang2Vec*+BLSTM-CRF com *fine-tuning* segue um padrão parecido, de $+11,7\%$ para a AB, e $+1,9\%$ para o F_1. Este experimento revela que as ENs raras, mesmo que específicas, influenciam no reconhecimento das que dependem do contexto, e que a melhoria no desempenho atinge em especial a função de perda voltada para o problema de desbalanceamento de classes.

5.5 Considerações Finais

Neste capítulo foram apresentados os resultados obtidos e a metodologia experimental aplicada para responder as três questões de pesquisa. Os resultados demonstraram melhorias significativas para todas as métricas de avaliação em relação ao *baseline*. As soluções objetivas para cada uma das questões de pesquisa estão descritas a seguir:

- 1ª *Word embeddings* são sim suficientes para capturar as propriedades semânticas e sintáticas do vocabulário CVP. A abordagem mais eficiente consiste no ajuste das *word embeddings* pré-treinadas, e na concatenação entre uma camada de *embedding* com vetores pré-treinados e uma com vetores específicos do domínio. O uso dessas duas técnicas compõem o *CVP2Vec*, com os vetores pré-treinados dos modelos *FastText* e *Wang2Vec*.
- 2ª O uso da função de *loss* CB-CCE para o tratamento do desbalanceamento de dados garante um aumento significativo na qualidade e no desempenho, reconhecendo eficientemente os tipos de ENs raras.
- 3ª A remoção da anotação dos tipos de ENs minoritárias aumenta a performance e a capacidade dos modelos em reconhecer as que possuem uma maior dependência de contexto.

6 CONCLUSÃO

A sistematização dos dados sobre crimes patrimoniais é fundamental para a construção de conhecimento e de políticas públicas de segurança. Neste trabalho foi proposto o modelo CVP para o reconhecimento de entidades nomeadas em narrativas de roubos, tendo como base a arquitetura de rede neural BLSTM-CRF. Em conjunto com a aplicação de *word embeddings* pré-treinadas, tratamento de OOV, e técnicas para o desbalanceamento de nível arquitetural e dos dados, o modelo apresentou alta assertividade na extração de 36 tipos de entidades nomeadas importantes para o domínio, automatizando o processo de obtenção de informações referentes aos crimes violentos contra o patrimônio.

Este trabalho também descreve a construção do conjunto de dados anotado CVP e do *CVP2Vec*, duas contribuições importantes, uma vez que o corpus anotado pode ser usado em outros sistemas de mesmo domínio, e o conjunto de técnicas do *CVP2Vec* geram *word embeddings* que capturam as especificidades da linguagem dos textos de roubos, criando representações vetoriais das palavras que podem ser reconhecidas como ENs a partir de diferentes modelos. Inicialmente, este trabalho visava investigar o NER para apenas 11 rótulos predefinidos e com o modelo *baseline spaCy* NER. Após uma demanda no aumento da quantidade dos rótulos e da presença do problema de desbalanceamento de dados, tornou-se necessário a definição de questões de pesquisa para investigar novas soluções.

Os experimentos seguiram três questões de pesquisa para avaliação do modelo CVP. Em relação ao *baseline*, foram obtidas melhorias significativas para as métricas Acurácia Balanceada (com ganhos entre $9,2\%$ e 15%) e F_1-*score*. A avaliação detalhada de cada questão de pesquisa também é uma contribuição deste trabalho. O protocolo experimental criado forneceu respostas para qual a melhor representação de *embedding* para o problema; se as soluções a nível arquitetural para lidar com o problema de desbalanceamento melhoram a qualidade do modelo; e, o quanto os rótulos minoritários interferem na performance.

Uma outra contribuição é que o modelo *spaCy* NER utilizado como *baseline* faz hoje parte do HNERD, auxiliando na captura de informações dos registros CVP. Todo o processo de elaboração e desenvolvimento deste trabalho, desde o pré-processamento, estratificação, e o conjunto dos modelos criados, serão adaptados para a liberação de uma nova versão da ferramenta interativa HNERD.

Para trabalhos futuros, pretende-se explorar técnicas de Aprendizagem Ativa (*Active Learning*) aplicáveis para o Aprendizado Profundo, para redução do esforço do processo de

anotação dos dados, de forma que mesmo que poucos dados sejam usados para treinamento, o modelo possa ter a opção de escolher os exemplos a partir dos quais irá aprender. O estudo do problema de desbalanceamento e a definição do esquema de anotação também podem ser estendidos para a adição de mais tipos de entidades nomeadas.

Também planeja-se investigar a adoção de *embeddings* (pré-treinados ou não) gerados com modelos de linguagem contextualizada, como BERT (DEVLIN *et al.*, 2018) e ELMo (PETERS *et al.*, 2018). Outras oportunidades são unir esses *embeddings* de modelos de linguagem com a arquitetura de rede neural *Transformer* de Vaswani *et al.* (2017), como substituição para o codificador de rótulo BLSTM. Ambos os modelos BERT, ELMo, além do recente BIGBIRD (ZAHEER *et al.*, 2020), utilizam *Transformers*, arquitetura que está se tornando um novo paradigma para a geração de modelos NER (LI *et al.*, 2020).

REFERÊNCIAS

ADORNO, S.; PASINATO, W. Violência e Impunidade Penal: da Criminalidade Detectada à Criminalidade Investigada. **Dilemas-Revista de Estudos de Conflito e Controle Social**, Rio de Janeiro, Brasil, v. 3, n. 7, p. 51–84, 2010.

AGUILAR, G.; KAR, S.; SOLORIO, T. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. **arXiv preprint arXiv:2005.04322**, Cornell University, Ithaca, Nova Iorque, EUA, 2020.

AYRES, R. L. **Crime and Violence as Development Issues in Latin America and the Caribbean**. Washington, D.C., EUA: The World Bank, 1998.

AZEVEDO, F. A.; CARVALHO, L. R.; GRINBERG, L. T.; FARFEL, J. M.; FERRETTI, R. E.; LEITE, R. E.; FILHO, W. J.; LENT, R.; HERCULANO-HOUZEL, S. Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain. **Journal of Comparative Neurology**, Wiley Online Library, v. 513, n. 5, p. 532–541, 2009.

BEZERRA, E. Introdução à Aprendizagem Profunda. **Artigo–31º Simpósio Brasileiro de Banco de Dados–SBBD2016–Salvador**, Salvador, Bahia, Brasil, 2016.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, Massachusetts, EUA, v. 5, p. 135–146, 2017.

CEARÁ, G. do Estado do. **O Cenário da Violência e da Criminalidade no Brasil e no Ceará**. Fortaleza, Ceará, Brasil, 2019. Disponível em: https://www.ceara.gov.br/ceara-pacifico/. Acesso em: 15 de Set. de 2019.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Special Issue on Learning from Imbalanced Data Sets. **ACM SIGKDD explorations newsletter**, ACM New York, Nova Iorque, Nova York, EUA, v. 6, n. 1, p. 1–6, 2004.

CHEN, C.; LIAW, A.; BREIMAN, L. *et al.* Using Random Forest to Learn Imbalanced Data. **University of California, Berkeley**, Berkeley, Califórnia, EUA, v. 110, n. 1-12, p. 24, 2004.

CHEN, H.; CHUNG, W.; XU, J. J.; WANG, G.; QIN, Y.; CHAU, M. Crime Data Mining: a General Framework and Some Examples. **computer**, IEEE, Piscataway, Nova Jersey, EUA, n. 4, p. 50–56, 2004.

CHOLLET, F. **Deep Learning with Python**. 1st. ed. EUA: Manning Publications Co., 2017. ISBN 1617294438.

COELHO DA SILVA, T. L.; ARAUJÚ, N. da S.; MACÊDO, J. A. F. de; ARAÚJO, D.; SOARES, F. M.; REGO, P. A. L.; NETO, A. V. L. Novel approach for label disambiguation via deep learning. In: PERNER, P. (Ed.). **Machine Learning and Data Mining in Pattern Recognition, 15th International Conference on Machine Learning and Data Mining, MLDM 2019, July 20-25, 2019, Proceedings, Volume II**. Nova Iorque, Nova York, EUA: ibai publishing, 2019. p. 431–442.

COELHO DA SILVA, T. L.; MAGALHÃES, R. P.; MACÊDO, J. A. F. de; ABREU, D. A.; ARAÚJO, N. d. S.; MELO, V. T. de; PINHEIRO, P. O.; REGO, P. A. L.; LIRA NETO, A. V. Improving Named Entity Recognition using Deep Learning with Human in the Loop. **EDBT/ICDT 2019 Joint Conference**, Lisboa, Portugal, 2019.

CUI, Y.; JIA, M.; LIN, T.-Y.; SONG, Y.; BELONGIE, S. Class-Balanced Loss Based on Effective Number of Samples. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. San Juan, Porto Rico, EUA: IEEE, 2019. p. 9268–9277.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv preprint arXiv:1810.04805**, Cornell University, Ithaca, Nova Iorque, EUA, 2018.

GARNEAU, N.; LEBOEUF, J.-S.; LAMONTAGNE, L. Predicting and interpreting embeddings for out of vocabulary words in downstream tasks. In: **Proceedings of the 2018 EMNLP Workshop BlackboxNLP**: Analyzing and interpreting neural networks for NLP. Bruxelas, Bélgica: Association for Computational Linguistics, 2018. p. 331–333.

Google Developers. **Embeddings**: Translating to a lower-dimensional space. Mountain View, Califórnia, Estados Unidos da América, 2020. Machine Learning Crash Course with TensorFlow APIs. Disponível em: https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space. Acesso em: 21 de Jun. de 2020.

HARTMANN, N. S.; FONSECA, E. R.; SHULBY, C. D.; TREVISO, M. V.; RODRIGUES, J. S.; ALUíSIO, S. M. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: **Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, Rio Grande do Sul, Brasil: SBC, 2017. p. 122–131.

HAYKIN, S. **Redes neurais: Princípios e Prática. 2ª Edição**. Porto Alegre, Rio Grande do Sul, Brasil: Editora Bookman, 2007.

HONNIBAL, M.; MONTANI, I. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. **To appear**, 2017.

HUANG, Z.; XU, W.; YU, K. Bidirectional LSTM-CRF Models for Sequence Tagging. **arXiv preprint arXiv:1508.01991**, Cornell University, Ithaca, Nova Iorque, EUA, 2015.

JACKSON, P.; MOULINIER, I. **Natural Language Processing for Online Applications**: Text Retrieval, Extraction and Categorization. Amsterdã, Países Baixos: John Benjamins Publishing, 2007. v. 5.

J.R.R. Tolkien. **O Senhor dos Anéis: A Sociedade do Anel**. São Paulo, Brasil: Martins Fontes Editora, 2017. v. 1.

JURAFSKY, D.; MARTIN, J. H. Regular Expressions, Text Normalization, Edit Distance. In: **Speech and Language Processing**. Stanford, Califórnia, EUA: Stanford University, 2017. p. 10–33.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: **Proc. 18th International Conf. on Machine Learning**. Burlington, Massachusetts, EUA: Morgan Kaufmann Publishers, 2001. p. 282–289.

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, Califórnia, EUA: Association for Computational Linguistics, 2016. p. 260–270.

LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. Fine-Grained Named Entity Recognition in Legal Documents. In: SPRINGER. **International Conference on Semantic Systems**. [*S. l.*], 2019. p. 272–287.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In: **Soviet physics doklady**. [*S. l.*: *s. n.*], 1966. v. 10, n. 8, p. 707–710.

LI, J.; SUN, A.; HAN, J.; LI, C. A Survey on Deep Learning for Named Entity Recognition. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, San Juan, Porto Rico, EUA, 2020.

LI, X.; SUN, X.; MENG, Y.; LIANG, J.; WU, F.; LI, J. Dice Loss for Data-imbalanced NLP Tasks. **arXiv preprint arXiv:1911.02855**, Cornell University, Ithaca, Nova Iorque, EUA, 2019.

LIDDY, E. D. Natural Language Processing. In: **Encyclopedia of Library and Information Science**. Nova Iorque, Nova York, EUA: Marcel Decker, Inc., 2001.

LING, W.; DYER, C.; BLACK, A. W.; TRANCOSO, I. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics**: Human language technologies. [*S. l.*: *s. n.*], 2015. p. 1299–1304.

LIU, Y.; LOH, H. T.; SUN, A. Imbalanced Text Classification: A Term Weighting Approach. **Expert systems with Applications**, Elsevier, Amsterdã, Países Baixos, v. 36, n. 1, p. 690–701, 2009.

MA, X.; HOVY, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. **arXiv preprint arXiv:1603.01354**, Cornell University, Ithaca, Nova Iorque, EUA, 2016.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, Massachusetts, EUA: MIT Press, 1999.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. **arXiv preprint arXiv:1301.3781**, Cornell University, Ithaca, Nova Iorque, EUA, 2013.

MOSLEY, L. A balanced approach to the multi-class imbalance problem. In: **Iowa State University Capstones, Theses and Dissertations**. Ames, Iowa, EUA: Iowa State University, 2013.

NADEAU, D.; SEKINE, S. A Survey of Named Entity Recognition and Classification. **Lingvisticae Investigationes**, John Benjamins, Amsterdã, Países Baixos, v. 30, n. 1, p. 3–26, 2007.

NGUYEN, T.; NGUYEN, D.; RAO, P. Adaptive Name Entity Recognition under Highly Unbalanced Data. **arXiv preprint arXiv:2003.10296**, Cornell University, Ithaca, Nova Iorque, EUA, 2020.

PARK, A.; CLARE, J.; SPICER, V.; BRANTINGHAM, P. L.; CALVERT, T.; JENION, G. Examining Context-specific Perceptions of Risk: Exploring the Utility of "Human-in-the-Loop" Simulation Models for Criminology. **Journal of experimental criminology**, Springer, Nova Iorque, Nova York, EUA, v. 8, n. 1, p. 29–47, 2012.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. GloVe: Global Vectors for Word Representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [*S. l.*: *s. n.*], 2014. p. 1532–1543.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. **arXiv preprint arXiv:1802.05365**, Cornell University, Ithaca, Nova Iorque, EUA, 2018.

PINTER, Y.; GUTHRIE, R.; EISENSTEIN, J. Mimicking Word Embeddings using Subword RNNs. **arXiv preprint arXiv:1707.06961**, Cornell University, Ithaca, Nova Iorque, EUA, 2017.

RAJA, R. A.; LAY-KI, S.; SU-CHENG, H. Exploring Edit Distance for Normalising Out-of-Vocabulary Malay Words on Social Media. In: EDP SCIENCES. **MATEC Web of Conferences**. [*S. l.*], 2019. v. 255, p. 03001.

SACHAN, D. S.; XIE, P.; SACHAN, M.; XING, E. P. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. **arXiv preprint arXiv:1711.07908**, Cornell University, Ithaca, Nova Iorque, EUA, 2017.

SECHIDIS, K.; TSOUMAKAS, G.; VLAHAVAS, I. On the Stratification of Multi-Label Data. In: SPRINGER. **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. [*S. l.*], 2011. p. 145–158.

SHABAT, H.; OMAR, N.; RAHEM, K. Named Entity Recognition in Crime Using Machine Learning Approach. In: SPRINGER. **Asia Information Retrieval Symposium**. [*S. l.*], 2014. p. 280–288.

SONG, F. Y. Y. Z. S.; XIAO, A. S. J. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. **arXiv preprint arXiv:1506.03365**, Cornell University, Ithaca, Nova Iorque, EUA, 2015.

SPECK, R.; NGOMO, A.-C. N. Ensemble Learning for Named Entity Recognition. In: SPRINGER. **International Semantic Web Conference**. Riva del Garda, Trento, Itália, 2014. p. 519–534.

SSPDS. **ESTATÍSTICAS**. Fortaleza, Ceará Brasil, 2020. Disponível em: https://www.sspds.ce.gov.br/estatisticas-2/. Acesso em: 20 de Set. de 2020.

TALON. **Deep Learning Book**. 2019. Disponível em: http://www.deeplearningbook.com.br/. Acesso em: 30 de Set. de 2019.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention Is All You Need. In: **Advances in neural information processing systems**. [*S. l.*: *s. n.*], 2017. p. 5998–6008.

WESTON, L.; TSHITOYAN, V.; DAGDELEN, J.; KONONOVA, O.; TREWARTHA, A.; PERSSON, K. A.; CEDER, G.; JAIN, A. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. **Journal of chemical information and modeling**, ACS Publications, Washington, D.C., EUA, v. 59, n. 9, p. 3692–3702, 2019.

ZAHEER, M.; GURUGANESH, G.; DUBEY, A.; AINSLIE, J.; ALBERTI, C.; ONTANON, S.; PHAM, P.; RAVULA, A.; WANG, Q.; YANG, L. *et al.* Big Bird: Transformers for Longer Sequences. **arXiv preprint arXiv:2007.14062**, Cornell University, Ithaca, Nova Iorque, EUA, 2020.

ZHANG, Y.; WALLACE, B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. **arXiv preprint arXiv:1510.03820**, Cornell University, Ithaca, Nova Iorque, EUA, 2015.

APÊNDICE A – VARIAÇÕES DOS MODELOS BLSTM-CRF DO EXPERIMENTO #1 POR RÓTULO

As tabelas abaixo exibem os resultados dos modelos BLSTM-CRF do Cenário de Experimentação #1 detalhados por rótulo. As linhas representam um tipo de EN ligada a uma variação do modelo BLSTM-CRF (coluna), com os valores para as métricas *Precision*, *Recall* e F_1-*score*. Os resultados destacados em negrito indicam os maiores valores para o F_1 de acordo com os rótulos. Os valores em $0,00\%$ indicam que o modelo não reconheceu a entidade nomeada.

Cada variação do modelo com uma função de perda possui quatro tabelas com resultados, em que as duas primeiras contém os valores para os rótulos majoritários, e as duas últimas os valores para os minoritários: Tabelas 6, 7, 8 e 9 para o modelo BLSTM-CRF+CE-CCE; Tabelas 10, 11, 12 e 13 para o modelo BLSTM-CRF; e, Tabelas 14, 15, 16 e 17 para o modelo BLSTM-CRF+DL. Os rótulos frequentes e raros são divididos em quatro tabelas porque ultrapassam a margem do documento se colocados em apenas duas. A legenda de cada tabela referencia qual a primeira e a segunda (continuação) partes.

Tabela 6 – Parte 1: resultados dos modelos BLSTM-CRF+CB-CCE para os rótulos majoritários

Rótulos majoritários	**Modelos BLSTM-CRF+CB-CCE – Parte 1**														
	WE do domínio			***FastText***			**GloVe**			***Wang2Vec***			***Word2Vec***		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
ADF	69,43	83,68	75,90	53,42	90,53	67,19	49,12	87,90	63,02	61,57	91,05	73,46	51,39	87,37	64,72
AMEAÇA	39,78	69,94	50,71	23,83	79,74	36,69	24,55	71,24	36,52	23,66	81,05	36,63	15,19	70,59	25,00
BICICLETA	68,27	72,45	70,30	57,46	78,57	66,38	66,38	78,57	71,96	74,04	78,57	76,24	50,70	73,47	60,00
CARRO	65,08	69,49	67,21	53,63	90,68	67,40	61,15	86,02	71,48	60,41	88,56	71,82	48,04	88,14	62,18
COMPARSA	54,07	73,05	62,14	49,86	74,51	59,74	50,98	75,79	60,96	49,28	79,23	60,76	46,83	71,07	56,46
EMPRESA	40,00	82,00	53,77	28,53	87,00	42,96	36,77	82,00	50,77	37,66	87,00	52,57	22,74	83,00	35,70
ESTAB_COMERCIAL	56,13	65,41	60,42	33,44	78,20	46,85	35,86	78,20	49,17	42,50	76,69	54,69	29,19	70,68	41,32
ESTAB_PÚBLICO	83,33	88,50	85,84	27,51	92,04	42,36	47,20	89,38	61,77	36,50	88,50	51,68	25,38	89,38	39,53
LOC_DE_EMB_DESEMB	71,85	88,10	79,14	72,83	79,76	76,14	76,29	88,10	81,77	82,56	84,52	83,53	66,67	80,95	73,12
LOC_DE_RES	66,67	90,32	76,71	51,38	90,32	65,50	61,45	82,26	70,35	73,91	82,26	77,86	37,98	79,03	51,31
MOTO	68,38	93,66	79,05	68,24	96,83	80,06	69,13	95,42	80,18	71,20	95,78	81,68	60,49	96,48	74,36
OBJ_SUB	74,86	85,49	79,82	76,42	81,10	78,69	74,66	84,47	79,26	73,76	84,31	78,68	71,73	80,98	76,07
OUTROS_LOC	61,11	76,74	68,04	30,95	90,70	46,15	46,49	82,17	59,38	48,67	85,27	61,97	37,12	86,05	51,87
ROUBO	92,91	96,54	94,69	79,10	96,77	87,05	77,76	95,85	85,86	72,92	96,77	83,17	66,56	92,17	77,30
SIMULACRO	55,84	77,25	64,82	45,86	79,64	58,21	44,30	79,04	56,77	43,85	79,04	56,41	32,77	69,46	44,53
TRANSEUNTE	74,67	86,15	80,00	48,33	89,23	62,70	49,12	86,15	62,57	55,24	89,23	68,24	30,00	87,69	44,71
USO_FORÇA_FÍSICA	44,19	65,52	52,78	16,93	74,14	27,56	27,45	72,41	39,81	22,22	75,86	34,38	13,48	74,14	22,81
VEIC_GRANDE	57,80	72,41	64,29	48,70	86,21	62,24	60,33	83,91	70,19	57,78	89,66	70,27	51,02	86,21	64,10
VIA	76,17	92,09	83,38	66,52	87,01	75,40	76,39	93,22	83,97	77,01	91,81	83,76	69,52	85,03	76,49

Fonte: Elaborado pela autora.

Tabela 7 – Parte 2: resultados dos modelos BLSTM-CRF+CB-CCE para os rótulos majoritários

Rótulos majoritários	Modelos BLSTM-CRF+CB-CCE – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
ADF	72,61	92,11	81,21	76,71	88,42	**82,15**	76,02	88,42	81,75	70,83	89,47	79,07
AMEAÇA	32,53	79,74	46,21	48,11	66,67	**55,89**	41,38	70,59	52,17	43,03	68,63	52,90
BICICLETA	70,59	73,47	72,00	84,52	72,45	**78,02**	70,19	74,49	72,28	75,79	73,47	74,61
CARRO	60,49	83,05	70,00	70,16	76,70	**73,28**	78,26	68,64	73,14	64,98	76,27	70,18
COMPARSA	64,44	67,04	**65,71**	57,11	74,08	64,50	55,70	75,02	63,94	56,88	73,13	63,99
EMPRESA	60,43	84,00	**70,29**	51,59	81,00	63,04	40,87	85,00	55,20	50,59	86,00	63,70
ESTAB_COMERCIAL	67,15	69,17	68,15	68,15	69,17	68,66	72,27	64,66	68,25	68,57	72,18	**70,33**
ESTAB_PÚBLICO	91,07	90,27	**90,67**	73,91	90,27	81,28	79,39	92,04	85,25	76,12	90,27	82,59
LOC_DE_EMB_DESEMB	83,72	85,71	**84,71**	78,72	88,10	83,15	73,20	84,52	78,45	80,44	88,10	84,09
LOC_DE_RES	57,84	95,16	71,95	72,50	93,55	81,69	82,35	90,32	**86,15**	75,34	88,71	81,48
MOTO	87,42	90,49	**88,93**	78,25	91,20	84,23	75,72	92,25	83,18	79,33	91,90	85,16
OBJ_SUB	77,75	84,71	81,08	77,78	85,26	**81,35**	76,20	84,24	80,02	76,94	85,45	80,97
OUTROS_LOC	64,29	69,77	66,91	70,37	73,64	**71,97**	72,50	67,44	69,88	69,60	67,44	68,50
ROUBO	92,95	97,24	95,05	93,82	97,93	**95,83**	92,54	97,24	94,83	92,97	97,47	95,16
SIMULACRO	73,74	79,04	76,30	77,52	78,44	77,98	84,28	80,24	**82,21**	81,01	76,65	78,77
TRANSEUNTE	86,15	86,15	86,15	87,50	86,15	86,82	83,82	87,69	85,71	88,89	86,15	**87,50**
USO_FORÇA_FÍSICA	36,28	70,69	47,95	49,33	63,79	**55,64**	25,84	79,31	38,98	31,11	72,41	43,52
VEIC_GRANDE	54,96	82,76	66,06	76,92	80,46	**78,65**	66,67	78,16	71,96	62,61	82,76	71,29
VIA	75,34	93,22	83,33	80,40	91,53	85,60	84,38	91,53	**87,81**	78,80	92,37	85,05

Fonte: Elaborado pela autora.

Tabela 8 – Parte 1: resultados dos modelos BLSTM-CRF+CB-CCE para os rótulos minoritários

Rótulos minoritários	Modelos BLSTM-CRF+CB-CCE – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
AB	59,68	68,52	63,79	42,70	70,37	53,15	43,43	79,63	56,21	48,10	70,37	57,14	37,37	68,52	48,37
AMB_VIRTUAL	9,52	9,09	9,30	18,33	50,00	26,83	34,38	50,00	**40,74**	15,91	31,82	21,21	10,42	45,46	16,95
ESTAB_ESPORTIVO	40,00	40,00	40,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_MÉDICO	41,67	41,67	41,67	25,00	66,67	36,36	72,73	66,67	**69,57**	38,89	58,33	46,67	29,17	58,33	38,89
ESTAB_RELIGIOSO	66,67	66,67	66,67	50,00	66,67	57,14	100,00	66,67	**80,00**	37,50	100,00	54,55	50,00	33,33	40,00
ESTAC	55,56	62,50	58,82	55,56	62,50	58,82	66,67	50,00	57,14	71,43	62,50	66,67	41,67	62,50	50,00
EVENTO	10,00	33,33	15,39	7,69	33,33	12,50	50,00	33,33	**40,00**	18,18	66,67	28,57	3,45	33,33	6,25
FAVELA	50,00	50,00	50,00	28,57	28,57	28,57	54,55	42,86	48,00	77,78	50,00	60,87	27,27	42,86	33,33
FURTO	55,56	71,43	62,50	35,85	90,48	51,35	48,39	71,43	57,69	28,00	66,67	39,44	19,51	76,19	31,07
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	28,57	55,56	37,74	26,00	72,22	38,24	30,77	66,67	42,11	22,22	66,67	33,33	27,59	44,44	34,04
LOC_DE_ENSINO	73,68	58,33	65,12	53,85	87,50	66,67	57,90	91,67	70,97	57,58	79,17	66,67	45,65	87,50	60,00
LOC_DE_LAZER	39,54	68,00	50,00	44,05	74,00	55,22	51,95	80,00	62,99	51,43	72,00	60,00	24,83	72,00	36,92
MDI	45,71	76,19	57,14	26,61	78,57	39,76	27,78	71,43	40,00	30,53	69,05	42,34	20,87	57,14	30,57
OUTROS_MEIOS_TRANSP	64,71	75,86	69,84	42,86	72,41	53,85	48,65	62,07	54,55	30,65	65,52	41,76	36,07	75,86	48,89
TERRENO	60,00	69,23	64,29	56,25	69,23	62,07	56,25	69,23	62,07	66,67	61,54	64,00	53,33	61,54	57,14
TRANSP_ALT	66,67	96,55	78,87	39,26	91,38	54,92	61,11	94,83	74,32	30,06	89,66	45,02	31,68	87,93	46,58

Fonte: Elaborado pela autora.

Tabela 9 – Parte 2: resultados dos modelos BLSTM-CRF+CB-CCE para os rótulos minoritários

Rótulos minoritários	Modelos BLSTM-CRF+CB-CCE – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
AB	70,00	77,78	**73,68**	67,27	68,52	67,89	72,00	66,67	69,23	67,27	68,52	67,89
AMB_VIRTUAL	66,67	27,27	38,71	38,89	31,82	35,00	66,67	27,27	38,71	37,50	27,27	31,58
ESTAB_ESPORTIVO	15,79	60,00	25,00	50,00	40,00	44,44	36,36	80,00	**50,00**	30,00	60,00	40,00
ESTAB_MÉDICO	30,00	50,00	37,50	53,33	66,67	59,26	75,00	50,00	60,00	72,73	66,67	**69,57**
ESTAB_RELIGIOSO	18,18	66,67	28,57	28,57	66,67	40,00	50,00	66,67	57,14	100,00	66,67	**80,00**
ESTAC	75,00	75,00	75,00	100,00	62,50	76,92	100,00	75,00	**85,71**	83,33	62,50	71,43
EVENTO	25,00	33,33	28,57	12,50	33,33	18,18	10,00	33,33	15,39	20,00	33,33	25,00
FAVELA	88,89	57,14	**69,57**	63,64	50,00	56,00	75,00	64,29	69,23	72,73	57,14	64,00
FURTO	68,18	71,43	69,77	82,35	66,67	**73,68**	71,43	71,43	71,43	73,68	66,67	70,00
HOSPEDARIA	100,00	12,50	**22,22**	0,00	0,00	0,00	50,00	12,50	20,00	0,00	0,00	0,00
INST_FINANCEIRA	23,91	61,11	34,38	33,33	44,44	38,10	52,38	61,11	**56,41**	36,00	50,00	41,86
LOC_DE_ENSINO	58,07	75,00	65,46	64,00	66,67	65,31	43,18	79,17	55,88	77,27	70,83	**73,91**
LOC_DE_LAZER	61,82	68,00	64,76	64,29	72,00	67,93	55,22	74,00	63,25	67,27	74,00	**70,48**
MDI	50,79	76,19	60,95	57,14	76,19	**65,31**	42,17	83,33	56,00	49,23	76,19	59,81
OUTROS_MEIOS_TRANSP	76,67	79,31	77,97	59,46	75,86	66,67	91,67	75,86	**83,02**	70,97	75,86	73,33
TERRENO	75,00	69,23	**72,00**	56,25	69,23	62,07	64,29	69,23	66,67	69,23	69,23	69,23
TRANSP_ALT	86,44	87,93	87,18	89,66	89,66	**89,66**	72,60	91,38	80,92	80,00	89,66	84,55

Fonte: Elaborado pela autora.

Tabela 10 – Parte 1: resultados dos modelos BLSTM-CRF para os rótulos majoritários

Rótulos majoritários	Modelos BLSTM-CRF – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADF	74,76	82,63	78,50	86,49	84,21	85,33	85,79	85,79	85,79	85,08	81,05	83,02	82,01	81,58	81,79
AMEAÇA	72,88	56,21	63,47	79,27	42,48	55,32	78,65	45,75	57,85	80,00	41,83	54,94	84,21	31,37	45,71
BICICLETA	83,33	66,33	73,86	85,71	73,47	**79,12**	90,67	69,39	78,61	88,31	69,39	77,71	84,62	67,35	75,00
CARRO	81,22	67,80	73,90	86,64	79,66	**83,00**	87,93	64,83	74,63	85,17	75,42	80,00	80,91	75,42	78,07
COMPARSA	52,93	75,19	62,13	79,29	57,85	66,90	77,89	64,12	70,34	78,19	64,64	**70,77**	77,85	62,75	69,49
EMPRESA	73,40	69,00	71,13	71,72	71,00	71,36	66,09	76,00	70,70	81,32	74,00	**77,49**	59,06	75,00	66,08
ESTAB_COMERCIAL	73,08	57,14	64,14	63,36	62,41	62,88	68,29	63,16	65,63	71,17	59,40	64,75	60,63	57,90	59,23
ESTAB_PÚBLICO	87,50	86,73	87,11	93,20	84,96	**88,89**	95,75	79,65	86,96	78,99	83,19	81,03	85,35	87,61	86,46
LOC_DE_EMB_DESEMB	80,23	82,14	81,18	85,53	77,38	81,25	94,03	75,00	83,44	85,14	75,00	79,75	85,51	70,24	77,12
LOC_DE_RES	82,09	88,71	**85,27**	88,89	77,42	82,76	84,62	70,97	77,19	83,64	74,19	78,63	90,91	64,52	75,47
MOTO	82,26	89,79	85,86	85,58	94,01	**89,60**	89,93	88,03	88,97	84,49	90,14	87,22	81,76	88,38	84,94
OBJ_SUB	75,71	84,47	79,85	87,61	80,16	**83,72**	80,88	85,92	83,32	86,93	80,35	83,51	85,30	79,41	82,25
OUTROS_LOC	62,90	60,47	61,66	70,59	74,42	72,45	76,80	74,42	**75,59**	66,88	79,85	72,79	65,96	72,09	68,89
ROUBO	90,83	95,85	93,27	93,86	95,16	94,51	93,15	94,01	93,58	94,73	95,16	94,94	92,34	94,47	93,39
SIMULACRO	86,71	74,25	80,00	87,81	64,67	74,48	83,93	56,29	67,38	83,33	65,87	73,58	76,15	49,70	60,15
TRANSEUNTE	83,33	76,92	80,00	96,15	76,92	85,47	95,75	69,23	80,36	94,12	73,85	82,76	88,68	72,31	79,66
USO_FORÇA_FÍSICA	60,98	43,10	50,51	69,70	39,66	50,55	66,67	34,48	45,46	71,88	39,66	51,11	43,33	22,41	29,55
VEIC_GRANDE	67,37	73,56	70,33	78,21	70,12	73,94	77,53	79,31	78,41	82,28	74,71	78,31	72,45	81,61	76,76
VIA	85,12	87,29	86,19	90,91	90,40	90,65	92,15	89,55	**90,83**	91,40	87,01	89,15	88,59	83,33	85,88

Fonte: Elaborado pela autora.

Tabela 11 – Parte 2: resultados dos modelos BLSTM-CRF para os rótulos majoritários

Rótulos majoritários	Modelos BLSTM-CRF – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADF	87,93	80,53	84,07	91,72	81,58	**86,35**	84,78	82,11	83,42	89,29	78,95	83,80
AMEAÇA	77,48	56,21	**65,15**	63,57	58,17	60,75	70,40	57,52	63,31	69,49	53,60	60,52
BICICLETA	90,54	68,37	77,91	90,41	67,35	77,19	94,12	65,31	77,11	87,84	66,33	75,58
CARRO	76,47	66,10	70,91	78,03	73,73	75,82	71,10	65,68	68,28	83,43	63,98	72,42
COMPARSA	66,58	67,21	66,89	70,63	64,81	67,59	67,31	65,75	66,52	67,04	67,38	67,21
EMPRESA	69,75	83,00	75,80	68,87	73,00	70,87	68,81	75,00	71,77	72,64	77,00	74,76
ESTAB_COMERCIAL	63,33	71,43	67,14	62,09	71,43	66,43	62,50	67,67	64,98	66,42	68,42	**67,41**
ESTAB_PÚBLICO	85,71	90,27	87,93	75,76	88,50	81,63	81,30	88,50	84,75	83,47	89,38	86,33
LOC_DE_EMB_DESEMB	77,42	85,71	81,36	80,44	88,10	**84,09**	80,22	86,91	83,43	81,82	85,71	83,72
LOC_DE_RES	76,92	80,65	78,74	79,10	85,48	82,17	73,97	87,10	80,00	79,66	75,81	77,69
MOTO	84,79	92,25	88,36	83,39	88,38	85,81	79,88	90,85	85,01	86,55	88,38	87,46
OBJ_SUB	78,18	80,94	79,54	74,69	83,65	78,91	77,76	80,90	79,30	75,78	81,73	78,64
OUTROS_LOC	65,19	68,22	66,67	72,41	65,12	68,57	67,18	68,22	67,69	67,44	67,44	67,44
ROUBO	95,05	97,24	**96,13**	94,82	97,01	95,90	93,74	96,54	95,12	94,77	96,08	95,42
SIMULACRO	88,41	73,05	80,00	87,59	71,86	78,95	85,32	73,05	78,71	95,20	71,26	**81,51**
TRANSEUNTE	93,22	84,62	88,71	88,53	83,08	85,71	88,89	86,15	87,50	90,48	87,69	**89,06**
USO_FORÇA_FÍSICA	66,67	62,07	64,29	71,70	65,52	68,47	44,44	62,07	51,80	79,55	60,35	**68,63**
VEIC_GRANDE	73,33	75,86	74,58	71,43	74,71	73,03	78,02	81,61	**79,78**	74,39	70,12	72,19
VIA	85,64	90,96	88,22	90,80	86,44	88,57	87,96	88,70	88,33	87,82	87,57	87,69

Fonte: Elaborado pela autora.

Tabela 12 – Parte 1: resultados dos modelos BLSTM-CRF para os rótulos minoritários

Rótulos minoritários	Modelos BLSTM-CRF – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
AB	73,47	66,67	69,90	88,24	55,56	68,18	82,86	53,70	65,17	86,11	57,41	68,89	87,88	53,70	66,67
AMB_VIRTUAL	0,00	0,00	0,00	76,92	45,46	**57,14**	55,56	22,73	32,26	80,00	18,18	29,63	100,00	9,09	16,67
ESTAB_ESPORTIVO	23,08	60,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_MÉDICO	57,14	33,33	42,11	50,00	50,00	50,00	75,00	50,00	**60,00**	75,00	50,00	**60,00**	57,14	33,33	42,11
ESTAB_RELIGIOSO	100,00	33,33	50,00	0,00	0,00	0,00	0,00	0,00	0,00	100,00	33,33	50,00	0,00	0,00	0,00
ESTAC	50,00	25,00	33,33	100,00	12,50	22,22	100,00	37,50	54,55	0,00	0,00	0,00	0,00	0,00	0,00
EVENTO	0,00	0,00	0,00	0,00	0,00	0,00	100,00	33,33	**50,00**	50,00	33,33	40,00	0,00	0,00	0,00
FAVELA	87,50	50,00	63,64	100,00	14,29	25,00	100,00	35,71	52,63	83,33	35,71	50,00	100,00	14,29	25,00
FURTO	68,75	52,38	59,46	100,00	14,29	25,00	100,00	19,05	32,00	83,33	23,81	37,04	66,67	9,52	16,67
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	43,75	38,89	41,18	36,36	44,44	40,00	42,86	16,67	24,00	50,00	33,33	40,00	38,89	38,89	38,89
LOC_DE_ENSINO	59,09	54,17	56,52	85,71	75,00	**80,00**	65,39	70,83	68,00	62,07	75,00	67,93	62,50	62,50	62,50
LOC_DE_LAZER	68,75	66,00	67,35	60,00	66,00	62,86	68,42	78,00	72,90	78,26	72,00	**75,00**	73,53	50,00	59,52
MDI	69,05	69,05	**69,05**	58,07	42,86	49,32	64,29	42,86	51,43	70,37	45,24	55,07	56,25	42,86	48,65
OUTROS_MEIOS_TRANSP	64,52	68,97	66,67	88,24	51,72	65,22	100,00	48,28	65,12	93,75	51,72	66,67	89,47	58,62	70,83
TERRENO	72,73	61,54	66,67	75,00	46,15	57,14	87,50	53,85	66,67	83,33	38,46	52,63	85,71	46,15	60,00
TRANSP_ALT	78,33	81,03	79,66	90,00	77,59	83,33	90,74	84,48	87,50	90,39	81,03	85,46	78,95	77,59	78,26

Fonte: Elaborado pela autora.

Tabela 13 – Parte 2: resultados dos modelos BLSTM-CRF para os rótulos minoritários

Rótulos minoritários	Modelos BLSTM-CRF – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
AB	67,44	53,70	59,79	72,55	68,52	70,48	84,62	61,11	**70,97**	76,19	59,26	66,67
AMB_VIRTUAL	57,14	18,18	27,59	100,00	13,64	24,00	30,00	13,64	18,75	100,00	9,09	16,67
ESTAB_ESPORTIVO	75,00	60,00	**66,67**	75,00	60,00	**66,67**	100,00	40,00	57,14	50,00	20,00	28,57
ESTAB_MÉDICO	80,00	33,33	47,06	57,14	33,33	42,11	83,33	41,67	55,56	50,00	33,33	40,00
ESTAB_RELIGIOSO	100,00	66,67	**80,00**	50,00	66,67	57,14	66,67	66,67	66,67	100,00	33,33	50,00
ESTAC	71,43	62,50	66,67	83,33	62,50	**71,43**	71,43	62,50	66,67	83,33	62,50	**71,43**
EVENTO	50,00	33,33	40,00	0,00	0,00	0,00	50,00	33,33	40,00	50,00	33,33	40,00
FAVELA	85,71	42,86	57,14	87,50	50,00	63,64	55,56	35,71	43,48	88,89	57,14	**69,57**
FURTO	58,82	47,62	52,63	68,42	61,91	65,00	64,71	52,38	57,90	72,22	61,91	**66,67**
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	100,00	12,50	**22,22**	0,00	0,00	0,00
INST_FINANCEIRA	54,55	33,33	**41,38**	29,41	27,78	28,57	37,50	33,33	35,29	50,00	33,33	40,00
LOC_DE_ENSINO	85,71	75,00	**80,00**	77,78	58,33	66,67	60,00	62,50	61,22	72,22	54,17	61,91
LOC_DE_LAZER	67,86	76,00	71,70	67,86	76,00	71,70	66,07	74,00	69,81	71,43	70,00	70,71
MDI	61,22	71,43	65,93	62,22	66,67	64,37	60,87	66,67	63,64	60,42	69,05	64,44
OUTROS_MEIOS_TRANSP	80,77	72,41	76,36	91,30	72,41	**80,77**	79,17	65,52	71,70	83,33	68,97	75,47
TERRENO	56,25	69,23	62,07	81,82	69,23	75,00	81,82	69,23	75,00	90,00	69,23	**78,26**
TRANSP_ALT	90,91	86,21	**88,50**	86,44	87,93	87,18	84,75	86,21	85,47	89,09	84,48	86,73

Fonte: Elaborado pela autora.

Tabela 14 – Parte 1: resultados dos modelos BLSTM-CRF+DL para os rótulos majoritários

Rótulos majoritários	Modelos BLSTM-CRF+DL – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADF	73,79	80,00	76,77	69,71	64,21	66,85	0,00	0,00	0,00	70,94	87,37	78,30	0,00	0,00	0,00
AMEAÇA	43,08	36,60	39,58	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
BICICLETA	0,00	0,00	0,00	76,14	68,37	72,04	74,47	71,43	72,92	70,53	68,37	69,43	58,88	64,29	61,46
CARRO	65,43	52,12	58,02	85,00	79,24	**82,02**	72,02	66,53	69,16	76,11	79,66	77,85	73,46	65,68	69,35
COMPARSA	71,81	53,56	61,36	76,27	63,18	69,11	79,31	67,47	72,91	75,11	70,99	**72,99**	78,28	61,89	69,13
EMPRESA	52,52	73,00	61,09	51,33	77,00	61,60	61,79	76,00	68,16	60,83	73,00	66,36	63,39	71,00	66,98
ESTAB_COMERCIAL	55,56	63,91	59,44	51,18	65,41	57,43	45,71	60,15	51,95	52,53	62,41	57,05	53,62	55,64	54,61
ESTAB_PÚBLICO	90,10	80,53	**85,05**	80,41	69,03	74,29	83,00	73,45	77,93	77,60	85,84	81,51	84,11	79,65	81,82
LOC_DE_EMB_DESEMB	68,27	84,52	75,53	79,27	77,38	78,31	81,01	76,19	78,53	74,73	80,95	77,71	74,36	69,05	71,61
LOC_DE_RES	70,77	74,19	72,44	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
MOTO	79,50	88,73	83,86	85,02	91,90	88,33	83,78	89,09	86,35	81,76	91,55	86,38	82,84	88,38	85,52
OBJ_SUB	78,67	82,71	80,64	83,62	85,69	**84,64**	81,52	84,59	83,03	81,95	87,06	84,43	79,28	85,80	82,41
OUTROS_LOC	53,79	60,47	56,93	50,79	75,19	60,63	62,26	76,74	68,75	59,43	80,62	68,42	53,25	69,77	60,40
ROUBO	94,77	96,08	95,42	89,13	94,47	91,72	87,21	94,24	90,59	92,19	95,16	93,65	89,93	92,63	91,26
SIMULACRO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRANSEUNTE	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
USO_FORÇA_FÍSICA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
VEIC_GRANDE	62,65	59,77	61,18	61,32	74,71	67,36	69,79	77,01	73,22	65,39	78,16	71,20	67,68	77,01	72,04
VIA	83,61	86,44	85,00	90,61	84,46	87,43	93,55	90,11	**91,80**	88,25	91,24	89,72	88,53	85,03	86,74

Fonte: Elaborado pela autora.

Tabela 15 – Parte 2: resultados dos modelos BLSTM-CRF+DL para os rótulos majoritários

Rótulos majoritários	Modelos BLSTM-CRF+DL – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADF	83,33	76,32	79,67	85,56	81,05	**83,24**	80,10	82,63	81,35	83,78	81,58	82,67
AMEAÇA	74,51	49,67	**59,61**	73,53	49,02	58,82	67,24	50,98	57,99	65,52	49,67	56,51
BICICLETA	80,00	73,47	**76,60**	83,33	66,33	73,86	70,83	69,39	70,10	62,61	73,47	67,61
CARRO	80,68	60,17	68,93	80,93	66,53	73,02	70,05	64,41	67,11	71,50	64,83	68,00
COMPARSA	76,97	60,26	67,60	70,76	64,81	67,65	70,74	65,15	67,83	69,42	63,52	66,34
EMPRESA	70,64	77,00	**73,68**	69,44	75,00	72,12	65,87	83,00	73,45	50,63	81,00	62,31
ESTAB_COMERCIAL	65,56	74,44	**69,72**	59,09	68,42	63,42	67,14	70,68	68,86	64,23	66,17	65,19
ESTAB_PÚBLICO	78,91	89,38	83,82	76,15	87,61	81,48	75,19	88,50	81,30	65,16	89,38	75,37
LOC_DE_EMB_DESEMB	85,71	85,71	**85,71**	81,11	86,91	83,91	70,59	85,71	77,42	71,29	85,71	77,84
LOC_DE_RES	78,33	75,81	77,05	72,86	82,26	**77,27**	70,15	75,81	72,87	73,13	79,03	75,97
MOTO	90,11	89,79	89,95	89,55	90,49	**90,02**	85,15	90,85	87,91	85,03	88,03	86,51
OBJ_SUB	78,20	85,80	81,83	79,10	82,51	80,77	77,11	85,61	81,14	78,74	81,33	80,02
OUTROS_LOC	81,82	62,79	**71,05**	69,23	62,79	65,85	62,59	71,32	66,67	68,00	65,89	66,93
ROUBO	94,42	97,47	**95,92**	93,39	97,70	95,50	93,56	97,01	95,25	92,27	96,31	94,25
SIMULACRO	89,68	67,67	**77,13**	85,29	69,46	76,57	78,95	71,86	75,24	0,00	0,00	0,00
TRANSEUNTE	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
USO_FORÇA_FÍSICA	0,00	0,00	0,00	3,23	1,72	**2,25**	0,00	0,00	0,00	0,00	0,00	0,00
VEIC_GRANDE	69,07	77,01	72,83	72,17	80,46	76,09	79,07	78,16	**78,61**	55,74	78,16	65,07
VIA	91,54	85,59	88,47	87,50	88,98	88,24	86,25	90,40	88,28	84,14	88,42	86,23

Fonte: Elaborado pela autora.

Tabela 16 – Parte 1: resultados dos modelos BLSTM-CRF+DL para os rótulos minoritários

Rótulos minoritários	**Modelos BLSTM-CRF+DL – Parte 1**														
	WE do domínio			***FastText***			**GloVe**			***Wang2Vec***			***Word2Vec***		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
AB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
AMB_VIRTUAL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_ESPORTIVO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_MÉDICO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_RELIGIOSO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAC	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EVENTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
FAVELA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
FURTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LOC_DE_ENSINO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LOC_DE_LAZER	44,44	64,00	52,46	44,68	42,00	43,30	42,35	72,00	53,33	54,84	68,00	60,71	51,61	32,00	39,51
MDI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
OUTROS_MEIOS_TRANSP	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TERRENO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRANSP_ALT	30,71	74,14	43,43	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Fonte: Elaborado pela autora.

Tabela 17 – Parte 2: resultados dos modelos BLSTM-CRF+DL para os rótulos minoritários

Rótulos minoritários	**Modelos BLSTM-CRF+DL – Parte 2**											
	Concat*FastText*			**ConcatGloVe**			**Concat*Wang2Vec***			**Concat*Word2Vec***		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
AB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
AMB_VIRTUAL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_ESPORTIVO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_MÉDICO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_RELIGIOSO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAC	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EVENTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
FAVELA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
FURTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	27,59	44,44	**34,04**	26,32	27,78	27,03	0,00	0,00	0,00	0,00	0,00	0,00
LOC_DE_ENSINO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LOC_DE_LAZER	71,19	84,00	**77,06**	49,25	66,00	56,41	46,75	72,00	56,69	44,44	72,00	54,96
MDI	0,00	0,00	0,00	0,00	0,00	0,00	24,76	61,91	**35,37**	0,00	0,00	0,00
OUTROS_MEIOS_TRANSP	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TERRENO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRANSP_ALT	49,43	74,14	59,31	74,14	74,14	**74,14**	0,00	0,00	0,00	0,00	0,00	0,00

Fonte: Elaborado pela autora.

APÊNDICE B – VARIAÇÕES DOS MODELOS BLSTM-CRF DO EXPERIMENTO #2 POR RÓTULO

As tabelas abaixo exibem os resultados dos modelos BLSTM-CRF com *fine-tuning* do Cenário de Experimentação #2 detalhados por rótulo, e possuem as mesmas especificações do Apêndice A. Cada variação do modelo com uma função de perda possui quatro tabelas com seus resultados: Tabelas 18, 19, 20 e 21 para o modelo BLSTM-CRF+CE-CCE; Tabelas 22, 23, 24 e 25 para o modelo BLSTM-CRF; e, Tabelas 26, 27, 28 e 29 para o modelo BLSTM-CRF+DL.

Tabela 18 – Parte 1: resultados dos modelos BLSTM-CRF+CB-CCE com *fine tuning* para os rótulos majoritários

Rótulos majoritários	**Modelos BLSTM-CRF+CB-CCE – Parte 1**														
	WE do domínio			***FastText***			**GloVe**			***Wang2Vec***			***Word2Vec***		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
ADF	69,43	83,68	75,90	55,25	94,21	69,65	73,86	93,68	82,60	76,58	89,47	82,52	73,93	91,05	81,60
AMEAÇA	39,78	69,94	50,71	43,17	64,05	51,58	53,68	66,67	59,48	50,50	66,01	57,22	46,79	66,67	54,99
BICICLETA	68,27	72,45	70,30	78,13	76,53	77,32	76,34	72,45	74,35	78,95	76,53	77,72	82,56	72,45	77,17
CARRO	65,08	69,49	67,21	71,04	77,97	74,34	73,90	77,97	75,88	68,07	68,64	68,35	69,15	78,81	73,66
COMPARSA	54,07	73,05	62,14	52,26	76,57	62,12	64,81	74,76	**69,43**	61,79	72,45	66,69	55,47	76,57	64,34
EMPRESA	40,00	82,00	53,77	71,43	75,00	**73,17**	58,74	84,00	69,14	62,12	82,00	70,69	60,43	84,00	70,29
ESTAB_COMERCIAL	56,13	65,41	60,42	58,75	70,68	64,16	64,29	74,44	68,99	65,33	73,68	69,26	62,33	68,42	65,23
ESTAB_PÚBLICO	83,33	88,50	85,84	70,35	90,27	79,07	73,76	92,04	81,89	69,86	90,27	78,76	80,80	89,38	84,87
LOC_DE_EMB_DESEMB	71,85	88,10	79,14	86,91	86,91	86,91	84,44	90,48	87,36	86,21	89,29	**87,72**	79,57	88,10	83,62
LOC_DE_RES	66,67	90,32	76,71	82,35	90,32	**86,15**	79,69	82,26	80,95	79,71	88,71	83,97	78,08	91,94	84,44
MOTO	68,38	93,66	79,05	78,72	95,07	86,12	85,20	91,20	88,10	80,50	91,55	85,67	85,16	92,96	**88,89**
OBJ_SUB	74,86	85,49	79,82	74,63	87,69	80,64	76,17	87,73	81,54	75,57	88,67	81,60	76,13	88,20	81,72
OUTROS_LOC	61,11	76,74	68,04	65,09	85,27	73,83	72,73	80,62	**76,47**	66,67	82,17	73,61	66,03	79,85	72,28
ROUBO	92,91	96,54	94,69	92,56	97,47	94,95	93,58	97,47	95,49	90,75	97,24	93,88	91,15	97,24	94,09
SIMULACRO	55,84	77,25	64,82	80,42	68,86	74,19	82,12	74,25	77,99	79,04	79,04	79,04	69,40	76,05	72,57
TRANSEUNTE	74,67	86,15	80,00	93,65	90,77	**92,19**	86,36	87,69	87,02	86,36	87,69	87,02	89,06	87,69	88,37
USO_FORÇA_FÍSICA	44,19	65,52	52,78	27,27	72,41	39,62	48,00	62,07	54,14	55,70	75,86	**64,23**	48,72	65,52	55,88
VEIC_GRANDE	57,80	72,41	64,29	83,15	85,06	**84,09**	74,75	85,06	79,57	74,00	85,06	79,14	70,30	81,61	75,53
VIA	76,17	92,09	83,38	84,20	91,81	87,84	88,62	92,37	**90,46**	85,79	92,09	88,83	87,09	89,55	88,30

Fonte: Elaborado pela autora.

Tabela 19 – Parte 2: resultados dos modelos BLSTM-CRF+CB-CCE com *fine tuning* para os rótulos majoritários

Rótulos majoritários	**Modelo BLSTM-CRF+CB-CCE**											
	Concat*FastText*			**ConcatGloVe**			**Concat*Wang2Vec***			**Concat*Word2Vec***		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
ADF	64,05	90,00	74,84	82,04	88,95	**85,35**	77,21	87,37	81,98	67,60	88,95	76,82
AMEAÇA	48,15	67,97	56,37	61,39	63,40	**62,38**	46,70	64,71	54,25	37,41	67,97	48,26
BICICLETA	80,68	72,45	76,34	82,76	73,47	**77,84**	80,22	74,49	77,25	79,55	71,43	75,27
CARRO	70,44	81,78	75,69	59,39	81,78	68,81	78,11	77,12	**77,61**	67,69	74,58	70,97
COMPARSA	58,80	72,02	64,74	65,53	69,19	67,31	57,64	73,48	64,60	59,03	71,85	64,81
EMPRESA	61,77	84,00	71,19	55,10	81,00	65,59	53,29	81,00	64,29	52,56	82,00	64,06
ESTAB_COMERCIAL	61,84	70,68	65,97	68,75	66,17	67,43	72,73	66,17	**69,29**	63,87	74,44	68,75
ESTAB_PÚBLICO	83,74	91,15	87,29	77,86	90,27	83,61	87,93	90,27	**89,08**	80,80	89,38	84,87
LOC_DE_EMB_DESEMB	82,96	86,91	84,88	76,29	88,10	81,77	80,68	84,52	82,56	84,88	86,91	85,88
LOC_DE_RES	67,05	95,16	78,67	74,36	93,55	82,86	77,47	88,71	82,71	69,74	85,48	76,81
MOTO	80,85	93,66	86,79	77,45	91,90	84,06	79,64	92,25	85,48	80,06	90,49	84,96
OBJ_SUB	76,75	85,18	80,74	80,11	83,41	**81,73**	77,83	84,55	81,05	76,96	84,63	80,61
OUTROS_LOC	70,99	72,09	71,54	74,02	72,87	73,44	74,60	72,87	73,73	73,11	67,44	70,16
ROUBO	93,76	97,01	95,36	94,18	97,01	**95,57**	93,96	96,77	95,35	92,75	97,24	94,94
SIMULACRO	85,32	73,05	78,71	75,00	80,84	77,81	87,18	81,44	**84,21**	86,21	74,85	80,13
TRANSEUNTE	90,32	86,15	88,19	90,16	84,62	87,30	87,50	86,15	86,82	90,16	84,62	87,30
USO_FORÇA_FÍSICA	42,86	67,24	52,35	49,37	67,24	56,93	35,43	77,59	48,65	33,33	70,69	45,30
VEIC_GRANDE	63,89	79,31	70,77	67,31	80,46	73,30	71,13	79,31	75,00	63,03	86,21	72,82
VIA	83,89	92,66	88,05	81,05	91,81	86,09	83,80	93,50	88,39	81,06	90,68	85,60

Fonte: Elaborado pela autora.

Tabela 20 – Parte 1: resultados dos modelos BLSTM-CRF+CB-CCE com *fine tuning* para os rótulos minoritários

Rótulos minoritários	Modelos BLSTM-CRF+CB-CCE – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
AB	59,68	68,52	63,79	62,86	81,48	70,97	74,00	68,52	71,15	68,85	77,78	73,04	71,70	70,37	71,03
AMB_VIRTUAL	9,52	9,09	9,30	50,00	27,27	35,29	41,18	31,82	35,90	36,36	18,18	24,24	45,46	22,73	30,30
ESTAB_ESPORTIVO	40,00	40,00	40,00	33,33	40,00	36,36	40,00	40,00	40,00	40,00	40,00	40,00	33,33	40,00	36,36
ESTAB_MÉDICO	41,67	41,67	41,67	54,55	50,00	52,17	57,14	66,67	61,54	33,33	58,33	42,42	60,00	50,00	54,55
ESTAB_RELIGIOSO	66,67	66,67	66,67	100,00	66,67	80,00	66,67	66,67	66,67	100,00	100,00	**100,00**	50,00	66,67	57,14
ESTAC	55,56	62,50	58,82	100,00	62,50	76,92	100,00	62,50	76,92	80,00	50,00	61,54	83,33	62,50	71,43
EVENTO	10,00	33,33	15,39	33,33	33,33	33,33	33,33	33,33	33,33	25,00	33,33	28,57	33,33	33,33	33,33
FAVELA	50,00	50,00	50,00	77,78	50,00	60,87	100,00	42,86	60,00	85,71	42,86	57,14	100,00	57,14	**72,73**
FURTO	55,56	71,43	62,50	81,25	61,91	70,27	77,78	66,67	**71,80**	77,78	66,67	**71,80**	55,56	71,43	62,50
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	100,00	12,50	**22,22**	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	28,57	55,56	37,74	50,00	61,11	**55,00**	33,33	55,56	41,67	40,00	55,56	46,51	25,53	66,67	36,92
LOC_DE_ENSINO	73,68	58,33	65,12	61,29	79,17	69,09	61,29	79,17	69,09	86,36	79,17	**82,61**	53,13	70,83	60,71
LOC_DE_LAZER	39,54	68,00	50,00	57,14	72,00	63,72	63,08	82,00	**71,30**	49,43	86,00	62,77	59,72	86,00	70,49
MDI	45,71	76,19	57,14	46,77	69,05	55,77	60,42	69,05	64,44	46,77	69,05	55,77	39,47	71,43	50,85
OUTROS_MEIOS_TRANSP	64,71	75,86	69,84	77,42	82,76	**80,00**	84,00	72,41	77,78	75,00	72,41	73,68	81,48	75,86	78,57
TERRENO	60,00	69,23	64,29	81,82	69,23	**75,00**	75,00	69,23	72,00	81,82	69,23	**75,00**	75,00	69,23	72,00
TRANSP_ALT	66,67	96,55	78,87	68,83	91,38	78,52	80,00	89,66	84,55	80,30	91,38	85,48	89,47	87,93	88,70

Fonte: Elaborado pela autora.

Tabela 21 – Parte 2: resultados dos modelos BLSTM-CRF+CB-CCE com *fine tuning* para os rótulos minoritários

Rótulos minoritários	Modelos BLSTM-CRF+CB-CCE – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
AB	82,00	75,93	**78,85**	65,46	66,67	66,06	70,91	72,22	71,56	73,08	70,37	71,70
AMB_VIRTUAL	75,00	27,27	**40,00**	31,82	31,82	31,82	71,43	22,73	34,48	54,55	27,27	36,36
ESTAB_ESPORTIVO	23,08	60,00	33,33	33,33	60,00	42,86	36,36	80,00	**50,00**	25,00	40,00	30,77
ESTAB_MÉDICO	38,10	66,67	48,49	77,78	58,33	**66,67**	85,71	50,00	63,16	61,54	66,67	64,00
ESTAB_RELIGIOSO	40,00	66,67	50,00	50,00	66,67	57,14	66,67	66,67	66,67	100,00	66,67	80,00
ESTAC	83,33	62,50	71,43	100,00	75,00	**85,71**	85,71	75,00	80,00	85,71	75,00	80,00
EVENTO	33,33	33,33	33,33	100,00	33,33	**50,00**	16,67	33,33	22,22	12,50	33,33	18,18
FAVELA	100,00	50,00	66,67	70,00	50,00	58,33	81,82	64,29	72,00	100,00	57,14	**72,73**
FURTO	71,43	71,43	71,43	77,78	66,67	**71,80**	70,00	66,67	68,29	73,68	66,67	70,00
HOSPEDARIA	50,00	12,50	20,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	30,30	55,56	39,22	28,13	50,00	36,00	40,00	66,67	50,00	33,33	50,00	40,00
LOC_DE_ENSINO	66,67	83,33	74,07	87,50	58,33	70,00	60,00	75,00	66,67	59,26	66,67	62,75
LOC_DE_LAZER	70,00	70,00	70,00	55,07	76,00	63,87	68,52	74,00	71,15	63,79	74,00	68,52
MDI	61,54	76,19	68,09	67,44	69,05	**68,24**	51,61	76,19	61,54	53,23	78,57	63,46
OUTROS_MEIOS_TRANSP	60,00	82,76	69,57	64,87	82,76	72,73	84,62	75,86	**80,00**	70,00	72,41	71,19
TERRENO	75,00	69,23	72,00	69,23	69,23	69,23	69,23	69,23	69,23	69,23	69,23	69,23
TRANSP_ALT	75,00	87,93	80,95	89,83	91,38	**90,60**	80,00	89,66	84,55	80,95	87,93	84,30

Fonte: Elaborado pela autora.

Tabela 22 – Parte 1: resultados dos modelos BLSTM-CRF com *fine tuning* para os rótulos majoritários

Rótulos majoritários	Modelo BLSTM-CRF – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADF	74,76	82,63	78,50	83,92	87,90	85,86	84,65	90,00	87,25	86,67	88,95	**87,79**	81,59	86,32	83,89
AMEAÇA	72,88	56,21	63,47	85,42	53,60	65,86	85,71	58,82	**69,77**	76,58	55,56	64,39	82,35	54,90	65,88
BICICLETA	83,33	66,33	73,86	94,37	68,37	79,29	89,74	71,43	79,55	91,46	76,53	**83,33**	93,15	69,39	79,53
CARRO	81,22	67,80	73,90	84,98	76,70	**80,62**	81,77	66,53	73,36	81,61	77,12	79,30	84,50	71,61	77,52
COMPARSA	52,93	75,19	62,13	76,13	66,27	70,86	72,20	68,67	70,39	75,79	67,73	**71,53**	73,20	65,41	69,08
EMPRESA	73,40	69,00	71,13	67,26	76,00	71,36	63,93	78,00	70,27	62,41	83,00	71,25	64,34	83,00	72,49
ESTAB_COMERCIAL	73,08	57,14	64,14	67,86	71,43	69,60	70,59	72,18	**71,38**	65,77	73,68	69,50	69,17	69,17	69,17
ESTAB_PÚBLICO	87,50	86,73	87,11	94,50	91,15	**92,79**	89,19	87,61	88,39	85,00	90,27	87,55	87,07	89,38	88,21
LOC_DE_EMB_DESEMB	80,23	82,14	81,18	91,36	88,10	**89,70**	86,59	84,52	85,54	84,44	90,48	87,36	85,71	78,57	81,99
LOC_DE_RES	82,09	88,71	85,27	82,09	88,71	85,27	86,89	85,48	86,18	88,71	88,71	**88,71**	85,25	83,87	84,55
MOTO	82,26	89,79	85,86	88,89	90,14	89,51	89,58	90,85	**90,21**	86,24	90,49	88,32	89,32	88,38	88,85
OBJ_SUB	75,71	84,47	79,85	81,85	85,61	83,69	77,71	86,67	81,94	84,35	83,49	**83,92**	79,08	85,22	82,03
OUTROS_LOC	62,90	60,47	61,66	69,49	63,57	66,40	76,80	74,42	**75,59**	69,78	75,19	72,39	62,50	73,64	67,62
ROUBO	90,83	95,85	93,27	96,10	96,54	**96,32**	93,99	97,24	95,58	94,62	97,24	95,91	94,62	97,24	95,91
SIMULACRO	86,71	74,25	80,00	82,39	70,06	75,73	87,22	69,46	77,33	88,11	75,45	**81,29**	84,93	74,25	79,23
TRANSEUNTE	83,33	76,92	80,00	84,85	86,15	85,50	90,32	86,15	88,19	90,16	84,62	87,30	89,06	87,69	88,37
USO_FORÇA_FÍSICA	60,98	43,10	50,51	76,32	50,00	60,42	65,39	58,62	61,82	75,00	62,07	**67,93**	76,19	55,17	64,00
VEIC_GRANDE	67,37	73,56	70,33	80,49	75,86	78,11	80,90	82,76	**81,82**	80,00	73,56	76,65	77,38	74,71	76,02
VIA	85,12	87,29	86,19	87,43	88,42	87,92	91,98	90,68	91,32	91,83	92,09	**91,96**	91,59	86,16	88,79

Fonte: Elaborado pela autora.

Tabela 23 – Parte 2: resultados dos modelos BLSTM-CRF com *fine tuning* para os rótulos majoritários

Rótulos majoritários	Modelo BLSTM-CRF – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
ADF	85,79	82,63	84,18	92,90	82,63	87,47	90,17	82,11	85,95	80,73	81,58	81,15
AMEAÇA	53,93	67,32	59,88	73,83	51,63	60,77	56,60	58,82	57,69	62,41	57,52	59,86
BICICLETA	86,59	72,45	78,89	92,11	71,43	80,46	91,43	65,31	76,19	79,52	67,35	72,93
CARRO	79,41	68,64	73,64	81,95	71,19	76,19	79,69	64,83	71,50	80,21	65,25	71,96
COMPARSA	66,58	68,58	67,57	65,81	70,56	68,10	68,35	64,89	66,58	62,70	70,99	66,59
EMPRESA	73,45	83,00	**77,93**	75,76	75,00	75,38	67,54	77,00	71,96	74,23	72,00	73,10
ESTAB_COMERCIAL	62,76	68,42	65,47	66,91	69,93	68,38	60,81	67,67	64,06	64,23	66,17	65,19
ESTAB_PÚBLICO	85,25	92,04	88,51	83,19	87,61	85,35	78,30	89,38	83,47	81,10	91,15	85,83
LOC_DE_EMB_DESEMB	81,52	89,29	85,23	84,88	86,91	85,88	84,27	89,29	86,71	77,90	88,10	82,68
LOC_DE_RES	76,39	88,71	82,09	80,00	90,32	84,85	87,30	88,71	88,00	74,60	75,81	75,20
MOTO	84,37	91,20	87,65	82,58	90,14	86,20	82,69	90,85	86,58	87,37	90,14	88,74
OBJ_SUB	79,56	81,37	80,46	77,54	82,43	79,91	77,30	81,18	79,19	76,09	82,47	79,15
OUTROS_LOC	65,56	76,74	70,71	72,36	68,99	70,64	65,15	66,67	65,90	64,96	68,99	66,92
ROUBO	92,95	97,24	95,05	95,24	96,77	96,00	94,61	97,01	95,79	93,96	96,77	95,35
SIMULACRO	80,65	74,85	77,64	87,23	73,65	79,87	86,21	74,85	80,13	81,46	73,65	77,36
TRANSEUNTE	91,94	87,69	**89,76**	85,08	87,69	86,36	82,61	87,69	85,08	86,36	87,69	87,02
USO_FORÇA_FÍSICA	56,90	56,90	56,90	72,00	62,07	66,67	60,66	63,79	62,19	62,96	58,62	60,71
VEIC_GRANDE	79,78	81,61	80,68	70,53	77,01	73,63	79,76	77,01	78,36	72,63	79,31	75,82
VIA	88,73	88,98	88,86	90,17	88,14	89,14	90,31	89,55	89,93	85,48	89,83	87,60

Fonte: Elaborado pela autora.

Tabela 24 – Parte 1: resultados dos modelos BLSTM-CRF com *fine tuning* para os rótulos minoritários

Rótulos minoritários	Modelo BLSTM-CRF – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
AB	73,47	66,67	69,90	84,09	68,52	75,51	74,55	75,93	75,23	90,91	55,56	68,97	80,00	66,67	72,73
AMB_VIRTUAL	0,00	0,00	0,00	33,33	9,09	14,29	55,56	22,73	**32,26**	55,56	22,73	**32,26**	40,00	9,09	14,82
ESTAB_ESPORTIVO	23,08	60,00	33,33	66,67	40,00	50,00	50,00	20,00	28,57	50,00	40,00	44,44	100,00	40,00	57,14
ESTAB_MÉDICO	57,14	33,33	42,11	66,67	33,33	44,44	72,73	66,67	**69,57**	54,55	50,00	52,17	50,00	33,33	40,00
ESTAB_RELIGIOSO	100,00	33,33	50,00	100,00	66,67	**80,00**	100,00	66,67	**80,00**	100,00	66,67	**80,00**	66,67	66,67	66,67
ESTAC	50,00	25,00	33,33	83,33	62,50	71,43	100,00	12,50	22,22	100,00	37,50	54,55	75,00	37,50	50,00
EVENTO	0,00	0,00	0,00	50,00	33,33	40,00	50,00	33,33	40,00	25,00	33,33	28,57	100,00	33,33	**50,00**
FAVELA	87,50	50,00	63,64	100,00	35,71	52,63	85,71	42,86	57,14	100,00	50,00	66,67	100,00	50,00	66,67
FURTO	68,75	52,38	59,46	76,92	47,62	58,82	68,75	52,38	59,46	86,67	61,91	**72,22**	76,47	61,91	68,42
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	43,75	38,89	41,18	60,00	50,00	54,55	50,00	27,78	35,71	55,56	55,56	**55,56**	37,50	33,33	35,29
LOC_DE_ENSINO	59,09	54,17	56,52	85,00	70,83	77,27	72,73	66,67	69,57	79,17	79,17	**79,17**	65,00	54,17	59,09
LOC_DE_LAZER	68,75	66,00	67,35	67,86	76,00	71,70	75,00	72,00	73,47	69,64	78,00	**73,59**	69,64	78,00	**73,59**
MDI	69,05	69,05	69,05	68,57	57,14	62,34	60,53	54,76	57,50	67,44	69,05	68,24	60,47	61,91	61,18
OUTROS_MEIOS_TRANSP	64,52	68,97	66,67	88,46	79,31	83,64	100,00	68,97	81,63	95,65	75,86	**84,62**	82,61	65,52	73,08
TERRENO	72,73	61,54	66,67	75,00	69,23	72,00	75,00	69,23	72,00	60,00	69,23	64,29	66,67	61,54	64,00
TRANSP_ALT	78,33	81,03	79,66	80,33	84,48	82,35	89,09	84,48	86,73	80,95	87,93	84,30	80,95	87,93	84,30

Fonte: Elaborado pela autora.

Tabela 25 – Parte 2: resultados dos modelos BLSTM-CRF com *fine tuning* para os rótulos minoritários

Rótulos minoritários	Modelo BLSTM-CRF – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
AB	62,50	64,82	63,64	72,34	62,96	67,33	94,60	64,82	**76,92**	80,44	68,52	74,00
AMB_VIRTUAL	44,44	18,18	25,81	50,00	9,09	15,39	42,86	13,64	20,69	100,00	13,64	24,00
ESTAB_ESPORTIVO	100,00	40,00	57,14	75,00	60,00	**66,67**	100,00	40,00	57,14	60,00	60,00	60,00
ESTAB_MÉDICO	71,43	41,67	52,63	54,55	50,00	52,17	66,67	50,00	57,14	71,43	41,67	52,63
ESTAB_RELIGIOSO	100,00	66,67	**80,00**	66,67	66,67	66,67	50,00	66,67	57,14	66,67	66,67	66,67
ESTAC	100,00	62,50	76,92	83,33	62,50	71,43	85,71	75,00	**80,00**	85,71	75,00	**80,00**
EVENTO	33,33	33,33	33,33	50,00	33,33	40,00	100,00	33,33	**50,00**	16,67	33,33	22,22
FAVELA	58,33	50,00	53,85	100,00	57,14	**72,73**	100,00	35,71	52,63	87,50	50,00	63,64
FURTO	62,50	71,43	66,67	73,68	66,67	70,00	65,00	61,91	63,42	68,42	61,91	65,00
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	47,37	50,00	48,65	44,44	44,44	44,44	41,18	38,89	40,00	53,85	38,89	45,16
LOC_DE_ENSINO	80,95	70,83	75,56	72,22	54,17	61,91	65,22	62,50	63,83	82,35	58,33	68,29
LOC_DE_LAZER	63,93	78,00	70,27	67,24	78,00	72,22	63,64	70,00	66,67	74,47	70,00	72,17
MDI	55,36	73,81	63,27	70,73	69,05	**69,88**	56,52	61,91	59,09	58,33	66,67	62,22
OUTROS_MEIOS_TRANSP	67,86	65,52	66,67	95,46	72,41	82,35	76,92	68,97	72,73	95,00	65,52	77,55
TERRENO	75,00	69,23	72,00	75,00	69,23	72,00	81,82	69,23	**75,00**	75,00	69,23	72,00
TRANSP_ALT	91,23	89,66	**90,44**	89,47	87,93	88,70	89,09	84,48	86,73	94,34	86,21	90,09

Fonte: Elaborado pela autora.

Tabela 26 – Parte 1: resultados dos modelos BLSTM-CRF+DL com *fine tuning* para os rótulos majoritários

Rótulos majoritários	Modelo BLSTM-CRF+DL – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADF	73,79	80,00	76,77	82,38	83,68	83,03	82,01	81,58	81,79	77,73	86,32	81,80	84,62	81,05	82,80
AMEAÇA	43,08	36,60	39,58	58,41	43,14	49,62	68,22	57,52	62,41	0,00	0,00	0,00	57,36	48,37	52,48
BICICLETA	0,00	0,00	0,00	79,07	69,39	73,91	80,90	73,47	77,01	85,23	76,53	**80,65**	83,33	66,33	73,86
CARRO	65,43	52,12	58,02	81,59	82,63	82,11	82,22	62,71	71,15	75,77	72,88	74,30	84,62	83,90	**84,26**
COMPARSA	71,81	53,56	61,36	70,97	71,76	71,36	75,42	65,84	70,30	69,66	73,31	**71,44**	74,32	67,81	70,92
EMPRESA	52,52	73,00	61,09	69,09	76,00	72,38	64,71	77,00	70,32	60,77	79,00	68,70	66,38	77,00	71,30
ESTAB_COMERCIAL	55,56	63,91	59,44	67,42	66,92	67,17	58,33	73,68	65,12	66,90	72,93	69,78	68,35	71,43	**69,85**
ESTAB_PÚBLICO	90,10	80,53	85,05	87,72	88,50	88,11	81,90	84,07	82,97	78,13	88,50	82,99	84,87	89,38	87,07
LOC_DE_EMB_DESEMB	68,27	84,52	75,53	80,46	83,33	81,87	81,18	82,14	81,66	85,06	88,10	86,55	72,45	84,52	78,02
LOC_DE_RES	70,77	74,19	72,44	89,47	82,26	85,71	55,68	79,03	65,33	83,33	88,71	**85,94**	85,25	83,87	84,55
MOTO	79,50	88,73	83,86	85,20	91,20	88,10	87,97	90,14	89,04	88,93	87,68	88,30	87,08	90,14	88,58
OBJ_SUB	78,67	82,71	80,64	78,57	88,55	83,26	80,06	84,86	82,39	80,04	86,94	**83,35**	79,19	86,98	82,90
OUTROS_LOC	53,79	60,47	56,93	75,00	74,42	**74,71**	67,38	73,64	70,37	74,05	75,19	74,62	56,77	68,22	61,97
ROUBO	94,77	96,08	95,42	92,49	96,54	94,48	94,58	96,54	95,55	92,04	95,85	93,91	93,58	97,47	95,49
SIMULACRO	0,00	0,00	0,00	89,71	73,05	**80,53**	84,03	72,46	77,81	71,08	70,66	70,87	64,29	70,06	67,05
TRANSEUNTE	0,00	0,00	0,00	77,14	83,08	80,00	78,08	87,69	82,61	68,42	80,00	73,76	84,38	83,08	**83,72**
USO_FORÇA_FÍSICA	0,00	0,00	0,00	0,00	0,00	0,00	44,78	51,72	**48,00**	0,00	0,00	0,00	0,00	0,00	0,00
VEIC_GRANDE	62,65	59,77	61,18	74,49	83,91	78,92	79,31	79,31	79,31	78,16	78,16	78,16	65,79	86,21	74,63
VIA	83,61	86,44	85,00	89,39	90,40	89,89	89,49	93,79	91,59	91,81	91,81	**91,81**	89,27	89,27	89,27

Fonte: Elaborado pela autora.

Tabela 27 – Parte 2: resultados dos modelos BLSTM-CRF+DL com *fine tuning* para os rótulos majoritários

Rótulos majoritários	Model BLSTM-CRF+DL – Parte 2											
	Concat*FastText*			ConcatGloVe			Concat*Wang2Vec*			Concat*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADF	83,33	78,95	81,08	80,30	83,68	81,96	79,25	88,42	**83,58**	82,05	84,21	83,12
AMEAÇA	77,48	56,21	**65,15**	77,66	47,71	59,11	76,47	50,98	61,18	75,00	50,98	60,70
BICICLETA	86,05	75,51	80,44	79,12	73,47	76,19	83,72	73,47	78,26	67,93	73,47	70,59
CARRO	79,24	61,44	69,21	84,18	63,14	72,16	76,55	73,31	74,89	81,82	57,20	67,33
COMPARSA	70,03	66,78	68,37	71,43	68,24	69,80	69,20	67,12	68,15	72,41	64,89	68,45
EMPRESA	69,16	74,00	71,50	69,64	78,00	**73,59**	66,13	82,00	73,21	63,08	82,00	71,30
ESTAB_COMERCIAL	61,15	72,18	66,21	60,67	68,42	64,31	68,66	69,17	68,91	66,19	69,17	67,65
ESTAB_PÚBLICO	71,13	89,38	79,22	80,00	88,50	84,03	91,59	86,73	**89,09**	87,18	90,27	88,70
LOC_DE_EMB_DESEMB	73,53	89,29	80,65	84,27	89,29	**86,71**	71,57	86,91	78,50	82,35	83,33	82,84
LOC_DE_RES	72,86	82,26	77,27	68,06	79,03	73,13	65,06	87,10	74,48	64,10	80,65	71,43
MOTO	88,37	93,66	**90,94**	88,46	89,09	88,77	85,76	89,09	87,39	83,93	90,14	86,93
OBJ_SUB	80,75	83,06	81,89	80,45	80,35	80,40	81,47	83,65	82,55	77,77	83,57	80,57
OUTROS_LOC	73,91	65,89	69,67	67,50	62,79	65,06	70,44	62,79	66,39	73,87	63,57	68,33
ROUBO	95,02	96,77	95,89	95,48	97,24	**96,35**	94,40	97,01	95,68	94,64	97,70	96,15
SIMULACRO	89,47	71,26	79,33	85,61	71,26	77,78	94,26	68,86	79,59	85,51	70,66	77,38
TRANSEUNTE	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
USO_FORÇA_FÍSICA	0,00	0,00	0,00	36,67	56,90	44,60	19,23	43,10	26,60	0,00	0,00	0,00
VEIC_GRANDE	80,68	81,61	**81,14**	72,00	82,76	77,01	81,71	77,01	79,29	72,04	77,01	74,44
VIA	90,52	88,98	89,74	89,43	88,42	88,92	84,28	92,37	88,14	83,29	92,94	87,85

Fonte: Elaborado pela autora.

Tabela 28 – Parte 1: resultados dos modelos BLSTM-CRF+DL com *fine tuning* para os rótulos minoritários

Rótulos minoritários	Modelo BLSTM-CRF+DL – Parte 1														
	WE do domínio			*FastText*			GloVe			*Wang2Vec*			*Word2Vec*		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
AB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
AMB_VIRTUAL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_ESPORTIVO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_MÉDICO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_RELIGIOSO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAC	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EVENTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
FAVELA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
FURTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	0,00	0,00	0,00	5,71	22,22	9,09	24,14	38,89	29,79	0,00	0,00	0,00	8,20	27,78	12,66
LOC_DE_ENSINO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LOC_DE_LAZER	44,44	64,00	52,46	53,85	84,00	65,63	66,04	70,00	67,96	63,38	90,00	74,38	70,18	80,00	**74,77**
MDI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	38,33	54,76	45,10	0,00	0,00	0,00
OUTROS_MEIOS_TRANSP	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TERRENO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRANSP_ALT	30,71	74,14	43,43	83,64	79,31	**81,42**	0,00	0,00	0,00	57,90	75,86	65,67	0,00	0,00	0,00

Fonte: Elaborado pela autora.

Tabela 29 – Parte 2: resultados dos modelos BLSTM-CRF+DL com *fine tuning* para os rótulos minoritários

Rótulos minoritários	**Model BLSTM-CRF+DL – Parte 2**											
	Concat*FastText*			**ConcatGloVe**			**Concat*Wang2Vec***			**Concat*Word2Vec***		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
AB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
AMB_VIRTUAL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_ESPORTIVO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_MÉDICO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAB_RELIGIOSO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ESTAC	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EVENTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
FAVELA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	13,95	42,86	**21,05**
FURTO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	37,04	55,56	**44,44**	24,00	33,33	27,91	30,00	33,33	31,58	19,05	44,44	26,67
LOC_DE_ENSINO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
LOC_DE_LAZER	53,17	84,00	65,12	55,39	72,00	62,61	66,67	76,00	71,03	70,00	70,00	70,00
MDI	0,00	0,00	0,00	0,00	0,00	0,00	57,14	66,67	**61,54**	0,00	0,00	0,00
OUTROS_MEIOS_TRANSP	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TERRENO	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TRANSP_ALT	52,22	81,03	63,51	63,24	74,14	68,25	61,64	77,59	68,70	58,90	74,14	65,65

Fonte: Elaborado pela autora.

APÊNDICE C – VARIAÇÕES DOS MELHORES MODELOS BLSTM-CRF E DO *SPACY* NER POR RÓTULO

As Tabelas 30 e 31 exibem os resultados dos melhores modelos BLSTM-CRF e do *spaCy* NER detalhados por rótulo. As linhas representam um tipo de EN ligada a um modelo (coluna), com os valores para as métricas *Precision*, *Recall* e F_1-*score*. Os resultados destacados em negrito indicam os maiores valores para o F_1 de acordo com o rótulo. Os valores em $0,00\%$ indicam que o modelo não reconheceu a entidade nomeada.

Tabela 30 – Resultados dos modelos BLSTM-CRF e *spaCy* NER para os rótulos majoritários

Rótulos majoritários	**Modelo *spaCy* NER**			**Modelo BSLTM-CRF**											
				CB-CCE *Loss*						**CRF *Loss***					
				FastText			**Concat*Wang2Vec****			***Wang2Vec***			***Wang2Vec****		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
ADF	88,55	77,37	82,58	53,42	90,53	67,19	77,21	87,37	81,98	85,08	81,05	83,02	86,67	88,95	**87,79**
AMEAÇA	79,10	34,64	48,18	23,83	79,74	36,69	46,70	64,71	54,25	80,00	41,83	54,94	76,58	55,56	**64,39**
BICICLETA	89,66	79,59	84,32	57,46	78,57	66,38	80,22	74,49	77,25	88,31	69,39	77,71	91,46	76,53	**83,33**
CARRO	86,94	73,31	79,54	53,63	90,68	67,40	78,11	77,12	77,61	85,17	75,42	**80,00**	81,61	77,12	79,30
COMPARSA	78,79	61,55	69,11	49,86	74,51	59,74	57,64	73,48	64,60	78,19	64,64	70,77	75,79	67,73	**71,53**
EMPRESA	87,18	68,00	76,40	28,53	87,00	42,96	53,29	81,00	64,29	81,32	74,00	**77,49**	62,41	83,00	71,25
ESTAB_COMERCIAL	86,84	74,44	80,16	33,44	78,20	46,85	72,73	66,17	69,29	71,17	59,40	64,75	65,77	73,68	**69,50**
ESTAB_PÚBLICO	90,10	80,53	85,05	27,51	92,04	42,36	87,93	90,27	**89,08**	78,99	83,19	81,03	85,00	90,27	87,55
LOC_DE_EMB_DESEMB	93,10	96,43	94,74	72,83	79,76	76,14	80,68	84,52	82,56	85,14	75,00	79,75	84,44	90,48	**87,36**
LOC_DE_RES	89,29	80,65	84,75	51,38	90,32	65,50	77,47	88,71	82,71	83,64	74,19	78,63	88,71	88,71	**88,71**
MOTO	84,24	92,25	88,07	68,24	96,83	80,06	79,64	92,25	85,48	84,49	90,14	87,22	86,24	90,49	**88,32**
OBJ_SUB	86,87	80,43	83,53	76,42	81,10	78,69	77,83	84,55	81,05	86,93	80,35	83,51	84,35	83,49	**83,92**
OUTROS_LOC	77,78	70,54	73,98	30,95	90,70	46,15	74,60	72,87	**73,73**	66,88	79,85	72,79	69,78	75,19	72,39
ROUBO	95,37	94,93	95,15	79,10	96,77	87,05	93,96	96,77	95,35	94,73	95,16	94,94	94,62	97,24	**95,91**
SIMULACRO	94,55	62,28	75,09	45,86	79,64	58,21	87,18	81,44	**84,21**	83,33	65,87	73,58	88,11	75,45	81,29
TRANSEUNTE	89,83	81,54	85,48	48,33	89,23	62,70	87,50	86,15	86,82	94,12	73,85	82,76	90,16	84,62	**87,30**
USO_FORÇA_FÍSICA	60,00	31,03	40,91	16,93	74,14	27,56	35,43	77,59	48,65	71,88	39,66	51,11	75,00	62,07	**67,93**
VEIC_GRANDE	88,41	70,12	78,21	48,70	86,21	62,24	71,13	79,31	75,00	82,28	74,71	**78,31**	80,00	73,56	76,65
VIA	92,71	89,83	91,25	66,52	87,01	75,40	83,80	93,50	88,39	91,40	87,01	89,15	91,83	92,09	**91,96**

Fonte: Elaborado pela autora.
(*): Este símbolo indica que os modelos tiveram a camada das *word embeddings* pré-treinadas atualizada durante o treino.

Tabela 31 – Resultados dos modelos BLSTM-CRF e *spaCy* NER para os rótulos minoritários

Rótulos minoritários	**Modelo *spaCy* NER**			**Modelo BSLTM-CRF**											
				CB-CCE *Loss*						**CRF *Loss***					
				FastText			**Concat*Wang2Vec****			***Wang2Vec***			***Wang2Vec****		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
AB	100,00	75,93	86,32	42,70	70,37	53,15	70,91	72,22	**71,56**	86,11	57,41	68,89	90,91	55,56	68,97
AMB_VIRTUAL	36,36	18,18	24,24	18,33	50,00	26,83	71,43	22,73	**34,48**	80,00	18,18	29,63	55,56	22,73	32,26
ESTAB_ESPORTIVO	100,00	40,00	57,14	0,00	0,00	0,00	36,36	80,00	**50,00**	0,00	0,00	0,00	50,00	40,00	44,44
ESTAB_MÉDICO	37,50	50,00	42,86	25,00	66,67	36,36	85,71	50,00	**63,16**	75,00	50,00	60,00	54,55	50,00	52,17
ESTAB_PÚBLICO	90,10	80,53	85,05	27,51	92,04	42,36	87,93	90,27	**89,08**	78,99	83,19	81,03	85,00	90,27	87,55
ESTAB_RELIGIOSO	100,00	66,67	80,00	50,00	66,67	57,14	66,67	66,67	66,67	100,00	33,33	50,00	100,00	66,67	**80,00**
ESTAC	100,00	12,50	22,22	55,56	62,50	58,82	85,71	75,00	**80,00**	0,00	0,00	0,00	100,00	37,50	54,55
EVENTO	33,33	33,33	33,33	7,69	33,33	12,50	16,67	33,33	22,22	50,00	33,33	**40,00**	25,00	33,33	28,57
FAVELA	100,00	57,14	72,73	28,57	28,57	28,57	81,82	64,29	**72,00**	83,33	35,71	50,00	100,00	50,00	66,67
FURTO	75,00	42,86	54,55	35,85	90,48	51,35	70,00	66,67	68,29	83,33	23,81	37,04	86,67	61,91	**72,22**
HOSPEDARIA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
INST_FINANCEIRA	57,14	44,44	50,00	26,00	72,22	38,24	40,00	66,67	50,00	50,00	33,33	40,00	55,56	55,56	**55,56**
LOC_DE_ENSINO	100,00	66,67	80,00	53,85	87,50	66,67	60,00	75,00	66,67	62,07	75,00	67,93	79,17	79,17	**79,17**
LOC_DE_LAZER	71,43	70,00	70,71	44,05	74,00	55,22	68,52	74,00	71,15	78,26	72,00	**75,00**	69,64	78,00	73,59
MDI	44,44	66,67	53,33	26,61	78,57	39,76	51,61	76,19	61,54	70,37	45,24	55,07	67,44	69,05	**68,24**
TERRENO	88,89	61,54	72,73	56,25	69,23	62,07	69,23	69,23	**69,23**	83,33	38,46	52,63	60,00	69,23	64,29
TRANSP_ALT	98,04	86,21	91,74	39,26	91,38	54,92	80,00	89,66	84,55	90,39	81,03	**85,46**	80,95	87,93	84,30

Fonte: Elaborado pela autora.
(*): Este símbolo indica que os modelos tiveram a camada das *word embeddings* pré-treinadas atualizada durante o treino.