UNIVERSIDADE FEDERAL DE SANTA CATARINA, PARA A OBTENÇÃO DO TÍTULO
DE MESTRE EM CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Rogerio de Aquino Silva

UMA METODOLOGIA PARA CRIAÇÃO DE UM CORPUS TEXTUAL ADEQUADA AO RECONHECIMENTO DE ENTIDADES NOMEADAS EM PORTUGUÊS

Florianópolis
2021

Rogerio de Aquino Silva

UMA METODOLOGIA PARA CRIAÇÃO DE UM CORPUS TEXTUAL ADEQUADA AO RECONHECIMENTO DE ENTIDADES NOMEADAS EM PORTUGUÊS

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Santa Catarina, para a obtenção do título de Mestre em Ciência da Informação.
Orientador: Prof. Dr. Gustavo Medeiros de Araújo

Florianópolis
2021

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

silva, rogerio
UMA METODOLOGIA PARA CRIAÇÃO DE UM CORPUS TEXTUAL ADEQUADAAO RECONHECIMENTO DE ENTIDADES NOMEADAS EM PORTUGUÊS / rogerio silva ; orientador, Gustavo Medeiros de Araújo, 2021.
116 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós Graduação em Ciência da Informação, Florianópolis, 2021.

Inclui referências.

1. Ciência da Informação. 2. Processamento de texto. I. Medeiros de Araújo, Gustavo . II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Informação. III. Título.

ROGÉRIO DE AQUINO SILVA

Título: UMA METODOLOGIA PARA CRIAÇÃO DE UM CORPUS TEXTUAL ADEQUADAAO RECONHECIMENTO DE ENTIDADES NOMEADAS EM PORTUGUÊS

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Dr. Gustavo Medeiros de Araújo
Instituição PGCIN/UFSC

Dr. Moisés Lima Dutra
Instituição PGCIN/UFSC

Dr. Cristian Cechinel
Instituição (PPGTIC/UFSC

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em CIÊNCIA DA INFORMÇÃO.

Documento assinado digitalmente
Edgar Bisset Alvarez
Data: 16/08/2021 17:55:31-0300
CPF: 236.516.228-27
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Gustavo Medeiros de Araujo
Data: 16/08/2021 15:56:22-0300
CPF: 828.754.991-49
Verifique as assinaturas em https://v.ufsc.br

Dr. Gustavo Medeiros de Araújo
Orientador(a)

FLORIANÓPOLIS, 2021.

RESUMO

A extração de entidades nomeadas é a tarefa de recuperação de informações presentes em um texto e de classificação dessas informações em categorias predefinidas, tais como pessoas, empresas, locais, valores monetários, porcentagens e datas. Diante da grande quantidade de dados não estruturados, por exemplo, documentos de texto, postagens e e-mail, que são gerados a todo momento durante a utilização dos meios digitais, torna-se necessária a criação de ferramentas de mineração de texto que possibilitem a transformação de dados em informação. Hoje grande parte dos modelos que possuem acurácia acima de 90% no processo de extração de entidades, são criados a partir do idioma inglês. Isso ocorre, em parte, devido à quantidade de dados disponíveis para treinamento de um modelo, pois, para sua criação, é necessário que exista um conjunto de documentos que são, conhecidos como corpus, com trechos de textos que possuam as anotações das entidades contidas. Hoje, parte dos corpora públicos que existem na língua portuguesa não possuem anotações. Sendo assim, esta dissertação propõe uma metodologia para a criação de um corpus anotado em português para o reconhecimento de entidades nomeadas. O objetivo da metodologia proposta é a criação de um corpus adequadamente anotado para treinar modelos no reconhecimento de entidades nomeadas. Dessa forma, este trabalho visa aproximar a acurácia de extração de entidades nomeadas dos modelos encontrados na literatura em outros idiomas que possuem resultados próximos a 90%. Acurácia é quantidade de entidades corretas extraídas pelo modelo em relação à quantidade total de entidades existentes. Os resultados preliminares do modelo proposto neste trabalho mostram que a utilização de um conjunto de técnicas, como a limpeza e a padronização dos dados de treino e o uso de redes neurais recorrentes, permite chegar a 85,63% de acurácia. A metodologia proposta abordada os aspectos da arquitetura implementada, bem como a metodologia de testes do projeto. No projeto, são utilizados corpora contendo trechos de textos e anotações de palavras a partir notícias jornalísticas. As entidades extraídas são nomes próprios do tipo Local, Pessoa e Organização.

Palavras-chave: Corpus; Extração de Entidades Nomeadas; Redes Neurais Artificiais; Processamento de Linguagem Natural; Recuperação da Informação.

ABSTRACT

The extraction of named entities is the task of retrieving information present in a text and classifying it in predefined categories, such as people, companies, places, monetary values, percentages and dates. In view of the large amount of unstructured data, such as text documents, posts and e-mail, which are generated at all times during the use of digital media, it is necessary to create text mining tools that enable the transformation of data into information. Today, most of the models that have an accuracy above 90 % in the entity extraction process, are the models created from the English language. This occurs, in part, due to the amount of data available for training a model, since for its creation it is necessary to have a set of documents, which are known as corpus, with excerpts of texts that have the notes of the entities contained in them. Today, part of the public corpora that exist in the Portuguese language have no notes. Therefore, this dissertation proposes a methodology for creating a corpus in Portuguese annotated for the recognition of named entities. The objective of the proposed methodology is to create a properly annotated corpus to train models in the recognition of named entities. Thus, the aim is to approach the accuracy of extracting named entities from the models found in the literature in other languages that have results close to 90 %. Accuracy is the number of correct entities extracted by the model in relation to the total number of existing entities. The preliminary results of the model proposed in this work show that the use of a set of techniques, such as the cleaning and standardization of training data and the use of recurrent neural networks, is possible to reach 85.63 % accuracy. The proposed methodology addressed aspects of the implemented architecture, as well as the project's testing methodology. The project uses corpus containing excerpts from texts and annotations of words from journalistic news. The extracted entities are proper names of the type Local, Person and Organization.

Keywords: Corpus; Extraction of named entities Artificial Neural Networks; Natural Language Processing; Information Retrieval.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A recuperação de informação (RI) surgiu a partir de esforços para facilitar a manipulação de dados em grandes bases (MOOERS, 1960). Mesmo com esforços empregados na areá de recuperação de informação ainda existem grandes desafios no que se refere ao processamento de linguagem natural (PLN). A extração de informações a partir de textos em português ainda constitui um campo de investigação aberto, devido à baixa acurácia no reconhecimento de entidades, na análise semântica, sintática e na extração de relacionamentos (AMARAL; VIEIRA, 2014). A acurácia é determinada pela quantidade de classificações corretas. Somado ao desafio do processamento de linguagem natural para o português, há uma escala crescente de dados gerados, devido ao sucessivo aumento no número de usuários da internet.

Uma vez que a existência de bases (corpus) para criação de modelos é menor, muitas vezes desatualizadas e com estruturas complexas, quando comparadas a outros idiomas, como o inglês, há uma escala crescente de dados gerados, devido ao sucessivo aumento no número de usuários da internet (IBGE, 2017). Estima-se que a cada segundo são gerados milhares de dados em uma diversidade de formatos (MOSLEY *et al.*, 2017), tais como imagens, textos, vídeos e áudios.

Um estudo realizado pelo Data Management Association (Dama) constatou que ainda em 2005 existiam mais de 500 quatrilhões de Megabytes em dados armazenadas no universo digital. O estudo também aponta que, a cada dois anos, a produção de dados dobra, com previsão de chegar a 350 zettabytes em 2020 (MOSLEY *et al.*, 2017). Grande parte dos dados gerados, cerca de 80%, são dados não estruturados, como, por exemplo, texto em linguagem natural. A crescente escala de produção de textos, juntamente com a baixa acurácia para o reconhecimento de entidades nomeadas em português, envolve questões que podem ser enfrentadas com técnicas modernas de aprendizagem de máquina, como redes neurais artificiais, árvore de decisão, máquinas de vetor de suporte e redes bayesianas.

Com a utilização de técnicas de aprendizagem de máquina, é possível criar modelos com capacidade de extrair entidades a partir de textos. Na língua inglesa, por exemplo, existem modelos que possuem acurácia de 92,6%, criados a partir da utilização do framework spaCy; e 91,7%, com o framework ClearNLP. Segundo Amaral e Vieira (2014), em português, a acurácia fica em torno de 80,77%, utilizando como base de treinamento o corpus HAREM e redes neurais do tipo *Conditional random fields*(CRF). Para um modelo atingir um alto desempenho, é necessário um conjunto de dados previamente classificados com notações sobre a sua estrutura gramatical e suas entidades. Esse conjunto é conhecido como corpus textual, e a atividade de classificar os trechos de textos é realizada por linguistas conhecedores da estrutura do idioma. Entretanto, como cada corpus é criado para uma finalidade específica e

nem sempre possui a mesma estrutura, não existe um padrão. Além disso, (VILLALVA; MATEUS, 2008) ressalta que a morfologia e a sintaxe da língua portuguesa possuem características próprias criando complexidades quando essa língua é comparada com a língua inglesa, que possui menos elementos em sua notação gramatical quanto à conjugação de verbos.

O domínio da informação também é uma questão a ser considerada, pois conforme Dias (2015) uma determinada comunidade pode possuir hábitos em relação ao uso da informação, isto é, ao modo como seus membros realizam buscas e organizam novos conhecimentos. Esses hábitos afetam a escrita e a estrutura sintática, mesmo que o idioma seja igual em comunidades diferentes. A linguagem jornalística é o gênero textual que possui melhor aderência à contemporaneidade do idioma, algumas das características da escrita jornalística, como a objetividade, a simplicidade, a imparcialidade e o referencial, evitam termos em desuso, pois a informação deve ser transmitida de forma clara ao leitor (PRETTO, 2009). Por esse motivo, foram escolhido o corpora textuais baseados em textos jornalísticos a fim de criar uma base de treinamento para o reconhecimento de entidades nomeadas.

1.1 JUSTIFICATIVA

A partir do emprego de redes neurais artificiais, é possível realizar a extração de entidades nomeadas, podendo extrair entidades como: nome, documento, telefone, conta bancária, nome de instituição etc. Também é possível determinar o relacionamento existente entre as entidades, por exemplo, a qual entidade "NOME"pertence a entidade "DOCUMENTO"localizada em um texto. Dessa forma, é viável estruturar dados que poderão ser utilizados em diversas outras áreas e aplicações como, comunicação entre sistemas de órgãos públicos governamentais.

Com a automatização da extração de entidades por meio de técnicas de aprendizagem de máquina, os benefícios serão: i) redução do tempo necessário para que a informação chegue do ponto de origem ao ponto de destino; ii) a redução de erros nos cadastros que hoje são realizados manualmente. Além disso, o serviço pode ser utilizado ininterruptamente a qualquer hora, diferente dos cadastros manuais, que são realizados por pessoas; iii) redução de gastos pois um servidor pode realizar mais extrações que varias pessoas juntas por um valor bem menor; iv) cadastramento automático de formulários, enviando documentos ou textos através de uma interface de programação de aplicações (API), posteriormente processados pelo modelo gerado através do processamento de linguagem natural e receber os dados estruturados para o preenchimento automático.

Hoje existe uma carência de dados que possuam uma base com estrutura linguística em larga escala, muitas vezes obrigando a criação de modelos e treinamentos específicos para uma única problemática ou domínio (não existindo uma base padrão

para utilização e treinamento). Um exemplo de corpus que pode ser citado para criação de um modelo de extração de entidade é o Corpora em língua portuguesa, porém, em alguns aspectos, ele não é aderente ao português falado, devido ao domínio da linguagem, uma vez que usa trechos de notícias de jornais, que reduz a eficácia do modelo em documentos usados no dia a dia ou na análise de textos de redes sociais, que possuem uma linguagem mais contemporânea próximo ao português falado (IBPAD, 2018).

1.2 DELIMITAÇÃO DO PROBLEMA

Para o presente estudo, formulou-se a seguinte questão problema:

É possível criar uma metodologia de extração de entidades nomeadas, a partir de um corpus jornalístico escrito no idioma português, que apresente uma acurácia próxima aos valores gerados em modelos baseados no idioma inglês?

Dessa forma, o problema a ser trabalhado será o desenvolvimento de uma metodologia contendo um novo corpus para a extração de entidades nomeadas em português. Não serão trabalhadas questões como identificação sintática e semântica e outras características de mineração de texto.

1.3 HIPÓTESE

A partir da extração e organização de trechos de textos contidos no corpus CETENFolha, é possível criar um modelo de extração de entidades, por meio do uso de redes neurais artificiais, que poderá aumentar a acurácia no Reconhecimento de Entidades Nomeadas (REN). A escolha do corpus CETENFolha é justificada por ser de domínio público e por possuir anotações relacionadas a entidades contidas nos trechos. Além disso, o corpus CETENFolha possui o volume de 340.947 trechos e 25.475.272 palavras de artigos jornalísticos dos anos 1994 e 1995, extraídos do jornal Folha de São Paulo (LINGUATECA, 2018).

1.4 OBJETIVOS

1.4.1 Objetivo geral

O objetivo geral do projeto é a criação de uma metodologia que guie a geração de um modelo de extração de entidades nomeadas no idioma português. Dessa forma, o modelo pode ser treinado a partir de um corpus de domínio público e gratuito, a fim de que detenha à acurácia próxima a acurácia encontrada nos modelos já existentes no idioma inglês.

1.4.2 Objetivos específicos

a) Criar uma base de dados contendo trechos de textos em português com *tags* que indiquem as entidades e sua devida posição a partir do processamento do corpus CETENFolha.

b) Criar um modelo base de extração de entidades utilizando o domínio jornalístico.

c) Analisar, por meio das métricas; acurácia, F1 e *Recall*, quais arquiteturas de redes neurais e parâmetros se adéquam melhor à tarefa de criação do modelo.

d) Criar uma arquitetura que possibilite a reutilização dos modelos já criados e permita extrair entidades de novos trechos de textos não existentes na base.

2 REVISÃO DE LITERATURA

Nas próximas seções, serão abordadas algumas linhas de pensamento encontradas na literatura, em relação aos aspectos gerais, dados e tecnologia empregadas durante o decorrer do estudo.

2.1 CORPUS

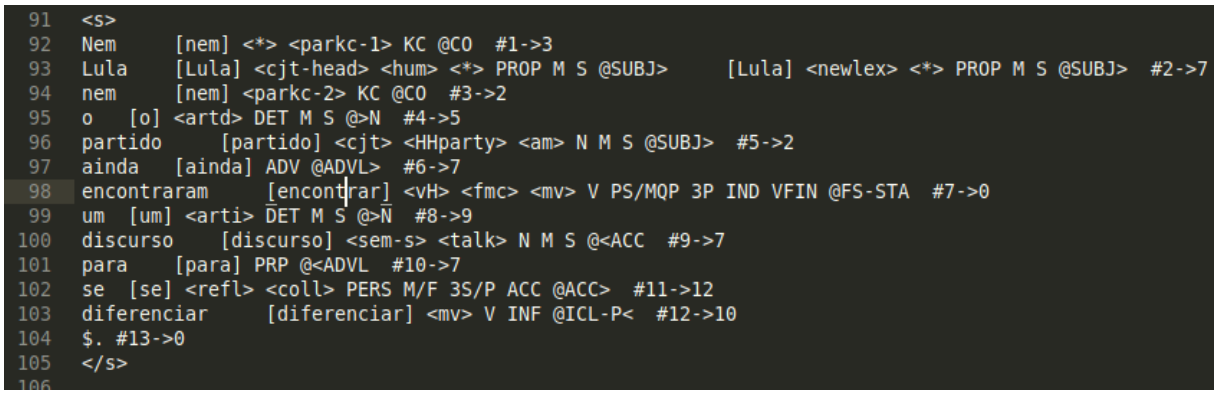
Corpus pode ser definido, conforme o dicionário Aurélio, como um conjunto de documentos ou textos (CORPUS, 2020). O primeiro corpus eletrônico público conhecido foi o *Brown University Standard Corpus of Present-Day American English* lançado em 1964, era escrito na língua inglesa e possuía um milhão de palavras (SARDINHA, 2000).

Os corpora são construídos de forma minuciosa e escritos de acordo com o objetivo ao qual são destinados, conforme sugerem Tagnin e Teixeira (2004). A língua é um sistema probabilístico, pois, mesmo que um determinado termo possua a sintaxe correta, como em *amigo próximo*, é mais provável que ocorra com maior frequência o uso de *amigo íntimo* no idioma português brasileiro, dentro do mesmo contexto.

Para criação de um modelo de extração de entidades, é necessário que o volume de trechos seja representativo, pois a rede neural artificial aprende conforme encontra os padrões de escrita. Uma vez que não existe um consenso que defina um tamanho mínimo ou máximo de um corpus (SARDINHA, 2000), é necessário observar o objetivo da pesquisa e o idioma utilizado, pois, quanto maior a variação da escrita, maior deverá ser o volume de trechos.

Entre as anotações utilizadas em um corpus, podemos citar como a mais utilizada a morfossintática, conhecida pelo termo *part-of-speech* (POS) (TAGNIN; TEIXEIRA, 2004). Na anotação morfossintática, cada palavra possui uma anotação denominada *POS tagging*, com a informação da classe gramatical da palavra no contexto, podendo identificar, por exemplo, palavras estrangeiras, entidades nomeadas, termos técnicos, cor, sentimentos e roupas. As anotações são criadas conforme o objetivo do corpus.

Figura 1 – Trecho do corpus CETENFolha



Fonte: Elaborada pelo autor (2020)

Conforme "Figura 1 - Trecho do Corpus CETENFolha"é possível ver um trecho extraído do corpus CETENFolha com anotações de classe gramatical para cada palavra, conjugação, infinitivo da palavra e identificação de plural ou singular. Quando existem entidades, elas são classificadas conforme o Quadro 1 - Tipos de entidades.

Quadro 1 – Tipos de entidades

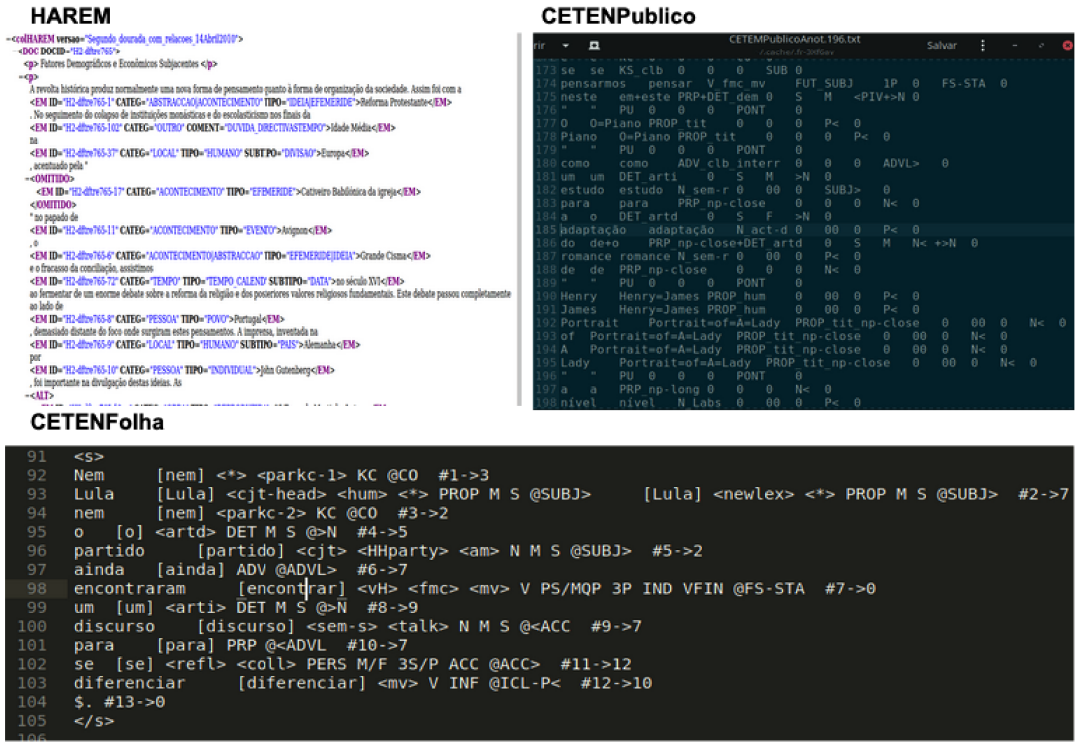
ANOTAÇÃO	ENTIDADE	CONTEÚDO
<hum>	PER	nome de pessoas
<civ>	LOC	paises, estados, cidades, bairros, ruas
<temp>	DATE	datas, dias da semana
<org>	ORG	organizações, empresas

Fonte: Elaborado pelo autor

Devido as anotações dos trechos em um corpus, serem anotadas conforme o objetivo do corpus, muitas vezes se faz necessário realizar o tratamento para padronizá-lo. Para o caso do corpus em questão, conforme "Quadro 1 - Tipos de Entidades", são localizadas as seguintes anotações: <hum>, <civ>, <temp> e <org>, que são as anotações criadas para as entidades pessoa, local, data e organização. Durante o tratamento das informações do corpus, é realizada uma conversão para o tipo de anotação mais utilizado nas tarefas de extração de entidades nomeadas. Dessa forma, passaram a ser anotados como PER, LOC, DATE e ORG, assim, caso seja necessário inserir novos trechos de uma nova origem que possuam outro tipo de anotação, o processo será repetido.

Conforme a Figura 2 - Comparativo entre a estrutura dos Corpus existentes, é possível notar a diferença na estrutura de três corpus, HAREM, CETENPublico e CETENFolha. No HAREM as anotações são realizadas em estrutura XML *(Extensible Markup Language)*, possui a anotação de entidade e tipo de entidade. No corpus CE-

Figura 2 – Comparativo entre a estrutura dos Corpus existentes



Fonte: Elaborada pelo autor (2021)

Tabela 1 – Quantidade de trechos por Corpus

CORPUS	TAMANHO
CETENFolha	340.947
CETENPublico	234.483.623
HAREM	290.001

Fonte: Elaborada pelo autor

TENPublico, os trechos possuem anotações de classe gramátical, porem não existem anotações de entidades nomeadas. Já o corpus CETENFolha, tem como diferencial, anotações tanto de classe gramatical quanto de entidades.

Devido ter uma grande quantidade de trechos para realizar o treinamento da rede neural artificial conforme Tabela 1 - Quantidade de trechos por Corpus, e possuir uma estrutura mapeada e simplificada conforme Figura 2, com as devidas anotações de entidades, foi escolhido para o estudo o corpus CETENFolha.

2.2 MINERAÇÃO DE TEXTO

O termo mineração de texto, como é conhecido hoje, foi descrito inicialmente como descoberta de conhecimento em textos (FELDMAN; DAGAN, 1995). A ideia central é a utilização de técnicas semiautomáticas para extração do conhecimento a partir de dados não estruturados em formato de textos escritos em linguagem natural. A mineração de texto é uma subaérea da mineração de dados, basicamente o que a difere da mineração de dados é o tipo de dado em que são empregadas as técnicas de mineração, serão estruturados ou não estruturados (REZENDE, 2003).

O processo de mineração de dados possui 5 etapas principais, são elas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento (REZENDE, 2003). Essas etapas são apresentadas de forma mais detalhada a seguir.

1. **Identificação do problema**: são identificadas as necessidades de automatização de tarefas relacionadas ao tratamento do texto.

2. **Pré-processamento**: nesta etapa é realizada a normalização do texto, são extraídas as palavras com menor peso semântico, por exemplo, a, e, de e ou, que são conhecidas como *stopwords*. Em alguns casos são retirados os acentos, além disso, as letras são alteradas de forma que sejam todas maiúsculas ou minusculas e são retirados os caracteres especiais.

3. **Extração de padrões**: são utilizados métodos para definir padrões a fim de automatizar a tarefa, por exemplo, extrair um código ou uma palavra que apareça sempre em alguma posição do texto e que possua o mesmo padrão inicial ou final que possa identificar sua posição para extração. Um exemplo de técnica usada para realizar a extração é o *regex*.

4. **Pós-processamento**: são realizadas validações para verificar e validar a veracidade do dado, se foi extraído da forma correta, ou normalizar o dado para inserção em uma nova base de dados já estruturada.

5. **Utilização do conhecimento**: nesta etapa os dados já foram processados e já são utilizados como informação, por exemplo, uma data de nascimento, um valor monetário uma localização geográfica.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Processamento de linguagem natural (PLN) pode ser definido como o emprego de técnicas computacionais para criação de modelos capazes de extrair informações de algum tipo de linguagem natural (MICHAEL, 2009). Conforme definido por Vieira e

Lopes (2010), é a área da computação que estuda o desenvolvimento de tecnologias para reconhecer ou gerar textos em linguagens humanas, que possui como foco cinco níveis de estudo da linguagem, são eles:

1. **Sonético ou Fonológico**: responsável pelo processamento da língua falada ou sons.
2. **Morfológico**: responsável pelo processamento das palavras contidas em um texto e sua classe gramatical.
3. **Sintático**: responsável pelo processamento da estrutura do texto.
4. **Semântico**: responsável pelo significado do textos processado.
5. **Pragmático**: responsável pelo processamento do contexto em que o texto está inserido.

Conforme definido por Carvalho (2012), o processamento de linguagem natural tem como objetivo aproximar o homem da máquina, desenvolvendo ferramentas que possibilitem uma comunicação mais natural.

Dentro da temática de processamento de linguagem natural, existe a técnica de reconhecimento de entidades nomeadas que realiza o reconhecimento de entidades presentes em um texto e viabiliza o processamento da linguagem humana para criação de modelos.

2.3.1 Reconhecimento de entidades nomeadas

O reconhecimento de entidades nomeadas (REN), também encontrado no idioma inglês como *Named-entity recognition (NER)* ou Reconhecimento de entidade nomeada em tradução livre, é descrito como a tarefa de análise semântica de um texto que busca localizar e classificar trechos em categorias já predefinidas, como, nome de pessoas, lugares, organizações, datas, eventos ou outras categorias que possam ser desejadas (CARVALHO, 2012). A partir de técnicas de processamento de linguagem natural (PLN), é realizado o processamento automático, por meio de abordagens estatísticas, quantitativas e probabilísticas, possibilitando que a entidade seja classificada de acordo com seu significado no contexto geral do texto (CARVALHO, 2012).

Tomando como exemplo o seguinte trecho "Elon Musk deu início à transmissão ao vivo da Neuralink. Visivelmente nervoso, o fundador e CEO da Tesla começou sua fala a mais de 145 mil pessoas reforçando que aquele evento tinha como único objetivo recrutar profissionais", sabemos que uma pessoa foi citada *"Elon Musk"*; *"Neuralink"* pode ser classificado como um evento; e *"Tesla"*, como uma organização.

Segundo Carvalho (2012), a extração de entidades é considerada uma tarefa crucial para áreas que necessitam de extração de informação, como a mineração de

texto. O conhecimento obtido possibilita a execução de tarefas mais complexas, uma vez que os dados não estruturados são transformados em dados estruturados.

Em alguns casos, podem ser criados modelos para extração de entidades de domínios específicos, como na medicina, especializados em classificar vírus, proteínas e genes (CARVALHO, 2012), ou na área jurídica em busca de legislação e jurisprudência (ARAUJO *et al.*, 2018).

2.3.1.1 *POS-Tagging*

O termo *POS-Tagging* pode ser traduzido livremente como marcação da parte da falada, *POS* é a redução de *Part-of-Speech*, ou seja, parte do discurso. É o termo utilizado para definir o processo de classificação gramatical ou sintática dos trechos de um texto ou fala dentro do contexto em que o trecho é inserido no texto (DANIEL; MARTIN, 2008). Na anotação inserida em cada trecho, são inseridas as classificações gramaticais conforme as oito partes da fala, são elas: substantivo, verbo, partes da fala, pronome, preposição, advérbio, conjunção, particípio e artigo (DANIEL; MARTIN, 2008), além de possíveis entidades existentes.

Segundo Daniel e Martin (2008), o papel do *POS-Tagging* para o processamento de linguagem natural é de grande importância, uma vez que através da classe gramatical de uma palavra é possível verificar informações sobre a estrutura sintática das palavras vizinhas, pois, substantivos são precedidos por determinantes e adjetivos, verbos podem ser precedidos substantivos e substantivos são geralmente parte de sintagmas nominais. Sendo assim, é possível identificar trechos de um texto que podem ser recursos úteis para serem rotulados como as entidades nomeadas, como, pessoas e organizações (DANIEL; MARTIN, 2008);

No quadro abaixo é possível visualizar uma relação de anotações encontradas em um corpus. O padrão abaixo é chamado de *Penn Treebank* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993) e contém originalmente 45 anotações gramaticais possíveis, nem todas são aplicáveis ao idioma português, muitos corpora já anotados seguem este padrão quando são realizadas anotações gramaticais, que são inseridas após cada trecho do texto. Porém, podem existir anotações de entidades, em que são inseridas anotações em trechos de acordo com categorias de entidades predefinidas. Para as entidades que são normalmente encontradas em textos, por exemplo, pessoa, organização, data e local, já existem alguns padrões que são adotados, porém nada impede que sejam criados novos tipos de anotações para outros tipos de entidades.

Quadro 2 – Relação de anotações

TAG	DESCRIÇÃO	EXEMPLO	TAG	DESCRIÇÃO	EXEMPLO
CC	conjunção coordenativa	e, porém, ou	POS	final possessivo	s (em inglês)
CD	número cardinal	um, dois, três...	PRP	pronome pessoal	Eu, você, nós...
DT	determinante	a, o	PRP$	pronome possessivo	Meu, Minha, Seu, Sua
EX	existencial	há	RB	advérbio	rapidamente
FW	palavra estrangeira	mea culpa	RBR	advérbio comparativo	melhor, menos, mais
IN	preposição / conj. subordinada	de, em, por	RBS	superlatv. adverb	tão, como, depois
JJ	adjetivo	legal, amável	RP	partícula	se, da, do, de
JJR	Adj. comparativo	maior que, menor que	SYM	símbolo	"+",%, &
JJS	superlativo Adj.	melhor, mais	UH	interjeição	ah, oops
LS	marcador de item da lista	1, 2, 3...	VB	forma de base do verbo	comer, falar
MD	modal	poderia, deveria	VBD	pretérito do verbo	comi, bebi, dirigi
NN	substantivo de massa	àgua, cadeira	VBG	verbo gerúndio	comendo, pulando, brincando
NNS	substantivo, plural	cadeiras, cidades	VBN	verbo particípio passado	comido, ido, bebido
NNP	substantivo próprio	IBM, Brasil, Tesla	VBP	verb 1ª SG pres.	como
NNPS	substantivo próprio, plu.	Carolinas	VBZ	verb 3ª SG pres.	come
PDT	predeterminador	todos, ambos			

Fonte: Adaptado de Daniel e Martin (2008)

2.4 APRENDIZAGEM DE MÁQUINA

Conforme tradução extraída do livro *Machine Learning* de Mitchell (1997), a definição de aprendizagem de máquina é: “Diz-se que um programa de computador aprende com a experiência E com respeito para alguma classe de tarefas T e medida de desempenho P, se o seu desempenho em tarefas em T, medido por P, melhora com a experiência E.” (MITCHELL, 1997, p. 2).

Conforme sugere COPPIN (2004), existem vários métodos de aprendizagem de máquina, entre os dois tipos mais utilizados, é possível citar a aprendizagem por hábito, que basicamente aprende com as opções já classificadas anteriormente; e a aprendizagem por conceito, que analisa todas as hipóteses possíveis e busca a mais provável. Os dois métodos citados utilizam algoritmos matemáticos e probabilísticos em forma de funções.

No momento da escolha de um algoritmo ou método para aplicação em um projeto, deve-se levar em consideração alguns fatores, como, dados de entrada e objetivos a serem atingidos. Quanto aos tipos de aprendizado disponíveis, pode-se dividi-los em três tipos, são eles: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço (GRUS, 2016).

Quadro 3 – Tipos de aprendizagem de máquina e aplicações

Aprendizagem supervisionada	Aprendizagem não supervisionada	Aprendizagem por reforço
Reconhecimento de imagens	Elicitação de atributos	Decisões em tempo real
Retenção de clientes	Visualização de atributos	Tarefas de constante aprendizagem
Diagnósticos	Sistema de recomendação	Aquisição de conhecimento
Previsões de mercado	Segmentação de clientes	Navegação

Fonte: Elaborado pelo autor (2020)

Conforme descrito no "Quadro 3 - Tipos de aprendizagem de máquina e aplicações"o aprendizado de máquina pode ser aplicado em diversos projetos das mais diversas áreas. Uma vez delimitado o escopo, os dados disponíveis e o objetivo, é possível escolher a melhor técnica a ser utilizada para o projeto em questão.

2.5 REDE NEURAL ARTIFICIAL

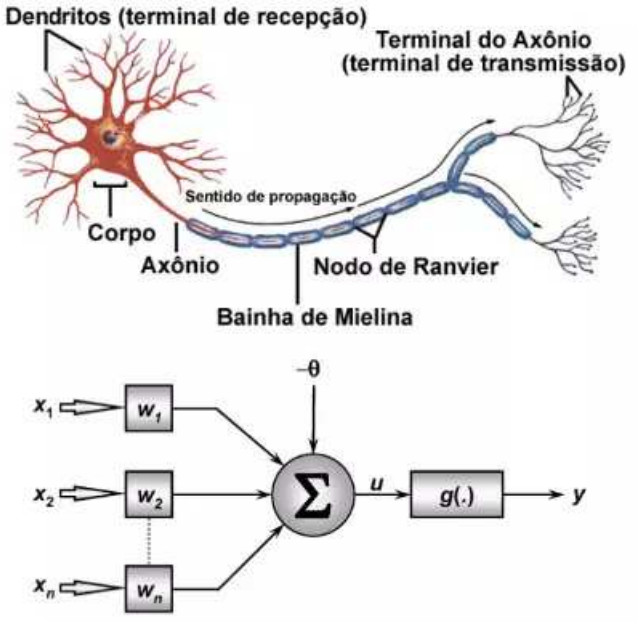
Rede neural artificial é o termo utilizado para definir um conjunto de neurônios artificiais baseados em algoritmos matemáticos e inspirados nas funcionalidades biológicas dos neurônios. Essa definição foi introduzida no meio científico por Mcculloch e Pitts (1943).

Atualmente, entre as técnicas de aprendizagem de máquina mais utilizadas estão as redes neurais artificiais, que possuem como inspiração o funcionamento da estrutura neural de organismos inteligentes (COPPIN, 2004). Possibilitando a criação de grandes redes neurais com centenas ou até mesmo milhares de neurônios artificiais, que são interligados em camadas, onde a saída de uma unidade neural é interligada à entrada de outra.

Um neurônio artificial busca simular o mais próximo possível a atividade de um neurônio biológico. O neurônio biológico é uma célula do sistema nervoso responsável pela condução de impulsos nervosos (BORGES *et al.*, 2015).

Na figura abaixo, é possível ver uma comparação entre a representação dos componentes básicos de um neurônio biológico e um neurônio artificial.

Figura 3 – Comparação de neurônio Artificial x neurônio biológico



Fonte: Rocha (2017)

No neurônio biológico, existem os dendritos são responsáveis pela recepção e transmissão dos estímulos nervosos para a região do corpo celular. Essa região é conhecida como soma e contém o núcleo do neurônio e o citoplasma onde são produzidas as proteínas (BORGES *et al.*, 2015). O axônio é coberto pela mielina, uma camada isolante que auxilia na transmissão do impulso nervoso que passa da base do axônio até a extremidade final. A extremidade do axônio desenvolve dilatações conhecidas como terminais axônicos ou terminais nervosos, que são responsáveis pela transmissão dos estímulos e são conectadas a dendritos de outros neurônios criando assim uma rede neural (BORGES *et al.*, 2015).

Um neurônio artificial é definido por Haykin (2005) como uma unidade de processamento de informação fundamental para o funcionamento de uma rede neural artificial. Existem diversos tipos de neurônios artificiais, entre os primeiros modelos matemáticos que deram origem a neurônios artificiais, podemos citar o *Perceptron* que foi criado por Frank Rosenblatt em 1957 (HAYKIN, 2005). O exemplo do neurônio artificial apresentado na figura acima pode ser detalhado da seguinte forma:

1. As entradas são representadas por (x1, x2, xn), que são multiplicados com os respectivos pesos sinápticos associados (w1, w2, wn).

2. O somador representado pelo simbolo Σ somará os sinais de entrada, ponderados pelas sinapses do neurônio, e aplicara a equação (HAYKIN, 2005).

3. O resultado da somatória das entradas ponderadas será somado ao limiar de ativação (θ) para ser passado como argumento para a função de ativação. Em alguns casos é possível observar a anotação de b_k, que é conhecido como *bias* externo e tem basicamente o mesmo objetivo do limiar de ativação, com o efeito de aumentar ou diminuir a entrada líquida da função de ativação (HAYKIN, 2005).

4. A função de ativação ($g(.)$), que também é referida como função restritiva, é aplicada com o objetivo de restringir a amplitude de saída do neurônio para um valor finito normalmente entre -1 e 1 (HAYKIN, 2005).

Em termos matemáticos, um neurônio artificial pode ser descrito por duas equações:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{1}$$

$$y_k = \varphi(u_k + b_k) \tag{2}$$

A maneira que a arquitetura de uma rede neural é desenhada depende diretamente do algoritmo utilizado para o treinamento, uma vez que o tipo de algoritmo

escolhido influenciará nas regras de aprendizagem da rede (HAYKIN, 2005). Basicamente podemos definir três tipos de arquiteturas de redes diferentes:

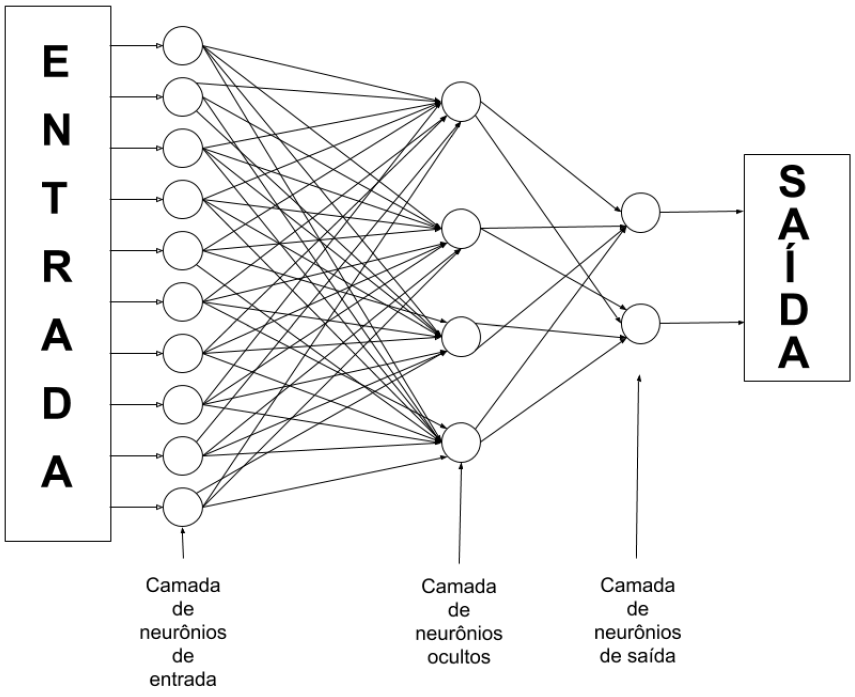
1. **Redes alimentadas adiantes com camada única**: possui uma camada de entrada que se projeta sobre uma camada de saída de neurônios (nós computacionais), nesse caso, contamos apenas a camada de neurônios, ou seja, a camada que realizará o processamento das informações, visto que a camada de entrada não realizará nenhuma computação (HAYKIN, 2005).

2. **Redes alimentadas diretamente com multicamadas**: esta arquitetura possui a presença de camadas ocultas de neurônios entre a camada de entrada e saída, onde a saída de uma camada é ligada na entrada de outra camada através das ligações sinápticas dos neurônios. Assim uma rede multicamada pode possuir mais de uma camada oculta. Sendo possível potencializar o aprendizado e contribuir na execução de tarefas estatísticas de ordem elevada. (HAYKIN, 2005).

3. **Redes recorrentes**: é uma rede neural capaz de se realimentar, pois possui laços de realimentação que podem ligar a saída novamente à entrada do neurônio até mesmo em uma rede de única camada e mesmo em redes recorrentes multicamadas tendo ou não camada oculta (HAYKIN, 2005). A adoção de laço de realimentação na arquitetura de uma rede neural artificial tem um impacto profundo em sua capacidade de aprendizado e desempenho (HAYKIN, 2005)

Na figura abaixo, é possível visualizar um esquema de uma rede neural multicamada, que utiliza neurônios ocultos, são conectados através de nós sinápticos.

Figura 4 – Rede neural artificial multicamada



Fonte: Adaptada de Haykin (2005)

2.6 *REDES NEURAIS RECORRENTES* (RNN)

Redes neurais recorrentes é um termo normalmente utilizado para se referir a duas classes de redes neurais, que são projetadas para trabalhar previsões de sequências; uma de impulso finito e outra de impulso infinito. As redes neurais recorrentes são neurônios artificiais organizados em camadas sucessivas, cada nó de uma camada é conectado a um nó de uma próxima camada sucessiva direcionada a todos os outros nós conectados.

Segundo Olah (2015), redes neurais recorrentes são indicadas para: dados de texto, dados de fala, problemas de previsão de classificação, problemas de previsão de regressão, modelos geradores. Porém, existem alguns problemas na utilização de redes recorrentes, quando há dependências de longo prazo. As dependências de longo prazo podem partir de *timesteps* anteriores a uma entrada mapeada atual, dessa forma, algumas vezes precisamos apenas de informações recentes e uma entrada mapeada para classificar, por exemplo, uma saída. Nesse tipo de problema, quando a lacuna é pequena, as redes neurais recorrentes podem aprender a utilizar informações do passado para realizar a predição da saída. Porém, existem casos em que a previsão necessita de mais informações sobre o contexto da frase, por exemplo, para prever a última palavra da frase “Eu cresci no Brasil... falo fluentemente português” o ponto

entre a predição da informação relevante da palavra de contexto “Brasil” e o *timestep* “cresci” acaba tornando-se distante (OLAH, 2015).

2.7 *LONG SHORT-TERM MEMORY*(LSTM)

Para tipos de problemas em que temos a necessidade de aprendizados em que existiram dependências de longo prazo, existem redes LSTM. As redes LSTM são redes neurais recorrentes que possuem memória de longo prazo. Todas as redes neurais recorrentes possuem módulos de repetições. em uma rede neural recorrente padrão, esse módulo possui a estrutura mais simplificada, como uma única camada de repetição, já o LSTM possui três portões, para proteger e controlar o estado da célula. Em redes LSTM, existem quatro camadas de repetições que facilitam na hora da utilização em problemas complexos de aprendizagem em que existe a necessidade de predição ou classificação baseada em uma entrada já mapeada em outra etapa (OLAH, 2015). Ao contrário do primeiro diagrama na LSTM, a informação percorre entre os módulos, o LSTM tem a capacidade de remover ou adicionar informações ao estado da célula, cuidadosamente reguladas por estruturas chamadas portas. Os portões são uma forma de, opcionalmente, deixar passar as informações. Eles são compostos de uma camada de rede neural sigmoide e uma operação de multiplicação pontual. A camada sigmoide produz números entre zero e um, descrevendo quanto de cada componente deve ser liberado. O valor de zero significa "não deixe nada passar", enquanto o valor de um significa "deixe tudo passar!"(BROWNLEE, 2017).

2.8 MÉTRICAS

Segundo Klubeck (2011), uma métrica é composta por informações, medidas e dados. E, dentro de uma métrica, é possível incluir outra métrica. Klubeck (2011) ainda menciona que uma métrica conta uma história completa por meio da representação de informações, que é obtida através da compilação de medidas que são utilizadas para transmitir o significado. As medidas são construídas a partir dos dados coletáveis, sendo eles valores ou números que, por sua vez, respondem uma questão raiz, não sendo possível existir uma boa métrica sem uma questão raiz bem formulada.

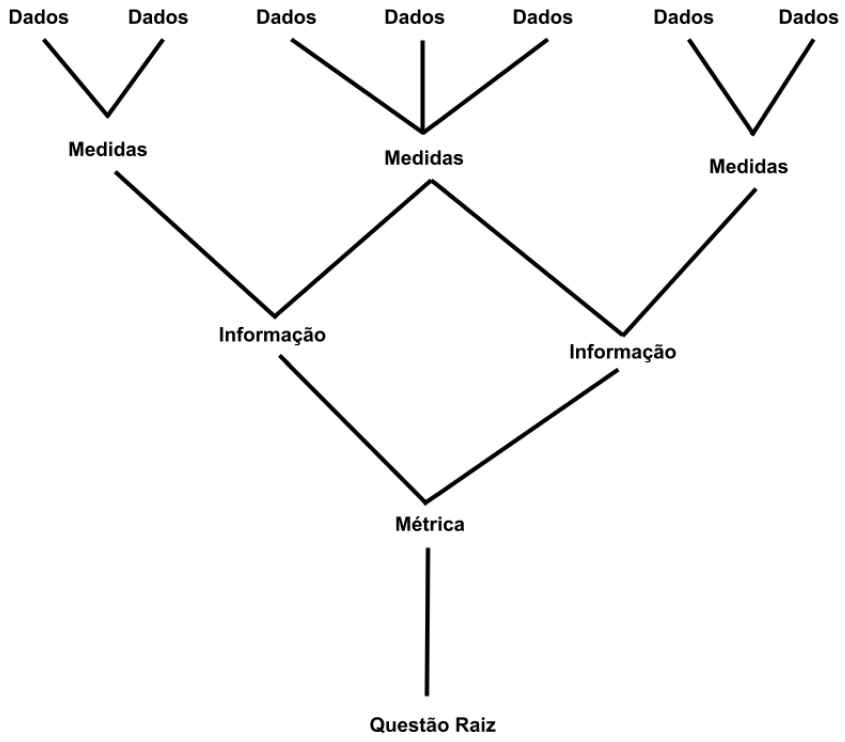
São exemplos de dados, medidas, informações, métricas e pergunta raiz:

- **Dados**: 9 e 17
- **Medidas**: 9KM/L e 17KM/L
- **Informações**: quilômetro por litro rodados usando gasolina sem chumbo em um carro compacto: 9KM/L na cidade, 17KM/L na rodovia.

- **Métrica**: seria uma imagem, sendo ela tabela ou gráfico, que conta uma historia. A história nesse caso seria a comparação entre a eficiência levando em consideração quilômetros por litro de gasolina de diferentes modelos de veículos compactos.
- **Pergunta**: qual carro é mais econômico?

O autor exemplifica uma métrica criando uma analogia com uma árvore conforme ilustrado na Figura 5.

Figura 5 – Componentes de uma métrica



Fonte: Adaptada de Klubeck (2011)

Na Figura 5, os dados são representados pelas folhas, as medidas e informações são os galhos e a métrica é o tronco da árvore. Toda a existência da árvore mencionada só é possível com um bom conjunto de raízes, que é representado pela questão raiz.

As métricas utilizadas para avaliação dos objetivos deste trabalho são melhor descritas na seção 4 chamada Trabalhos relacionados, em que é feito um comparativo com relação aos trabalhos que possuem objetivo ou método de desenvolvimento relacionados a temática do trabalho.

3 REVISÃO SISTEMÁTICA DE LITERATURA

A revisão sistemática de literatura foi elaborada seguindo a metodologia proposta por Sampaio e Mancini (2007), conforme Tabela 4.

Quadro 4 – Passos metodologia revisão sistemática de literatura

PASSO	DESCRIÇÃO
Passo 1 – Pergunta	É possível a criação de um modelo padrão de extração de entidades nomeadas em documentos escritos na língua portuguesa com o emprego de redes neurais recorrentes e com base em corpus disponíveis na internet?
Passo 2 – Busca	**Bases de dados:** SCORPUS, IEEE, Library and Information Science Abstracts (LISA) e WEB OF SCIENCE **Palavras-chave:** Named Entity Recognition; corpus; recurrent neural networks **Idioma:** Inglês e português **Tipo de publicação:** Artigos científicos 2017 a 2019
Passo 3 – Revisão e seleção de documentos	Critérios de inclusão: (1. Artigos com textos completos; 2. Artigos revisados por especialistas; 3. Artigos nos idiomas inglês e português; 4. Tipo de documento: Artigos científicos) Critérios de exclusão: (1. Artigos com temática diferente; 2. Artigos com idioma que não seja inglês e espanhol; 3. Apresentam redes neurais que não sejam LSTM ou variações; 4. Não possui palavras-chaves no resumo; 5. Artigos em duplicidade; 6. Documentos que não sejam artigos científicos; 7. Artigos que não possuem corpora disponíveis na internet; 8. Artigos incompletos)
Passo 4 – Analisando a qualidade metodológica dos documentos	Leitura completa dos documentos, aplicando os critérios de inclusão e exclusão
Passo 5 – Apresentação de resultados	Descrição das características e resultados dos, após aplicação dos critérios.

Fonte: Elaborado pelo autor

A busca realizada com base na metodologia descrita no quadro 4, resultou na recuperação de 169 documentos, sendo aplicados os critérios de exclusão e inclusão para seleção dos documentos utilizados neste artigo.

Os detalhes relacionados à Revisão Sistemática de Literatura são descritos de forma integral no Apêndice A.

4 TRABALHOS RELACIONADOS

Na seção a seguir são apresentados alguns trabalhos relacionados à extração de entidades nomeadas que foram encontrados na literatura e foram utilizados com o fim comparativo entre a metodologia criada para execução deste projeto e estruturas já existentes.

Um grande número de propostas com o objetivo de desenvolver novos corpora para classificadores pode ser encontrado na literatura. Os corpora podem ser baseados em dados de mídia social, *opendata* de feed de notícias, sites ou mesmo dados de empresas (OLIVEIRA *et al.*, 2017).

Os autores de Holthaus *et al.* (2016) desenvolveram um corpus para ambientes inteligentes. O corpus inclui materiais de áudio e vídeo, reações de robôs e apartamentos, bem como informações extraídas de sensores e atuadores. Os dados foram coletados de 62 voluntários e podem ser usados para o treinamento de robôs automatizados. Conforme argumentado pelos autores, esses dados são valiosos para análises aprofundadas das interações das pessoas com dispositivos, inteligência ambiental e robôs nos ambientes cotidianos. O nosso trabalho se diferencia do trabalho realizado por Holthaus *et al.* (2016), pois utilizamos dados de texto para fornecer corpora para tarefas de Processamento de linguagem natural (PNL).

Em Cavalin *et al.* (2016) é apresentado um Corpus de Perguntas e Respostas (QA-Corpus) específico de domínio construído com tweets e notícias em português. Ao usar a mídia social, os autores poderiam reunir respostas candidatas e confiáveis para possíveis perguntas do usuário, tornando o conjunto de dados mais real. Eles usaram o aprendizado profundo para combinar as perguntas e obter respostas dos candidatos. Tanto este artigo como Cavalin *et al.* (2016) criaram um corpora para a língua portuguesa, que carece de bons conjuntos de dados/corpora, a fim de comparar os resultados. Ao contrário de Cavalin *et al.* (2016), que se concentra nos sistemas de controle de qualidade, o presente trabalho se concentra na tarefa de Processamento de linguagem natural (PNL).

O trabalho realizado por Spoustová, Spousta e Pecina (s.d.) teve a iniciativa de criar um corpus nacional na língua tcheca. Os autores descrevem um projeto para construir um corpus maior composto por textos em tcheco extraídos de páginas da web. A motivação por trás desse projeto reside no fato de que os autores acreditam que grandes corpora são essenciais para os métodos modernos de linguística computacional e processamento de linguagem natural. A diferença entre o presente trabalho e Spoustová, Spousta e Pecina (s.d.) é que não são usadas páginas da Web, mas melhora a escolha existente para produzir um corpus único, unificado e comparável.

Um esforço árabe para criar um corpus ocorreu baseando-se em jornais publicados on-line de diferentes idiomas de países árabes. Os autores Abdelali, Cowie e

Soliman (2005) criaram o corpus para melhorar diferentes pesquisas em Recuperação de Informação, Tradução Automática e Processamento de Língua Árabe, em geral. Assim como Abdelali, Cowie e Soliman (2005), nosso artigo é uma tentativa de padronizar um corpus grande e comparável para o processamento da língua portuguesa em várias áreas de pesquisa. Além disso, ambos os trabalhos usam o texto de jornais para construir o conjunto de dados/corpus.

O autor Araujo *et al.* (2018) e colaboradores fizeram um corpus para um domínio específico da linguagem, o vocabulário jurídico. O corpus contém entidades como "TEMPO", "JURISPRUDENCIA"e "LEGISLACAO". O modelo criado utilizou as redes neurais LSTM-CRF e LSTM-CNN, com uma precisão de 90,01%. Por outro lado, o estudo utiliza o corpus HAREM em dois cenários diferentes. No primeiro, o técorpus é utilizado integralmente, com as entidades "PESSOA", "ORGANIZACAO", "LOCAL", "VALOR", "TEMPO", "ABSTRACAO", "OBRA", "ACONTECIMENTO", "COISA"e "OUTRO". No segundo cenário, foram consideradas apenas as entidades "PESSOA", "LOCAL", "ORGANIZACAO", "DATA"e "VALOR". Para o primeiro cenário, a precisão obtida foi de 74,91%. Por sua vez, o segundo cenário obteve 83,38% de precisão. Os modelos usaram uma variação de um conjunto de redes neurais, como LSTMs recorrentes, CNNs e redes convolucionais, porém o volume de dados utilizado, tanto para treinamento quanto para teste, foi baixo.

Com o desenvolvimento do projeto, foi possível destacar seis características que diferem entre o modelo proposto e trabalhos relacionados, conforme mostrado na quadro 5 :

Quadro 5 – Trabalhos relacionados

Trabalho	E.B.	M.H.E.	I.L.2.L.	A.E.	P.V.	G.A.
MODELO PROPOSTO	SIM	SIM	SIM	SIM	SIM	SIM
(HOLTHAUS et al., 2016)	NÃO	NÃO INFORMADO	NÃO	NÃO	NÃO INFORMADO	SIM
(CAVALIN et al., 2016)	NÃO	NÃO	NÃO	NÃO	NÃO	SIM
(SPOUSTOVÁ;SPOUSTA; PECINA, 2010)	NÃO	NÃO	NÃO	NÃO	NÃO	SIM
(ABDELALI; COWIE; SOLIMAN, 2005)	NÃO	NÃO	NÃO	NÃO	NÃO	SIM
(LUZ DE ARAUJO et al., 2018)	NÃO	NÃO	NÃO	SIM	NÃO	SIM
(SANTOS et al., 2019)	NÃO	NÃO	NÃO INFORMADO	SIM	NÃO	SIM

Fonte: Elaborado pelo autor

A descrição de cada métrica utilizada como comparativo são:

1. **Entity Balancing (E.B.)**: Antes do pré-processamento do treinamento, são selecionadas as seções que possuem quantidades representativas de entidades. Dessa forma, é possível obter com maior acerto a acurácia isolada por entidade, pois evita distorções na medição, uma vez que todas as entidades possuem quantidades aproximadas ao seu percentual geral encontrado no corpus inteiro.

2. **Models History Evolution (M.H.E.)**: Um relatório analítico é gerado com o número de iterações para orientar a parametrização do modelo.

3. **Pre-Processing indexing letter to letter of each token (I.L.2.L.)**: Duas abordagens foram geradas durante o pré-processamento do modelo, indexando para cada token existente no vocabulário e indexando cada letra existente no token.

4. **Accuracy by Entity (A.E.)**: verifica por classe de entidade a quantidade recuperada pelo modelo e a quantidade realmente existente na base.

5. **The creation of a database with processed vectors (P.V)**: Após a análise, os vetores são gerados com informações separadas, token, entidade, seção, posição e tamanho. Esses vetores são armazenados para uso posterior, reduzindo a etapa de pré-processamento.

6. **General Accuracy (G.A.)**: Verifica a quantidade total recuperada pelo modelo e a quantidade total realmente existente na base.

5 ARQUITETURA DE IMPLEMENTAÇÃO

Nesta seção são abordados, as técnicas e ferramentas utilizadas para criação da arquitetura que possibilitou treinar a rede neural artificial com base nos trechos do corpus CETENFolha e obter os resultados reacionados a extração de entidades nomeadas.

5.1 PYTHON

Para desenvolvimento do estudo foi utilizada a linguagem de programação Python, a linguagem de programação Python é muito utilizada em projetos de processamento de linguagem natural devido a possuir vários módulos organizados em bibliotecas contendo funções e métodos que auxiliam na limpeza e processamento do texto (VAN ROSSUM; DRAKE, 2009).

No Quadro 6 - Bibliotecas e versões, segue o versionamento usado em cada biblioteca.

Quadro 6 – Bibliotecas e versões

Biblioteca	**Versão**
Keras	2.2.4
Keras_Preprocessing	1.0.9
nltk	3.4.4
regex	2019.4.14
sklearn	0.0
pymongo	3.8.0
mongoquery	1.3.5

Fonte: Elaborado pelo autor

Nas próximas subseções serão detalhadas as funções e uso das principais bibliotecas usadas no desenvolvimento do projeto.

5.1.1 NLTK

A biblioteca NLTK *Natural Language Toolkit* é usada para processamento simbólico e estatístico de linguagem natural, em processamento de textos, possuindo suporte para diversos idiomas, incluindo o inglês e português. A biblioteca possui um conjunto de funções que permitem realizar algumas atividades, como, *tokenizar* sentenças, resumir texto, clusterização de palavras, mapeamento de n-gramas e *stopwords*, lógica de primeira ordem, avaliação de modelos, precisão, recall, concordância de coeficientes e distribuições de frequência entre outras funções (BIRD; KLEIN; LOPER, 2009).

Para o presente trabalho foi utilizada a função para mapeamento de *stopwords*, sendo gerado uma lista de palavras sem peso semântico presentes no idioma portu-

Figura 6 – Extração de caracteres especiais

```
import re
text = "Rua Conselheiro Mafra, nº 656 - 4º Andar - Sala 403.
Centro - CEP: 88010-914. Telefone: (48) 3251-6107"
text = re.sub(r'[^\w\s]','',text)
print(text)
>> Rua Conselheiro Mafra n 656  4 Andar  Sala 403
```

Fonte: Elaborado pelo autor

Figura 7 – Extração de tags HTML

```
import re
text = "<h1>Rua Conselheiro Mafra</h1>"
text = re.sub(r"<[^>]*>",'',text)
print(text)
>> Rua Conselheiro Mafra
```

Fonte: Elaborado pelo autor

guês, por exemplo, o, a, os, as, de, da, do, esse e essa. Estas palavras são retiradas do texto para possibilitar a geração do dicionário, sem a presença de palavras que possam criar distorções no resultado do reconhecimento de entidades nomeadas.

5.1.2 Regex

O regex é uma biblioteca muito utilizada para localizar cadeias de caracteres, uma vez que seja necessário realizar, mineração de texto ou limpeza de dados (AHO, 1991), por exemplo, caracteres especiais em determinadas partes de um texto ou números.

Conforme o exemplo na Figura 6 - Extração de caracteres especiais, na primeira sentença foi utilizado o regex que resulta na retirada de qualquer carácter que não seja dígito ou letra. No caso abaixo foram retirados os caracteres, º, -, ., : e ,.

Na segunda sentença conforme Figura 7 - Extração de tags HTML, foi usado para retirar qualquer *tag* em HTML e XML do texto procurando o padrão com abertura «"e fechamento da *tag* "/>".

5.1.3 Keras

O Keras é uma biblioteca de código aberto para experimentação de redes neurais artificiais de aprendizagem profunda, ela possibilita a criação de diferentes redes neurais de forma rápida, preenchendo apenas seus parâmetros (CHOLLET *et al.*, 2015). Entre as redes neurais artificiais disponíveis para implementação existem as LSTM que foram utilizadas neste projeto. Conforme Figura 8 - Módulo de treinamento do projeto, é possível visualizar o código python usado na criação do módulo de treinamento e os parametrôs usados.

Figura 8 – Módulo de treinamento do projeto

```
from keras.models import Model, Input
from keras.layers import LSTM, Embedding, Dense, TimeDistributed,
Dropout, Conv1D
from keras.layers import Bidirectional, concatenate, SpatialDropout1D,
GlobalMaxPooling1D

epochs = 40
batch_size = 32
tamanho_test = 0.2

max_len = 150
max_len_char = 10

TimeDistributed

# input and embedding for words
word_in = Input(shape=(max_len,))
emb_word = Embedding(input_dim=n_words + 2, output_dim=20,
                     input_length=max_len, mask_zero=True)(word_in)

# input and embeddings for characters
char_in = Input(shape=(max_len, max_len_char,))
emb_char = TimeDistributed(Embedding(input_dim=n_chars + 2,
output_dim=10,input_length=max_len_char, mask_zero=True))(char_in)
# character LSTM to get word encodings by characters
char_enc = TimeDistributed(LSTM(units=20, return_sequences=False,
                                recurrent_dropout=0.5))(emb_char)

# main LSTM
x = concatenate([emb_word, char_enc])
x = SpatialDropout1D(0.3)(x)
main_lstm = Bidirectional(LSTM(units=50, return_sequences=True,
                               recurrent_dropout=0.6))(x)
out = TimeDistributed(Dense(n_tags + 1, activation="softmax"))(main_lstm)

model = Model([word_in, char_in], out)


model.compile(optimizer="adam", loss="sparse_categorical_crossentropy",
metrics=['mae'])
model.summary()

history = model.fit([X_word_tr,
                     np.array(X_char_tr).reshape((len(X_char_tr), max_len,
                     max_len_char))],np.array(y_tr).reshape(len(y_tr),
                     max_len, 1),
                     batch_size=batch_size, epochs=epochs,
                     validation_split=0.1, verbose=1)
```

Tabela 2 – Parâmetros da rede neural artificial

Parâmetro	Valor	Descrição
units	50	Dimensionalidade do espaço de saída.
recurrent_dropout	0.6	Flutua entre 0 e 1. Fração das unidades a cair para a transformação linear do estado recorrente.
optimizer	adam	Algoritmo de otimização
loss	sparse_categorical _crossentropy	Parâmetro que avalia de desempenho do algoritimo de aprendizado, como perda logarítmica ou erro quadrático médio, em cada epoch realizada pelo algoritimo
activation	softmax	Função de ativação a ser usada.
metrics	mae	Função usada para julgar o desempenho do seu modelo
epochs	20	Hiperparâmetro que define o número de vezes que o algoritmo de aprendizado funcionará em todo o conjunto de dados de treinamento.
batch_size	32	Hiperparâmetro que define o número de amostras a serem trabalhadas antes de atualizar os parâmetros do modelo interno.
tamanho_test	0.2	Tamanho da base de teste em relação a base principal
max_len	150	Tamanho máximo de cada vetor que representa o trecho de entrada
max_len_char	10	Tamanho máximo de cada vetor que representa o token

Fonte: Elaborada pelo autor

Os valores aplicados nos parâmetros da rede neural artificial, podem ser visualizados na Tabela 2 - Parâmetros da rede neural artificial, que possui a descrição de cada parâmetro usado.

5.1.4 Scikit-learn

A Scikit-learn é uma biblioteca de aprendizagem de máquina de código aberto, ela possui diversos algorítimos de regressão, classificação e clusterização (PEDREGOSA *et al.*, 2011). Conta também um módulo com funções que calculam as métricas mais utilizadas na exploração e análises de resultados de treinamentos e aprendizagem dos algorítimos. Este módulo é utilizado para gerar as métricas contidas na sessão de Resultados deste projeto, são elas, acurácia geral, acurácia por entidade e acurácia isolada que retira a entidade "O", que são *tokens* que não possuem entidades anotadas.

Conforme a Figura 9 - Cálculo de métricas, é possível aferir o resultado da

acurácia comparando os dois vetores gerados a partir das colunas de "Previsto"que é a previsão realizada pela rede neural artificial para os trechos de teste e "Verdadeiro"que são as entidades conforme anotadas no corpus.

Figura 9 – Cálculo de métricas

```
print('====================Acurácia Geral=========================')
y_true = df.Verdadeiro.to_list()
y_pred = df.Previsto.to_list()
print('Acuracia Geral: ',accuracy_score(y_true,y_pred))

relatorio['acuracia_geral'] = accuracy_score(y_true,y_pred)

print('=================Acurácia - Real (Filtrada) ================')

#Retira a Classe que possui maior incidência que cria distorções
#na acurácia

y_true_2 = list()
y_pred_2 = list()
for i in df.values:
    if i[1]!='O' and i[2]!='O':
        y_true_2.append(i[1])

    if i[1]=='O' and i[2]!='O':
        y_true_2.append(i[1])
        y_pred_2.append(i[2])

    if i[1]!='O' and i[2]=='O':
        y_true_2.append(i[1])
        y_pred_2.append(i[2])

print('Acurácia: ',accuracy_score(y_true_2, y_pred_2))
```

Fonte: Elaborado pelo autor

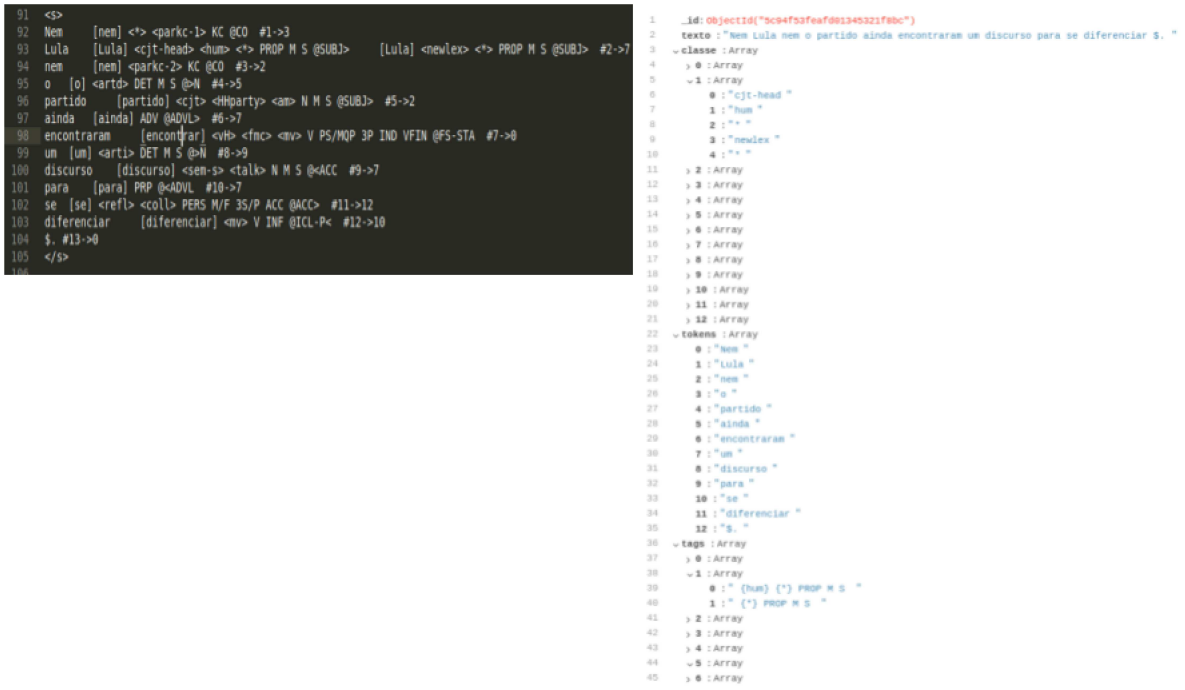
5.2 MONGODB

MongoDB é um banco de dados no-sql orientado a documentos, ou seja, um banco para armazenamento de dados não tabulares, diferente dos bancos de dados tradicionais como SQL Server, Oracle ou Postgresql, de licença livre e código aberto. O MongoDB usa documentos JSON (*JavaScript Object Notation*) para armazenamento e criação do *Schema* que compõe a base, chamada de coleção. A representação para abstração da coleção de documentos, gera uma forma hierarquia para chave e valor contida na coleção de documentos (CHODOROW; DIROLF, 2010).

Foram criadas duas coleções de documentos no MongoDB, a primeira com os dados dos trechos dos corpus processados em vetores, e a segunda com os parâmetros da rede neural artificial e do pipeline, os identificadores de cada documento usado no treino e teste, tempo de execução dos testes e resultados os resultados para as métricas de acurácia.

Conforme Figura 10 - Transformação e armazenamento de um trecho do corpus, a esquerda da figura o trecho original do corpus, que é convertido para um formato de chave valor, dentro de vetores de um arquivo JSON e armazenado na coleção do banco de dados MongoDB, conforme a parte direita da figura, no momento do armazenamento, é gerado um identificador para o trecho inserido, com a chave "_id". A coleção que armazena os vetores já processados é chamada de "CORPUS".

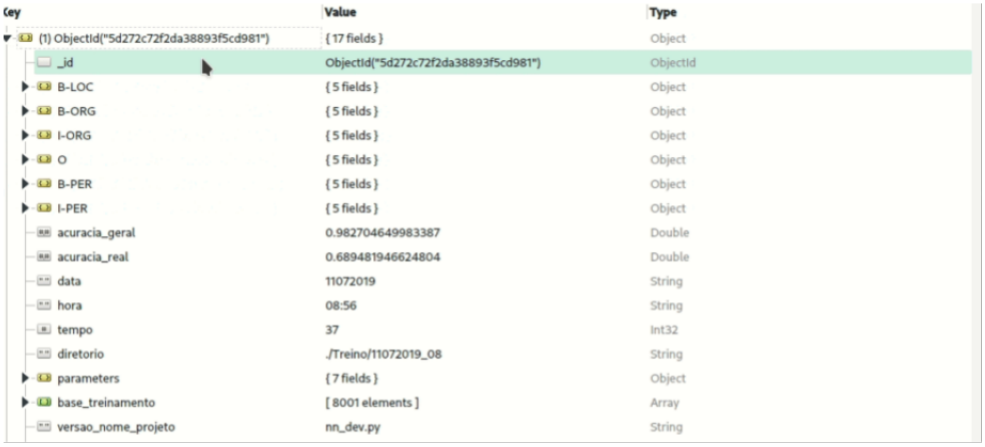
Figura 10 – Transformação e armazenamento de um trecho do corpus



Fonte: Elaborada pelo autor

A saída do pipeline que gera o modelo e extrai as entidades nomeadas presentes no texto, possui o mesmo tipo de saída de chave e valor, porém com os dados de *log* do treinamento da rede neural artificial, sendo eles, os parâmetros usados na rede e as métricas de acurácia isolada, acurácia geral, acurácia por entidade, tempo de execução, diretório de onde é salvo o arquivo binário contendo os pesos do aprendizado para ser usado posteriormente, data e hora da execução. Salvando o documento na coleção "LOG", conforme Figura 11 - Estrutura da saída do pipeline da extração de entidades nomeadas.

Figura 11 – Estrutura da saída do pipeline da extração de entidades nomeadas



Fonte: Elaborada pelo autor

5.3 BERT

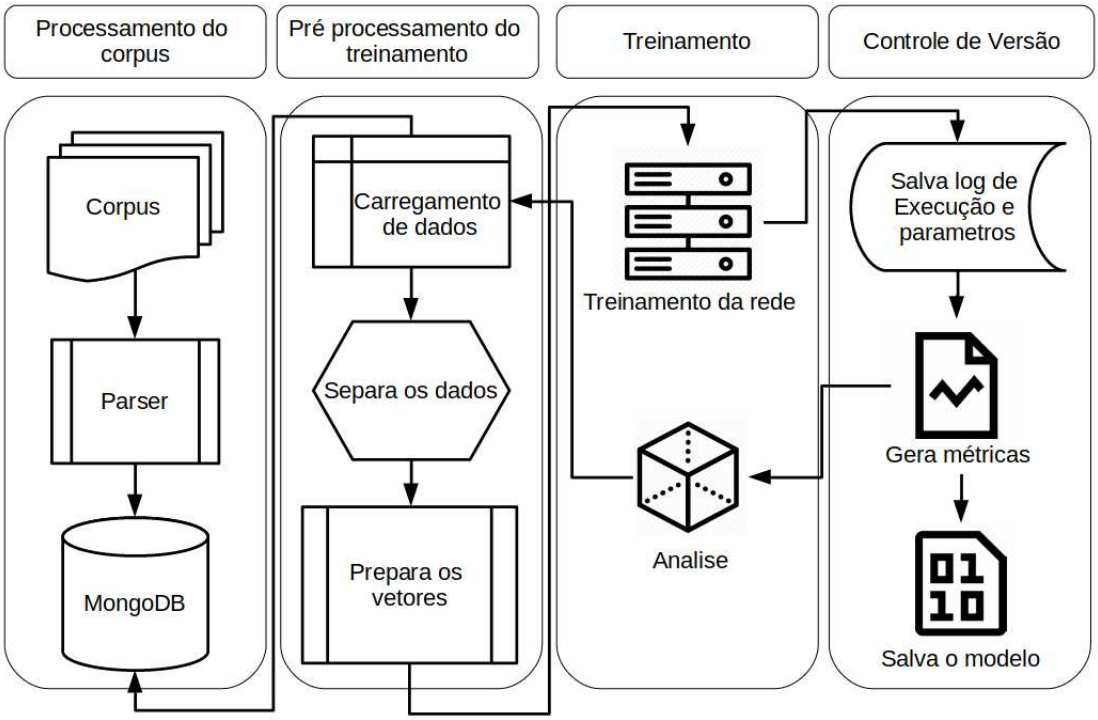
Devido este projeto ter iniciado no ano de 2018, foi utilizado redes neurais artificiais do tipo LSTM *(Long short-term memory)*, usando a biblioteca python Keras para implementação, conforme descrito no capitulo 5 - Arquitetura de implementação. Ainda no mês de outubro do ano de 2018 foi publicado pela primeira vez o artigo *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* produzidos por Devlin *et al.* (2019) sobre a biblioteca BERT ainda em fase exploratória no idioma inglês. O BERT também utiliza redes neurais LSTM *(Long short-term memory)*, porém seu diferencial esta na conversão dos vetores processo conhecido como *embeddings*, uma vez que o BERT já armazena os pesos dos vetores convertidos pré treinados, obtendo bons resultados em diversos idiomas para tarefas de extração de entidades nomeadas, classificação de texto e analise de sentimento (DEVLIN *et al.*, 2019). Por tratar-se de um artigo em fase exploratória no inicio do projeto e a inexistência em 2018 de suporte para o idioma português, não foram realizados testes nem desen-

volvimentos utilizando a biblioteca BERT. Optando assim pela utilização da biblioteca Keras.

6 METODOLOGIA

Nesta seção, será apresentada a metodologia utilizada para padronizar o corpus CETENFolha, a fim de produzir um único corpus a ser utilizado na criação do modelo estatístico de extração de entidades nomeadas. A metodologia foi desenvolvida em quatro etapas, como mostra a Figura 12 – Arquitetura do modelo proposto.

Figura 12 – Arquitetura do modelo proposto



Fonte: Elaborada pelo autor

6.1 PROCESSAMENTO DO CORPUS

A primeira etapa teve como objetivo encontrar e analisar corpora textuais disponíveis na internet e que possuíssem melhor estrutura para treinamento do modelo. Levamos em consideração todas as propriedades que um corpus necessita para que seja possível realizar o treinamento, são elas: delimitação de início e fim do parágrafo, *tokens*, tag com a classe gramatical e o pos tag, que é a classificação da entidade. Foi encontrado o corpus CETENFolha disponibilizado pelo projeto Linguateca, que possui 24 milhões de palavras em 340.947 extratos de trechos de artigos do jornal Folha de São Paulo.

Após a escolha do corpus, foi necessário criar o método que permitisse realizar o *parsing* (análise sintática) do arquivo, possibilitando que o corpus fosse processado e disponibilizado em um arquivo de saída, que é um dicionário que mantém todas as

informações relevantes para o treino e é inserido no banco de dados MongoDB Figura 12 - Arquitetura do modelo proposto. Uma das preocupações em relação à criação da estrutura foi o tipo de base de dados, pois, devido ao volume de informação, é necessário que seja possível selecionar de forma rápida as informações relevantes para o treino, visto que nem toda informação em um corpus é relevante.

6.2 PRÉ-PROCESSAMENTO DO TREINAMENTO

foram criados métodos para validar textos, esses métodos serviriam para garantir que os textos não tivessem erros que pudessem atrapalhar o aprendizado ou causar erros durante o processo de treinamento. Essas etapas estão representadas na Figura 1 como, Pré Processamento do Treino e Separação dos dados. O Conversor de *tags* (Figura 12 como pré-Processamento do treino e separação dos dados, Conversão das *tags*) recebe a Tag original do corpus e é padronizado em relação a um dicionário de *tags* controlado para que não exista uma *tag* representada de duas formas diferentes. Além disso, essa etapa gera dicionários de palavras, *tags*, caracteres e entidades por meio de vetores para treinamento e validação, gerando um vetor com as palavras que são as variáveis independentes e outro com as *tags* que são as variáveis dependentes.

Para criar a variável dependente, deve-se converter as entidades de cada número do *token*. A nova lista de *tokens* numéricos será relacionada ao dicionário "tag2idx". Cada entidade nessa lista está relacionada a uma entidade na lista de vetores variáveis independentes.

Quando o *token* não possui uma entidade localizada, é classificado como "O". Se o token tiver uma entidade localizada, ele será classificado com uma letra de acordo com sua posição. Se o *token* for o primeiro *token* da entidade, ele terá a letra “B” (início) + “-” + “*tag* da entidade”. Se o *token* não for o primeiro, ele terá a letra “I” + “-” + “*tag* de entidade”. Os *tokens* de entidades definidos são:

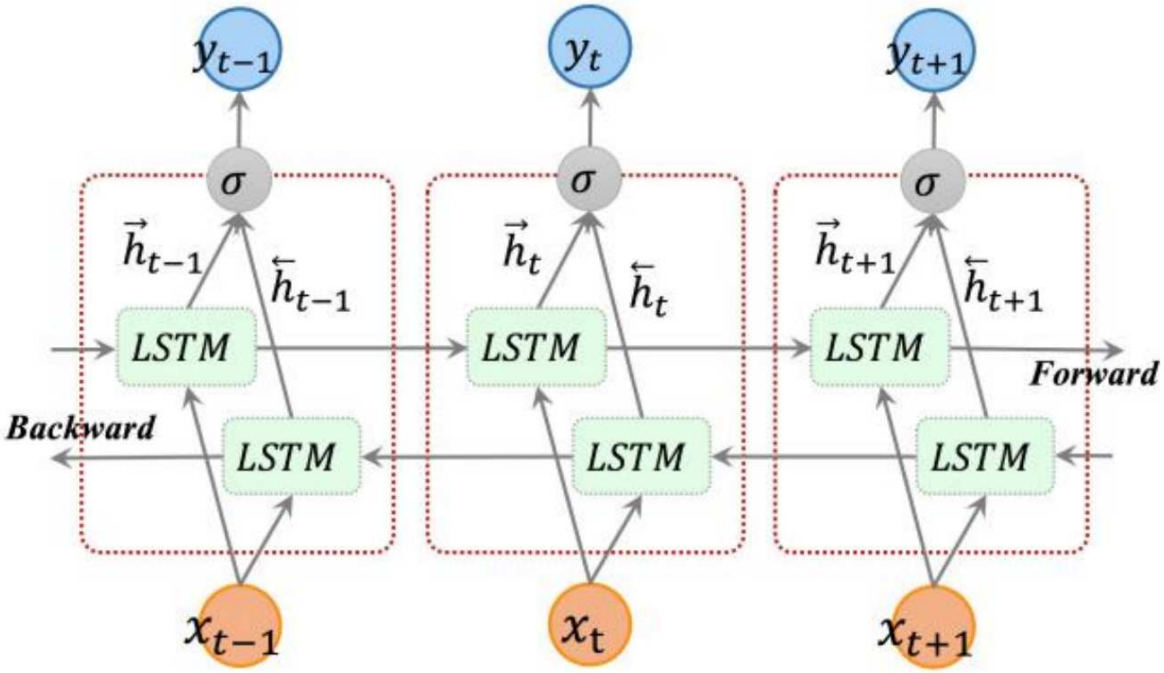
1. B-LOC: primeiro *token* de entidade local, ou seja: "São"; 2. I-LOC: *token* de localização restantes, ou seja: "Paulo"; 3. B-ORG: Organização do primeiro *token* de entidade, ou seja: "Banco"; 4. I-ORG: *token* restantes da organização, ou seja: "Votorantim"; 5. B-PER: *token* de entidade da primeira pessoa, i. e .: "Paulo"; 6. I-PER: *token* de pessoa restante, i. e .: "de", "Tarso"; 7. O: *token* restantes não classificáveis, i. e .: ”Transferência”.

Finalmente, foi preciso padronizar o tamanho das listas presentes no vetor, uma vez que a rede neural artificial precisa que todos os vetores tenham o mesmo tamanho. Para a padronização de vetores, verificou-se que o tamanho máximo de *tokens* existentes nos textos era de 75 caracteres. Assim, *tokens* menores que 75 caracteres foram preenchidos com zeros.

6.3 MÓDULO DE TREINAMENTO

Nesta etapa, as técnicas escolhidas para a criação do modelo foram as redes neurais *Long short-term memory* (LSTM), com os algoritmos *Conditional Random Fields* (CRF) e máquina de vetores. O estudo apresentado por Suzgun, Belinkov e Shieber (2018), mostra a flexibilidade das redes neurais recorrentes em generalizar modelos para processamento de linguagem natural. O estudo avaliou variações da rede LSTM para a generalização de modelos em idiomas formais.

Figura 13 – Rede neural recorrente com memória de longo prazo bidirecional



Fonte: Adaptada de Olah (2015)

A rede neural escolhida é composta por unidades *Long short-term memory* (LSTM), que é uma variação das redes tradicionais *Recurrent neural network* (RNN's). As unidades *Long short-term memory* (LSTM) possuem a capacidade de persistir as informações por um tempo arbitrário. Esse tipo de unidade neural é muito utilizado nas áreas de reconhecimento de fala, tradução, legenda de imagens e processamento de linguagem natural. Essas áreas possuem como característica comum a existência de padrões em relação de longo prazo. Para que os padrões dispersos ao longo do tempo possam ser capturados pela rede neural, o uso de unidades LSTM é essencial, devido a sua capacidade de conectar informações de estados anteriores com estados correntes.

A rede neural escolhida interliga várias unidades LSTM nos dois sentidos (forward e backward). Essa arquitetura de rede neural é conhecida como *Bidirectional Long*

Short-Term Memory (BLSTM), que possui a capacidade de predição em dois sentidos. A predição bidirecional tem como objetivo maximizar a entropia.

O módulo responsável pelo treinamento recebe os vetores do pré-processamento e realiza o treino. Os resultados do treinamento são: i) o modelo em arquivo binário para ser reutilizado sem a necessidade de novo treinamento; ii) um relatório de métricas de saída em relação à base de validação; iii) o tempo de execução; iv) os parâmetros usados para construção do modelo; v) o código de identificação de cada trecho de texto da base de dados, que foram usados durante o processo; vi) o histórico de cada iteração da rede com as métricas de Loss; vii) o erro; viii) um arquivo contendo: todos os *tokens* usados na validação, o valor verdadeiro e o classificado pelo modelo.

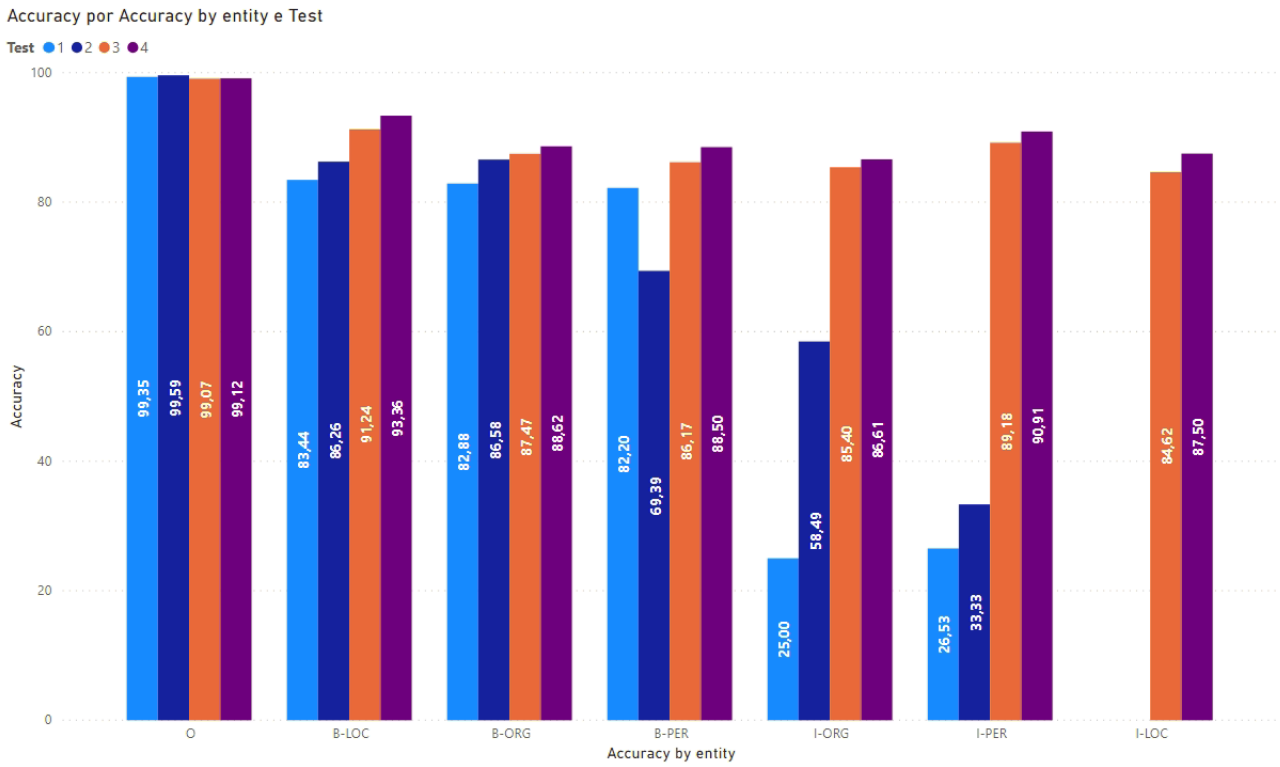
6.4 CONTROLE DE VERSÃO

O módulo de controle de versão é responsável pelo carregamento dos dados salvos, que incluem: o modelo, o dicionário de palavras, o dicionário de caracteres, o dicionario da etiqueta, os dados convertidos em número e a conversão para o formato de texto novamente. A principal função desse módulo é manter um histórico de versão dos modelos criados no processo de treinamento. Portanto, é possível verificar a evolução do modelo em relação à precisão e, assim, identificar as mudanças que foram significativas. Portanto, é possível verificar a evolução do modelo, em relação à precisão, e identificar as mudanças que foram significativas.

7 RESULTADOS

Para treinamento do modelo, foram utilizados 45000 trechos do CETENFolha. Os trechos foram selecionados de modo a ter ao menos uma entidade do tipo Organização, Pessoa ou Local, podendo ou não ter outras entidades no mesmo trecho. As entidades poderiam ter um ou mais *tokens* (termos), sendo o primeiro classificado como “B”+“entidade” e os subsequentes como “I”+“entidade”. Dentre o conjunto de textos, 80% foram usados para treino do modelo e 20% usados para validação. Também foi gerado um vocabulário com 95518 *tokens* únicos.

Figura 14 – Acurácia por execução



Fonte: Elaborada pelo autor

Conforme Figura 14 - Acurácia por execução, conforme é realizado uma nova rodada de aprendizagem, são inseridas maiores quantidades de trechos, inicialmente n resulta no aumento da acurácia, mesmo com a redução de *epochs* que são as quantidades de iterações realizadas para aprendizagem. No caso do teste 2 e 3, é possível observar que quantidade de trechos inseridos influência na acurácia até um certo limite, foram inseridos mais 9 mil trechos entre os dois testes representando um aumento de 47.36% de trechos novos, obtendo uma acurácia isolada de 83,65% para o teste 3, em relação ao teste 2 com acurácia isolada de 76,28%, houve um aumento de 7,37% na acurácia isolada. Já entre o teste 3 e 4, foram adicionados 22 mil novos

trechos representando um aumento de 107.14% de trechos novos inseridos, porém o aumento da acurácia isolada foi apenas 1,99%, evidenciando que a quantidade de trechos é suficiente para o aprendizado dentro do domínio da linguagem escolhido.

Com a redução dos *epochs*, houve por consequência a diminuição do tempo de execução para aprendizagem. Os parâmetros para cada rodada de aprendizagem podem ser visualizados na Tabela 3 - Tabela parâmetros de execução por teste.

Tabela 3 – Tabela parâmetros de execução por teste

Teste	Quantidade de trechos	Epochs	Tempo de execução (min)	Acurácia	Acurácia isolada
1	18001	40	1334	98,52%	72,65%
2	19001	20	43	98,69%	76,28%
3	28001	20	74	97,23%	83,65%
4	50001	20	147	97,54%	85,64%

Fonte: Elaborada pelo autor

As métricas utilizadas para avaliação do modelo, conforme apresentadas na tabela 4, são: i) Acurácia, que representa o percentual de acertos do modelo em relação à coluna tokens de validação; ii) o Erro, que é o percentual de *tokens* que não foram classificados corretamente pelo modelo; iii) o F-measure, representado como F1, é o total de *tokens* de validação dividido pelo total que foram classificados corretamente, mais o total de falsos positivos; iv) o Recall, representado pela coluna *tokens* validação, que é o total classificado corretamente dividido pelo total que deveria ser classificado, em relação a cada tipo de entidade presente no conjunto.

Tabela 4 – Avaliações experimentais

Entidade	Tokens Treinamento	Tokens Validação	Acertos	Erros	Acurácia	Erro	F1	Recall
B-PER	102537	20295	17961	2334	88.49963045084996	11.500369549150037	93.898996	88.49963
I-PER	77276	15152	13774	1378	90.90549102428722	9.094508975712777	95.23612	90.905491
B-ORG	91538	18320	16236	2084	88.62445414847161	11.375545851528384	93.969209	88.624454
I-ORG	60155	11919	10323	1596	86.60961490057892	13.390385099421092	92.824386	86.609615
B-LOC	74784	14886	13897	989	93.35617358591965	6.643826414080343	96.563944	93.356174
I-LOC	21498	4271	3737	534	87.49707328494497	12.502926715055022	93.331668	87.497073
O	2191533	432558	428738	3820	99.11688143555315	0.8831185644468488	99.556482	99.116881

Fonte: Elaborada pelo autor

Conforme Tabela 4 - Avaliações experimentais, pode-se notar que as duas entidades com as menores acurácias isoladas são a “I-ORG” e “I-LOC”, com 86,60% e 87,49% respectivamente. O resultado da acurácia para as entidades “I-ORG” e “I-LOC” se deve ao fato de que elas somente classificam *tokens* intermediários, e neste caso,

há uma menor quantidade de *tokens* representados nos conjuntos de treinamento e validação. Dessa forma, espera-se que o valor de entidades do tipo “I” tenha acurácia inferior ao do tipo “B”.

7.1 CONTRIBUIÇÕES

Dentre as contribuições geradas a partir do desenvolvimento deste projeto, podemos citar:

1. A redução do tempo de pré-processamento com a criação de uma metodologia que possibilita a armazenagem dos vetores já processados.
2. O armazenamento do resultado do calculo de métricas e parâmetros utilizados durante a criação do modelo, possibilita uma análise histórica do modelo contribuindo para a melhora na evolução. Com isso, é possível mapear os resultados em relação às alterações de parâmetros.
3. Aumento na acurácia em relação a outros trabalhos encontrados na literatura, conforme visto no capítulo "Trabalhos Relacionados".

8 CONCLUSÃO

Os resultados alcançados foram promissores. As evidências indicam que, por meio da restruturação do corpus, foi possível criar modelos relativamente próximos, em termos de assertividade, aos modelos baseados na língua inglesa.

Antes da criação da metodologia proposta neste estudo, era necessário, a cada tentativa de criação de um modelo de extração de entidades, realizar o processamento de todos os trechos presentes no corpus, utilizando a seguinte sequência de eventos: i) realizar o *parser*; ii) gerar vetores, contendo: *tokens* de palavras, classificação gramatical e *pos-taggings*; iii) converter as anotações do *pos-tagging* para que sejam normalizadas. Esse conjunto de atividades demanda muito tempo de processamento. Após a abordagem proposta, todos os vetores processados são armazenados, sem a necessidade de refazer o trabalho a cada nova tentativa. Para exemplificar o ganho no tempo de processamento, foram necessárias 4 horas para processar 2 milhões de trechos, e apenas 1 hora para processar com a abordagem metodológica proposta, ou seja, uma redução de 75% no tempo.

Outro beneficio gerado durante o desenvolvimento, foi o módulo de controle de versão, que armazena todas as informações de parâmetros, tempo de execução, resultados, métricas e cada identificador do texto usado no treinamento e no teste do modelo. Dessa forma, temos um painel consolidado com gráficos que mostram a evolução de cada versão de modelo gerado, a partir dos parâmetros escolhidos.

Consequentemente, os dados gerados durante o desenvolvimento deste projeto podem auxiliar os profissionais da área de Processamento de Linguagem Natural que utilizam a língua portuguesa para extração de entidades nomeadas. A criação de um método que facilita o planejamento do volume de exercícios para treinamento; uma nova estrutura de corpus; possibilidade de selecionar apenas as entidades mais relevantes no momento da formação; número de iterações necessárias para atingir um bom nível de precisão sem causar *overfitting* (quando o modelo se ajusta muito bem aos dados de treino, porem, não é eficaz em novos dados), serão os principais pontos de contribuição deste trabalho.

Apesar dos desafios encontrados ao longo deste trabalho, pode-se afirmar que o objetivo geral foi alcançado, uma vez que foi possível criar uma metodologia eficaz capaz de extrair entidades, pois, tanto a acurácia geral quanto a acurácia por entidade apresentaram resultados superiores a 80%.

9 PUBLICAÇÕES REALIZADAS

Durante o desenvolvimento dessa dissertação foram alcançadas as seguintes publicações:

1. Reconhecimento de Entidades Nomeadas Em Relatórios de Inteligência Financeira, no evento. In: WIDAT 2019: Workshop de informação, dados e tecnologia. João Pessoa: Editora UFPB, 2018 (SANTANA *et al.*, 2018).

2. *A New Entity Extraction Model Based on Journalistic Brazilian Portuguese Language to Enhance Named Entity Recognition*, na revista *DIONE 2020: Data and Information in Online Environments*, apresentado na conferência *International Conference on Data and Information in Online* (AQUINO SILVA *et al.*, 2020a).

3. *An Improved NER Methodology to the Portuguese Language* na revista, *Mobile Networks and Applications* (AQUINO SILVA *et al.*, 2020b).

4. Grau de pertencimento como insumo para classificação automática de textos: uma abordagem sintática. na revista CIÊNCIA DA INFORMAÇÃO (ONLINE): Editora IBCT, 2020 (DYCK *et al.*, 2020)

5. *Evaluating the Effect of Corpus Normalisation in Topics Coherence*, na revista *DIONE 2021: Data and Information in Online Environments*, apresentado na conferência *International Conference on Data and Information in Online* (SILVA SOUSA *et al.*, 2021)

10 PRÓXIMOS PASSOS

A partir do resultado obtido nesta pesquisa planeja-se outros objetivos para trabalhos futuros, sendo eles: aumentar o número de trechos para cobrir todas as entidades durante todas as iterações de teste, uma vez que entidades com menores quantidades podem ser prejudicadas conforme os resultados de acurácia por entidade; a criação de novos modelos utilizando outras variações de redes neurais e comparar os modelos com trechos que não sejam do domínio da linguagem jornalística. Também pretende-se construir uma plataforma que possibilite utilizar modelos criados por outros sistemas, passando apenas os parâmetros de nome do modelo e trecho a ser processado criando.

REFERÊNCIAS

ABDELALI, Ahmed; COWIE, James; SOLIMAN, H. Building a modern standard Arabic corpus. *In*: WORKSHOP on computational modeling of lexical acquisition. [*S.l.*: *s.n.*], 2005. P. 25–28.

AHO, Alfred V. **Algorithms for finding patterns in strings, Handbook of theoretical computer science (vol. A): algorithms and complexity**. [*S.l.*]: MIT Press, Cambridge, MA, 1991.

AMARAL, Daniela Oliveira F do; VIEIRA, Renata. NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de Conditional Random Fields. **Linguamática**, v. 6, n. 1, p. 41–49, 2014.

AQUINO SILVA, Rogerio de *et al.* A New Entity Extraction Model Based on Journalistic Brazilian Portuguese Language to Enhance Named Entity Recognition. *In*: MUGNAINI, Rogério (Ed.). **Data and Information in Online Environments**. Cham: Springer International Publishing, 2020. P. 53–63.

AQUINO SILVA, Rogerio de *et al.* An Improved NER Methodology to the Portuguese Language. *In*: MOBILE Networks and Applications. Cham: Springer International Publishing, 2020.

ARAUJO, Pedro H. Luz de *et al.* LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text. *In*: INTERNATIONAL Conference on the Computational Processing of Portuguese (PROPOR). Canela, RS, Brazil: Springer, set. 2018. (Lecture Notes on Computer Science (LNCS)), p. 313–323. DOI: 10.1007/978-3-319-99722-3_32. Disponível em: https://cic.unb.br/~teodecampos/LeNER-Br/.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with Python: analyzing text with the natural language toolkit**. [*S.l.*]: "O'Reilly Media, Inc.", 2009.

BORGES, Rafael R. *et al.* Sincronização de disparos em redes neuronais com plasticidade sináptica. **Revista Brasileira de Ensino de Física**, v. 37, n. 2, p. 313–330, 2015. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-11172015000200011&lng=pt&tlng=pt#aff01. Acesso em: 18 set. 2020.

BROWNLEE, Jason. **How to Develop a Bidirectional LSTM For Sequence Classification in Python with Keras.** 2017. Disponível em: https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classificati%20on-python-keras/. Acesso em: 27 jul. 2019.

CARVALHO, Wesley Seidel. Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina. **Universidade de São Paulo**, São Paulo, 2012. DOI: 10.11606/D.45.2012.tde-23052013-104248. Disponível em: https://www.teses.usp.br/teses/disponiveis/45/45134/tde-23052013-104248/publico/dissertacao_rem_wesley_seidel.pdf. Acesso em: 10 ago. 2020.

CAVALIN, Paulo *et al.* Building a question-answering corpus using social media and news articles. *In*: SPRINGER. INTERNATIONAL Conference on Computational Processing of the Portuguese Language. [*S.l.*: *s.n.*], 2016. P. 353–358.

CHODOROW, Kristina; DIROLF, Michael. **MongoDB: The Definitive Guide**. 1st. [*S.l.*]: O'Reilly Media, Inc., 2010. ISBN 1449381561.

CHOLLET, François *et al.* **Keras**. [*S.l.*: *s.n.*], 2015. https://keras.io.

COPPIN, Ben. **Artificial intelligence illuminated**. [*S.l.*]: Jones e Bartlett Publishers, 2004. 1 ed. ISBN 0-7637-3230-3. Disponível em: http://library.lol/main/291D68C3706BFBAD466258CAB359DCD8. Acesso em: 10 ago. 2020.

CORPUS. **In: DICIO, Dicionário Online de Português.** 2020. Disponível em: https://www.dicio.com.br/corpus/. Acesso em: 20 jun. 2020.

DANIEL, Jurafsky; MARTIN, James H. **Speech and Language Processing**. 2. ed. [*S.l.*]: Prentice Hall, 2008. ISBN 9780131873216,0131873210. Disponível em: http://gen.lib.rus.ec/book/index.php?md5=6ead868d4b528f58eba01315ab7c7390.

DEVLIN, Jacob *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. [*S.l.*: *s.n.*], 2019. arXiv: 1810.04805 [cs.CL].

DIAS, Célia da Consolação. A análise de domınio, as comunidades discursivas, a garantia de literatura e outras garantias. **Informação & Sociedade**, Universidade Federal da Paraıba-Programa de Pós-Graduação em Ciência da . . ., v. 25, n. 2, 2015.

DYCK, A. F. *et al.* Grau de pertencimento como insumo para classificação automática de textos: uma abordagem sintática. IBICT, v. 49, p. 19–33, 2020.

FELDMAN, R.; DAGAN, I. Knowledge Discovery in Textual Databases (KDT). *In*: KDD. [*S.l.*: *s.n.*], 1995.

GRUS, Joel. **Data Science do Zero**: Primeiras Regras com o Python. [*S.l.*]: Editora Alta Books, jun. 2016. ISBN 978-8576089988.

HAYKIN, Simon. **Redes neurais**. 2nd. [*S.l.*]: Bookman, 2005. ISBN 9788573077186. Disponível em:

http://gen.lib.rus.ec/book/index.php?md5=8b97272ccefcbd87bf869fd0c292d7c2. Acesso em: 18 set. 2020.

HOLTHAUS, Patrick *et al.* How to address smart homes with a social robot? A multi-modal corpus of user interactions with an intelligent environment. *In*: PROCEEDINGS of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). [*S.l.*: *s.n.*], 2016. P. 3440–3446.

IBGE. **Acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal**. 2017. Disponível em: https://biblioteca.ibge.gov.br/index2.php/biblioteca-catalogo?view=detalhes%5C&id=2101631. Acesso em: 10 jul. 2019.

IBPAD. **O que é Linguística de Corpus? – Veja 5 aplicações**. 2018. Disponível em: https://www.ibpad.com.br/blog/comunicacao-digital/o-que-e-linguistica-de-corpus-vej%20a-5-aplicacotes/. Acesso em: 22 jun. 2019.

KLUBECK, Martin. **Metrics**: How to Improve Key Business Results. 1. ed. [*S.l.*]: Apress, 2011. ISBN 1430237260,9781430237266. Disponível em: http://gen.lib.rus.ec/book/index.php?md5=762f8d2bea1a14ed30346730618a27e9.

LINGUATECA. **CETENFolha**. 2018. Disponível em: https://www.linguateca.pt/cetenfolha/index_info.html. Acesso em: 20 jun. 2020.

MARCUS, Mitchell P.; SANTORINI, Beatrice; MARCINKIEWICZ, Mary Ann. Building a Large Annotated Corpus of English: The Penn Treebank. **Computational Linguistics**, v. 19, n. 2, p. 313–330, 1993. Disponível em: https://www.aclweb.org/anthology/J93-2004.

MCCULLOCH, Warren; PITTS, Walter. A Logical Calculus of Ideas Immanent in Nervous Activity. **Bulletin of Mathematical Biophysics**, v. 5, p. 127–147, 1943.

MICHAEL, Covington A. **Natural Language Processing for Prolog Programmers + Revision 2009**. [*S.l.*: *s.n.*], 2009. Disponível em: http://gen.lib.rus.ec/book/index.php?md5=7d5ecc8af42048caa6e1481204f6bced. Acesso em: 10 ago. 2020.

MITCHELL, T.M. **Machine Learning**. [*S.l.*]: McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: https://books.google.com.br/books?id=EoYBngEACAAJ.

MOOERS, Calvin N. The next twenty years in information retrieval; some goals and predictions. **American Documentation**, Wiley Online Library, v. 11, n. 3, p. 229–236, 1960.

MOSLEY, Mark *et al.* **DAMA guide to the data management body of knowledge**. [*S.l.*]: Technics Publications, 2017.

OLAH, Christopher. Understanding lstm networks, 2015. Disponível em: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Acesso em: 22 jun. 2019.

OLIVEIRA, Maxwell Guimarães de *et al.* A gold-standard social media corpus for urban issues. *In*: ACM. PROCEEDINGS of the Symposium on Applied Computing. [*S.l.*: *s.n.*], 2017. P. 1011–1016.

PEDREGOSA, F. *et al.* Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PRETTO, Juliana Regina. O estilo jornalıstico. **ESTUDOS LINGUÍSTICOS**, v. 38, n. 3, p. 481–491, 2009.

REZENDE, S.O. **Sistemas inteligentes**: fundamentos e aplicações. [*S.l.*]: Manole, 2003. ISBN 9788520416839. Disponível em: https://books.google.com.br/books?id=UsJe%5C_PlbnWcC. Acesso em: 10 ago. 2020.

ROCHA, Vinicius Roggério da. **Perceptron – redes neurais**. Jun. 2017. Disponível em: https://www.monolitonimbus.com.br/perceptron-redes-neurais/. Acesso em: 10 ago. 2020.

SAMPAIO, RF; MANCINI, MC. Estudos de revisão sistemática: um guia para sıntese criteriosa da evidência cientıfica. **Revista Brasileira de Fisioterapia**, FapUNIFESP (SciELO), v. 11, n. 1, p. 83–89, fev. 2007. DOI: 10.1590/s1413-35552007000100013. Disponível em: https://doi.org/10.1590/s1413-35552007000100013.

SANTANA, J. *et al.* Reconhecimento de entidades nomeadas em relatórios de inteligência financeira. UFPB, v. 2, p. 312–318, 2018.

SARDINHA, Tony Berber. Lingüística de Corpus: Histórico e Problemática. pt. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, scielo, v. 16, p. 323–367, 2000. ISSN 0102-4450. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&nrm=iso.

SILVA SOUSA, Luana da *et al.* Evaluating the Effect of Corpus Normalisation in Topics Coherence. *In*: BISSET ÁLVAREZ, Edgar (Ed.). **Data and Information in Online Environments**. Cham: Springer International Publishing, 2021. P. 197–208.

SPOUSTOVÁ, Drahomíra; SPOUSTA, Miroslav; PECINA, Pavel. Building a web corpus of czech.

SUZGUN, Mirac; BELINKOV, Yonatan; SHIEBER, Stuart M. On Evaluating the Generalization of LSTM Models in Formal Languages. **arXiv preprint arXiv:1811.01001**, 2018.

TAGNIN, Stella; TEIXEIRA, Elisa. Lingüística de Corpus e Tradução Técnica - Relato da montagem de um corpus multivarietal de culinária. **Tradterm**, v. 10, p. 313–358, dez. 2004. DOI: 10.11606/issn.2317-9511.tradterm.2004.47184. Disponível em: http://www.revistas.usp.br/tradterm/article/view/47184.

VAN ROSSUM, Guido; DRAKE, Fred L. **Python 3 Reference Manual**. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697.

VIEIRA, Renata; LOPES, Lucelene. Processamento de linguagem natural e o tratamento computacional de linguagens científicas: Linguagens Especializadas em Corpora: modos de dizer e interfaces de pesquisa. EDIPUCRS, Porto Alegre, p. 183, 2010.

VILLALVA, Alina; MATEUS, Maria Helena Mira. **Morfologia do português**. [*S.l.*]: Universidade Aberta Lisboa, 2008. ISBN 9789726744870.

APÊNDICE A – REVISÃO SISTEMÁTICA DE LITERATURA

O documento a seguir trata-se de uma revisão sistemática de literatura, realizada durante do processo de confecção da metodologia atual. O mesmo será atualizado durante o processo de desenvolvimento devido o avanço de pesquisas que estão em produção na área de tecnologia.

Como será abordado a no documento de Revisão Sistemática de literatura foram consultadas 3 bases SCORPUS, LISA e IEEE com os filtros Período: 2017 a 2019 nos idiomas Português e Inglês. Foi retornado o total de 169 documentos a serem analisados com os critérios aplicados, que foram incluídos e excluídos da revisão de acordo com os critérios de inclusão e exclusão apresentados no decorrer do documento.

SUMÁRIO

1 INTRODUÇÃO

A recuperação de informação (RI) surgiu a partir de esforços para facilitar a manipulação de dados em grandes bases (FERNEDA, 2018). Os esforços para RI ainda encontram grandes desafios quando se trata de processamento de linguagem natural (PLN). A extração de informação a partir de texto em português ainda constitui um campo aberto de investigação devido a baixa acurácia em vários aspectos como reconhecimento de entidades, análise semântica e sintática, extração de relacionamentos, dentre outros (Amaral e Vieira 2013). Somado ao desafio de PLN para português, há uma escala crescente de dados gerados, devido ao sucessivo aumento no número de usuários da internet (IBGE, 2017). Estima-se que a cada segundo são gerados milhares de dados em uma diversidade de formatos de dados nas como imagens, textos, vídeos e áudios.

Um estudo realizado pelo Data Management Association (Dama) constatou que atualmente existam mais de 500 quatrilhões em dados armazenadas no universo digital. O estudo também aponta, que a cada dois anos, a produção de dados dobra, com previsão de chegar a 350 zettabytes em 2020 (DMBOK et al. 2017). Grande parte dos dados gerados, cerca de 80%, são dados não estruturados, como textos. A crescente escala de produção de textos, juntamente com a baixa acurácia para o reconhecimento de entidades nomeadas em português, são questões que podem ser enfrentadas com técnicas modernas de inteligência artificial.

Com a utilização de técnicas de aprendizagem de máquina é possível criar modelos com capacidade de extrair entidades a partir de textos. Na língua inglesa, por exemplo, existem modelos que possuem acurácia de 92,6% criados a partir da utilização do framework spaCy e 91,7% com o framework ClearNLP, que voltados ao treinamento de modelos para extração de entidades. Segundo (Amaral e Vieira 2013), em português, a acurácia fica em torno de 80,77%, utilizando como base de treinamento o corpus HAREM e redes neurais CRF (Amaral, et. al 2014), pois para o modelo atingir um alto desempenho é necessário um conjunto de dados já classificado com notações sobre a estrutura gramatical e entidades existentes nele. Esse conjunto é conhecido como corpus e normalmente a atividade de classificar os trechos é realizada por linguistas que são aptos a realizar essa tarefa, por conhecerem a estrutura do idioma e suas propriedades. Entretanto, como cada corpus é criado para uma finalidade, nem sempre possuem a mesma estrutura, não existindo assim um padrão. Além disso, (Villalva 2007) ressalta que a morfologia e a sintaxe da língua portuguesa possui características próprias criando complexidades quando comparada a língua inglesa que possui menos elementos em sua notação gramatical quanto à conjugação de verbos.

O domínio da informação também é uma questão a ser considerada, pois conforme (Dias 2015) uma determinada comunidade pode possuir hábitos em relação ao uso da informação, como realizar suas buscas e como organizam seus novos conhecimentos. Estes hábitos afetam a escrita e a estrutura sintática de alguns termos, mesmo que o idioma seja igual em comunidades diferentes. Sendo a linguagem jornalística, o gênero textual que possui melhor aderência a contemporaneidade do idioma, algumas das características da escrita jornalística como a objetividade, a simplicidade, a imparcialidade e o referencial, evitam termos em desuso pois a informação deve ser transmitida de forma clara ao leitor (PRETTO, 2009). Por esse motivo, escolhemos o corpus baseados em textos jornalísticos para criar a base de treinamento para o reconhecimento de entidades nomeadas.

Existem diversas técnicas de aprendizagem de máquina que possibilitam a criação de modelos capazes de classificar as entidades, sendo o mais utilizado para este fim, redes neurais recorrentes (RNN) (OLAH, 2015). Geralmente as RNNs são aplicadas em problemas onde existam a necessidade de reconhecimento de padrões que variam em uma determinada série temporal (NELSON, 2017).

A presente pesquisa apresenta uma Revisão Sistemática da Literatura (RSL) sobre "extração de entidades nomeadas e relacionamentos de documentos com a utilização de aprendizagem de máquina", realizada no mês de Julho de 2019. O presente estudo foi organizado com a seguinte disposição: metodologia (seção 2); resultados preliminares (seção 3) com o protocolo de pesquisa, revisão e seleção dos estudos recuperados, análise dos estudos selecionados, resultados e sumarização dos resultados; os trabalhos relacionados (seção 4); as considerações finais (seção 5), e um apêndice contendo os critérios de exclusão dos artigos recuperados que não foram incluídos nos trabalhos relacionados.

2 METODOLOGIA

A metodologia utilizada no processo de construção desta RSL emprega o uso do método descrito por Sampaio e Mancini (2007), no qual é descrito cinco etapas para constituir uma revisão sistemática. A primeira etapa prevê a criação do protocolo de pesquisa, que estabelece os termos de busca bem como os critérios de inclusão, exclusão que devem focar na busca pela qualidade metodológica dos documentos recuperados.

Conforme passos definidos por Sampaio e Mancini (2007), o primeiro passo é definição da pergunta, a mesma deve ser clara, com a descrição do estudo, população e contexto. O segundo passo é a busca de evidências que tragam todos documentos que

sejam relevantes para a temática abordada que vão ser analisados durante a RSL, que vão ser localizados a partir dos termos de busca e demais estratégias nas bases de dados. O terceiro passo é analisado todos os títulos e resumos dos documentos retornados e realizado a inclusão e exclusão de acordo com o critérios propostos no protocolo de pesquisa. Os documentos que forem incluídos devem ser lidos em sua totalidade e novamente aplicado os critérios de inclusão e exclusão. Por último, no quinto passo são apresentados os resultados em relação aos documentos incluídos no estudo após a validação dos critérios de inclusão e exclusão, contendo as principais características, métodos e resultados de cada um.

Após as etapas abordadas, são criados relatórios contendo os documentos, incluídos, excluídos e os critérios utilizados na análise.

Quadro 1 - Passos metodologia Sampaio e Mancini (2007)

PASSO	**DESCRIÇÃO**
Passo 1 – Pergunta	É possível a criação de um modelo padrão de extração de entidades nomeadas em documentos escritos na língua portuguesa com o emprego de redes neurais recorrentes e com base em corpus disponíveis na internet?
Passo 2 – Busca	Bases de dados: SCORPUS, IEEE, *Library and Information Science Abstracts* (LISA) e WEB OF SCIENCE Palavras-chave: Named Entity Recognition;corpus;recurrent neural networks Idioma: Inglês e português Tipo de publicação: Artigos científicos Período: 2017 a 2019
Passo 3 – Revisão e seleção de documentos	Critérios de inclusão: (1. Artigos com textos completos; 2. Artigos revisados por especialistas; 3. Artigos nos idiomas inglês e português; 4. Tipo de documento Artigos científicos; 5.) Critérios de exclusão: (1. Artigos com temática diferente; 2. Artigos com idioma que não seja inglês e espanhol; 3. Apresentam redes neurais que não sejam LSTM ou variações; 4. Não possui palavras-chaves no resumo;5. Artigos em duplicidade; 6. Documentos que não sejam artigos científicos; 7. Artigos que não possuem corpus disponíveis na internet; 8. Artigos incompletos)

Passo 4 – Analisando a qualidade metodológica dos documentos	Leitura completa dos documentos aplicando os critérios de inclusão e exclusão
Passo 5 – Apresentação de Resultados	Descrição das características e resultados dos documentos que resultarem após aplicação dos critérios.

Fonte: Dados da pesquisa, 2019.

3 RESULTADOS PRELIMINARES

Na presente seção serão abordados os principais passos realizados durante a RSL, com o protocolo de pesquisa que contém os elementos norteadores e estratégia de busca utilizados nas bases, seguindo para a revisão e seleção dos documentos recuperados onde são aplicados os critérios de inclusão e exclusão e por último a leitura completa dos documentos que passaram pelos critérios de inclusão.

3.1 PROTOCOLO DE PESQUISA

Com base no objetivo geral da pesquisa, foram criados os critérios para seleção dos documentos nas bases de dados: SCOPUS, IEEE, Web of Science e LISA. À escolha das bases buscou à maior variedade possível de publicações visto que são multidisciplinares. O termo de busca foi: "Named Entity Recognition" AND corpus AND "recurrent neural networks", limitando a busca pelo período de 2017 a 2019, a fim de obter os estudos mais recentes em relação à problemática. À string de busca utilizada foi em língua inglesa devido ao grande volume de documentos nas bases utilizar o idioma inglês no texto ou *abstract*.

O tipo de documento selecionado foi artigos científicos, que disponibilizam o texto completo e que estejam em ingles ou portugues.

Quadro 2 - Protocolo de Revisão Sistemática de Literatura

ELEMENTOS NORTEADORES	**DESCRIÇÃO**

Objetivo		Criar um modelo base de extração de entidades nomeadas baseados em corpus extraídos da internet com o emprego de redes neurais recorrentes
Problema da Pesquisa		É possível a criação de um modelo padrão de extração de entidades nomeadas em documentos escritos na língua portuguesa com o emprego de redes neurais recorrentes e com base em corpus disponíveis na internet?
Palavras-chave	Português	Reconhecimento de entidades nomeadas; corpus; redes neurais recorrentes
	Inglês	Named Entity Recognition;corpus;recurrent neural networks
Bases		SCORPUS, IEEE, LISA e WEB OF SCIENCE
Tipos de Documentos		Artigos científicos
Idioma dos documentos		Inglês e Português
Campos de Busca		“texto completo” e “qualquer campo”
Operadores de busca		AND
Termo de busca		"Named Entity Recognition" AND corpus AND "recurrent neural networks"
Período		2017 a 2019

Fonte: Dados da pesquisa, 2019.

Inicialmente foram retornados, 155 documentos na base SCORPUS, 7 na base LISA, 2 na base *Institute of Electrical and Electronics Engineers* (IEEE) e 5 na Web of science.

Quadro 3 - Documentos retornados

Base de Dados	**Campos Selecionados**	**Período**	**Termo de Busca**	**Resultados preliminares**
SCORPUS		2017 a 2019		155
LISA		2017 a 2019		7
IEEE		2017 a 2019		2

Web of Science		2017 a 2019		5
Total				169

Fonte: Dados da pesquisa, 2019.

Ao serem executadas as buscas notasse que nem toda base possui todos campos para seleção dos filtros, sendo necessário à leitura e aplicação dos critérios para inclusão e exclusão.

3. 2 REVISÃO E SELEÇÃO DOS DOCUMENTOS RECUPERADOS

Após à recuperação dos documentos ocorre à análise para seleção dos documentos que conforme aplicação dos critérios de inclusão e exclusão, conforme o Quadro 4.

Quadro 4 - Critérios de inclusão e exclusão

Critérios	Incluir	Excluir
1. Artigos com textos completos	X	
2. Artigos revisados por especialistas	X	
3. Artigos nos idiomas inglês e português	X	
4. Tipo de documento Artigos científicos	X	
5. Artigos com temática diferente		X
6. Artigos com idioma que não seja inglês e espanhol		X
7. Apresentam redes neurais que não sejam LSTM		X
8. Não possui palavras-chaves no resumo		X
9. Artigos em duplicidade		X
10. Documentos que não sejam artigos científicos		X
11. Artigos que não possuem corpus disponíveis na internet		X
12. Artigos incompletos		X

Fonte: Dados da pesquisa, 2019.

Optou-se pela criação de critérios que possam melhorar à avaliação dos documentos recuperados. Os critérios de exclusão são: Artigos com temática diferente que são artigos encontrados que possuem objetivo diferente ou divergem em relação ao tema proposto neste estudo; Apresentam redes neurais que não sejam do tipo LSTM ou variações ou seja empregam o uso de redes neurais diferentes de redes recorrentes; Não possui palavras-chave conforme as palavras definidas no protocolo; Artigos em Duplicidade; Documentos que não sejam artigos científicos e Artigos que não possuem corpus disponíveis na internet, pois dessa forma não é possível reproduzir os resultados do estudo uma vez que à base de dados usada não é de acesso público, problema este que ocorre em situações onde base de treinamento é relacionada à alguma empresa ou possui dados sensíveis.

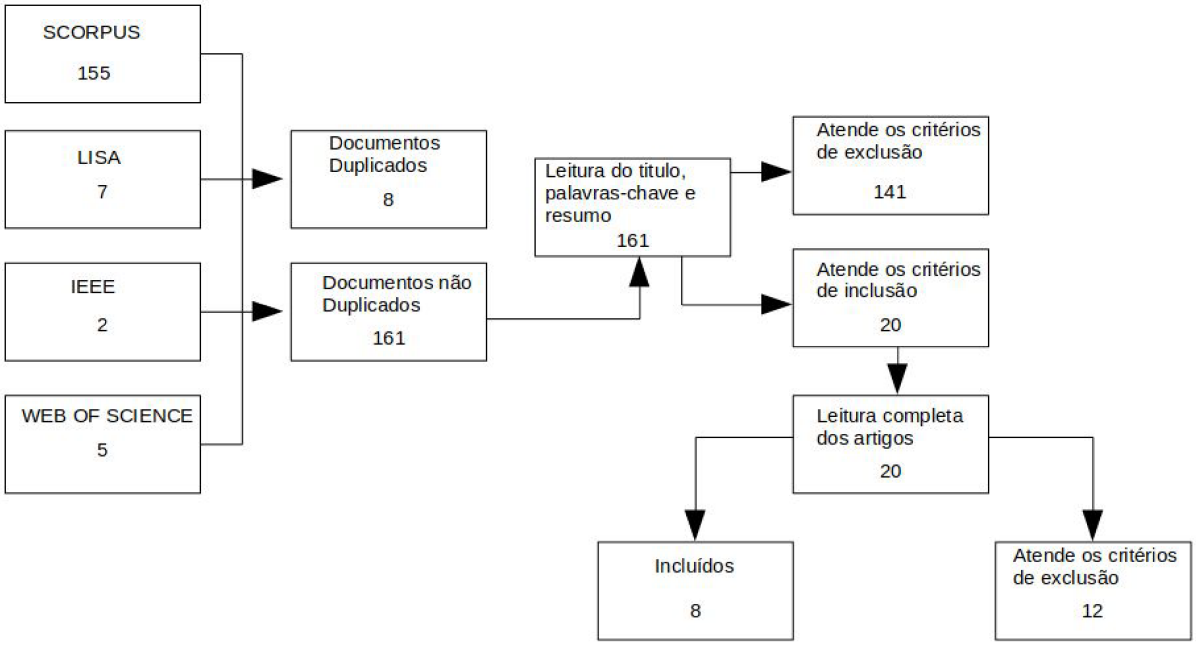
Para aplicação dos critérios, inicialmente optou-se pela leitura dos títulos, palavras-chave e resumos em todos 161 documentos recuperados. No Apêndice I são apresentados todos os artigos que foram excluídos devido aos critérios de exclusão com ano de publicação, autor, título, tipo de publicação, base de dados, local e o critério de exclusão aplicado.

3.3 ANÁLISE DOS DOCUMENTOS RECUPERADOS

Após à recuperação dos 169 documentos foram retirados 8 documentos duplicados, resetando 161 documentos, nos quais foram realizados à leitura dos títulos, palavras-chave e resumo, aplicando os critérios de inclusão e exclusão, obteve-se 20 aplicando os critérios de inclusão e 141 em relação aos critérios de exclusão.

Ao final foram lidos os 20 artigos em sua totalidade e aplicados novamente os critérios. Restando 8 artigos que atendem os critérios de inclusão e 12 aos critérios de exclusão. À análise foi executada conforme Figura 1.

Figura 1 - Etapas da análise dos documentos



Fonte: Dados da pesquisa, 2019.

Os artigos incluídos foram lidos em sua totalidade, sendo descritos na seção 4 trabalhos relacionados, com objetivo e resultados, extraindo uma síntese de cada artigo.

4 TRABALHOS RELACIONADOS

Na realização desta RSL foram obtidos 8 artigos nas bases de dados: SCORPUS, LISA, IEEE e Web of Science, que atendem os critérios de inclusão em sua totalidade.

Os trabalhos não atendem ou atendem parcialmente os critérios, foram excluídos por abordarem assuntos que não são relacionados à temática abordada ou tecnologia empregada.

Abaixo, é apresentado os 8 artigos que foram incluídos e usados como trabalhos relacionados, por existirem pontos de convergência em relação ao tema proposto (Quadro 5).

Quadro 5 - Trabalhos Relacionados

Item	Base	Tipo de Documento	Ano	Autor(es)	Título	Local	Idioma

1	SCORPUS	Artigo	2019	Saimaiti, A.; Wang, L.; Yibulayin, T.	Learning subword embedding to improve Uyghur named-entity recognition	Information (Switzerland)	Inglês
2	SCORPUS	Artigo	2019	Thomas, A.; Sangeetha, S.	An innovative hybrid approach for extracting named entities from unstructured text data	Computational Intelligence	Inglês
3	SCORPUS	Artigo	2018	Lee, H.-G.; Park, G.; Kim, H.	Effective integration of morphological analysis and named entity recognition based on a recurrent neural network	Pattern Recognition Letters	Inglês
4	SCORPUS	Artigo	2018	Wang, L.; Li, S.; Yan, Q.; Zhou, G.	Domain-specific named entity recognition with document-level optimization	ACM Transactions on Asian and Low-Resource Language Information Processing	Inglês
5	SCORPUS	Artigo	2018	Liu, J.; Wang, L.; Zhou, M.; Wang, J.; Lee, S.	Fine-grained entity type classification with adaptive context	Soft Computing	Inglês
6	SCORPUS	Artigo	2018	Kadari, R.; Zhang, Y.; Zhang, W.; Liu, T.	CCG supertagging with bidirectional long short-Term memory networks	Natural Language Engineering	Inglês

7	LISA	Artigo	2019	William Paulo Ducca Fernandes; Schirmer Silva, Luiz José; Isabella Zalcberg Frajhof; Guilherme da Franca Couto Fernandes de Almeida; Carlos Nelson Konder; Rafael Barbosa Nasser; Gustavo Robichez de Carvalho; Simone Diniz Junqueira Barbosa; Hélio Côrtes Vieira Lopes	Appellate Court Modifications Extraction for Portuguese	Artificial Intelligence and Law	Inglês
8	LISA	Artigo	2019	Helwe, Chadi; Elbassuoni, Shady	Arabic named entity recognition via deep co-learning	The Artificial Intelligence Review	Inglês

Fonte: Dados da pesquisa, 2019.

Saimaiti, Wang e Yibulayin (2019) comparam os métodos BiLSTM, SRILM-Ngram e MaxMatch, para realizar extrações de entidades em textos escritos no idioma Uyghur. Para o experimento foi utilizado 39.027 trechos com anotações manuais processados à partir de um corpus multi idioma criado pelo Laboratório de tecnologia da informação da Universidade de Xinjiang. Foram utilizados 29.270 trechos para treinamento de cada modelo, 3.902 para validação cruzada e 5.855 para teste. As entidades presentes nos trechos são: Pessoa (PER), Local (LOC) e Organização (ORG). À métrica utilizada para avaliação dos modelos foi F1 *score*. Os resultados foram: 89.02% no modelo BiLSTM, 88.42% com o modelo SRILM-Ngram e 88.78 utilizando o modelo MaxMatch. Sendo o melhor resultado observado 89.02% gerado pelo modelo BiLSTM.

Thomas e Sangeetha (2019) utilizam uma abordagem de redes neurais híbridas BiLSTM-CNN-CRF, com utilização do corpus CoNLL 2003 com 203.621 trechos e um corpus judicial com 500.406 trechos chamado Judicial Corpora. As entidades mapeadas são Pessoa, Local e Organização. Foram separados 14.986 do corpus CoNLL 2003 e 12.110 do corpus Judicial para testes. Os autores adotam duas abordagem de pré processamento

onde são comparadas as técnicas para envio das informações para criar o modelo. São elas *Knowledge-Based Clustering Model* (KB + Clustering) e *Knowledge-Based Deep Learning* (KB + DL). A métricas utilizada para avaliação final dos métodos foi F1 *score*. Os F1 = 91.91% para o método KB + DL e F1 = 85.2% para o método KB + Clustering.

Lee, Park e Kim (2018) Realizam à comparação do emprego de uma rede neural recorrente híbrida do tipo *bidirectional gated recurrent unit model with a conditional random field layer* (BI-GRU-CRF) adicionando uma técnica denominada *MorpheNE* que à partir de uma equação que avalia a probabilidade de sequências de *tokens* para uma determinada tag usando como princípio à cadeia de markov. O estudo é comparado com outro modelo que possui as mesmas características em relação às redes neurais apresentadas inicialmente, construído por um dos autores em 2016 denominado LEE-2016. As métricas utilizadas para comparar os modelos são: Acurácia, Precisão, Recall e F1 *score.* O resultado apresentado foi Acurácia = 93.94%, Precisão = 78.98%, Recall = 69.21, F1 score = 73.77 para o modelo LEE-2016 e Acurácia = 94.65%, Precisão = 83.43%, Recall = 88.01%, F1 score = 85.66% para o modelo com emprego do *MorpheNE*.

Wang et al. (2018) utiliza uma abordagem baseada em rede neural recorrente do tipo LSTM, CRF e LSTM-CRF. Um corpus público com dados de contratos de casamento, com o total de 48.502 trechos para treino e com à base de teste dividida em 1.976 para contrato e 1.651 para casamento. As métricas de avaliação são Precisão, Recall e F1 *score*. O resultado foi, Precisão = 79.29%, Recall = 70.70%, F1 score = 74.75% com o emprego de CRF na base de contrato, Precisão = 62.35%, Recall = 48.10%, F1 score = 54.31% na base de teste de casamento; Precisão = 91.99%, Recall = 71.89%, F1 score = 80.71% com o emprego de LSTM-CRF na base de contrato, Precisão = 69.47%, Recall = 64.05%, F1 score = 66.65% na base de teste de casamento.

Liu et al. (2018) são utilizados dois corpus FIGER e OntoNotes com o total de 13.109 trechos e uma base de testes de 77 trechos anotados manualmente, e para criação do modelo redes neurais BiLSTM e LSTM. Os autores afirmam que utilizaram 840 bilhões de *tokens* no pré treinamento sendo utilizado o algoritmo de *word embeddings* conhecido como *Glove*. Não foi mencionado à tentativa de testes utilizando outro algoritmo como *Word2Vec* na comparação. Como métrica de avaliação foram utilizados Strict, Loose macro e Loose micro. À métrica Loose micro é visualizada como F1 *score* pelos autores. Os resultados obtidos foram F1 = 75.35% para o corpus FIGER e F1 = 65.35% com os dados de teste do corpus OntoNotes.

Kadari et al. (2018) sugerem um modelo baseado em BiLSTM com o uso corpus CCGBank. Houve experimentos em relação à notação do corpus tanto manual quanto

automática. Também foram gerados alguns modelos com base nos experimentos.. Sendo que o melhor resultado apresentou acurácia de 94.09%.

Fernandes et al. (2019) e autores empregam redes neurais BiLSTM-CRF e BIGRU–CRF. Para treino e validação um corpus composto pelas decisões Judiciais do Tribunal de Justiça Federal do Estado do Rio de Janeiro. O corpus foi dividido com 2.568 decisões separadas para treino do modelo e 454 separados para teste do modelo, as entidades foram anotadas manualmente não passando por nenhum processo de classificação automática no pré processamento. Os textos são escritos em língua portuguesa porém as entidades anotadas foram: *IncMorDm* sendo o aumento do valor de danos morais, *InitMonCo* sendo à data inicial da correção monetária, *InitIntArr* sendo a data inicial de interesse dos atrasados; *DecMorDm* é a diminuição do dano moral valor e *ValLegFee* é o valor da taxa legal devida pela, as entidades nesse sentido não são necessariamente *tokens* com anotações de nomes próprios como vimos nos outros estudos. Os resultados foram: Precisão = 95.71% Recall = 93.89% e F1 = 94.79%.

Helwe e Elbassuoni (2019) . O corpus utilizado possui 25000 trechos aleatórios de artigos do wikipédia sendo de domínio público, em idioma Árabe. Pelo que foi notado apenas entidades são anotadas, não possuindo anotações de postaging para outros tokens que não sejam entidades. Para testes foram utilizados outros corpus, 292 notícias extraídas do google news do feed of the Arabic (Egito), O corpus AQMAR, que possui 2456 sentenças, construído durante um estudo de Mohit et al. (2012) e postagens do twitter, um corpus construído por Darwish (2013) com 982 postagens. As entidades anotadas também foram PER, LOC e ORG. Foi empregada uma rede neural BiLSTM-CRF, porém outros modelos também foram propostos como LSTM unidirecional e algumas variações de pré processamento. Como resultado, cada corpus em que foi testado os modelos possui uma acurácia superior em uma entidade utilizando uma técnica. Nos textos do twitter os melhores resultados foram usando BiLSTM-CRF com dados semi anotados à acurácia média ficou em 59.2%. Já no textos de notícias e no corpus AQMAR os melhores resultados foram utilizando redes BiLSTM com dados semi anotados, à média da acurácia para o corpus de notícias foi 74.1% e 62.8% para o corpus AQMAR.

5 CONSIDERAÇÕES FINAIS

À pesquisa realizada recuperou diversos documentos contendo as palavras *Named Entity Recognition*, *corpus, recurrent neural networks*, todos resultados recuperados nas quatro bases selecionadas foram em inglês, para melhorar os resultados foram retirados

estudos que não tinham ligação com à temática proposta.

Como é possível notar uma grande quantidade de documentos possuem relação com estudos ligados à área da medicina, por sua vez foram excluídos por não terem ligação direta ao tema, pois o emprego da extração de entidades ligados à medicina normalmente é devido à necessidades relacionadas à processamento de linguagem natural em diagnósticos.

Os 8 trabalhos incluídos no final da RSL por sua vez foram os que possuem melhor aderência à temática proposta e trouxeram diversas contribuições para o estudo, diversas questões foram levantadas quanto ao pré processamento das *features* e combinações de redes neurais para atingir melhores resultados.

6 REFERÊNCIAS

FERNANDES, W.P.D.; SILVA, L.J.S.; FRAJHOF, I.Z.; *et al*. Appellate Court Modifications Extraction for Portuguese. **Artificial Intelligence and Law**, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068924728&doi=10.1007%2fs10506-019-09256-x&partnerID=40&md5=9dc415baa6308db9c5fe55bdfa437882>. Acesso em: 10 jul. 2019.

HELWE, C.; ELBASSUONI, S. Arabic named entity recognition via deep co-learning. **Artificial Intelligence Review**, v. 52, n. 1, p. 197–215, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061188066&doi=10.1007%2fs10462-019-09688-6&partnerID=40&md5=c084b1b907acee9804b440852347f2b4>. Acesso em: 20 jul. 2019.

KADARI, R.; ZHANG, Y.; ZHANG, W.; *et al*. CCG supertagging with bidirectional long short-Term memory networks. **Natural Language Engineering**, v. 24, n. 1, p. 77–90, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038415671&doi=10.1017%2fS1351324917000250&partnerID=40&md5=1ae9e424b8aabbe6ad2bec8599d724ff>. Acesso em: 09 jul. 2019.

LEE, H.-G.; PARK, G.; KIM, H. Effective integration of morphological analysis and named entity recognition based on a recurrent neural network. **Pattern Recognition Letters**, v. 112, p. 361–365, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052451041&doi=10.1016%2fj.patrec.2018.08.015&partnerID=40&md5=e5e32a5875d5659e27b1e30f5539168a>. Acesso em: 06 jul. 2019.

LIU, J.; WANG, L.; ZHOU, M.; *et al*. Fine-grained entity type classification with adaptive context. **Soft Computing**, v. 22, n. 13, p. 4307–4318, 2018. Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85040948961&doi=10.1007%2fs00500-017-2963-2&partnerID=40&md5=5ff8be377f05b00d3b4aafc2061a0db3>. Acesso em: 03 jul. 2019.

SAIMAITI, A.; WANG, L.; YIBULAYIN, T. Learning subword embedding to improve Uyghur named-entity recognition. **Information (Switzerland)**, v. 10, n. 4, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065882713&doi=10.3390%2finfo10040139&partnerID=40&md5=36b2b39a1d8dbcbdd855e8660a65d6f9>. Acesso em: 01 jul. 2019.

SAMPAIO, R. F.; MANCINI, M. C. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. **Revista brasileira de Fisioterapia**, São Carlos, v. 11, n. 1, p. 83-89, 2007. Disponível em: http://www.scielo.br/pdf/rbfis/v11n1/12.pdf. Acesso em: 15 jul. 2019.

THOMAS, A.; SANGEETHA, S. An innovative hybrid approach for extracting named entities from unstructured text data. **Computational Intelligence**, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065029748&doi=10.1111%2fcoin.12214&partnerID=40&md5=0a37553bb46a9f84f078fdef3095a910>. Acesso em: 06 jul. 2019.

WANG, L.; LI, S.; YAN, Q.; *et al*. Domain-specific named entity recognition with document-level optimization. **ACM Transactions on Asian and Low-Resource Language Information Processing**, v. 17, n. 4, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053426743&doi=10.1145%2f3213544&partnerID=40&md5=3f358b88db3d930fc042f74999b9582b>. Acesso em: 08 jul. 2019.

APÊNDICE I – DOCUMENTOS EXCLUÍDOS E CRITÉRIOS DE EXCLUSÃO

ITEM	BASE	ANO	TÍTULO	LOCAL DE PUBLICAÇÃO	CRITÉRIO APLICADO
1	SCORPUS	2019	Character convolutions for Arabic Named Entity Recognition with Long Short-Term Memory Networks	Computer Speech and Language	7. Apresentam redes neurais que não sejam LSTM

2	SCORPUS	2019	Interpretable deep learning to map diagnostic texts to ICD-10 codes	International Journal of Medical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
3	SCORPUS	2019	Precursor-induced conditional random fields: Connecting separate entities by induction for improved clinical named entity recognition	BMC Medical Informatics and Decision Making	7. Apresentam redes neurais que não sejam LSTM
4	SCORPUS	2019	Supervised methods to extract clinical events from cardiology reports in Italian	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
5	SCORPUS	2019	Contextual label sensitive gated network for biomedical event trigger extraction	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
6	SCORPUS	2019	Semantic vector learning for natural language understanding	Computer Speech and Language	5. Artigos com temática diferente. Artigos com temática diferente
7	SCORPUS	2019	Distant supervision for treatment relation extraction by leveraging MeSH subheadings	Artificial Intelligence in Medicine	5. Artigos com temática diferente. Artigos com temática diferente

8	SCORPUS	2019	Incorporating User Generated Content for Drug Drug Interaction Extraction Based on Full Attention Mechanism	IEEE Transactions on Nanobioscience	5. Artigos com temática diferente. Artigos com temática diferente
9	SCORPUS	2019	Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification	Artificial Intelligence in Medicine	7. Apresentam redes neurais que não sejam LSTM
10	SCORPUS	2019	Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter	Future Generation Computer Systems	5. Artigos com temática diferente. Artigos com temática diferente
11	SCORPUS	2019	Arabic named entity recognition using deep learning approach	International Journal of Electrical and Computer Engineering	12. Artigos incompletos
12	SCORPUS	2019	Arabic named entity recognition via deep co-learning	Artificial Intelligence Review	9. Artigos em duplicidade
13	SCORPUS	2019	A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente

14	SCORPUS	2019	Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning	Expert Systems with Applications	5. Artigos com temática diferente. Artigos com temática diferente
15	SCORPUS	2019	Deep neural network for hierarchical extreme multi-label text classification	Applied Soft Computing Journal	5. Artigos com temática diferente. Artigos com temática diferente
16	SCORPUS	2019	CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
17	SCORPUS	2019	Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN	PLoS ONE	5. Artigos com temática diferente. Artigos com temática diferente
18	SCORPUS	2019	Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition	Computers in Biology and Medicine	5. Artigos com temática diferente. Artigos com temática diferente
19	SCORPUS	2019	Identifying clinical terms in medical text using ontology-guided machine learning	Journal of Medical Internet Research	5. Artigos com temática diferente. Artigos com temática diferente

20	SCORPUS	2019	Parsing clinical text using the state-of-the-art deep learning based parsers: A systematic comparison	BMC Medical Informatics and Decision Making	5. Artigos com temática diferente. Artigos com temática diferente
21	SCORPUS	2019	Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation	Sensors (Switzerland)	5. Artigos com temática diferente. Artigos com temática diferente
22	SCORPUS	2019	A Knowledge-Based Recommendation System That Includes Sentiment Analysis and Deep Learning	IEEE Transactions on Industrial Informatics	5. Artigos com temática diferente. Artigos com temática diferente
23	SCORPUS	2019	Extraction of risk factors for cardiovascular diseases from Chinese electronic medical records	Computer Methods and Programs in Biomedicine	5. Artigos com temática diferente. Artigos com temática diferente
24	SCORPUS	2019	How to utilize syllable distribution patterns as the input of LSTM for Korean morphological analysis	Pattern Recognition Letters	5. Artigos com temática diferente. Artigos com temática diferente
25	SCORPUS	2019	Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network	Journal of Visual Communication and Image Representation	5. Artigos com temática diferente. Artigos com temática diferente

26	SCORPUS	2019	Neural architectures for open-type relation argument extraction	Natural Language Engineering	5. Artigos com temática diferente. Artigos com temática diferente
27	SCORPUS	2019	Unsupervised concept extraction from clinical text through semantic composition	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
28	SCORPUS	2019	Hit or miss? Evaluating the potential of a research niche: A case study in the field of virtual quality management	Sustainability (Switzerland)	5. Artigos com temática diferente. Artigos com temática diferente
29	SCORPUS	2019	DeepBioWSD: Effective deep neural word sense disambiguation of biomedical text data	Journal of the American Medical Informatics Association	5. Artigos com temática diferente. Artigos com temática diferente
30	SCORPUS	2019	Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein–protein interaction	Knowledge-Based Systems	5. Artigos com temática diferente. Artigos com temática diferente
31	SCORPUS	2019	From POS tagging to dependency parsing for biomedical event extraction	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
32	SCORPUS	2019	Information extraction from historical handwritten document images with a context-aware neural model	Pattern Recognition	5. Artigos com temática diferente. Artigos com temática diferente

33	SCORPUS	2019	Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
34	SCORPUS	2019	Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding	Drug Safety	5. Artigos com temática diferente. Artigos com temática diferente
35	SCORPUS	2019	Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0)	Drug Safety	5. Artigos com temática diferente. Artigos com temática diferente
36	SCORPUS	2019	A neural classification method for supporting the creation of BioVerbNet	Journal of Biomedical Semantics	5. Artigos com temática diferente. Artigos com temática diferente
37	SCORPUS	2019	Forward-looking element recognition based on the LSTM-CRF model with the integrity algorithm	Future Internet	5. Artigos com temática diferente. Artigos com temática diferente

38	SCORPUS	2019	LSTMVoter: Chemical named entity recognition using a conglomerate of sequence labeling tools	Journal of Cheminformatics	9. Artigos em duplicidade
39	SCORPUS	2019	BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
40	SCORPUS	2019	Learning Morpheme Representation for Mongolian Named Entity Recognition	Neural Processing Letters	7. Apresentam redes neurais que não sejam LSTM
41	SCORPUS	2019	A new type of eye movement model based on recurrent neural networks for simulating the gaze behavior of human reading	Complexity	5. Artigos com temática diferente. Artigos com temática diferente
42	SCORPUS	2019	A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages	Natural Language Engineering	5. Artigos com temática diferente. Artigos com temática diferente
43	SCORPUS	2019	A Bootstrapping Approach with CRF and Deep Learning Models for Improving the Biomedical Named Entity Recognition in Multi-Domains	IEEE Access	7. Apresentam redes neurais que não sejam LSTM

44	SCORPUS	2019	A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining	IEEE Access	5. Artigos com temática diferente. Artigos com temática diferente
45	SCORPUS	2019	The effect of morphology in named entity recognition with sequence tagging	Natural Language Engineering	5. Artigos com temática diferente. Artigos com temática diferente
46	SCORPUS	2019	Multifeature Named Entity Recognition in Information Security Based on Adversarial Learning	Security and Communication Networks	5. Artigos com temática diferente. Artigos com temática diferente
47	SCORPUS	2019	Exploiting the concept level feature for enhanced name entity recognition in Chinese EMRs	Journal of Supercomputing	5. Artigos com temática diferente. Artigos com temática diferente
48	SCORPUS	2019	Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches	IEEE Access	7. Apresentam redes neurais que não sejam LSTM
49	SCORPUS	2019	Appellate Court Modifications Extraction for Portuguese	Artificial Intelligence and Law	9. Artigos em duplicidade
50	SCORPUS	2019	Low-resource neural character-based noisy text normalization	Journal of Intelligent and Fuzzy Systems	5. Artigos com temática diferente. Artigos com temática diferente

51	SCORPUS	2019	Supervised word sense disambiguation using new features based on word embeddings	Journal of Intelligent and Fuzzy Systems	5. Artigos com temática diferente. Artigos com temática diferente
52	SCORPUS	2019	Deep Learning versus Conventional Machine Learning for Detection of Healthcare-Associated Infections in French Clinical Narratives	Methods of Information in Medicine	5. Artigos com temática diferente. Artigos com temática diferente
53	SCORPUS	2019	Learning Chinese word segmentation based on bidirectional GRU-CRF and CNN network model	International Journal of Technology and Human Interaction	7. Apresentam redes neurais que não sejam LSTM
54	SCORPUS	2019	Arabic Word Segmentation With Long Short-Term Memory Neural Networks and Word Embedding	IEEE Access	7. Apresentam redes neurais que não sejam LSTM
55	SCORPUS	2019	Deep Learning for Multi-Class Identification from Domestic Violence Online Posts	IEEE Access	5. Artigos com temática diferente. Artigos com temática diferente
56	SCORPUS	2019	Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss	Entropy	5. Artigos com temática diferente. Artigos com temática diferente
57	SCORPUS	2019	Cross-Lingual Word Embeddings	Synthesis Lectures on Human Language	5. Artigos com temática diferente. Artigos com temática

				Technologies	diferente
58	SCORPUS	2018	Joint entity recognition and relation extraction as a multi-head selection problem	Expert Systems with Applications	7. Apresentam redes neurais que não sejam LSTM
59	SCORPUS	2018	A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
60	SCORPUS	2018	EHR phenotyping via jointly embedding medical concepts and words into a unified vector space	BMC Medical Informatics and Decision Making	5. Artigos com temática diferente. Artigos com temática diferente
61	SCORPUS	2018	SBLC: A hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields	BMC Medical Informatics and Decision Making	5. Artigos com temática diferente. Artigos com temática diferente
62	SCORPUS	2018	MER: A shell script and annotation server for minimal named entity recognition and linking	Journal of Cheminformatics	5. Artigos com temática diferente. Artigos com temática diferente
63	SCORPUS	2018	Effective hate-speech detection in Twitter data using recurrent neural networks	Applied Intelligence	5. Artigos com temática diferente. Artigos com temática diferente

64	SCORPUS	2018	Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules	Journal of Cheminformatics	9. Artigos em duplicidade
65	SCORPUS	2018	Improving NER tagging performance in low-resource languages via multilingual learning	ACM Transactions on Asian and Low-Resource Language Information Processing	7. Apresentam redes neurais que não sejam LSTM
66	SCORPUS	2018	Using Vector Representation of Propositions and Actions for STRIPS Action Model Learning	Journal of Beijing Institute of Technology (English Edition)	7. Apresentam redes neurais que não sejam LSTM
67	SCORPUS	2018	Information extraction from scientific articles: a survey	Scientometrics	7. Apresentam redes neurais que não sejam LSTM
68	SCORPUS	2018	Bidirectional Grid Long Short-Term Memory (BiGridLSTM): A method to address context-sensitivity and vanishing gradient	Algorithms	5. Artigos com temática diferente. Artigos com temática diferente
69	SCORPUS	2018	Predicting of anaphylaxis in big data EMR by exploring machine learning approaches	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
70	SCORPUS	2018	Geo-text data and data-driven geospatial semantics	Geography Compass	5. Artigos com temática diferente.

					Artigos com temática diferente
71	SCORPUS	2018	D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information	Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
72	SCORPUS	2018	Conditional random fields for clinical named entity recognition: A comparative study using Korean clinical texts	Computers in Biology and Medicine	5. Artigos com temática diferente. Artigos com temática diferente
73	SCORPUS	2018	Hybrid LSTM/MaxEnt Networks for Arabic Syntactic Diacritics Restoration	IEEE Signal Processing Letters	5. Artigos com temática diferente. Artigos com temática diferente
74	SCORPUS	2018	Drug-drug interaction extraction from biomedical texts using long short-term memory network	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
75	SCORPUS	2018	Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task	Journal of the American Medical Informatics Association	5. Artigos com temática diferente. Artigos com temática diferente

76	SCORPUS	2018	Chinese event extraction based on attention and semantic features: A bidirectional circular neural network	Future Internet	5. Artigos com temática diferente. Artigos com temática diferente
77	SCORPUS	2018	Toward sustainable virtualized healthcare: Extracting medical entities from chinese online health consultations using deep neural networks	Sustainability (Switzerland)	5. Artigos com temática diferente. Artigos com temática diferente
78	SCORPUS	2018	Identifying bacterial biotope entities using sequence labeling: Performance and feature analysis	Journal of the Association for Information Science and Technology	5. Artigos com temática diferente. Artigos com temática diferente
79	SCORPUS	2018	EDM-JBW: A novel event detection model based on JS-ID′Forder and Bikmeans with word embedding for news streams	Journal of Computational Science	5. Artigos com temática diferente. Artigos com temática diferente
80	SCORPUS	2018	Evaluating Twitter as a complementary data source for pharmacovigilance	Expert Opinion on Drug Safety	5. Artigos com temática diferente. Artigos com temática diferente

81	SCORPUS	2018	Autoregressive Neural F0 Model for Statistical Parametric Speech Synthesis	IEEE/ACM Transactions on Audio Speech and Language Processing	5. Artigos com temática diferente. Artigos com temática diferente
82	SCORPUS	2018	A machine learning based approach to identify protected health information in Chinese clinical text	International Journal of Medical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
83	SCORPUS	2018	The Improved Model for word2vec Based on Part of Speech and Word Order [基于词性与词序的相关因子训练的word2vec改进模型]	Tien Tzu Hsueh Pao/Acta Electronica Sinica	5. Artigos com temática diferente. Artigos com temática diferente
84	SCORPUS	2018	Extracting psychiatric stressors for suicide from social media using deep learning	BMC Medical Informatics and Decision Making	9. Artigos em duplicidade
85	SCORPUS	2018	Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text	Applied Sciences (Switzerland)	5. Artigos com temática diferente. Artigos com temática diferente
86	SCORPUS	2018	Semi-Supervised Recurrent Neural Network for Adverse Drug Reaction mention extraction	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente

87	SCORPUS	2018	Entity highlight generation as statistical and neural machine translation	IEEE/ACM Transactions on Audio Speech and Language Processing	5. Artigos com temática diferente. Artigos com temática diferente
88	SCORPUS	2018	Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts	Artificial Intelligence and Law	9. Artigos em duplicidade
89	SCORPUS	2018	Using word embeddings in Twitter election classification	Information Retrieval Journal	5. Artigos com temática diferente. Artigos com temática diferente
90	SCORPUS	2018	Machine transliteration and transliterated text retrieval: a survey	Sadhana - Academy Proceedings in Engineering Sciences	5. Artigos com temática diferente. Artigos com temática diferente
91	SCORPUS	2018	What matters in a transferable neural network model for relation classification in the biomedical domain?	Artificial Intelligence in Medicine	5. Artigos com temática diferente. Artigos com temática diferente
92	SCORPUS	2018	Position-aware deep multi-task learning for drug–drug interaction extraction	Artificial Intelligence in Medicine	5. Artigos com temática diferente. Artigos com temática diferente

93	SCORPUS	2018	Spoken Language Understanding with a Novel Simultaneous Recognition Technique for Intelligent Personal Assistant Software	International Journal on Artificial Intelligence Tools	5. Artigos com temática diferente. Artigos com temática diferente
94	SCORPUS	2018	Deep learning for Arabic NLP: A survey	Journal of Computational Science	5. Artigos com temática diferente. Artigos com temática diferente
95	SCORPUS	2018	An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition	Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
96	SCORPUS	2018	How does neural machine translation (NMT) translate proper names and numbers? [Wie übersetzt NMT Eigennamen und Zahlen?]	Lebende Sprachen	5. Artigos com temática diferente. Artigos com temática diferente
97	SCORPUS	2018	Recognizing irregular entities in biomedical text via deep neural networks	Pattern Recognition Letters	5. Artigos com temática diferente. Artigos com temática diferente
98	SCORPUS	2018	Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning	Journal of Medical Internet Research	5. Artigos com temática diferente. Artigos com temática diferente

99	SCORPUS	2018	Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths	Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
100	SCORPUS	2018	CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines	Journal of the American Medical Informatics Association	5. Artigos com temática diferente. Artigos com temática diferente
101	SCORPUS	2018	A cyclic self-learning Chinese word segmentation for the geoscience domain	Geomatica	5. Artigos com temática diferente. Artigos com temática diferente
102	SCORPUS	2018	A Sequential Neural Encoder with Latent Structured Description for Modeling Sentences	IEEE/ACM Transactions on Audio Speech and Language Processing	5. Artigos com temática diferente. Artigos com temática diferente
103	SCORPUS	2018	Automatic microblog-oriented unknown word recognition with unsupervised method	Chinese Journal of Electronics	5. Artigos com temática diferente. Artigos com temática diferente
104	SCORPUS	2018	LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks	IEEE Transactions on Visualization and Computer Graphics	5. Artigos com temática diferente. Artigos com temática diferente

105	SCORPUS	2018	Combine Factual Medical Knowledge and Distributed Word Representation to Improve Clinical Named Entity Recognition	AMIA ... Annual Symposium proceedings. AMIA Symposium	5. Artigos com temática diferente. Artigos com temática diferente
106	SCORPUS	2018	Classifying semantic clause types with recurrent neural networks: Analysis of attention, context & genre characteristics	Revue Traitement Automatique des Langues	12. Artigos incompletos
107	SCORPUS	2018	A comparative study of word representation methods with conditional random fields and maximum entropy markov for bio-named entity recognition	Malaysian Journal of Computer Science	7. Apresentam redes neurais que não sejam LSTM
108	SCORPUS	2018	Transfer learning for biomedical named entity recognition with neural networks	Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
109	SCORPUS	2018	Modeling speech acts in asynchronous conversations: A neural-CRF approach	Computational Linguistics	5. Artigos com temática diferente. Artigos com temática diferente
110	SCORPUS	2018	Extracting chemical-protein relations using attention-based neural networks	Database	5. Artigos com temática diferente. Artigos com temática diferente

111	SCORPUS	2018	An end-to-end deep learning architecture for extracting protein-protein interactions affected by genetic mutations	Database	5. Artigos com temática diferente. Artigos com temática diferente
112	SCORPUS	2018	Improving the learning of chemical-protein interactions from literature using transfer learning and specialized word embeddings	Database	5. Artigos com temática diferente. Artigos com temática diferente
113	SCORPUS	2018	Extracting medical events from clinical records using conditional random fields and parameter tuning for hidden Markov models	Journal of Intelligent and Fuzzy Systems	5. Artigos com temática diferente. Artigos com temática diferente
114	SCORPUS	2018	An Attentive Neural Sequence Labeling Model for Adverse Drug Reactions Mentions Extraction	IEEE Access	5. Artigos com temática diferente. Artigos com temática diferente
115	SCORPUS	2018	Asynchronous Speech Recognition Affects Physician Editing of Notes	Applied Clinical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
116	SCORPUS	2018	Code mixed cross script factoid question classification-A deep learning approach	Journal of Intelligent and Fuzzy Systems	12. Artigos incompletos

117	SCORPUS	2018	Capsules based Chinese word segmentation for ancient Chinese medical books	IEEE Access	5. Artigos com temática diferente. Artigos com temática diferente
118	SCORPUS	2018	Automatic quality estimation for ASR system combination	Computer Speech and Language	5. Artigos com temática diferente. Artigos com temática diferente
119	SCORPUS	2018	Natural Language Processing for Social Media, Second Edition	Synthesis Lectures on Human Language Technologies	5. Artigos com temática diferente. Artigos com temática diferente
120	SCORPUS	2018	Opportunities and obstacles for deep learning in biology and medicine	Journal of the Royal Society Interface	5. Artigos com temática diferente. Artigos com temática diferente

121	SCORPUS	2017	Disease named entity recognition from biomedical literature using a novel convolutional neural network	BMC Medical Genomics	5. Artigos com temática diferente. Artigos com temática diferente
122	SCORPUS	2017	Dependency-based long short term memory network for drug-drug interaction extraction	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
123	SCORPUS	2017	A multiple distributed representation method based on neural network for biomedical event extraction	BMC Medical Informatics and Decision Making	5. Artigos com temática diferente. Artigos com temática diferente
124	SCORPUS	2017	Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
125	SCORPUS	2017	Multi-level cross-lingual attentive neural architecture for low resource name tagging	Tsinghua Science and Technology	5. Artigos com temática diferente. Artigos com temática diferente
126	SCORPUS	2017	Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach	BMC Medical Informatics and Decision Making	5. Artigos com temática diferente. Artigos com temática diferente

127	SCORPUS	2017	A Tidy data model for natural language processing using cleanNLP	R Journal	5. Artigos com temática diferente. Artigos com temática diferente
128	SCORPUS	2017	De-identification of clinical notes via recurrent neural network and conditional random field	Journal of Biomedical Informatics	5. Artigos com temática diferente. Artigos com temática diferente
129	SCORPUS	2017	Long short-term memory RNN for biomedical named entity recognition	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
130	SCORPUS	2017	A method for named entity normalization in biomedical articles: Application to diseases and plants	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
131	SCORPUS	2017	An attention-based effective neural model for drug-drug interactions extraction	BMC Bioinformatics	9. Artigos em duplicidade
132	SCORPUS	2017	Named entity recognition model based on neural networks using parts of speech probability and Gazetteer features	Advanced Science Letters	12. Artigos incompletos
133	SCORPUS	2017	A neural network multi-task learning approach to biomedical named entity recognition	BMC Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente

134	SCORPUS	2017	A transition-based joint model for disease named entity recognition and normalization	Bioinformatics	5. Artigos com temática diferente. Artigos com temática diferente
135	SCORPUS	2017	Entity recognition from clinical texts via recurrent neural network	BMC Medical Informatics and Decision Making	5. Artigos com temática diferente. Artigos com temática diferente
136	SCORPUS	2017	Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts	Journal of the American Medical Informatics Association	5. Artigos com temática diferente
137	SCORPUS	2017	Spoken language understanding for a nutrition dialogue system	IEEE/ACM Transactions on Audio Speech and Language Processing	5. Artigos com temática diferente
138	SCORPUS	2017	Character-level neural network for biomedical named entity recognition	Journal of Biomedical Informatics	5. Artigos com temática diferente
139	SCORPUS	2017	Greedy transition-based dependency parsing with stack LSTMS	Computational Linguistics	5. Artigos com temática diferente
140	SCORPUS	2017	Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN	Expert Systems with Applications	7. Apresentam redes neurais que não sejam LSTM

141	SCORPUS	2017	Improving the automatic segmentation of subtitles through conditional random field	Speech Communication	7. Apresentam redes neurais que não sejam LSTM
142	SCORPUS	2017	Data-centric and logic-based models for automated legal problem solving	Artificial Intelligence and Law	9. Artigos em duplicidade
143	SCORPUS	2017	Coupled POS tagging on heterogeneous annotations	IEEE/ACM Transactions on Audio Speech and Language Processing	7. Apresentam redes neurais que não sejam LSTM
144	SCORPUS	2017	Combination of Deep Recurrent Neural Networks and Conditional Random Fields for Extracting Adverse Drug Reactions from User Reviews	Journal of Healthcare Engineering	7. Apresentam redes neurais que não sejam LSTM
145	SCORPUS	2017	Clinical Named Entity Recognition Using Deep Learning Models	AMIA ... Annual Symposium proceedings. AMIA Symposium	5. Artigos com temática diferente
146	SCORPUS	2017	De-identification of patient notes with recurrent neural networks	Journal of the American Medical Informatics Association	5. Artigos com temática diferente
147	SCORPUS	2017	Neural Network Methods for Natural Language Processing	Synthesis Lectures on Human Language Technologies	5. Artigos com temática diferente

148	SCORPUS	2017	A two-stage joint model for domain-specific entity detection and linking leveraging an unlabeled corpus	Information (Switzerland)	5. Artigos com temática diferente
149	SCORPUS	2017	A feature vector representation approach for short text based on rnnlm and pooling computation	Academic Journal of Manufacturing Engineering	5. Artigos com temática diferente
150	web of science	2019	LSTMVoter: chemical named entity recognition using a conglomerate of sequence labeling tools	JOURNAL OF CHEMINFORMATICS	5. Artigos com temática diferente
151	web of science	2018	Extracting psychiatric stressors for suicide from social media using deep learning	BMC MEDICAL INFORMATICS AND DECISION MAKING	5. Artigos com temática diferente
152	web of science	2018	Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules	JOURNAL OF CHEMINFORMATICS	5. Artigos com temática diferente
153	web of science	2018	LSTM Recurrent Neural Networks for Cybersecurity Named Entity Recognition	THIRTEENTH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING ADVANCES (ICSEA 2018)	12. Artigos incompletos

154	web of science	2017	End-to-End Deep Framework for Disease Named Entity Recognition Using Social Media Data	2017 IEEE 30TH NEUMANN COLLOQUIUM (NC)	12. Artigos incompletos
155	LISA	2019	A data-driven neural network architecture for sentiment analysis	Data Technologies and Applications	5. Artigos com temática diferente
156	LISA	2018	Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts	Artificial Intelligence and Law	5. Artigos com temática diferente
157	LISA	2017	Data-centric and logic-based models for automated legal problem solving	Artificial Intelligence and Law	5. Artigos com temática diferente
158	LISA	2018	Named Entity Recognition Modeling for the Thai Language from a Disjointedly Labeled Corpus	2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)	12. Artigos incompletos
159	LISA	2018	Improved Deep Persian Named Entity Recognition	2018 9th International Symposium on Telecommunications (IST)	12. Artigos incompletos
160	IEEE Xplore	2018	Named Entity Recognition Modeling for the Thai Language from a Disjointedly Labeled Corpus	2018 5th International Conference on Advanced Informatics:	12. Artigos incompletos

				Concept Theory and Applications (ICAICTA)	
161	IEEE Xplore	2018	Improved Deep Persian Named Entity Recognition	2018 9th International Symposium on Telecommunications (IST)	12. Artigos incompletos

Fonte: Dados da pesquisa, 2019.

REFERÊNCIA DOS ARTIGOS EXCLUÍDOS

ABDI, Maan Tareq; MOHD, Masnizah. A COMPARATIVE STUDY OF WORD REPRESENTATION METHODS WITH CONDITIONAL RANDOM FIELDS AND MAXIMUM ENTROPY MARKOV FOR BIO-NAMED ENTITY RECOGNITION. **MALAYSIAN JOURNAL OF COMPUTER SCIENCE**, v. 31, n. 5, SI, p. 15–30, 2018.

ABID, F.; ALAM, M.; YASIR, M.; *et al*. Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. **Future Generation Computer Systems**, v. 95, p. 292–308, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059958367&doi=10.1016%2fj.future.2018.12.018&partnerID=40&md5=4d1db8b31e7e46728b12ed1610de0f1f>.

ALAMI, N.; MEKNASSI, M.; EN-NAHNAHI, N. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. **Expert Systems with Applications**, v. 123, p. 195–211, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060099427&doi=10.1016%2fj.eswa.2019.01.037&partnerID=40&md5=89cc306be7d5cba7716cf44132ebc9f3>.

AL-AYYOUB, M.; NUSEIR, A.; ALSMEARAT, K.; *et al*. Deep learning for Arabic NLP: A survey. **Journal of Computational Science**, v. 26, p. 522–531, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85035362747&doi=10.1016%2fj.jocs.2017.11.011&partnerID=40&md5=2089cda22a84b6c65284936e568d24b5>.

ALMAGRO, M.; MARTÍNEZ, R.; MONTALVO, S.; *et al*. A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation. **Journal of Biomedical Informatics**, v. 94, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065661289&doi=10.1016%2fj.jbi.2019.103207&partnerID=40&md5=b0b4d74badc111e5f6565b5e47d09797>.

ALMUHAREB, A.; ALSANIE, W.; AL-THUBAITY, A. Arabic Word Segmentation With Long Short-Term Memory Neural Networks and Word Embedding. **IEEE Access**, v. 7, p.

12879–12887, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061317402&doi=10.1109%2fACCESS.2019.2893460&partnerID=40&md5=498e88f36e562acccf316cd9898047a1>.

ÁLVAREZ, A.; MARTÍNEZ-HINAREJOS, C.-D.; ARZELUS, H.; *et al*. Improving the automatic segmentation of subtitles through conditional random field. **Speech Communication**, v. 88, p. 83–95, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85011860353&doi=10.1016%2fj.specom.2017.01.010&partnerID=40&md5=bbfeb8f9be96e6c6139acdba549ae228>.

ARBABI, A.; ADAMS, D.R.; FIDLER, S.; *et al*. Identifying clinical terms in medical text using ontology-guided machine learning. **Journal of Medical Internet Research**, v. 21, n. 5, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067385912&doi=10.2196%2f12596&partnerID=40&md5=48cbe92c3a6648ba15413abdfbf3377c>.

ARNOLD, T. A Tidy data model for natural language processing using cleanNLP. **R Journal**, v. 9, n. 2, p. 248–267, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041170187&partnerID=40&md5=152e77d9f36b5bc51a7f30c4feee8cf2>.

ATUTXA, A.; DE ILARRAZA, A.D.; GOJENOLA, K.; *et al*. Interpretable deep learning to map diagnostic texts to ICD-10 codes. **International Journal of Medical Informatics**, v. 129, p. 49–59, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066502163&doi=10.1016%2fj.ijmedinf.2019.05.015&partnerID=40&md5=a02c5de40ed7caffad8c5e1fdb1620d2>.

BAI, T.; CHANDA, A.K.; EGLESTON, B.L.; *et al*. EHR phenotyping via jointly embedding medical concepts and words into a unified vector space. **BMC Medical Informatics and Decision Making**, v. 18, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058311197&doi=10.1186%2fs12911-018-0672-0&partnerID=40&md5=fa04e9f8effdaf5c29fea9ee4df0d807>.

BALLESTEROS, M.; DYER, C.; GOLDBERG, Y.; *et al*. Greedy transition-based dependency parsing with stack LSTMS. **Computational Linguistics**, v. 43, n. 2, p. 311–347, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021840909&doi=10.1162%2fCOLI_a_00285&partnerID=40&md5=0c9b7a3602a5193603960abf1f59b63a>.

BANERJEE, I.; LING, Y.; CHEN, M.C.; *et al*. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. **Artificial Intelligence in Medicine**, v. 97, p. 79–88, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057000826&doi=10.1016%2fj.artmed.2018.11.004&partnerID=40&md5=c8bf0590b623597d3a06699a6f40abf4>.

BANERJEE, S.; NASKAR, S.; ROSSO, P.; *et al*. Code mixed cross script factoid question classification-A deep learning approach. **Journal of Intelligent and Fuzzy Systems**, v. 34, n. 5, p. 2959–2969, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054934632&doi=10.3233%2fJIFS-169481&partnerID=40&md5=05b69bbe3aa7bc2ca33181e9c46df741>.

BECKER, M.; STANIEK, M.; NASTASE, V.; *et al*. Classifying semantic clause types with recurrent neural networks: Analysis of attention, context & genre characteristics. **Revue Traitement Automatique des Langues**, v. 59, n. 2, p. 15–48, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063407953&partnerID=40&md5=e1b17370d3c828f5e1a99ee906cb96ea>.

BEKOULIS, G.; DELEU, J.; DEMEESTER, T.; *et al*. Joint entity recognition and relation extraction as a multi-head selection problem. **Expert Systems with Applications**, v. 114, p. 34–45, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050283540&doi=10.1016%2fj.eswa.2018.07.032&partnerID=40&md5=5103edd528e4a91e49c7b4ecbca55916>.

BOKAEI, M. H.; MAHMOUDI, M. Improved Deep Persian Named Entity Recognition. *In*: **2018 9th International Symposium on Telecommunications (IST)**. [s.l.: s.n.], 2018, p. 381–386.

BRANTING, L.K. Data-centric and logic-based models for automated legal problem solving. **Artificial Intelligence and Law**, v. 25, n. 1, p. 5–27, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85014107808&doi=10.1007%2fs10506-017-9193-x&partnerID=40&md5=c153c119834817d2cae8af4b8a6e71cc>.

CHEN, T.; XU, R.; HE, Y.; *et al*. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. **Expert Systems with Applications**, v. 72, p. 221–230, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85006717242&doi=10.1016%2fj.eswa.2016.10.065&partnerID=40&md5=4de20c20004c1ba0b0cbc1a35e258da6>.

CHING, T.; HIMMELSTEIN, D.S.; BEAULIEU-JONES, B.K.; *et al*. Opportunities and obstacles for deep learning in biology and medicine. **Journal of the Royal Society Interface**, v. 15, n. 141, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85045190865&doi=10.1098%2frsif.2017.0387&partnerID=40&md5=7819fb5a55c7648fe86ee5936fd9454e>.

CHIU, B.; MAJEWSKA, O.; PYYSALO, S.; *et al*. A neural classification method for supporting the creation of BioVerbNet. **Journal of Biomedical Semantics**, v. 10, n. 1, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060142224&doi=10.1186%2fs13326-018-0193-x&partnerID=40&md5=b92605b9a34cfa21d2c698679c88962d>.

CHO, H.; CHOI, W.; LEE, H. A method for named entity normalization in biomedical articles: Application to diseases and plants. **BMC Bioinformatics**, v. 18, n. 1, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85031011134&doi=10.1186%2fs12859-017-1857-8&partnerID=40&md5=83f8191d59111ce373a7be6c9b55e45c>.

CHOWDHURY, S.; DONG, X.; QIAN, L.; *et al*. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. **BMC Bioinformatics**, v. 19, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059225298&doi=10.1186%2fs12859-018-2467-9&partnerID=40&md5=dc488c28325d53cee98c9753acec0871>.

CHUNG, H.-Y. How does neural machine translation (NMT) translate proper names and numbers? [Wie übersetzt NMT Eigennamen und Zahlen?]. **Lebende Sprachen**, v. 63, n. 1, p. 142–167, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85045685893&doi=10.1515%2fles-2018-0007&partnerID=40&md5=c11e01e8fea0212f053bc49607f97a09>.

COCOS, A.; FIKS, A.G.; MASINO, A.J. Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. **Journal of the American Medical Informatics Association**, v. 24, n. 4, p. 813–821, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85026398638&doi=10.1093%2fjamia%2focw180&partnerID=40&md5=c78eadbc293f4aa460d14357b9951cc3>.

CORBETT, P.; BOYLE, J. Improving the learning of chemical-protein interactions from literature using transfer learning and specialized word embeddings. **Database**, v. 2018, n. 2018, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85056033575&doi=10.1093%2fdatabase%2fbay066&partnerID=40&md5=fe4b96d921bb5e74709b435c673e52e0>.

COUTO, F.M.; LAMURIAS, A. MER: A shell script and annotation server for minimal named entity recognition and linking. **Journal of Cheminformatics**, v. 10, n. 1, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058935026&doi=10.1186%2fs13321-018-0312-9&partnerID=40&md5=80ed0198e19b2001719c41db86e002a1>.

CRICHTON, G.; PYYSALO, S.; CHIU, B.; *et al*. A neural network multi-task learning approach to biomedical named entity recognition. **BMC Bioinformatics**, v. 18, n. 1, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85027963784&doi=10.1186%2fs12859-017-1776-8&partnerID=40&md5=dcd33792fc7dce4e09188ef3755fbd6b>.

DANG, T.H.; LE, H.-Q.; NGUYEN, T.M.; *et al*. D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. **Bioinformatics**, v. 34, n. 20, p. 3539–3546, 2018. Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054891869&doi=10.1093%2fbioinformatics%2fbty356&partnerID=40&md5=00bf18d6d1b1a5e87e5fd09446486684>.

DEGEN, H.; JING, Z.; KAIYU, H. Automatic microblog-oriented unknown word recognition with unsupervised method. **Chinese Journal of Electronics**, v. 27, n. 1, p. 1–8, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048802374&doi=10.1049%2fcje.2017.11.004&partnerID=40&md5=316f8537613184c74c48cfec338a4a85>.

DERNONCOURT, F.; LEE, J.Y.; UZUNER, O.; *et al*. De-identification of patient notes with recurrent neural networks. **Journal of the American Medical Informatics Association**, v. 24, n. 3, p. 596–606, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019722825&doi=10.1093%2fjamia%2focw156&partnerID=40&md5=bdde3995906788e86058248046e96250>.

DING, P.; ZHOU, X.; ZHANG, X.; *et al*. An Attentive Neural Sequence Labeling Model for Adverse Drug Reactions Mentions Extraction. **IEEE Access**, v. 6, p. 73305–73315, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057201548&doi=10.1109%2fACCESS.2018.2882443&partnerID=40&md5=c1be8352840c68e974b84474905ffb27>.

DONG, X.; CHOWDHURY, S.; QIAN, L.; *et al*. Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN. **PLoS ONE**, v. 14, n. 5, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065567606&doi=10.1371%2fjournal.pone.0216046&partnerID=40&md5=9e50a103fec64ed1bd884efc3e35879b>.

DU, Jingcheng; ZHANG, Yaoyun; LUO, Jianhong; *et al*. Extracting psychiatric stressors for suicide from social media using deep learning. **BMC MEDICAL INFORMATICS AND DECISION MAKING**, v. 18, n. 2, 2018.

DU, L.; XIA, C.; DENG, Z.; *et al*. A machine learning based approach to identify protected health information in Chinese clinical text. **International Journal of Medical Informatics**, v. 116, p. 24–32, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047417805&doi=10.1016%2fj.ijmedinf.2018.05.010&partnerID=40&md5=08adfdc375c0565c8b66dfb4a1c15aa5>.

EL BAZI, I.; LAACHFOUBI, N. Arabic named entity recognition using deep learning approach. **International Journal of Electrical and Computer Engineering**, v. 9, n. 3, p. 2025–2032, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067952612&doi=10.11591%2fijece.v9i3.pp2025-2032&partnerID=40&md5=52a2517e3abc51b3f2fb02a742d1ce07>.

ERION ÇANO; MORISIO, Maurizio. A data-driven neural network architecture for sentiment analysis. **Data Technologies and Applications**, v. 53, n. 1, p. 2–19, 2019. Disponível em: <https://search.proquest.com/docview/2202097780?accountid=26642>.

FAHANDEZI SADI, M.; ANSARI, E.; AFSHARCHI, M. Supervised word sense disambiguation using new features based on word embeddings. **Journal of Intelligent and Fuzzy Systems**, v. 37, n. 1, p. 1467–1476, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069434194&doi=10.3233%2fJIFS-182868&partnerID=40&md5=28175f5ed2222ec1ef7bb48285a13949>.

FARZINDAR, A.; INKPEN, D. Natural Language Processing for Social Media, Second Edition. **Synthesis Lectures on Human Language Technologies**, v. 10, n. 2, p. 1–197, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048736178&doi=10.2200%2fS00809ED2V01Y201710HLT038&partnerID=40&md5=b9ba7e9e840a0451db763950b3b2a013>.

FEI, H.; TAN, F. Bidirectional Grid Long Short-Term Memory (BiGridLSTM): A method to address context-sensitivity and vanishing gradient. **Algorithms**, v. 11, n. 11, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85056841797&doi=10.3390%2fa11110172&partnerID=40&md5=dfadddae5ae4213fb88a75c548fcf545>.

FENG, X.; HUANG, L.; QIN, B.; *et al*. Multi-level cross-lingual attentive neural architecture for low resource name tagging. **Tsinghua Science and Technology**, v. 22, n. 6, p. 633–645, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039425079&doi=10.23919%2fTST.2017.8195346&partnerID=40&md5=83abdcc7e3da9499ea8ebb14fca03860>.

FERNANDES, W.P.D.; SILVA, L.J.S.; FRAJHOF, I.Z.; *et al*. Appellate Court Modifications Extraction for Portuguese. **Artificial Intelligence and Law**, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068924728&doi=10.1007%2fs10506-019-09256-x&partnerID=40&md5=9dc415baa6308db9c5fe55bdfa437882>.

FOCIL-ARIAS, Carolina; SIDOROV, Grigori; GELBUKH, Alexander; *et al*. Extracting medical events from clinical records using conditional random fields and parameter tuning for hidden Markov models. **JOURNAL OF INTELLIGENT & FUZZY SYSTEMS**, v. 34, n. 5, p. 2935–2947, 2018.

GAO, W.; CAI, D. Using Vector Representation of Propositions and Actions for STRIPS Action Model Learning. **Journal of Beijing Institute of Technology (English Edition)**, v. 27, n. 4, p. 485–492, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063695628&doi=10.15918%2fj.jbit1004-0579.18072&partnerID=40&md5=e38fb2c4e15ffe2624090fe96c7ecba3>.

GARGIULO, F.; SILVESTRI, S.; CIAMPI, M.; *et al*. Deep neural network for hierarchical extreme multi-label text classification. **Applied Soft Computing Journal**, v. 79, p. 125–138, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063663613&doi=10.1016%2fj.asoc.2019.03.041&partnerID=40&md5=5191390aaa22c4a4782b3695f5a789f4>.

GASMI, Houssem; BOURAS, Abdelaziz; LAVAL, Jannik. LSTM Recurrent Neural Networks for Cybersecurity Named Entity Recognition. *In*: LAVAZZA, L AND OBERHAUSER, R AND KOCI, R (Org.). **THIRTEENTH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING ADVANCES (ICSEA 2018)**. PO BOX 7827, WILMINGTON, DE 19803 USA: IARIA XPS PRESS, 2018, p. 1–6.

GIORGI, J.M.; BADER, G.D. Transfer learning for biomedical named entity recognition with neural networks. **Bioinformatics**, v. 34, n. 23, p. 4087–4094, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057200026&doi=10.1093%2fbioinformatics%2fbty449&partnerID=40&md5=528e97a710b7d27a365c6a56d8b9d3ff>.

GOLDBERG, Y. Neural Network Methods for Natural Language Processing. **Synthesis Lectures on Human Language Technologies**, v. 10, n. 1, p. 1–311, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85033223087&doi=10.2200%2fS00762ED1V01Y201703HLT037&partnerID=40&md5=73db6b5659c49b4a23d3c1d45e5ced9b>.

GRIDACH, M. Character-level neural network for biomedical named entity recognition. **Journal of Biomedical Informatics**, v. 70, p. 85–91, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019375942&doi=10.1016%2fj.jbi.2017.05.002&partnerID=40&md5=5889bf9fe87f10b3ce004107e467f589>.

GÜNGÖR, O.; GÜNGÖR, T.; ÜSKÜDARLI, S. The effect of morphology in named entity recognition with sequence tagging. **Natural Language Engineering**, v. 25, n. 1, p. 147–169, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052653771&doi=10.1017%2fS1351324918000281&partnerID=40&md5=9a3d08d03164ff0947791914ea07e30a>.

GUPTA, S.; PAWAR, S.; RAMRAKHIYANI, N.; *et al*. Semi-Supervised Recurrent Neural Network for Adverse Drug Reaction mention extraction. **BMC Bioinformatics**, v. 19, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048287010&doi=10.1186%2fs12859-018-2192-4&partnerID=40&md5=e4c110d504ee077af6953d581e6e69aa>.

HELWE, C.; ELBASSUONI, S. Arabic named entity recognition via deep co-learning. **Artificial Intelligence Review**, v. 52, n. 1, p. 197–215, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061188066&doi=10.1007%2fs10462-019-09688-6&partnerID=40&md5=c084b1b907acee9804b440852347f2b4>.

HEMATI, Wahed; MEHLER, Alexander. LSTMVoter: chemical named entity recognition using a conglomerate of sequence labeling tools. **JOURNAL OF CHEMINFORMATICS**, v. 11, 2019.

HERNANDEZ-SUAREZ, A.; SANCHEZ-PEREZ, G.; TOSCANO-MEDINA, K.; *et al*. Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. **Sensors (Switzerland)**, v. 19, n. 7, 2019. Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064806156&doi=10.3390%2fs19071746&partnerID=40&md5=b82403e92abbc215651c9ae2499424b0>.

HIFNY, Y. Hybrid LSTM/MaxEnt Networks for Arabic Syntactic Diacritics Restoration. **IEEE Signal Processing Letters**, v. 25, n. 10, p. 1515–1519, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052585404&doi=10.1109%2fLSP.2018.2865098&partnerID=40&md5=6687363c7dfc7558625b4698bea150e8>.

HU, Y. Geo-text data and data-driven geospatial semantics. **Geography Compass**, v. 12, n. 11, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053390084&doi=10.1111%2fgec3.12404&partnerID=40&md5=37bcdc6603df15bae92be71344152085>.

HUANG, J.; SUN, Y.; ZHANG, W.; *et al*. Entity highlight generation as statistical and neural machine translation. **IEEE/ACM Transactions on Audio Speech and Language Processing**, v. 26, n. 10, p. 1860–1872, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048176148&doi=10.1109%2fTASLP.2018.2845111&partnerID=40&md5=e5bdfab5cbeaeba00df27fd85b79c661>.

JAGANNATHA, A.; LIU, F.; LIU, W.; *et al*. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). **Drug Safety**, v. 42, n. 1, p. 99–111, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060139319&doi=10.1007%2fs40264-018-0762-z&partnerID=40&md5=1fb9e1a5ae28d5a0806edbc23d02e07d>.

JALALVAND, S.; NEGRI, M.; FALAVIGNA, D.; *et al*. Automatic quality estimation for ASR system combination. **Computer Speech and Language**, v. 47, p. 214–239, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85027552871&doi=10.1016%2fj.csl.2017.06.003&partnerID=40&md5=b1ae7281dd1799a1a7bd58d9bbfb3adb>.

JAUREGI UNANUE, I.; ZARE BORZESHI, E.; PICCARDI, M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. **Journal of Biomedical Informatics**, v. 76, p. 102–109, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85034622038&doi=10.1016%2fj.jbi.2017.11.007&partnerID=40&md5=a9a334df8705c8291b707a64807b2896>.

JIANG, Z.; GAO, S.; LI, M. A feature vector representation approach for short text based on rnnlm and pooling computation. **Academic Journal of Manufacturing Engineering**, v. 15, n. 2, p. 6–14, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85026535958&partnerID=40&md5=1845bd5deeb856c17ed17e81d0e19a0a>.

JOTY, S.; MOHIUDDIN, T. Modeling speech acts in asynchronous conversations: A neural-CRF approach. **Computational Linguistics**, v. 44, n. 4, p. 859–894, 2018.

Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059285200&doi=10.1162%2fcoli_a_00339&partnerID=40&md5=2e7ff44290ed9fba7fcb9dad7e23caec>.

JUNG, S. Semantic vector learning for natural language understanding. **Computer Speech and Language**, v. 56, p. 130–145, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061967793&doi=10.1016%2fj.csl.2018.12.008&partnerID=40&md5=e42a317493b0dbdb2f4aea3fd20822f8>.

KHALIFA, M.; SHAALAN, K. Character convolutions for Arabic Named Entity Recognition with Long Short-Term Memory Networks. **Computer Speech and Language**, v. 58, p. 335–346, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066838600&doi=10.1016%2fj.csl.2019.05.003&partnerID=40&md5=d9a1da5f264859025ef2dae4426f9a56>.

KHAN, W.; DAUD, A.; KHAN, K.; *et al*. Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches. **IEEE Access**, v. 7, p. 38918–38936, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065160185&doi=10.1109%2fACCESS.2019.2897327&partnerID=40&md5=cdf42e8f4b0c5405ef6f51baf445a2cd>.

KIM, D.; LEE, J.; SO, C.H.; *et al*. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. **IEEE Access**, v. 7, p. 73729–73740, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068313056&doi=10.1109%2fACCESS.2019.2920708&partnerID=40&md5=16d26966e22ca29310b37f76ff3e1ff7>.

KIM, H.; YANG, S.; KO, Y. How to utilize syllable distribution patterns as the input of LSTM for Korean morphological analysis. **Pattern Recognition Letters**, v. 120, p. 39–45, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059590063&doi=10.1016%2fj.patrec.2018.12.019&partnerID=40&md5=daf3663c460f1a6942abec2385f6467d>.

KIM, Juae; KO, Youngjoong; SEO, Jungyun. A Bootstrapping Approach With CRF and Deep Learning Models for Improving the Biomedical Named Entity Recognition in Multi-Domains. **IEEE ACCESS**, v. 7, p. 70308–70318, 2019.

KORPUSIK, M.; GLASS, J. Spoken language understanding for a nutrition dialogue system. **IEEE/ACM Transactions on Audio Speech and Language Processing**, v. 25, n. 7, p. 1450–1461, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020757683&doi=10.1109%2fTASLP.2017.2694699&partnerID=40&md5=0fd378e0775f089f29248e0ddbcaf50e>.

KORVIGO, Ilia; HOLMATOV, Maxim; ZAIKOVSKII, Anatolii; *et al*. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. **JOURNAL OF CHEMINFORMATICS**, v. 10, 2018.

LAMURIAS, A.; SOUSA, D.; CLARKE, L.A.; *et al*. BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies. **BMC Bioinformatics**, v. 20, n. 1, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059594576&doi=10.1186%2fs12859-018-2584-5&partnerID=40&md5=23a545dee0c29702af8ddb6e18a00a52>.

LARDON, J.; BELLET, F.; ABOUKHAMIS, R.; *et al*. Evaluating Twitter as a complementary data source for pharmacovigilance. **Expert Opinion on Drug Safety**, v. 17, n. 8, p. 763–774, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050922952&doi=10.1080%2f14740338.2018.1499724&partnerID=40&md5=f557813580705bd0038f9aefd5f16425>.

LEE, C.; KO, Y. Spoken Language Understanding with a Novel Simultaneous Recognition Technique for Intelligent Personal Assistant Software. **International Journal on Artificial Intelligence Tools**, v. 27, n. 3, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047307524&doi=10.1142%2fS0218213018500094&partnerID=40&md5=89a4966dc27f3d71b72dd51a1fe9f584>.

LEE, W.; CHOI, J. Precursor-induced conditional random fields: Connecting separate entities by induction for improved clinical named entity recognition. **BMC Medical Informatics and Decision Making**, v. 19, n. 1, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069160486&doi=10.1186%2fs12911-019-0865-1&partnerID=40&md5=8b6600bc8dd0d24ddb8fd999f9d4127e>.

LEE, W.; KIM, K.; LEE, E.Y.; *et al*. Conditional random fields for clinical named entity recognition: A comparative study using Korean clinical texts. **Computers in Biology and Medicine**, v. 101, p. 7–14, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050947903&doi=10.1016%2fj.compbiomed.2018.07.019&partnerID=40&md5=b9c5e3d9caf8b24867f4148e91b87b97>.

LI, F.; ZHANG, M.; TIAN, B.; *et al*. Recognizing irregular entities in biomedical text via deep neural networks. **Pattern Recognition Letters**, v. 105, p. 105–113, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020914981&doi=10.1016%2fj.patrec.2017.06.009&partnerID=40&md5=4a4b52041c1d89013236a99406371577>.

LI, L.; HUANG, M.; LIU, Y.; *et al*. Contextual label sensitive gated network for biomedical event trigger extraction. **Journal of Biomedical Informatics**, v. 95, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067064927&doi=10.1016%2fj.jbi.2019.103221&partnerID=40&md5=fe775e7d4ac7bab3fdcfee2bbaa539f9>.

LI, S.; LI, M.; XU, Y.; *et al*. Capsules based Chinese word segmentation for ancient Chinese medical books. **IEEE Access**, v. 6, p. 70874–70883, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85056607578&doi=10.1109%2fACCESS.2018.2881280&partnerID=40&md5=509eb67acdb31175718cf4d07eb7c530>.

LI, X.; WANG, H.; HE, H.; *et al*. Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks. **BMC Bioinformatics**, v. 20, n. 1, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060927132&doi=10.1186%2fs12859-019-2617-8&partnerID=40&md5=6ae3a65476740f29925957b2ed3c9dfa>.

LI, Z.; CHAO, J.; ZHANG, M.; *et al*. Coupled POS tagging on heterogeneous annotations. **IEEE/ACM Transactions on Audio Speech and Language Processing**, v. 25, n. 3, p. 557–571, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85012915982&doi=10.1109%2fTASLP.2016.2644262&partnerID=40&md5=e56d6846da52fad2730473df13c478a7>.

LIU, S.; SHEN, F.; KOMANDUR ELAYAVILLI, R.; *et al*. Extracting chemical-protein relations using attention-based neural networks. **Database**, v. 2018, n. 2018, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054456100&doi=10.1093%2fdatabase%2fbay102&partnerID=40&md5=00da479270940c64ab02f783945a283d>.

LIU, X.; ZHOU, Y.; WANG, Z. Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network. **Journal of Visual Communication and Image Representation**, v. 60, p. 1–15, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061562994&doi=10.1016%2fj.jvcir.2019.02.001&partnerID=40&md5=bbd2d134953a0030f2a3977f06266b9e>.

LIU, Z.; TANG, B.; WANG, X.; *et al*. De-identification of clinical notes via recurrent neural network and conditional random field. **Journal of Biomedical Informatics**, v. 75, p. S34–S42, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020419299&doi=10.1016%2fj.jbi.2017.05.023&partnerID=40&md5=31d3271dc801c8596eaf8b39d9f7a55d>.

LIU, Z.; YANG, M.; WANG, X.; *et al*. Entity recognition from clinical texts via recurrent neural network. **BMC Medical Informatics and Decision Making**, v. 17, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021702868&doi=10.1186%2fs12911-017-0468-7&partnerID=40&md5=7ca7bdc4f23bcdc064309303a40e1a08>.

LOU, Y.; ZHANG, Y.; QIAN, T.; *et al*. A transition-based joint model for disease named entity recognition and normalization. **Bioinformatics**, v. 33, n. 15, p. 2363–2371, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85026396744&doi=10.1093%2fbioinformatics%2fbtx172&partnerID=40&md5=e1721e60bfb77d683204ff5753defd12>.

LUO, Ling; YANG, Zhihao; YANG, Pei; *et al*. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. **BIOINFORMATICS**, v. 34, n. 8, p. 1381–1388, 2018.

LYBARGER, K.J.; OSTENDORF, M.; RISKIN, E.; *et al*. Asynchronous Speech Recognition Affects Physician Editing of Notes. **Applied Clinical Informatics**, v. 9, n. 4, p. 782–790,

2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055079097&doi=10.1055%2fs-0038-1673417&partnerID=40&md5=e658223a30658acdab5fc79da42aa272>.

LYU, Chen; CHEN, Bo; REN, Yafeng; *et al*. Long short-term memory RNN for biomedical named entity recognition. **BMC BIOINFORMATICS**, v. 18, 2017.

MAGER, M.; ROSALES, M.J.; ÇETINOĞLU, Ö.; *et al*. Low-resource neural character-based noisy text normalization. **Journal of Intelligent and Fuzzy Systems**, v. 36, n. 5, p. 4921–4929, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066428518&doi=10.3233%2fJIFS-179039&partnerID=40&md5=c05eca3cd3161e9046fe86f4e07dff6d>.

MAO, J.; CUI, H. Identifying bacterial biotope entities using sequence labeling: Performance and feature analysis. **Journal of the Association for Information Science and Technology**, v. 69, n. 9, p. 1134–1147, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052494291&doi=10.1002%2fasi.24032&partnerID=40&md5=fbc14154c12f8ec4bf60ea3893602678>.

MENGER, V.; SCHEEPERS, F.; SPRUIT, M. Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. **Applied Sciences (Switzerland)**, v. 8, n. 6, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048524949&doi=10.3390%2fapp8060981&partnerID=40&md5=198f1ab141c4535b5f1236e797964bf3>.

MIFTAHUTDINOV, Zulfat; TUTUBALINA, Elena. End-to-End Deep Framework for Disease Named Entity Recognition Using Social Media Data. *In*: **2017 IEEE 30TH NEUMANN COLLOQUIUM (NC)**. 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE, 2017, p. 47–52.

MUNKHDALAI, T.; LIU, F.; YU, H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning. **Journal of Medical Internet Research**, v. 20, n. 4, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047745010&doi=10.2196%2fpublichealth.9361&partnerID=40&md5=87a63b184b0517bb394e9b4e00254fe7>.

MURTHY, R.; KHAPRA, M.M.; BHATTACHARYYA, P. Improving NER tagging performance in low-resource languages via multilingual learning. **ACM Transactions on Asian and Low-Resource Language Information Processing**, v. 18, n. 2, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058809485&doi=10.1145%2f3238797&partnerID=40&md5=9231c4d6b66a322b83991fdab645ce35>.

NASAR, Z.; JAFFRY, S.W.; MALIK, M.K. Information extraction from scientific articles: a survey. **Scientometrics**, v. 117, n. 3, p. 1931–1990, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054527773&doi=10.1007%2fs11192-018-2921-5&partnerID=40&md5=1baa1cacfc309261cf08cdd6c0c7e82a>.

NGUYEN, D.Q.; VERSPOOR, K. From POS tagging to dependency parsing for biomedical event extraction. **BMC Bioinformatics**, v. 20, n. 1, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061502492&doi=10.1186%2fs12859-019-2604-0&partnerID=40&md5=5b40776f70615075a99a28456e5dbb11>.

NGUYEN, T.-S.; NGUYEN, L.-M.; TOJO, S.; *et al*. Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. **Artificial Intelligence and Law**, v. 26, n. 2, p. 169–199, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044355733&doi=10.1007%2fs10506-018-9225-1&partnerID=40&md5=345553aac7dc5b51e455dcbfa588164c>.

PAN, B.; YU, C.-C.; ZHANG, Q.-C.; *et al*. The Improved Model for word2vec Based on Part of Speech and Word Order [基于词性与词序的相关因子训练的word2vec改进模型]. **Tien Tzu Hsueh Pao/Acta Electronica Sinica**, v. 46, n. 8, p. 1976–1982, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85056518381&doi=10.3969%2fj.issn.0372-2112.2018.08.024&partnerID=40&md5=1a1c25024d40561bfcd38ddfd11f17c4>.

PARK, G.; LEE, H.-G.; KIM, H. Named entity recognition model based on neural networks using parts of speech probability and Gazetteer features. **Advanced Science Letters**, v. 23, n. 10, p. 9530–9533, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039456516&doi=10.1166%2fasl.2017.9740&partnerID=40&md5=7ba775b350d01e70a51bc7bf2f94861c>.

PESARANGHADER, A.; MATWIN, S.; SOKOLOVA, M.; *et al*. DeepBioWSD: Effective deep neural word sense disambiguation of biomedical text data. **Journal of the American Medical Informatics Association**, v. 26, n. 5, p. 438–446, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063713363&doi=10.1093%2fjamia%2focy189&partnerID=40&md5=e9680a86e8d7cecbc574ab25e2afc8b8>.

PITSILIS, G.K.; RAMAMPIARO, H.; LANGSETH, H. Effective hate-speech detection in Twitter data using recurrent neural networks. **Applied Intelligence**, v. 48, n. 12, p. 4730–4742, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050686216&doi=10.1007%2fs10489-018-1242-y&partnerID=40&md5=ffe49268af1d6b3969852a2bfaa42e8c>.

PRABHAKAR, D.K.; PAL, S. Machine transliteration and transliterated text retrieval: a survey. **Sadhana - Academy Proceedings in Engineering Sciences**, v. 43, n. 6, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048277898&doi=10.1007%2fs12046-018-0828-8&partnerID=40&md5=2ebe218b24923af0458677a11644c313>.

QIU, Q.; XIE, Z.; WU, L. A cyclic self-learning Chinese word segmentation for the geoscience domain. **Geomatica**, v. 72, n. 1, p. 16–26, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052376537&doi=10.1139%2fgeomat-2018-0007&partnerID=40&md5=a4dc84ab7ed9259e97f1ecb26906155b>.

RABHI, S.; JAKUBOWICZ, J.; METZGER, M.-H. Deep Learning versus Conventional Machine Learning for Detection of Healthcare-Associated Infections in French Clinical Narratives. **Methods of Information in Medicine**, v. 58, n. 1, p. 31–41, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068614175&doi=10.1055%2fs-0039-1677692&partnerID=40&md5=c932eeb6d4b69986906ce6b3d792d586>.

ROSA, R.L.; SCHWARTZ, G.M.; RUGGIERO, W.V.; *et al*. A Knowledge-Based Recommendation System That Includes Sentiment Analysis and Deep Learning. **IEEE Transactions on Industrial Informatics**, v. 15, n. 4, p. 2124–2135, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052695921&doi=10.1109%2fTII.2018.2867174&partnerID=40&md5=eb5ef38546c18fc464200583c7f12ff6>.

ROTH, B.; CONFORTI, C.; POERNER, N.; *et al*. Neural architectures for open-type relation argument extraction. **Natural Language Engineering**, v. 25, n. 2, p. 219–238, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058169526&doi=10.1017%2fS1351324918000451&partnerID=40&md5=2a0885013f5c8fc0f562bb31b7f2e767>.

RUAN, Y.-P.; CHEN, Q.; LING, Z.-H. A Sequential Neural Encoder with Latent Structured Description for Modeling Sentences. **IEEE/ACM Transactions on Audio Speech and Language Processing**, v. 26, n. 2, p. 231–242, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85034239403&doi=10.1109%2fTASLP.2017.2773198&partnerID=40&md5=aabbff0fd0e0dfe1cd6ea09662a61a97>.

SAHU, S.K.; ANAND, A. Drug-drug interaction extraction from biomedical texts using long short-term memory network. **Journal of Biomedical Informatics**, v. 86, p. 15–24, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052543403&doi=10.1016%2fj.jbi.2018.08.005&partnerID=40&md5=b8e58338705c20be14057452a2899cca>.

SAHU, S.K.; ANAND, A. What matters in a transferable neural network model for relation classification in the biomedical domain? **Artificial Intelligence in Medicine**, v. 87, p. 60–66, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85045337759&doi=10.1016%2fj.artmed.2018.03.006&partnerID=40&md5=67b0f196143aa78bb9bd2374a2d00b26>.

SARKER, A.; BELOUSOV, M.; FRIEDRICHS, J.; *et al*. Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task. **Journal of the American Medical Informatics Association**, v. 25, n. 10, p. 1274–1283, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054889806&doi=10.1093%2fjamia%2focy114&partnerID=40&md5=9a2ccdfd6edc1dfd60425cd7358178dd>.

SEGURA-BEDMAR, I.; COLÓN-RUÍZ, C.; TEJEDOR-ALONSO, M.Á.; *et al*. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. **Journal of Biomedical Informatics**, v. 87, p. 50–59, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054180297&doi=10.1016%2fj.jbi.2018.09.012&partnerID=40&md5=e47176347da699e6e82e7b83327882ce>.

SØGAARD, A.; VULIĆ, I.; RUDER, S.; *et al*. Cross-Lingual Word Embeddings. **Synthesis Lectures on Human Language Technologies**, v. 12, n. 2, p. 1–132, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066947466&doi=10.2200%2fS00920ED2V01Y201904HLT042&partnerID=40&md5=76f24fd206b3f56a09ec964f94db9f61>.

SOYSAL, E.; WANG, J.; JIANG, M.; *et al*. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. **Journal of the American Medical Informatics Association**, v. 25, n. 3, p. 331–336, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85043312385&doi=10.1093%2fjamia%2focx132&partnerID=40&md5=03c8eb8a46da67e2874f8d0007e3e858>.

STROBELT, H.; GEHRMANN, S.; PFISTER, H.; *et al*. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. **IEEE Transactions on Visualization and Computer Graphics**, v. 24, n. 1, p. 667–676, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028690255&doi=10.1109%2fTVCG.2017.2744158&partnerID=40&md5=531b7e64584708cc6b371ebd66d75c46>.

SU, Jia; HU, Jinpeng; JIANG, Jingchi; *et al*. Extraction of risk factors for cardiovascular diseases from Chinese electronic medical records. **COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE**, v. 172, p. 1–10, 2019.

SUBRAMANI, S.; MICHALSKA, S.; WANG, H.; *et al*. Deep Learning for Multi-Class Identification from Domestic Violence Online Posts. **IEEE Access**, v. 7, p. 46210–46224, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064761572&doi=10.1109%2fACCESS.2019.2908827&partnerID=40&md5=f0376c3b67222c3dc725de59e1c23eea>.

SUN, X.; DONG, K.; MA, L.; *et al*. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. **Entropy**, v. 21, n. 1, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060395260&doi=10.3390%2fe21010037&partnerID=40&md5=6ef139ebcd549de88eabd38d27c5e3a2>.

SURIYACHAY, K.; SORNLERTLAMVANICH, V. Named Entity Recognition Modeling for the Thai Language from a Disjointedly Labeled Corpus. *In*: **2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)**. [s.l.: s.n.], 2018, p. 30–35.

TOLEDO, J.I.; CARBONELL, M.; FORNÉS, A.; *et al*. Information extraction from historical handwritten document images with a context-aware neural model. **Pattern Recognition**, v.

86, p. 27–36, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052926884&doi=10.1016%2fj.patcog.2018.08.020&partnerID=40&md5=400e9a5c0fa75e8c2703ca485136b9aa>.

TRAN, T.; KAVULURU, R. An end-to-end deep learning architecture for extracting protein-protein interactions affected by genetic mutations. **Database**, v. 2018, n. 2018, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057262452&doi=10.1093%2fdatabase%2fbay092&partnerID=40&md5=d8524fabcab03b0e5382772d5ee88af0>.

TRAN, T.; KAVULURU, R. Distant supervision for treatment relation extraction by leveraging MeSH subheadings. **Artificial Intelligence in Medicine**, v. 98, p. 18–26, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068258914&doi=10.1016%2fj.artmed.2019.06.002&partnerID=40&md5=d040e7b36d1d83dafd86416c993d0e7f>.

TULKENS, S.; ŠUSTER, S.; DAELEMANS, W. Unsupervised concept extraction from clinical text through semantic composition. **Journal of Biomedical Informatics**, v. 91, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061734465&doi=10.1016%2fj.jbi.2019.103120&partnerID=40&md5=cff5f7d09bf63adb0df79249a60755c6>.

TUTUBALINA, E.; NIKOLENKO, S. Combination of Deep Recurrent Neural Networks and Conditional Random Fields for Extracting Adverse Drug Reactions from User Reviews. **Journal of Healthcare Engineering**, v. 2017, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85029668695&doi=10.1155%2f2017%2f9451342&partnerID=40&md5=756985a3cb41cf9897dac67d4454c1e0>.

VIANI, N.; MILLER, T.A.; NAPOLITANO, C.; *et al*. Supervised methods to extract clinical events from cardiology reports in Italian. **Journal of Biomedical Informatics**, v. 95, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066986896&doi=10.1016%2fj.jbi.2019.103219&partnerID=40&md5=51cdb5a25e12e71ec8679e661a5c6dd3>.

WANG, A.; WANG, J.; LIN, H.; *et al*. A multiple distributed representation method based on neural network for biomedical event extraction. **BMC Medical Informatics and Decision Making**, v. 17, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038939080&doi=10.1186%2fs12911-017-0563-9&partnerID=40&md5=de035e0fc6d7c2158966ce284baac34a>.

WANG, W.; BAO, F.; GAO, G. Learning Morpheme Representation for Mongolian Named Entity Recognition. **Neural Processing Letters**, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065433423&doi=10.1007%2fs11063-019-10044-6&partnerID=40&md5=baf73395be038fe9b46bdcb63d297f7d>.

WANG, W.; YANG, X.; YANG, C.; *et al*. Dependency-based long short term memory network for drug-drug interaction extraction. **BMC Bioinformatics**, v. 18, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039751643&doi=10.1186%2fs12859-017-1962-8&partnerID=40&md5=53b2371e62f08feaf56587892df19c57>.

WANG, X.; TAKAKI, S.; YAMAGISHI, J. Autoregressive Neural F0 Model for Statistical Parametric Speech Synthesis. **IEEE/ACM Transactions on Audio Speech and Language Processing**, v. 26, n. 8, p. 1406–1419, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85045722500&doi=10.1109%2fTASLP.2018.2828650&partnerID=40&md5=06611e313f4dae1499055ca5da96176e>.

WANG, X.; ZHAO, X.; REN, J.; *et al*. A new type of eye movement model based on recurrent neural networks for simulating the gaze behavior of human reading. **Complexity**, v. 2019, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064414256&doi=10.1155%2f2019%2f8641074&partnerID=40&md5=0a2095f66bb72e789f08734137e46a92>.

WECKENMANN, A.; BODI, Ş.; POPESCU, S.; *et al*. Hit or miss? Evaluating the potential of a research niche: A case study in the field of virtual quality management. **Sustainability (Switzerland)**, v. 11, n. 5, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062918352&doi=10.3390%2fsu11051450&partnerID=40&md5=fde6b3ec0c3ae3ae4c59c52161c36bac>.

WENG, W.-H.; WAGHOLIKAR, K.B.; MCCRAY, A.T.; *et al*. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. **BMC Medical Informatics and Decision Making**, v. 17, n. 1, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85037036815&doi=10.1186%2fs12911-017-0556-8&partnerID=40&md5=53e4de92e9a6411b6546a86e8ed1da25>.

WU, Y.; JIANG, M.; XU, J.; *et al*. Clinical Named Entity Recognition Using Deep Learning Models. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2017, p. 1812–1819, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058766468&partnerID=40&md5=304d586673abdf44ac9fe4dd1b71748e>.

WU, Y.; YANG, X.; BIAN, J.; *et al*. Combine Factual Medical Knowledge and Distributed Word Representation to Improve Clinical Named Entity Recognition. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2018, p. 1110–1117, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062376980&partnerID=40&md5=c6871d2e99711b0d802a1b1f322af136>.

WU, Y.; ZHANG, J. Chinese event extraction based on attention and semantic features: A bidirectional circular neural network. **Future Internet**, v. 10, n. 10, 2018. Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054557186&doi=10.3390%2ffi10100095&partnerID=40&md5=8078ecc06853ce2529ae1b36cdf3ae9f>.

WUNNAVA, S.; QIN, X.; KAKAR, T.; *et al*. Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding. **Drug Safety**, v. 42, n. 1, p. 113–122, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060165534&doi=10.1007%2fs40264-018-0765-9&partnerID=40&md5=b7c805e187ce97dd4240a3dfc3630f63>.

XU, B.; SHI, X.; YIN, Y.; *et al*. Incorporating User Generated Content for Drug Drug Interaction Extraction Based on Full Attention Mechanism. **IEEE Transactions on Nanobioscience**, v. 18, n. 3, p. 360–367, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068395252&doi=10.1109%2fTNB.2019.2919188&partnerID=40&md5=5614f6e8da9ab1ad9465bc7051c03cae>.

XU, D.; GE, R.; NIU, Z. Forward-looking element recognition based on the LSTM-CRF model with the integrity algorithm. **Future Internet**, v. 11, n. 1, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060217647&doi=10.3390%2ffi11010017&partnerID=40&md5=420a0d1ae22615930e953729a5bb2748>.

XU, K.; ZHOU, Z.; GONG, T.; *et al*. SBLC: A hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields. **BMC Medical Informatics and Decision Making**, v. 18, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058035767&doi=10.1186%2fs12911-018-0690-y&partnerID=40&md5=de7e7c12b12e79c8f4ec5e949487a447>.

XU, Kai; YANG, Zhenguo; KANG, Peipei; *et al*. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. **COMPUTERS IN BIOLOGY AND MEDICINE**, v. 108, p. 122–132, 2019.

YADAV, S.; EKBAL, A.; SAHA, S.; *et al*. Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein–protein interaction. **Knowledge-Based Systems**, v. 166, p. 18–29, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059847064&doi=10.1016%2fj.knosys.2018.11.020&partnerID=40&md5=e8d439454566148d19c310418b0d4f04>.

YANG, H.; GAO, H. Toward sustainable virtualized healthcare: Extracting medical entities from chinese online health consultations using deep neural networks. **Sustainability (Switzerland)**, v. 10, n. 9, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053393790&doi=10.3390%2fsu10093292&partnerID=40&md5=eac359cd84de057b2b52725b38399714>.

YANG, X.; MACDONALD, C.; OUNIS, I. Using word embeddings in Twitter election classification. **Information Retrieval Journal**, v. 21, n. 2–3, p. 183–207, 2018. Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85033438363&doi=10.1007%2fs10791-017-9319-5&partnerID=40&md5=43e2280c2113f31dc2d5c709fb7044d8>.

YOON, W.; SO, C.H.; LEE, J.; *et al*. CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition. **BMC Bioinformatics**, v. 20, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066302191&doi=10.1186%2fs12859-019-2813-6&partnerID=40&md5=ab11535693e8ad6cbd3d351c14d7b802>.

YU, C.; WANG, S.; GUO, J. Learning Chinese word segmentation based on bidirectional GRU-CRF and CNN network model. **International Journal of Technology and Human Interaction**, v. 15, n. 3, p. 47–62, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065074592&doi=10.4018%2fIJTHI.2019070104&partnerID=40&md5=56f445ca410e9e029d3ae91f902b42b7>.

ZENNAKI, O.; SEMMAR, N.; BESACIER, L. A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages. **Natural Language Engineering**, v. 25, n. 1, p. 43–67, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052605477&doi=10.1017%2fS1351324918000293&partnerID=40&md5=c1895d90fcfe934b2a30a62c75bae867>.

ZHANG, H.; GUO, Y.; LI, T.; *et al*. Multifeature Named Entity Recognition in Information Security Based on Adversarial Learning. **Security and Communication Networks**, v. 2019, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062789635&doi=10.1155%2f2019%2f6417407&partnerID=40&md5=d76b8ce96085a3e29d7a0c37e2baa6f7>.

ZHANG, H.; ZHANG, W.; HUANG, T.; *et al*. A two-stage joint model for domain-specific entity detection and linking leveraging an unlabeled corpus. **Information (Switzerland)**, v. 8, n. 2, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85020897437&doi=10.3390%2finfo8020059&partnerID=40&md5=243d0025f252fd1e255422d02f7b9407>.

ZHANG, Y.; TIRYAKI, F.; JIANG, M.; *et al*. Parsing clinical text using the state-of-the-art deep learning based parsers: A systematic comparison. **BMC Medical Informatics and Decision Making**, v. 19, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063965095&doi=10.1186%2fs12911-019-0783-2&partnerID=40&md5=c8ed3826bb22ecca9d1348572dcd1c22>.

ZHANG, Y.; ZHENG, W.; LIN, H.; *et al*. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. **Bioinformatics**, v. 34, n. 5, p. 828–835, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042940451&doi=10.1093%2fbioinformatics%2fbtx659&partnerID=40&md5=5cda1a00ec5a97128824147b2aa1aa5d>.

ZHAO, Q.; WANG, D.; LI, J.; *et al*. Exploiting the concept level feature for enhanced name entity recognition in Chinese EMRs. **Journal of Supercomputing**, 2019. Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067364895&doi=10.1007%2fs11227-019-02917-3&partnerID=40&md5=968c7ef087e463d7108520fb1b134dcd>.

ZHAO, Z.; YANG, Z.; LUO, L.; *et al*. Disease named entity recognition from biomedical literature using a novel convolutional neural network. **BMC Medical Genomics**, v. 10, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039701006&doi=10.1186%2fs12920-017-0316-8&partnerID=40&md5=e71bce8e80fb658d05e9a70fb8f11a08>.

ZHENG, W.; LIN, H.; LUO, L.; *et al*. An attention-based effective neural model for drug-drug interactions extraction. **BMC Bioinformatics**, v. 18, n. 1, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85030860534&doi=10.1186%2fs12859-017-1855-x&partnerID=40&md5=d226bd1b72c88e4bbc844dbcad215e9e>.

ZHOU, D.; MIAO, L.; HE, Y. Position-aware deep multi-task learning for drug–drug interaction extraction. **Artificial Intelligence in Medicine**, v. 87, p. 1–8, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044025805&doi=10.1016%2fj.artmed.2018.03.001&partnerID=40&md5=c9e78a871ffad8cadf34dc7154b0a4c5>.

ZHOU, P.; CAO, Z.; WU, B.; *et al*. EDM-JBW: A novel event detection model based on JS-ID′Forder and Bikmeans with word embedding for news streams. **Journal of Computational Science**, v. 28, p. 336–342, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85033589590&doi=10.1016%2fj.jocs.2017.11.002&partnerID=40&md5=eb0b317579d2687ee447d0e176b74dc7>.