Estudo de Modelo Deep Learning para Reconhecimento de Entidades Nomeadas na Segurança Pública

Karla Figueiredo
Instituto de Matemática e Estatística – Departamento de Ciência da Computação
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brasil
karlafigueiredo@ime.uerj.br

Yago Tomé
Instituto de Matemática e Estatística – Departamento de Ciência da Computação
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brasil
yagotome@gmail.com

Leonídia Barreto
Ciência da Computação
Instituto de Matemática e Estatística – Departamento de Ciência da Computação
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro – RJ – Brasil
leonidiabarreto@gmail.com

Walkir A.T. Brito
Disque-Denúncia
Rio de Janeiro – RJ – Brasil
walkir.brito@disquedenuncia.org.br

Abstract: *O Reconhecimento de Entidades Nomeadas é uma área relacionada à Extração da Informação, cujas soluções estão submetidas à área de Processamento de Linguagem Natural. As entidades são termos ou conjunto de termos que representam informações importantes para uso em diversas áreas, pois visam identificar em textos: pessoas, locais e tempo, por exemplo. Em bases criminais, indicaria quem, onde e quando um crime pode ter sido cometido. A rotulagem das entidades feita por meio de ferramentas, tais como a Google Cloud Natural Language[1] mostrou-se inadequada para o contexto de denúncias criminais, escritas em português coloquial, contendo diversos erros tanto de ortografia como gramaticais. Assim, este trabalho apresenta uma adaptação de modelo para identificação e reconhecimento de Entidades Nomeadas baseado em Deep Learning, em textos em português (Brasil) em linguagem popular, visando a extração de informações que auxiliem a segurança pública. Os resultados indicam que a metodologia adotada é promissora, alcançando valores de 79,72%, 81,52% e 80,50% nas métricas* $recall_{macro}$*,* $precision_{macro}$ *e* $F1\text{-}score_{macro}$*, respectivamente, além de apresentar novas alternativas para aperfeiçoamentos futuros*.

Keywords - Processamento de Linguagem Natural, Reconhecimento de Entidades Nomeadas, Segurança Pública, Português Coloquial

I. INTRODUÇÃO

O Reconhecimento de Entidades Nomeadas (REN) faz parte da área de Extração de Informação (EI), sendo uma tarefa de detecção e classificação de informações importantes e presentes em um texto, como por *exemplo*: pessoas, locais, organizações, datas, eventos ou qualquer outra informação relevante [1], que precise ser identificada ou extraída de textos. De acordo com Jurafsky [2], a detecção e classificação de entidades nomeadas em um texto são o ponto de partida para a maioria das aplicações de extração de informação.

Durante muito tempo os modelos para REN seguiam abordagens sem uso de técnicas de Aprendizado de Máquinas (ML do inglês Machine Learning), sendo comum os modelos baseados em expressões regulares (*bag-of-words*) [3]. No entanto, modelos desta natureza não levam em consideração a sintaxe da linguagem ou sequer o contexto no qual as palavras estão inseridas. Buscando suprir essa necessidade, os modelos baseados em Redes Neurais Profundas (do inglês, Deep Learning - DL) vêm ganhando popularidade e melhorando o estado-da-arte na tarefa de REN, e de forma geral para a área de Processamento de Linguagem Natural (PLN) [2].

Em 2015, Santos e Guimarães [4] propuseram uma abordagem para o REN para o português brasileiro, utilizando um modelo DL a partir de *corpus* constituídos por textos em linguagem formal da língua portuguesa e textos literários do HAREM [5].

Entretanto, com o imenso uso das mídias sociais, existem muitas informações não estruturadas, com textos em linguagem coloquial e popular, disponíveis para serem exploradas, surgindo a necessidade de modelos capazes de reconhecer entidades também nesse tipo de texto.

Por outro lado, a Segurança, além da Saúde e a Educação, é um dos pilares da estrutura social, e por isso é sempre incluída nos programas de governo. Especificamente, o Estado do Rio de Janeiro é um dos estados do Brasil que mais sofre com a insegurança, sem mencionar o atual momento de redução de receita evidenciado pelo Plano de Recuperação Fiscal, ao qual o Estado está submetido. Assim, apesar dos esforços dada a importância do tema, a informação e os dados dessa área ainda carecem de tratamento, documentação, organização, e finalmente, de desenvolvimento de modelos e métodos que permitam a extração de conhecimento para que seja possível o uso racional dos recursos públicos disponíveis.

Com o objetivo de melhorar as perspectivas de segurança, diversos órgãos relacionados à Segurança Pública recebem um grande número de denúncias por meio do aplicativo celular, em texto livre, efetuados por seus usuários de forma anônima, denunciando crimes de diversas

[1]https://cloud.google.com/naturalanguage/docs/reference/rest

naturezas (como roubos, homicídios e tráfico de drogas). Os textos encontrados nestes relatos apresentam tom coloquial, pois devido ao fato de que as denúncias são postados por meio de um aplicativo celular, traz consigo a forma de expressão comumente utilizada nas mídias sociais, como abreviações, termos populares e diversos tipos de erros ortográficos e gramaticais. Assim, a extração de informações valiosas, tais como: quem, quando e onde um crime pode ter sido praticado, potencializa a capacidade de combate ao crime, permitindo o mapeamento de crimes, identificando pessoas envolvidas, os principais locais, dentre outras entidades que podem ser extraídas de acordo com a necessidade.

Além disso, o processo de identificação das entidades em uma quantidade grande de textos é um trabalho árduo que necessita de automatização, inclusive para dar celeridade à solução de crimes, que dependem muitas vezes do momento oportuno indicado pelas informações que estão nas denúncias.

O uso de ferramentas disponíveis para essa tarefa [6] se mostram inadequadas devido ao específico contexto e estrutura do texto. Isso pôde ser comprovado ao se empregar a ferramenta *Google Cloud* Natural *Language* (GCNL) [7]. Após a verificação e correção manual (feita por dois pesquisadores) das classificações (rotulagem dos termos), a quantidade de correções, sobre as rotulagens feitas pelo GCLN, foi da ordem de 37%. Isso denota que, devido às características específicas do texto e contexto das denúncias, há a necessidade de modelo adequado para reconhecimento de entidades nesse contexto, gerando a motivação do desenvolvimento de um modelo específico para este domínio.

Assim, este trabalho teve por objetivo o desenvolvimento de uma modelagem baseada em DL para o reconhecimento de entidades, as três entidades são: Pessoa, Tempo e Local nomeadas, a partir de texto em linguagem coloquial/popular no português do Brasil. O trabalho viabiliza a extração de informações sobre crimes em textos escritos neste tipo de linguagem em relatos feitos em aplicativos de denúncias criminais.

Como trabalhos correlatos, pode-se citar Pires [9], que usou a ferramenta CoreNLP[2] com o *corpus* HAREM [5] para extrair EN de notícias, alcançando 86,86% em relação à F1-score. Já Fonseca e outros autores [10] utilizaram a ferramenta OpenNLP[3] e *corpus* Amazônia e HAREM com resultados de F1-score de 57,61% para a entidade Pessoa.

Os trabalhos de Silva e Vieira [11] , Araújo [12] e Carnaz e outros [13] foram os trabalhos identificados mais próximos do ponto de vista do contexto criminal e português, mas distinto do ponto de vista do vocabulário e da forma. Os primeiros autores extraíram entidades de bases da Polícia Federal, visando um crime específico (financeiro), utilizando a ferramenta spaCy[4] e base anotada do HAREM. O segundo realizou o REN em Boletins de ocorrência policial visando outros objetivos, pois nesse caso o crime já ocorreu. O terceiro trabalho [13] desenvolveu um framework visando à extração de REN em texto de diversas origens: jornais online, textos de investigações criminais, registros de ocorrências foram utilizados para extração de informações. De forma geral, os textos utilizados nesses trabalhos possuem melhor qualidade de escrita, o que permitiu o uso de ferramentas construídas a partir de textos literários, sem os problemas já mencionados sobre os textos utilizados por este trabalho. No entanto, todos os trabalhos têm por motivação que a extração e representação automática de dados para a investigação criminal policial que vêm sendo valorizada devido à dificuldade em processar manualmente um grande número de dados de criminosos e envolvidos em crimes [13].

Assim, não foi identificado, até a presente data, nenhum outro trabalho publicado, que tenha reconhecido entidades nomeadas a partir de textos escritos em português coloquial, diretamente por denunciantes, contendo muitos erros de ortografia, sintaxe e semântica.

O restante do trabalho inclui uma próxima seção com sucinta fundamentação teórica; a descrição da metodologia desenvolvida está na seção 3; o estudo de caso é apresentado na quarta seção; e, finalmente, na última seção apresenta-se a conclusão e perspectivas de novos trabalhos.

II. FUNDAMENTOS TEÓRICOS

A. *Reconhecimento de Entidades Nomeadas*

O Reconhecimento de Entidades Nomeadas tem como objetivo encontrar elementos de um texto, como palavras ou termos, e classificá-los dentro de um conjunto de categorias pré-definidas, tais como nome de pessoas ou organizações, tempo e lugares, por exemplo. O REN é também um ramo do Processamento de Linguagem Natural (PLN) e pode ser visto como um pré-requisito para a análise semântica de textos, além de ser uma tarefa essencial para sistemas de gerenciamento de documentos, mineração de textos, entre outros.

A solução tradicional para resolver o problema de entidades nomeadas é por meio de um classificador de sequência termo a termo [2], como pode ser visto na Tabela 1, onde são apresentados termos e sua respectiva categoria de entidade, das quais a categoria "O" representa a classificação do termo como não-entidade, pelo menos para o estudo que estiver em curso.

Destaca-se que há diversas opções de anotações para a identificação de entidades, tais como BIO (*Begin, Inside, Outside*) and BILOU (*Beginning, Inside, Last token of multi-token, Outside and Unit-length chunks*)[14].

2 https://nlp.stanford.edu/software/.

3 http://opennlp.apache.org

4 https://spacy.io/

TABELA I: EXEMPLO DE CLASSIFICAÇÃO DE SEQUÊNCIA TERMO A TERMO (EXEMPLO DE REN)

Termo/Palavra	Classificação (entidade)
Fernando	PESSOA
atualmente	TEMPO
mora	O
no	O
Rio de Janeiro	LUGAR

B. Deep Learning

As Redes Neurais (RN) são a base DL e de acordo com Goodfellow [15], elas podem representar funções de crescente complexidade por meio de camadas de representações simples do conhecimento sobre os dados.

Long Short-Term Memory (LSTM).

A Rede de Memória de Longo-Curto Prazo (do inglês *Long-Short Time Memory* - LSTM), definida como um tipo de Redes Neurais Recorrentes (do inglês *Recurrent Neural Network* - RNN), foi projetada para evitar o problema das longas dependências textuais das RNN [16]. A RNN é uma rede neural em que há laços em suas conexões, ou seja, os dados de saída de uma classificação são usados de forma recorrente nas classificações seguintes. Essa característica é interessante para dados sequenciais e que tenham dependências de dados anteriores ou temporais, como palavras em uma sentença [15]. Como o REN é uma tarefa que processa sequências de termos, essas redes se tornaram referências para processamentos textuais [17].

Convolutional Neural Network (CNN).

Ao contrário da LSTM, a CNN[13] não é uma RNN , mas uma rede neural *feed-forward*, isto é, os dados são propagados na rede em uma única direção. A principal propriedade de uma rede CNN é a sua habilidade em extrair características dos dados, ao mesmo tempo em que mapeia as características extraídas dos dados de entrada em uma saída observada (processamento supervisionados). Nesse trabalho, o processamento dessas redes é realizado em nível de caracteres [18], portanto, será usada para incorporar informação relativa aos caracteres presentes em palavras e expressões.

Embedding

Representar palavras ou expressão por vetores em um espaço vetorial representativo tem auxiliado fundamentalmente em tarefas da área de PLN [2]. *Word Embedding* (WE) é uma denominação para a representação vetorial de palavras [19], contendo muitas dimensões. Cada uma dessas dimensões representa uma característica que captura propriedades sintáticas e semânticas úteis da palavra [20].

O mapeamento desses termos ou expressões para a construção dos vetores é, em geral, realizado por algoritmos não supervisionados (não necessitam de qualquer anotação ou rótulo) baseados em redes neurais, em modelos capazes de aprender representações vetoriais de alta dimensionalidade para palavras que compõem grandes *corpus*. Entre outros algoritmos destacam-se Wang2Vec [21] e Word2Vec [22].

Os Char Embeddings (CE) são análogos aos Word Embedding, porém servem para representar características de caracteres que compõem as palavras.

WE e CE são portanto componentes fundamentais na solução proposta neste trabalho, uma vez que é preciso extrair características das palavras para mapear estas relações com as entidades, permitindo que o modelo seja capaz de generalizar, ou seja, classificar novas palavras ou expressões avaliadas em novas frases ou textos não utilizados durante o ajuste do modelo.

C. Janelas de Contexto

Janela de contexto diz respeito a uma sequência de palavras de tamanho fixo, em que a palavra que está na posição central da sequência é a palavra alvo da janela, ou seja, a janela de contexto contém um número de palavras anteriores e posteriores à palavra objetivo, além da própria palavra objetivo, que carrega consigo o seu contexto [2]. Com isso, é possível classificar uma palavra levando em consideração o contexto, no qual está inserida em uma frase ou texto. O conceito é análogo para a janela de contexto de caracteres. Neste trabalho são usadas janelas de contexto para palavras e caracteres como entradas de dados das redes neurais (RNs), baseando-se no trabalho proposto por Santos e Zadrozny [18]. O tamanho das janelas são parâmetros que devem ser avaliados no modelo.

D. Arquitetura do Modelo para REN

O modelo para REN é baseado em [18], que foi proposto para tarefa de *Part-of-Speech* (POS) e em [4] para REN. A Figura 1 ilustra o esquema da arquitetura, que pode ser dividida em três partes: pré-processamento, processamento e saída.

O pré-processamento recebe as palavras e caracteres (relativos às palavras), ambos sob janelas de contexto. A parte que recebe palavras passa por camada de WE e aguarda o pré-processamento dos caracteres da palavra. As janelas de caracteres da palavra, após passarem pela camada de CE, passam por uma camada de redimensionamento para serem processadas em um modelo CNN, que inclui uma camada de *max-pooling*, seguida de um novo redimensionamento para que estes dados processados possam ser concatenados ao *embedding*, que representa o vetor da palavra central da janela de contexto das palavras. A parte da arquitetura proposta definida como "camada escondida" é composta por camadas recorrentes de redes do

tipo LSTM, que inclui *dropout*, e a camada de saída é formada por uma camada densa.

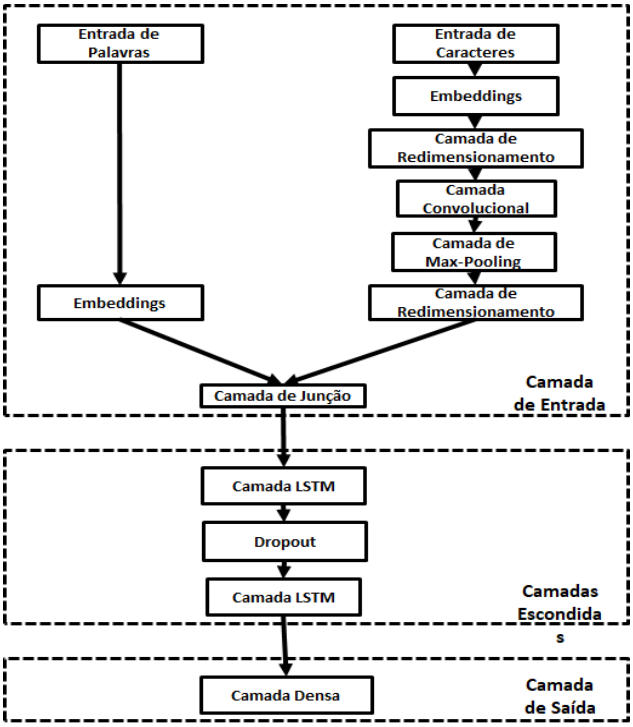

Fig. 1: Arquitetura do modelo

III. METODOLOGIA PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM DENÚNCIAS CRIMINAIS

Essa seção descreve a metodologia adotada no desenvolvimento do modelo para identificar, em textos de denúncias criminais feitas via aplicativo celular, as seguintes categorias de entidades nomeadas: pessoas, locais e tempo. Foi realizada uma filtragem na base de dados para utilizar apenas relatos dos seguintes assuntos selecionados: roubos em geral, roubos de carga/veículo, tráfico de drogas, tráfico de drogas/ armas, homicídios e armas. As informações extraídas da entidade "locais", por exemplo, são fundamentais para alimentar as ferramentas de geoprocessamento, as entidades "pessoas" podem ser usadas para identificar suspeitos em bancos de dados policiais e a identificação do "tempo" aponta o momento em que o crime pode ter sido praticado. Dessa forma, fica claro que essas três entidades são fundamentais e têm o objetivo de auxiliar na tomada de decisões na segurança pública.

Para classificar as entidades foi utilizado algoritmo supervisionado baseado na arquitetura apresentada na seção 2.5, que inclui redes RNN e CNN, com dados de entrada contendo as denúncias (texto), e gerando como saídas suas respectivas entidades. O texto com as denúncias foi obtido a partir de amostra da base de dados de aplicativo para celular, sem nenhuma restrição quanto ao vocabulário utilizado. No entanto, como essa base não estava previamente rotulada, foi necessária a classificação de cada palavra ou termo das frases [2] em uma entidade, dentre as que se tem interesse (Pessoa, Tempo e Local).

A Figura 2 apresenta um esquema, que será usado para detalhar a metodologia adotada neste trabalho. Na "Preparação dos Dados", visando medir o diferencial do uso de ferramentas disponíveis para a identificação de entidades, a base textual com as denúncias foi processada pela ferramenta GCNL, que classifica termo a termo, segundo as entendidas definidas na ferramenta. No entanto, conforme já mencionado na primeira seção deste trabalho, devido a linguagem popular dos das denúncias, contendo erros ortográficos, sintáticos e semânticos, a ferramenta para rotulagem de entidades do GCNL não foi eficaz, pois foi desenvolvida a partir de *corpus* constituído por textos literários, normalmente livre de erros ortográficos, sintáticos ou semânticos. Após essa etapa, foi feita uma validação manual das categorias de entidades informadas pela ferramenta (por dois pesquisadores) para identificar as falhas, medindo a capacidade da ferramenta, e corrigindo a rotulagem, de modo a aumentar a qualidade dos dados utilizados e, consequentemente, dos resultados obtidos.

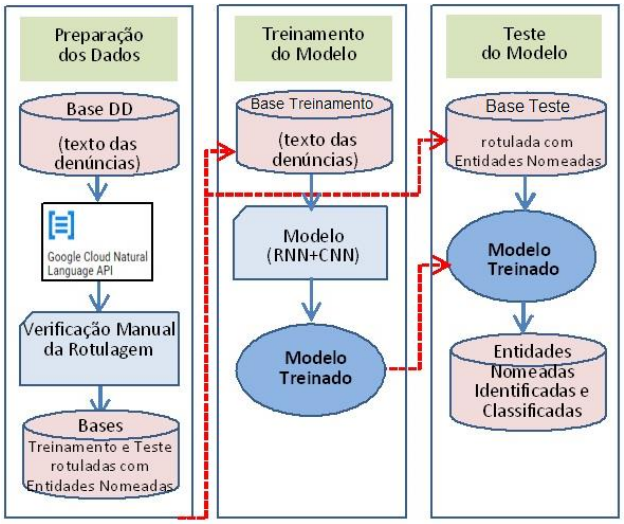

Fig. 2: Etapas da metodologia proposta

A partir da rotulagem dos dados, antes da base ser dividida em treinamento (80% da base) e teste (20% da base) para a realização do aprendizado, o texto foi "tokenizado". A parte usada para treinamento teve 20% dos dados separados para o processo de validação do modelo (RNN+CNN). Após o treinamento, os dados da base de testes foram avaliados no modelo treinado.

A arquitetura indicada na Figura 1 sofreu adaptações em algumas camadas das redes, hiperparâmetros e WE. O modelo foi desenvolvido em Python e o framework Keras[5].

[5] https://github.com/fchollet/keras

A. *Embeddings*

Para vetorização das palavras, levando em consideração a sintaxe, semântica e morfologia da palavra, foram utilizados os vetores de *WE* pré-treinados para português (repositório online[6]) desenvolvido a partir do método Skip-Gram [22] pelo algoritmo Wang2Vec [20].

Os vetores de caracteres, para o CE, foram obtidos durante o treinamento do modelo apresentado na Figura 1, na camada *Embedding* (abaixo da caixa Entrada de Caracteres e desenvolvido em Keras), cujos pesos dos neurônios dessa camada representam a dimensão do vetor.

B. *Avaliação do Modelo*

As métricas utilizadas para avaliar o modelo foram acurácia (= $\text{precison}_{\text{micro}}$ = $\text{recall}_{\text{micro}}$ = $\text{F1-score}_{\text{micro}}$) $\textit{precision}_{\text{macro}}$, $\textit{recall}_{\text{macro}}$ e $\textit{F1-score}_{\text{macro}}$, que são as métricas tradicionais da literatura para classificação de entidades nomeadas [6]. A acurácia é definida pela razão do acerto de predições, com relação a todas as entidades, sobre o total de termos. Já as métricas $\textit{precision}_{\text{macro}}$, $\text{recall}_{\text{macro}}$ e $\text{F1-}\textit{score}_{\text{macro}}$ são definidas pelas fórmulas 1, 2 e 3, respectivamente.

$$Precision_{macro} = \sum_{i=1}^{l} Precision_i / l \quad (1)$$

$$Precision_{macro} = \sum_{i=1}^{l} Recall_i / l \quad (2)$$

$$F1 - score_{macro} = \sum_{i=1}^{l} \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \Big/ l \quad (3)$$

Onde l é o total de entidades avaliadas. A métrica *precision* é a porcentagem de amostras positivas, classificadas corretamente, sobre o total de amostras classificadas como positivas, enquanto que *recall* diz respeito à porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas. O *recall* oferece informações sobre o desempenho de um classificador com relação a falsos negativos (quantos se perde), enquanto *precision* informa com respeito a falsos positivos (quantos se captura). O uso da métrica F1-score (média harmônica entre o *recall* e *precision*) é bastante adequado, principalmente quando as classes do problema não estão balanceadas e não há preferências específicas entre falso positivo ou negativo para redução das falhas.

Essas métricas foram usadas tanto para avaliar a classificação geral dos resultados, como para avaliar a classificação de cada entidade isoladamente. Para fazer essa avaliação separada por entidades, é necessário fazer uma interpretação clara quanto aos FP de uma entidade. Dessa forma, ressalta-se que os FP de uma entidade estão representando a quantidade de vezes que o modelo classificou de forma incorreta uma palavra como pertencendo a essa entidade.

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

IV. RESULTADOS

Nesta seção são apresentados os resultados obtidos com o modelo criado a partir dos relatos criminais, assim como todo o processo de avaliação adotado e a interpretação dos resultados obtidos.

A. *Preparação de Dados*

No aplicativo celular, o usuário efetua uma denúncia e a associa esta a um tipo de crime, indicando o assunto no aplicativo. Foi realizada uma filtragem nessa base de dados para utilizar apenas relatos dos seguintes assuntos: roubos em geral, roubos de carga/ veículo, tráfico de drogas, tráfico de drogas/ armas, homicídios e armas. Esses assuntos foram escolhidos preliminarmente por serem os que tinham maior frequência entre os relatos criminais e eram os que havia maior interesse do ponto de vista da segurança pública. Após essa filtragem foi efetuado o reconhecimento de entidades. Conforme mencionado na seção anterior, os relatos foram processados pelo GCNL para terem suas entidades identificadas e categorizadas, e, em seguida, essa categorização foi validada manualmente. Após esse procedimento, a base foi dividida em duas partes, sendo 80% para treinamento, dos quais 20% foram usados para validação do modelo, e 20% reservados para teste. A Tabela 2 indica a quantidade de categorias de entidades e seu respectivo percentual na base treinamento. Observa-se o alto volume do rótulo Outros, devido a soma de todos termos diferentes de Pessoa, Tempo ou Local que fazem parte do texto nas denúncias. No entanto, destaca-se que esse volume é normal em processos de REN.

TABELA II: TOTAL DE TERMOS PARA CADA ENTIDADE – BASE TREINAMENTO

Rótulo	Quantidade	Percentual (%)
Pessoa	**13.877**	**4,31**
Tempo	**4.044**	**1,25**
Local	**10.783**	**3,35**
Outros	**293.284**	**91,09**

B. *Hiperparâmetros*

Os hiperparâmetros da rede foram verificados de forma a maximizar a $\text{F1-score}_{\text{macro}}$ e com base nos trabalhos de Mendonça Jr. [23], Mikolov [24] e Bengio [25].

Os parâmetros de janela de contexto foram variados de 3 a 7; já os parâmetros de unidades LSTM foram variados de 200 a 440 com intervalos de variação de 10. As dimensões dos vetores de WE e CE foram variadas de 10 em 10, entre os valores 50, 100 e 300 e entre 10 e 50, respectivamente. Por fim, o número de épocas foi verificado para todos os

valores entre 1 e 40. Os parâmetros da Tabela 3 foram os que apresentaram o melhor resultado.

TABELA III: HIPERPARÂMETROS DO MODELO

Parâmetro	Valor
Número de épocas	20
Dimensão do vetor de *word embeddings*	50
Dimensão do vetor de *char embeddings*	20
Tamanho da janela de contexto de palavras	5
Tamanho da janela de contexto de caracteres	5
Unidades convolucionais	10
Unidades LSTM	420
Dropout rate	50%

C. Avaliação do modelo

A Tabela 4 apresenta o total de termos identificados por cada categoria de entidade na base teste, desde entidades compostas por um até cinco termos. Outros foi mantido para que se pudesse observar o volume de termos que não pertencem a nenhuma das três entidades nomeadas avaliadas nesse trabalho.

TABELA IV: TOTAL DE TERMOS PARA CADA ENTIDADE – BASE TESTE

Nº de palavras	Pessoa	Tempo	Local	Outros
1	320	24	147	212
2	18	14	48	139
3	8	9	21	120
4	4	7	18	87
5	4	8	24	744
Total de termos por entidade	416*	147	498	4918

*fórmula: 1x320 + 2x 18 + 3x8 + 4x4 + 5x4 = 416

Após avaliação dos resultados, considerando os hiperparâmetros indicados na seção 4.2, a melhor arquitetura (Tabela 3) apresentou a matriz de confusão exibida na Tabela 5 para base teste. Estes resultados observam o total de termos por entidade, desconsiderando formação de expressões.

Avaliando a Tabela 5, percebe-se que todas as entidades, naturalmente, apresentam maior confusão com a classe Outros, já que essa reúne todos os termos diferentes de Pessoa, Tempo ou Local, além de ser a que possui a maior quantidade total de termos. Também parece ser esperado que, entre essas entidades, as maiores dificuldades estejam entre as entidades Local e Pessoa, já que também possuem forma sintática de referência no texto semelhante.

A Tabela 6 apresenta as métricas produzidas a partir da matriz de confusão da Tabela 5. Os registros Pessoa, Tempo e Local da Tabela 6 indicam os valores das métricas, apresentadas na seção 3.B, por entidade. A última linha dessa Tabela aponta os resultados gerais ($precision_{macro}$, $recall_{macro}$ e $F1\text{-}score_{macro}$) para o modelo considerando todas as entidades. Essas métricas mostram uma boa generalização do modelo, considerando o fato de ter sido usada uma vetorização baseada em *word embedding*, construída a partir de textos literários, ou seja, com *corpus* distinto do *corpus* dos textos da denúncias. As métricas precision, recall e F1-score estão com bons resultados, exceto para a entidade Tempo, que certamente foi prejudica no processo de aprendizado, pois possui uma quantidade menor de exemplos na base de treinamento (Tabela 2).

As Tabelas 7, 8, 9, 10 e 11 têm o objetivo apresentar o desempenho do modelo, considerando a classificação das entidades compostas por um número de termos variando de 1 a 5 termos. Esses valores apontam quantos desses termos em cada expressão são corretamente indicados. Para facilitar a leitura, os valores da Tabela 4, que indicam o total de termos por expressão (contendo até 5 termos) para cada entidade, foram acrescentados entre parênteses ao lado das entidades indicadas em cada uma das Tabelas 7, 8, 9,10 e 11.

TABELA V: MATRIZ DE CONFUSÃO DO MODELO BASE TESTE

		Entidade predita			
		Pessoa	Tempo	Local	Outros
Entidade Real	Pessoa	310	1	9	96
	Tempo	0	105	0	42
	Local	13	2	381	102
	Outros	53	52	70	4743

TABELA VI: MÉTRICAS DO DESEMPENHO DO MODELO - BASE TESTE

	Recall (%)	*Precision (%)*	F1-score(%)
Pessoa	74,52	82,45	78,28
Tempo	71,43	65,63	68,40
Local	76,51	82,83	79,54
macro	74,15	76,97	75,41

TABELA VII – AVALIAÇÃO DE ENTIDADES COMPOSTAS POR EXPRESSÕES DE **1 TERMO** NA BASE TESTE

	Classificação	
Entidade	0/1	1/1
Pessoa(320)	80	240
Tempo(24)	14	10
Local(147)	35	112

TABELA VIII – AVALIAÇÃO DE ENTIDADES COMPOSTAS POR EXPRESSÕES DE **2 TERMOS** NA BASE TESTE

	Classificação		
Entidade	0/2	1/2	2/2
Pessoa(18)	1	4	13
Tempo(14)	4	5	5
Local(48)	6	10	32

TABELA IX – AVALIAÇÃO DE ENTIDADES COMPOSTAS POR EXPRESSÕES DE **3 TERMOS** NA

TABELA X – AVALIAÇÃO DE ENTIDADES COMPOSTAS POR EXPRESSÕES DE **4 TERMOS** NA BASE

BASE TESTE

	Classificação			
Entidade	0/3	1/3	2/3	3/3
Pessoa(8)	2	0	1	5
Tempo(9)	0	3	2	4
Local(21)	1	1	4	15

TESTE

	Classificação				
Entidade	**0/4**	**1/4**	**2/4**	**3/4**	**4/4**
Pessoa(4)	**2**	**0**	**0**	**0**	**2**
Tempo(7)	**0**	**0**	**1**	**3**	**3**
Local(18)	**0**	**3**	**0**	**5**	**10**

TABELA XI– AVALIAÇÃO DE ENTIDADES COMPOSTAS POR EXPRESSÕES DE **5 TERMOS** NA BASE TESTE

	Classificação					
Entidade	0/5	1/5	2/5	3/5	4/5	5/5
Pessoa(4)	0	0	0	1	3	0
Tempo(8)	0	0	0	0	2	6
Local(24)	0	2	6	2	7	7

As Tabelas 7, 8, 9, 10 e 11 apontam os erros e acertos de classificação para cada entidade composta por 1, 2, 3, 4 e 5 termos, respectivamente. Essa avaliação ajuda indicar se o modelo usado para classificação das entidades consegue identificar corretamente expressões

Assim, a primeira coluna (0/1, 0/2, 0/3, 0/4 e 0/5) de cada uma dessas tabelas indica o número de expressões em que nenhum dos termos, que compõe a expressão, conseguiu ser identificado pelo modelo como entidade, seja para pessoa, tempo e local. Nota-se que para entidades compostas por um único termo (Tabela 7), a entidade Tempo teve novamente o pior desempenho, acertando dez entidades de Tempo em 24, seguido pelas entidades Local e Pessoa. A Tabela 8 expõe a distribuição dos acertos das entidades constituídas por dois termos. Nela há evidência que um número maior de classificações corretas, quando o modelo acerta completamente a expressão, nesse caso os dois termos, indicando por 13 ocorrências. Apenas para dar dimensão do resultado, a Tabela 4 indica um total de 18 expressões com 2 termos para Pessoa, o que permite avaliar que das 18 expressões com 2 termos para a entidade Pessoa, o modelo acerta 13. Na Tabela 9 destaca-se que a entidade Local se saiu melhor, com expressões de 3 termos, pois acerta mais de 70% das expressões com 3 termos para a entidade Local. A Tabela 10 novamente aponta a entidade Local como a que teve maior quantidade de acertos para a expressão completa com 4 termos (55,6% das expressões). No entanto, a entidade tempo também pode ser mencionada, pois acerta 43% das vezes a expressão completa e 86%, considerando quaisquer 3 termos dos 4 termos da expressão. Finalmente, a Tabela 11 indica que a entidade Tempo teve o melhor desempenho acertado completamente os 5 termos em 6 expressões (de um total de 8 expressões) e acertando 4 termos em 5 para duas ocorrências de expressões da entidade Tempo no conjunto de teste. A Tabela 12 finaliza essa análise indicando os percentuais de acerto total de cada entidade, considerando o número de termos em que estas entidades são formadas. Destaque para o Tempo com 5 termos e Local para 3 termos e Pessoa para 2 termos. Apesar da entidade Tempo não ter tido uma amostra de entidades equivalente às entidades Pessoa e Local (ver Tabela 2), percebe-se que quanto mais termos a expressão da entidade Tempo possui, mais facilmente o modelo acerta. Para entidade Pessoa o melhor desempenho de classificação é para um termo na expressão que coincide com o maior volume de expressões na base de teste. A entidade Local também indica a melhor classificação para o caso de expressões unitárias.

TABELA XII– AVALIAÇÃO PERCENTUAL DO ACERTO TOTAL DE TERMOS PARA CADA ENTIDADE CONSTITUÍDA DE 1 A 5 TERMOS NA BASE TESTE

Nº de palavras	Pessoa	Tempo	Local
1	75,0	41,7	76,2
2	72,2	35,7	66,7
3	62,5	44,4	71,4
4	50,0	42,9	55,6
5	0,0	75,0	29,2

V. CONCLUSÕES

Esse trabalho teve como objetivo principal o desenvolvimento de um modelo baseado em Deep Learning para o reconhecimento de entidades nomeadas presentes em textos em linguagem coloquial/popular em português brasileiro, mais especificamente para o contexto de denúncias criminais.

Para tanto, foi desenvolvido uma metodologia que considerou textos da base de dados de aplicativo celular de órgão associado à Segurança Pública do Estado do Rio de Janeiro, que foram cedidos sob condições de sigilo de dados (por esse motivo não poderá ser divulgada a base de dados). Através dela, será possível a extração de informação em textos de relatos de crimes, auxiliando a segurança pública no combate à violência urbana.

Após toda a etapa de avaliação, considera-se que os resultados desse trabalho inicial alçaram as expectativas iniciais. O classificador obtido está com uma qualidade equivalente à proposta em Mendonça Jr. [23] e Santos e Guimarães [4], que apontaram bons resultados na tarefa de REN em português brasileiro. Entretanto, ressalta-se que suas bases de dados estavam em sintonia com corpus de textos literários (linguagem formal), usados para criar os métodos (por exemplo, o de vetorização das palavras) utilizados no pré-processamento de suas base de dados textual. Por outro lado, destaca-se que nos textos extraídos das denúncias, há muitos casos de erros léxicos, sintáticos e

semânticos, muitas abreviações, além de vocabulário específico.

Dessa forma, apesar das métricas obtidas com o modelo serem boas, principalmente a F1-score, percebeu-se que algumas categorias, como Tempo, não estão sendo reconhecidas de forma eficaz, tendo como principal motivo a distribuição dos registros de treinamento e teste, que não está equilibrada com relação à frequência de cada entidade (base desbalanceada), prejudica o aprendizado do modelo. Soma-se a isso, o fato de que as bibliotecas utilizadas nesse trabalho, que oferecem os vetores de palavras para português, foram criadas a partir de corpus literários, que se distanciam do contexto dos textos das denúncias.

Como trabalhos futuros, deseja-se estender a extração de entidades para textos de redes sociais. Também pode-se indicar, como perspectiva futura, a exploração e visualização dos vetores de palavras com o uso de algoritmos como t-SNE [26], visando entender mais profundamente a qualidade desses vetores, para possivelmente melhorá-los, por exemplo, ajustando nova vetorização, especificamente para essas bases de denúncias criminais.

REFERÊNCIAS

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification". *Lingvisticae Investigationes*, v. 30, n. 1, p. 3-26, 2007.

[2] D. Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

[3] A. Borthwick and R. Grishman. A maximum entropy approach to named entity recognition. Tese de Doutorado, New York University, Graduate School of Arts and Science, 1999.

[4] C.N. dos Santos and V. Guimarães. "Boosting named entity recognition with neural character embeddings", Proceedings of the Fifth Named Entity Workshop, Stroudsburg, PA, USA: Association for Computational Linguistics, 2015.

[5] D. Santos e N. Cardoso, "HAREM: Reconhecimento de Entidades Mencionadas em Português", Disponível em: https://www.linguateca.pt/HAREM/ Acesso em: outubro de 2018.

[6] W.. Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina", Tese de Doutorado, Universidade de São Paulo, 2012.

[7] J. Lavid. "Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics", *International journal of Translation*, 22(1):13–36, 2010.

[8] Lopes, F., Teixeira, C.e Oliveira, H.G. Comparing Different Methods for Named Entity Recognition in Portuguese Neurology Text. Journal of Medical Systems (2020) 44: 77 https://doi.org/10.1007/s10916-020-1542-8

[9] A. Pires. "Named entity extraction from Portuguese web text", Faculdade de Engenharia da Universidade do Porto, 2017.

[10] E. Fonseca, G. Chiele, R. Vieira e A. Vanin, "Reconhecimento de Entidades Nomeadas para o Português Usando o OpenNLP", Anais do Encontro Nacional de Inteligência Artificial e Computacional, 2015.

[11] F. Silva e R. Vieira, "Aplicação de reconhecimento de entidades nomeadas em investigação de crimes financeiros". Proceedings of the 2nd Symposium in information and human language technology, pp.134–143, 2019.

[12] N. Araújo, "Reconhecimento de Entidades Nomeadas em Textos de Boletins de Ocorrências", Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Quixadá da Universidade Federal do Ceará, 2019.

[13] Carnaz, G., Nogueira, V.B.e Antunes, M. A Graph Database Representation of Portuguese Criminal-Related Documents. Informatics 2021, 8, 37. https://doi.org/10.3390/informatics8020037

[14] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition", Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 147-155, 2009.

[15] I. Goodfellow, Y. Bengio, and A. Courville, "*Deep learning",* Cambridge: MIT press, 2016.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, v. 9, n. 8, p. 1735-1780, 1997.

[17] K. Tarwani and S. Edem, "Survey on Recurrent Neural Network in Natural Language Processing", *Int. J. Eng. Trends Technol*., vol. 48, no. 6, pp. 301–304, 2017.

[18] C. Santos, B. Zadrozny "Learning character-level representations for part-of-speech tagging", International Conference on Machine Learning. PMLR, p. 1818-1826, 2014.

[19] O. Levy and Y. Goldberg, "Dependency - based word embeddings", Proceedings of the 52nd Annual meeting of the association for computational linguistics, pp.302–308, 2014.

[20] J. Turian, L. Ratinov and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning", Proceedings of the 48th annual meeting of the association for computational linguistics. p. 384-394, 2010.

[21] W. Ling, C. Dyer, A. Black and I. Trancoso, "Two/too simple adaptations of word2vec for syntax problems", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space". Proceedings of International Conference on Learning Representations Workshop (ICLR), 2013.

[23] C. Mendonça Jr., L. Barbosa, H. Macedo. "Uma arquitetura híbrida LSTM-CNN para reconhecimento de entidades nomeadas em textos naturais em língua portuguesa", XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) SBC, 2016.

[24] T. Mikolov, "Statistical language models based on neural networks", Ph.D. Thesis, Brno University of Technology, Faculty of Information Technology, Brno, CZ, 2012.

[25] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures", *Neural networks: Tricks of the trade*. Springer, Berlin, Heidelberg, p. 437-478, 2012.

[26] L. van der Maaten, G. Hinton, "Visualizing Data using t-SNE". *Journal of Machine Learning Research* 9, pp. 2579-2605, 2008.