Avaliação de Reconhecimento de Entidades Nomeadas a partir de *Embedding Criado* para o Contexto de Segurança Pública

1st Leonidia Barreto
Programa de Pós-Graduação em Ciências Computacionais (PPG-CCOMP)
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro, Brasil
leonidia.barreto@pos.ime.uerj.br

2nd Pedro Jaber
Instituto de Matemática e Estatística
Departamento de Ciência da Computação
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro, Brasil
pedro.jaber@gmail.com

3rd Karla Figueiredo
Instituto de Matemática e Estatística
Departamento de Ciência da Computação
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro, Brasil
karlafigueiredo@ime.uerj.br

4nd Walkir A.T. Brito
Disque-Denúncia
Rio de Janeiro, Brasil
walkir.brito@disquedenuncia.org.br

***Abstract*—O Reconhecimento de Entidades Nomeadas tem sido de grande valia para a Segurança Pública. No entanto, as denúncias feitas pela população, escritas em língua portuguesa brasileira coloquial, além de possuírem vocabulário específico, contém muitos erros ortográficos e gramaticais, o que dificulta o trabalho de extração de informação a partir de ferramentas e bibliotecas disponíveis publicamente. Sendo assim, a construção de um modelo de *Word Embedding* contextualizado para Segurança Pública, torna acessíveis tais informações para modelos de *Machine Learning*. Desse modo, esse trabalho além de *corpus* e *Word Emnedding* voltado para Segurança Pública, apresenta um modelo para Reconhecimento de Entidades Nomeadas. O uso deste *Word Embedding* desenvolvido, aumenta em de 5,34% a acurácia média quando comparados ao uso de *Word Embedding* público, indicando ser um caminho promissor nesse contexto.**

***Index Terms*—Word Embeddings, Processamento de Linguagem Natural, Segurança Pública, Reconhecimento de Entidades Nomeadas**

I. Introdução

O crime e a violência mantiveram-se entre as cinco principais preocupações do mundo no último ano segundo o instituto de pesquisa Ipsos[1]. Observando-se os dados do Instituto de Segurança Pública[2] referentes somente a dezembro de 2022 do Estado do Rio de Janeiro (ERJ), foram registrados mais de dez mil roubos de diferentes categorias. Entre eles, os que tiveram os maiores números de denúncias foram o roubo a banco, a transeunte e de aparelho celular.

Segundo o professor José Ricardo Bandeira, presidente do Instituto de Criminalística e Ciências Policiais da América, o investimento em inteligência na segurança pública evitaria diferentes tipos de crime no ERJ [3]. Como evidência disso, no ano de 2022 a modernização e o investimento em equipamentos tecnológicos como drones trouxe reduções para crimes como roubo de rua, de carga, além da menor taxa de homicídio doloso dos últimos 31 anos [4].

Diante desse cenário, soluções envolvendo técnicas de *Deep Learning* (DL) possuem consideráveis chances de auxiliar no trabalho daqueles que se dedicam ao combate da criminalidade, já que essa área da Inteligência Artificial (IA) vem melhorando consideravelmente o estado-da-arte em Processamento de Linguagem Natural (PLN), entre outros domínios [1], e que pode ser empregada para extrair informação das inúmeras denúncias recebidas diariamente.

Entre as possibilidades, pode-se indicar o uso de Reconhecimento de Entidades Nomeadas (REN) — sendo possível associar Entidades Nomeadas (EN) como pessoas, locais e datas, entre outras — para extrair essas informações textuais de bases de denúncias criminais, visando automatizar e acelerar a tomada de decisão pela Segurança Pública. O REN, por sua vez, faz parte da área de Extração de Informação (EI), que dentro do PLN tem como tarefa a detecção e classificação de informações relevantes de uma base de interesse [2]. A relevância desses dados é estabelecida de acordo com os objetivos do segmento no qual eles serão utilizados, como o contexto da segurança pública.

Em [3] foi apresentado um modelo híbrido envolvendo Redes Neurais Convolucionais (RNC) e *Long Short-Term*

[1]https://www.ipsos.com/pt-br/what-worries-world-janeiro-de-2023
[2]https://www.isp.rj.gov.br/
[3]https://www.cnnbrasil.com.br/nacional/brasil-investe-r-160-bilhoes-em-seguranca-mas-so-r1-9-bilhao-em-inteligencia/
[4]https://diariodorio.com/seguranca-do-estado-recebe-mais-de-r-700-milhoes-em-equipamentos-tecnologicos/

Memory (LSTM) a partir de um dos *Word Embeddings* (WE) disponibilizados no Repositório de WE (RWE) do Núcleo Institucional de Linguística Computacional (NILC) e *character embeddings*, que foi ajustado com a finalidade de efetuar o REN em bases textuais de denúncias do Disque Denúncia RJ (DD RJ), cujo conteúdo está escrito em língua portuguesa brasileira coloquial. Um dos principais desafios apontados no estudo refere-se aos erros gramaticais e sintáticos presentes no texto, além do fato de que as palavras pertencem a um contexto diferente daquele que originou os WE utilizados. Em síntese, não há consonância entre o léxico das ferramentas disponíveis e o conteúdo textual das denúncias.

Desse modo, o objetivo central deste trabalho consiste em analisar a hipótese de que a utilização de WE, adquiridos a partir de um *corpus* textual mais específico — no caso, relacionado ao âmbito da segurança pública em linguagem coloquial em português brasileiro — teria o potencial de oferecer resultados aprimorados para um modelo de REN. Para esse fim, foram elaborados um *corpus* destinado à área de segurança pública, um vetorizador lexical específico para esse cenário e, adicionalmente, foram realizadas extrações de EN nesse mesmo contexto, comparando-se tais resultados com os obtidas anteriormente.

II. FUNDAMENTOS TEÓRICOS

Nesta seção serão apresentados, sucintamente, alguns conceitos importantes para o desenvolvimento da pesquisa, tais como Janela de Contexto, WE, modelos de veorização de palavras, REN e uma breve descrição do modelo para REN utilizado na tranalho anterior.

A. Janela de Contexto

A janela de contexto é um conceito fundamental no uso de WE. É uma técnica amplamente utilizada para representar palavras por meio de vetores com valores numéricos contínuos. Janela de contexto é composta por uma sequência de palavras vizinhas da palavra que se deseja contextualizar. Esta palavra, alvo da janela, pode estar na posição central da sequência, nesse caso em condição de simetria, com um determinado número de palavras anteriores e posteriores. Esta premissa é decorrente do fato de que os significados de uma palavra são influenciados pelo contexto em que ela ocorre no texto. O tamanho da janela em simetria ou assimetria afetam a forma como o modelo captura e representa as relações semânticas e sintáticas entre as palavras. [4].

B. Word Embeddings

WE é uma denominação para a representação de palavras através de vetores multidimensionais formados por números reais, obtidos a partir um *corpus* textual não rotulado [5]. Cada uma de suas dimensões representa características úteis da semântica e sintaxe das palavras [6]. A construção ou mapeamento da representação vetorial dessas palavras ou expressões é realizado por meio de algoritmos, em geral, não-supervisionados ou auto-supervisionados [7]. No treinamento de tais modelos, baseados em Redes Neurais (RN), a janela de contexto é deslocada ao longo do texto, permitindo que o modelo capture diferentes contextos em que as palavras ocorrem. Ao fazer isso, o modelo aprende a associar as palavras com base em sua coocorrência em diferentes contextos, mapeando-as em espaços vetoriais, onde palavras semanticamente relacionadas estão próximas umas das outras [8].

C. Continuous Bag-Of-Words e Skip-gram

Continuous Bag-Of-Words (CBOW) e *Skip-gram* (SG) são duas arquiteturas tradicionais de modelo de aprendizagem auto-supervisionada propostas para a construção de WE [9].

Na arquitetura CBOW a camada de projeção é compartilhada para todas as palavras [8], desse modo seus vetores são calculados em média. Esse modelo é chamado de CBOW porque, mesmo que a ordem das palavras não influencie na projeção, diferentemente do modelo padrão de *bag-of-words*, ele usa uma representação distribuída contínua do contexto. Na Fig. 1 é possível ver a arquitetura do modelo. Nota-se que a matriz de pesos entre a entrada e a projeção é compartilhada para todas as posições de palavras [8]. As palavras anteriores e posteriores à palavra alvo são dadas como entrada, e é esperado que seja dado como saída a palavra alvo dessa janela de contexto [9].

A arquitetura SG é similar à CBOW, contudo, diferentemente da CBOW, ao invés de predizer uma palavra pertencente a um contexto, a SG usa cada palavra como entrada de um classificador *log-linear* com uma camada de projeção contínua, então prediz-se um certo intervalo de palavras anteriores e posteriores ao "alvo" que está na entrada do modelo. O resultado dessa arquitetura foi um aumento de qualidade dos vetores resultantes, ou seja, com uma representação mais precisa do real significado das palavras. Contudo, a complexidade computacional também é maior [8]. A arquitetura SG está demonstrada na Fig. 1.

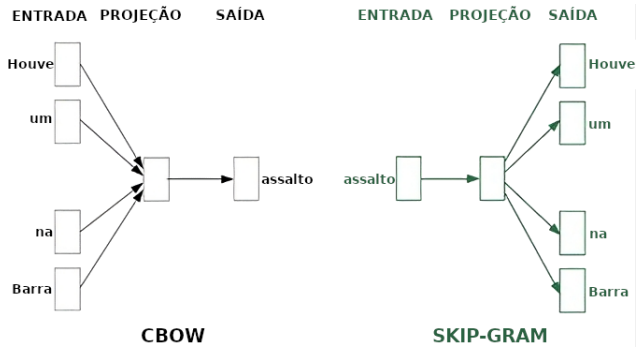

Fig. 1. A arquitetura CBOW prevê a palavra baseando-se no contexto dado, e a arquitetura Skip-gram prevê o contexto na vizinhaça da palavra alvo. Fonte: Adaptado de Mikolov et al. (2013) [8].

D. Reconhecimento de Entidades Nomeadas (REN)

O objetivo do REN é encontrar termos em um texto, formados por uma palavra ou expressões (conjunto de palavras), e classificá-los de acordo com um conjunto de categorias pré-definidas, tais como identificadores de pessoas, de lugares ou

de tempo. O REN é um ramo do PLN que pode ser declarado como um pré-requisito para a análise semântica de textos e como parte essencial para sistemas de gerenciamento de documentos, mineração de textos, entre outros [10]. A Tabela 1 apresenta exemplos de termos e suas respectivas categorias. Nela a categoria "O" representa a classificação do termo como não-entidade.

TABELA I
EXEMPLO DE CLASSIFICAÇÃO DE SEQUÊNCIA TERMO A TERMO

Termo/Palavra	Classificação/Entidade
O	O
suspeito	PESSOA
foi	O
visto	O
em	O
Niterói	LOCALIZAÇÃO
ontem	TEMPO

E. Arquitetura do Modelo para REN

O modelo para REN utilizado foi baseado em [14], que originalmente abordou a tarefa de *Part-of-Speech* e no trabalho [15], que se voltou para REN. No pré-processamento dessa arquitetura, tanto as palavras quanto os caracteres associados são considerados sob janelas de contexto. As palavras passam por uma camada de WE, enquanto os caracteres, após passar pela camada de *Char Embeddings*, são processados em um modelo de *Convolutional Neural Network* (CNN), seguido por *max-pooling* e redimensionamento para concatenação com o vetor de palavra central. A "camada escondida" é composta por camadas recorrentes do tipo LSTM com *dropout*, e a camada de saída é uma camada densa. A Fig. 2 ilustra a arquitetura descrita.

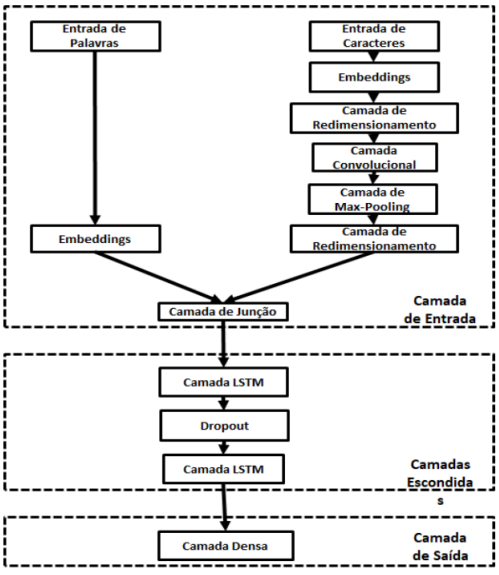

Fig. 2. Arquitetura do modelo de REN.

III. METODOLOGIA

Na Fig. 3 apresentamos uma visão geral da metodologia aplicada.

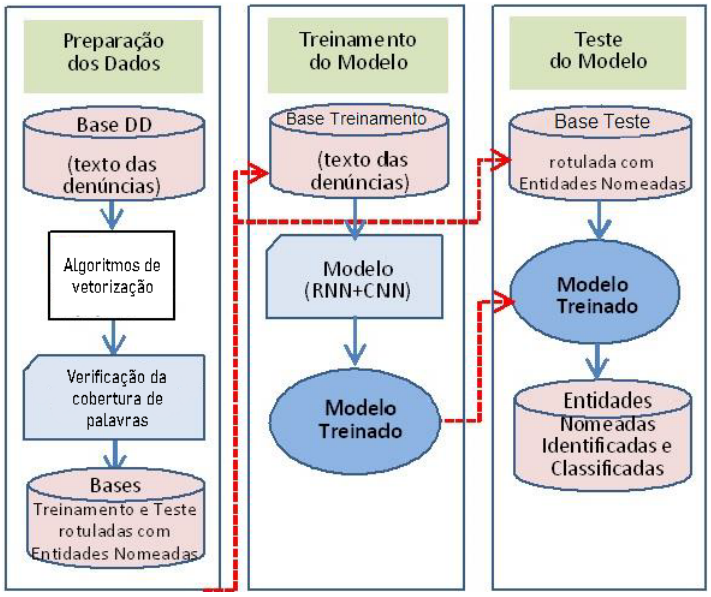

Fig. 3. Visão geral da metodologia aplicada.

A. Escolha da arquitetura do modelo de vetorização

Foi realizado um estudo prévio para avaliar desempenhos dos modelos CBOW e SG, disponibilizados na biblioteca GENSIM[5]. Essa comparação foi feita elegendo-se uma lista de doze palavras presentes no *corpus* do DD RJ e avaliando as distâncias medidas pelo cosseno entre elas. Os hiperparâmetros utilizados estão listados na Tabela 2. Os demais hiperparâmetros utilizados foram os definidos como padrão na implementação desses algoritmos da GENSIM.

TABELA II
HIPERPARÂMETROS UTILIZADOS NOS MODELOS CBOW E SKIP-GRAM DA GENSIM

Hiperparâmetro	Valor
Dimensão dos vetores de word embeddings	100
Tamanho da janela de contexto de palavras	5
min_count[a]	1
Sg[b]	1

[a]Ignora palavras com aparições inferiores a seu valor.
[b]Se igual a 1, o algoritmo será o Skip-gram.

B. Cobertura de palavras

Para calcular a cobertura de palavras foi escrito um algoritmo que constrói uma lista encadeada das palavras contidas na base. Esta lista encadeada é gerada de forma a manter apenas palavras únicas, ou seja, tanto os valores numéricos característicos do WE, quanto as duplicatas das palavras foram descartadas.

Posteriormente o algoritmo executa o mesmo tratamento de remoção de valores numéricos e de repetições de palavras para o WE contendo vetores de 50 dimensões disponibilizado no RWE do NILC, criando outra lista de palavras sem os valores numéricos do WE. Com isso, ao final deste processo,

[5]https://radimrehurek.com/gensim/auto_examples/index.html#documentation

haverá duas listas de palavras, que serão comparadas, e então o algoritmo apresenta os itens comuns entre elas e os itens não compartilhados.

C. Embedding de palavras

Para construir os embeddings de palavras candidatas foi utilizado o algoritmo *Word2Vec* em C disponibilizado no *Google Code Archive*[6].

Após o processo de criação dos *embeddings*, foi gerado um modelo com a arquitetura de REN para cada combinação exaustiva de tamanhos de janela de contexto (3, 5, 7, 9, 11) e dimensão vetorial (25, 50, 100, 200, 300). Portanto, ao final, foram avaliados 25 modelos, exibidos na Tabela 3.

TABELA III
TODAS AS ARQUITETURAS CANDIDATAS GERADAS NESSE PROJETO

Janela de Contexto (W)	**Tamanho da Dimensão (D)**				
	25	50	100	200	300
3	W3D25	W3D50	W3D100	W3D200	...
5	W5D25	W5D50	W5D100	W5D200	...
7	W7D25	W7D50	W7D100	W7D200	...
9	W9D25	W9D50	W9D100	W9D200	...
11	W11D25	...	...	...	...

D. Comparação de eficiência

Cada WE foi utilizado como parâmetro na arquitetura para o REN, desenvolvida em trabalho anterior [3]. Além dos vetores de palavras utilizados (WE), os hiper-parâmetros usados no modelo para REN foram fixados e estão indicados na Tabela 4.

TABELA IV
EXEMPLO DE CLASSIFICAÇÃO DE SEQUÊNCIA TERMO A TERMO

Hiperparâmetros	Valor
Número de épocas	20
Dimensão dos vetores 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%

Para cada modelo REN avaliado, obteve-se uma matriz de confusão, da qual foram extraídas as métricas: acurácia, precisão, *recall* e *F1-measure* nos resultados de classificação de cada entidade nomeada, além do caso base (WE do NILC).

Tanto na etapa de treinamento quanto na de teste foram utilizadas entidades rotuladas e verificadas manualmente, sendo 80% da base de dados usada para treinamento e 20% para teste. Dessa forma, foram obtidos resultados que elegeram a melhor modelagem.

Após a seleção da modelagem mais adequada de *word embedding* — de melhor rendimento geral — foram acrescentadas à base mais 285 entidades nomeadas e validadas manualmente para teste. Com essa base maior, foi executado novamente o modelo de REN tanto com o embedding eleito quanto com o embedding do NILC.

[6]https://github.com/tmikolov/word2vec

E. Base de dados

A base de dados utilizada foi fornecida pelo DD RJ. Ela é composta por denúncias efetuadas por usuários do aplicativo móvel do DD[7], sem nenhuma restrição quanto ao vocabulário utilizado, contendo assim abreviações de palavras, erros de digitação, ortográficos e gramaticais.

IV. RESULTADOS

A. Escolha do modelo

Conforme descrito na seção III-A, a escolha do algoritmo para avaliação do *word embedding* foi feita de forma preliminar, avaliando os dois modelos tradicionais, CBOW e Skip-gram, com configurações iguais (mesma janela e mesma dimensão). Nesse caso, foi tomada a distância de cossenos para 12 palavras selecionadas no *corpus* de segurança pública. Assim, embora a arquitetura Skip-gram tenha feito uma distinção um pouco melhor entre diferentes entidades, esta diferença não justificou a escolha por um modelo que apresentava maior tempo de processamento doque o modelo CBOW [8].

B. Cobertura de palavras

Como já citado no item III.B, foi desenvolvido um algoritmo para a avaliação da cobertura de palavras analisando a base do DD RJ e as palavras vetorizadas contidas no WE do NILC. A base do DD possui um vocabulário que contém uma especificidade oriunda de um contexto social composto majoritariamente por uma população menos favorecida, que expressa-se com jargões perintentes às comunidades. Soma-se a isso o uso de abreviações recorrentes no ambiente das redes sociais.

Em contrapartida, a base utilizada para treinar o WE do NILC é descrita em sua fonte como sendo composta por *corpus* de diferentes origens. Entre suas fontes estão textos escritos em português europeu e brasileiro extraídos de sites de notícias, *e-books*, entre outros. Dessa forma, entende-se que a fonte que gerou modelos de WE disponibilizados no NILC, não era aderente ao contexto do DD, conforme pode ser observado na Fig. 4.

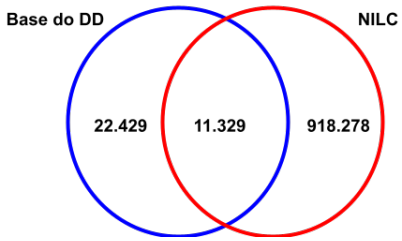

Fig. 4. Diagrama de Venn obtido a partir do algoritmo de cobertura de palavras.

[7]https://play.google.com/store/apps/details?id=br.org.disquedenuncia&hl=pt_BR&gl=US
https://apps.apple.com/br/app/disque-den%C3%BAcia-rj/id1147688588

A base do DD possui 33.758 palavras distintas e o *embedding* do NILC possui 929.607 palavras distintas. E, segundo o algoritmo de cobertura, as duas bases possuem 11.329 palavras em comum, o que resulta em uma cobertura de 33,56% da base do DD.

C. Levantamento das métricas

Conforme especificado nos itens III.C e III.D, foi gerada uma série de combinações de janelas de contexto com dimensão vetorial, utilizando-se uma parte da base rotulada para treinamento. Cada uma dessas combinações deu origem a um modelo. Para eleger a melhor dentre eles foi feita uma análise comparativa dos resultados das arquiteturas utilizando as métricas de acurácia, precisão, *recall* e *F1-measure*. As matrizes de confusão do *embedding* geradas pelo modelo WE selecionado (W3D50) pelo *embedding* disponibilizado no NILC, bem como os cálculos efetuados estão descritos abaixo:

TABELA V
MATRIZ DE CONFUSÃO DOS VALORES DE TESTE DO MODELO ELEITO

W3D50				
Entidades	O	LOCALIZAÇÃO	PESSOA	TEMPO
O	4769	69	48	32
LOCALIZAÇÃO[a]	108	375	15	0
PESSOA	135	26	251	4
TEMPO	59	1	0	87

[a]Exemplo os cálculos relacionados à entidade LOCALIZAÇÃO:
True positive: 375
True negative: 4769+251+87=5107
False positive: 69 + 26 + 1=96
False negative: 108 + 15 + 0=123

TABELA VI
MATRIZ RESULTADO DE TESTE DO MODELO ELEITO

Entidades	TP	TN	FP	FN
O	4769	713	302	149
LOCALIZAÇÃO	375	5107	96	123
PESSOA	251	5231	63	165
TEMPO	87	5395	36	60

TABELA VII
MATRIZ DE CONFUSÃO DOS VALORES DE TESTE DO RWE NILC

RWE				
Entidades	O	LOCALIZAÇÃO	PESSOA	TEMPO
O	4881	22	15	0
LOCALIZAÇÃO	468	28	2	0
PESSOA	397	4	15	0
TEMPO	142	5	0	0

TABELA VIII
MATRIZ DE RESULTADO DE TESTE DO RWE NILC

Entidades	TP	TN	FP	FN
O	4881	43	1007	37
LOCALIZAÇÃO	28	4896	31	470
PESSOA	15	4909	17	401
TEMPO	0	4924	0	147

Observando-se as Tabelas 5 e 7 nota-se que em ambas as combinações dos *word embeddings* com o modelo de REN, a maior confusão entre as entidades foi naturalmente com a não-entidade "O", que reúne todos os termos diferentes de Pessoa, Tempo e Localização, além de possuir a maior quantidade total de termos. Também há um esperado equívoco entre as entidades Localização e Pessoa, já que essas possuem estruturações textuais similares.

D. Treinamento e teste

Como já citado no item III.D, 285 entidades rotuladas foram acrescentadas à base de teste. As Tabelas 9 e 10 apresentam os resultados das métricas acurácia, precisão, *recall* e *F1-measure* nas etapas de treinamento e teste, comparando o *embedding* gerado a partir da base do DD (W3D50), com o embedding disponibilizado no repositório do NILC. Nessas tabelas "L" = "LOCALIZAÇÃO", "P" = "PESSOA", "T" = "TEMPO e "M" = "Macro" (resultado geral do modelo).

TABELA IX
RESULTADOS DAS MÉTRICAS DO EMBEDDING DA ARQUITETURA **W3D50** = "WD" E DO **RWE NILC** NO TREINO EM %.

	Acurácia		Precisão		*Recall*		*F1 score*	
	WD	RWE	WD	RWE	WD	RWE	WD	RWE
L	96,42	90,92	81,28	60,00	76,71	3,61	78,93	6,82
P	96,00	92,20	77,74	42,86	62,98	0,72	69,59	1,42
T	98,44	97,10	74,19	N/A	62,59	0,00	67,90	N/A
M	95,82	90,35	91,97	82,41	91,97	82,41	91,97	82,41

TABELA X
RESULTADOS DAS MÉTRICAS DO EMBEDDING DA ARQUITETURA **W3D50** E DO **RWE NILC** NO TESTE.

	Acurácia		Precisão		*Recall*		*F1 score*	
	WD	RWE	WD	RWE	WD	RWE	WD	RWE
L	96,16	90,76	79,62	47,46	75,30	5,62	77,40	10,05
P	96,01	92,18	79,94	46,88	60,34	3,61	68,77	6,70
T	98,28	97,10	70,73	N/A	59,18	0,00	64,44	N/A
M	95,66	90,32	91,69	82,35	91,69	82,35	91,69	82,35

E. Embedding de palavras

A partir da observação das Tabelas 11 e 12, pode-se indicar o modelo com janela de contexto formada por 3 palavras e dimensão igual a 50 (W3D50) como o melhor modelo para as três entidades, já que não seria viável usar um modelo para cada entidade. Esse modelo foi a que apresentou o melhor desempenho macro, considerando-se as métricas acurácias, precisão, *recall* e *F1-measure*.

Os resultados relacionados à não-entidade "O" não estão explicitados nas tabelas, contudo foram utilizados para o cálculo dos valores exibidos nelas. Isso porque tudo que é classificado como "O", por mais que não seja interpratado como uma entidade de interesse, é parte do texto que possibilita a classificação das outras entidades. Esses dados foram obtidos utilizando-se a base de treino.

F. Reconhecimento de Entidades Nomeadas (REN)

Observando-se os gráficos nas Figuras 5, 6, 7 e 8 — que apresentam as saídas relacionadas à base de teste —, percebe-se que o modelo de REN [3] combinado ao WE gerado neste trabalho (W3D50) apresentou resultados notoriamente superiores no reconhecimento das entidades LOCALIZAÇÃO,

TABELA XI: RESULTADOS DAS MÉTRICAS DE TREINO DAS DIFERENTES ARQUITETURAS PARA AS ENTIDADES **LOCALIZAÇÃO** E **PESSOA** EM % (Ac = Acurácia, Pr = Precisão, Re = *Recall*, F1 = *F1-score*).

Entidade	Localização				Pessoa			
Métrica	Ac	Pr	Re	F1	Ac	Pr	Re	F1
W3D25	95,99	79,35	73,29	76,20	96,08	79,88	62,02	69,82
W3D50	96,42	81,28	76,71	78,93	96,00	77,74	62,98	69,59
W3D100	96,03	78,57	75,10	76,80	95,85	74,25	65,87	69,81
W3D200	95,45	71,40	79,72	75,33	95,92	74,73	66,83	70,56
W3D300	96,13	80,98	72,69	76,61	96,21	80,86	62,98	70,81
W5D25	95,49	78,45	67,27	72,43	95,44	72,57	61,06	66,32
W5D50	94,76	67,75	79,32	73,08	95,29	69,40	67,07	68,22
W5D100	96,34	83,88	72,09	77,54	95,73	75,60	61,06	67,55
W5D200	96,26	81,88	73,49	77,46	95,74	75,00	62,02	67,89
W5D300	95,83	75,39	77,51	76,44	96,13	79,52	63,46	70,59
W7D25	95,50	75,74	72,09	73,87	94,88	63,92	68,99	66,36
W7D50	95,77	77,73	72,89	75,23	95,23	68,73	63,94	66,25
W7D100	96,19	80,26	75,10	77,59	95,48	71,24	63,70	67,26
W7D200	96,23	80,34	75,50	77,85	95,76	74,44	63,70	68,65
W7D300	96,47	84,62	72,89	78,32	95,93	78,86	60,10	68,21
W9D25	95,62	76,04	73,29	74,64	95,51	72,13	63,46	67,52
W9D50	95,23	70,83	77,51	74,02	95,56	73,50	62,02	67,28
W9D100	95,81	78,56	72,09	75,18	95,33	69,13	65,14	67,08
W9D200	95,91	78,99	72,49	75,60	95,69	74,29	62,50	67,89
W9D300	96,13	80,31	73,69	76,86	96,01	76,70	64,90	70,31
W11D25	95,36	74,17	72,09	73,12	95,39	70,70	63,22	66,75
W11D50	96,06	79,96	73,69	76,70	95,49	72,52	61,54	66,58
W11D100	96,25	80,91	74,90	77,79	95,53	74,10	59,13	65,78
W11D200	95,95	78,39	74,30	76,29	95,37	69,41	64,90	67,08
W11D300	96,35	83,56	72,49	77,63	95,82	74,93	63,94	69,00

TABELA XII: RESULTADOS DAS MÉTRICAS DE TREINO DAS DIFERENTES ARQUITETURAS PARA A ENTIDADE **TEMPO** E O RESULTADO MACRO EM %.

Entidade	Tempo				Macro			
Métrica	Ac	Pr	Re	F1	Ac	Pr	Re	F1
W3D25	98,18	68,85	57,14	62,45	95,47	91,34	91,34	91,34
W3D50	98,44	74,19	62,59	67,90	95,82	91,97	91,97	91,97
W3D100	98,43	72,06	66,67	69,26	95,57	91,52	91,52	91,52
W3D200	98,34	68,46	69,39	68,92	95,37	91,15	91,15	91,15
W3D300	98,40	71,97	64,63	68,10	95,72	91,79	91,79	91,79
W5D25	98,07	62,73	68,71	65,58	94,89	90,28	90,28	90,28
W5D50	96,00	38,89	85,71	53,50	93,60	87,97	87,97	87,97
W5D100	98,38	76,15	56,46	64,84	95,60	91,57	91,57	91,57
W5D200	98,42	72,87	63,95	68,12	95,68	91,72	91,72	91,72
W5D300	98,35	67,97	70,75	69,33	95,56	91,50	91,50	91,50
W7D25	98,23	67,86	64,63	66,20	94,82	90,15	90,15	90,15
W7D50	98,13	65,71	62,59	64,11	95,04	90,55	90,55	90,55
W7D100	98,36	74,14	58,50	65,40	95,43	91,25	91,25	91,25
W7D200	98,13	66,93	57,82	62,04	95,55	91,47	91,47	91,47
W7D300	98,30	70,00	61,90	65,70	95,71	91,77	91,77	91,77
W9D25	98,10	66,67	57,14	61,54	94,99	90,47	90,47	90,47
W9D50	98,38	73,77	61,22	66,91	94,98	90,43	90,43	90,43
W9D100	98,39	72,66	63,27	67,64	95,15	90,75	90,75	90,75
W9D200	98,31	74,77	54,42	62,99	95,44	91,27	91,27	91,27
W9D300	98,49	76,03	62,59	68,66	95,67	91,70	91,70	91,70
W11D25	98,37	69,13	70,07	69,59	95,13	90,72	90,72	90,72
W11D50	98,23	68,70	61,22	64,75	95,29	91,00	91,00	91,00
W11D100	98,43	74,19	62,59	67,90	95,52	91,42	91,42	91,42
W11D200	98,47	71,83	69,39	70,59	95,43	91,25	91,25	91,25
W11_D300	98,32	69,34	64,63	66,90	95,69	91,74	91,74	91,74

PESSOA e TEMPO, comparando-se com os resultados obtidos na combinação desse mesmo modelo com o WE do RWE do NILC.

G. Entidades formadas por mais de um termo

A Tabela 13 exibe o quantitativo da base teste para as entidades formadas com diferentes quantidades de termos (desde entidades formadas com apenas um termo até entendidades formadas por cinco termos). Pode-se tomar como um exemplo de entidade composta por três termos "Rio de Janeiro", que normalmente representa uma localização nos textos. Nesse caso, se o modelo classifica apenas parte dos termos: "Rio", "Janeiro", "Rio de" ou "de Janeiro", como Localização, o modelo teria arcertado apenas parcialmente a classificação dessa entidade em qualquer um desses casos, já que a expressão "Rio de Janeiro" possui três termos. Assim, para que essa expressão fosse corretamente classificada, todos os três termos teriam que ser classificados pelo modelo como Localização.

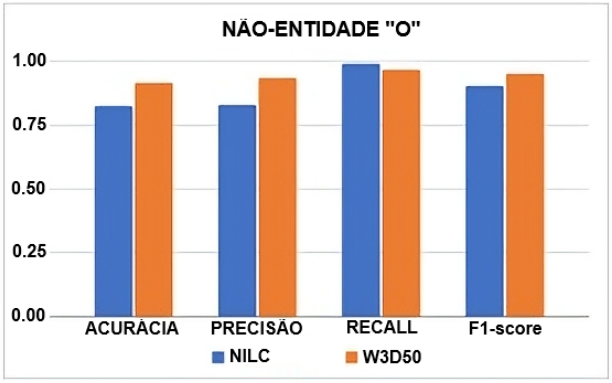

Fig. 5. Distribuição da acurácia, precisão, *recall* e *F1-measure* para a arquitetura eleita e para a arquitetura proposta em [3] verificados para a não-entidade "O".

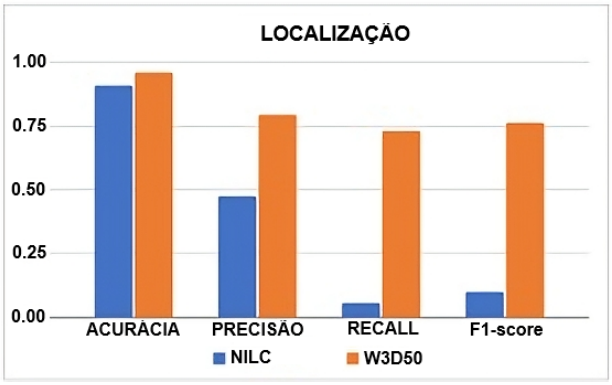

Fig. 6. Distribuição da acurácia, precisão, *recall* e *F1-measure* para a arquitetura eleita e para a arquitetura proposta em [3] verificados para a entidade "LOCALIZAÇÃO".

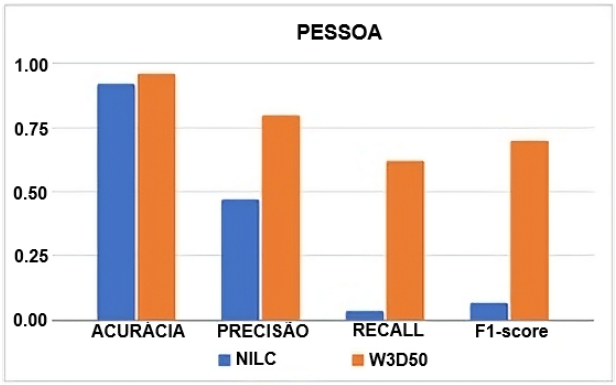

Fig. 7. Distribuição da acurácia, precisão, *recall* e *F1-measure* para a arquitetura eleita e para a arquitetura proposta em [3] verificados para a entidade "PESSOA".

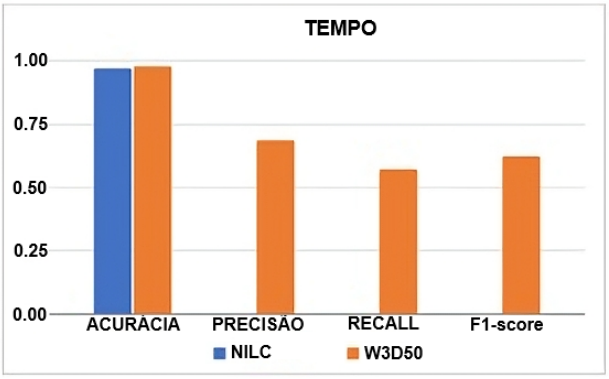

Fig. 8. Distribuição da acurácia, precisão, *recall* e *F1-measure* para a arquitetura eleita e para a arquitetura proposta em [3] verificados para a entidade "TEMPO".

A não-entidade "O" foi mantida na Tabela 13 para possibilitar a observação do volume de termos não pertencentes a nenhuma das três entidades nomeadas avaliadas neste trabalho.

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

Nº de palavras	Pessoa	Tempo	Localização	O
1	320	24	147	220
2	18	14	48	136
3	8	9	21	120
4	4	7	18	85
5	4	8	24	661
Total de termos por entidade	416[a]	147	498	4497

[a]fórmula: 1×320 + 2×18 + 3×8 + 4×4 + 5×4 = 416

As Tabelas 14, 15, 16, 17 e 18 têm como objetivo apresentar o desempenho do modelo de *embedding* W3D50 para a base de teste, considerando a classificação das entidades compostas desde um até cinco termos. Esses valores apontam quantos desses termos em cada expressão são corretamente preditos. Para facilitar a leitura, os valores da Tabela 13, que indicam o total de termos por expressão para cada entidade, foram acrescentados entre parênteses ao lado das entidades indicadas em cada uma das Tabelas 14 à 18. O conteúdo das tabelas 19, 20, 21, 22 e 23 têm esse mesmo objetivo, porém relativo ao modelo de *embedding* do NILC.

TABELA XIV – COMPOSTAS POR **1 TERMO**.

	Classificação	
Entidade	0/1	1/1
Pessoa (320)	123	197
Tempo (24)	12	12
Localização (147)	32	115

TABELA XV – COMPOSTAS POR **2 TERMOS**.

	Classificação		
Entidade	0/2	1/2	2/2
Pessoa (18)	3	10	5
Tempo (14)	5	5	4
Localização (48)	7	10	31

As Tabelas 14, 15, 16, 17, 18, 19, 20, 21, 22 e 23 apontam os acertos e erros de classificação para cada entidade composta por 1, 2, 3, 4 e 5 termos, respectivamente. Essa avaliação ajuda a indicar se o modelo usado para classificação das entidades consegue identificar corretamente expressões formadas por mais de uma palavra. Vale evidenciar que a base rotulada utilizada para treino e teste do modelo de REN, não

TABELA XVI – COMPOSTAS POR **3 TERMOS**.

	Classificação		
Entidade	0/3	1/3	2/3
Pessoa (8)	2	1	1
Tempo (9)	3	1	4
Localização (21)	3	2	1

TABELA XVII – COMPOSTAS POR **4 TERMOS**.

	Classificação			
Entidade	0/4	1/4	2/4	3/4
Pessoa (4)	1	2	0	0
Tempo (7)	0	1	1	2
Localização (18)	0	1	1	4

TABELA XVIII – COMPOSTAS POR **5 TERMOS**

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

TABELA XIX – COMPOSTAS POR **1 TERMO**.

	Classificação	
Entidade	0/1	1/1
Pessoa (320)	306	14
Tempo (24)	24	0
Localização (147)	131	16

TABELA XX – COMPOSTAS POR **2 TERMOS**.

	Classificação		
Entidade	0/2	1/2	2/2
Pessoa (18)	17	1	0
Tempo (14)	14	0	0
Localização (48)	44	4	0

TABELA XXI – COMPOSTAS POR **3 TERMOS**.

	Classificação		
Entidade	0/3	1/3	2/3
Pessoa (8)	8	0	0
Tempo (9)	9	0	0
Localização (21)	20	1	0

TABELA XXII – COMPOSTAS POR **4 TERMOS**.

	Classificação			
Entidade	0/4	1/4	2/4	3/4
Pessoa (4)	4	0	0	0
Tempo (7)	7	0	0	0
Localização (18)	16	2	0	0

TABELA XXIII – COMPOSTAS POR **5 TERMOS**

	Classificação					
Entidade	0/5	1/5	2/5	3/5	4/5	5/5
Pessoa (4)	4	0	0	0	0	0
Tempo (8)	8	0	0	0	0	0
Localização (24)	19	5	0	0	0	0

contém indicação de início e fim de palavra, conforme algumas metodologias de rotulagem (BIO e BILOU [11]).

Desse modo, 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 compõem a entidade, conseguiu ser identificado pelo modelo de maneira correta, seja para Pessoa, Tempo ou Localização.

Percebe-se que para entidades compostas por um único termo (Tabelas 14 e 19), a entidade Tempo foi a que apresentou menos acertos em ambos os casos. A partir das Tabelas 15 e 20, nota-se uma discrepância grande entre as combinações do modelo de REN com os *embeddings* da arquitetura W3D50 e do NILC. A taxa de acerto do embedding do NILC é bem menor que do embedding W3D50 considerando a característica da existência de entidades compostas por mais de um termo.

V. Conclusão

Este trabalho teve como principal objetivo avaliar o uso de modelo de Reconhecimento de Entidades Nomeadas, com um *word embedding* desenvolvido através de um *corpus* composto por denúncias escritas em língua portuguesa brasileira coloquial com muitos erros gramaticais, pertencente ao contexto de segurança pública. Os resultdos alcançados foram comparados aos resultdos obtidos com esse mesmo modelo de REN com um WE obtido a partir de *corpus* formato por textos com amplo espectro de contexto, escritos tanto em português brasileiro quanto europeu — obtidos de fontes consideradas bem escritas. Destaca-se que em ambos os casos a finalidade era extrair informações úteis para a segurança pública.

Com base nas métricas acurácia, precisão, *recall* e *F1-measure*, notou-se que o reconhecimento das entidades nomeadas PESSOA, LOCALIZAÇÃO e TEMPO, a partir da combinação do modelo com o WE gerado através da base do DD RJ apresentou as seguintes melhoras gerais: 5,34% em acurácia, 9,34% em precisão, *recall* e *F1-measure*.

Vale ressaltar que o WE do NILC apresentou uma quantidade de acerto grande nos verdadeiros negativos (TN), o que fez com que sua acurácia macro fosse de 82,35%, mesmo não conseguindo identificar a entidade TEMPO. Contudo, ao considerar a importante característica de que entidades podem ser compostas por mais de um termo, nota-se que o desempenho do WE do NILC foi bem menor que o do WE W3D50.

Reconhece-se a limitação na construção do WE W3D50 devido à limitação na quantidade de amostras utilizadas para gerá-lo. Faz-se necessário construir um *word embedding* que contenha mais palavras, com a finalidade de disponibilizar mais vetores para o modelo de REN. Isso ocorreu devido ao fato dele ter sido gerado utilizando-se uma base reduzida validada manualmente. Essa base reduzida seguramente prejudicou também o desempenho do WE do NILC.

Como trabalhos futuros, deseja-se expandir a extração de entidades para textos compartilhados em redes sociais, e em bases de dados pertencentes a outros órgãos relacionados à segurança pública. Acrescenta-se a melhoria no modelo de *word embedding* com um *corpus* mais adequado, além do uso de modelos baseados em *Tranformers* [12] [13], que incluem o aprendizado da vetorização em conjunto com o aprendizado do reconhecimento de entidades nomeadas.

References

[1] HARTMANN, N. et al. *Portuguese word embeddings: Evaluating on word analogies and natural language tasks.* arXiv preprint arXiv:1708.06025, 2017.

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

[3] FIGUEIREDO, K.; SOUSA, Yago G. T. de; BARRETO, Leonidia dos S.; BRITO, Walkir A. T. Estudo de Modelo Deep Learning para Reconhecimento de Entidades Nomeadas na Segurança Pública. CBIC 2021.

[4] D. Jurafsky. 2000. *Speech & Language Processing*. Pearson Education India.

[5] WANG, B.; WANG, A.; CHEN, F.; WANG, Y.; KUO, C. C. J.. *Evaluating Word Embedding Models: Methods and Experimental Results.* APSIPA transactions on signal and information processing, v. 8, 2019.

[6] 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.

[7] 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.

[8] MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. *Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR,* 2013.

[9] WANG, B.; WANG, A.; CHEN, F.; WANG, Y.; KUO, C. C. J.. *Evaluating Word Embedding Models: Methods and Experimental Results. APSIPA transactions on signal and information processing,* v. 8, 2019.

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

[11] Ratinov, L. e Roth, D. (2009) "Design Challenges and Misconceptions in Named Entity Recognition", In Proceedings of the 13th Conference on Computational Natural Language Learning.

[12] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia, Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota.

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

[15] 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.