UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS REITOR JOÃO DAVID FERREIRA LIMA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
NÍVEL MESTRADO

JAIRO BRANDÃO DE SANTANA

DESENVOLVIMENTO E ANÁLISE DE CORPUS PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM RELATÓRIOS DE INTELIGÊNCIA FINANCEIRA

FLORIANÓPOLIS
2020

Jairo Brandão de Santana

DESENVOLVIMENTO E ANÁLISE DE CORPUS PARA RECONHECIMENTO DE ENTIDADES NOMEADAS EM RELATÓRIOS DE INTELIGÊNCIA FINANCEIRA

Dissertação submetida ao Programa de Pós-graduação em Ciência da Informação da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Ciência da Informação.
Orientador: Prof. Dr. Gustavo Medeiros de Araújo
Coorientador: Prof. Dr. Vinicius Faria Culmant Ramos

Florianópolis
2020

Ficha de identificação da obra

Santana, Jairo Brandão de
Desenvolvimento e análise de corpus para reconhecimento de entidades nomeadas em relatórios de inteligência financeira / Jairo Brandão de Santana; orientador, Gustavo Medeiros de Araújo, coorientador, Vinicius Faria Culmant Ramos, 2020.
80 p.

Dissertação (mestrado) – Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-Graduação em Ciência da Informação, Florianópolis, 2020.

Inclui referências.

1. Ciência da Informação. 2. Lavagem de dinheiro. 3. Relatório de inteligência financeira. 4. Mineração de texto. 5. Processamento de linguagem natural. 6. Aprendizado de máquina. I. Araújo, Gustavo Medeiros de. II. Ramos, Vinicius Faria Culmant. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Informação. IV. Título.

Jairo Brandão de Santana

Desenvolvimento e Análise de Corpus para Reconhecimento de Entidades Nomeadas em Relatórios de Inteligência Financeira

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

Prof. Douglas Dyllon Jeronimo De Macedo, Dr.
Universidade Federal de Santa Catarina

Prof. Cristian Cechinel, Dr.
Universidade Federal de Santa Catarina

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 Informação.

Documento assinado digitalmente
Adilson Luiz Pinto
Data: 26/10/2020 19:46:58-0300
CPF: 184.218.848-83

_____

Prof. Adilson Luiz Pinto, Dr.
Coordenador do Programa de Pós-Graduação em Ciência da Informação

Documento assinado digitalmente
Gustavo Medeiros de Araujo
Data: 26/10/2020 18:30:02-0300
CPF: 828.754.991-49

_____

Prof. Gustavo Medeiros de Araújo, Dr.
Orientador

Florianópolis, 2020

Este trabalho é dedicado ao meu filho Victor e aos meus queridos pais, que sempre me incentivaram. Obrigado!

AGRADECIMENTOS

Agradeço à minha família, meu filho, meus pais e meus irmãos por sempre estarem ao meu lado e que tiveram que conviver com os momentos de ausência durante esse período de estudo.

Também sou muito grato à Polícia Federal, em especial à Academia Nacional de Polícia, por proporcionarem essa oportunidade de crescimento intelectual.

Aos colegas de turma, sou grato pela rica troca de conhecimentos e pela convivência agradável durante esse período.

À Coordenação, professores e servidores da Pós-Graduação em Ciência da Informação da UFSC, também sou grato por todo apoio, organização e pelas aulas de excelente nível que nos proporcionaram.

De forma especial agradeço aos orientadores, Dr. Gustavo Medeiros de Araujo e Dr. Vinicius Faria Culmant Ramos, por toda a paciência que tiveram comigo durante o desenvolvimento desse trabalho. Pela competência e elevado conhecimento para me conduzirem e ajudarem da melhor maneira na minha evolução na Ciência da Informação.

E, finalmente e mais importante, minha gratidão a Deus, criador e doador da vida. Pela capacidade concedida e por tão boas companhias nessa caminhada. À Ele todo glória, todo o louvor e toda a adoração.

RESUMO

Uma das competências da Polícia Federal é analisar os Relatórios de Inteligência Financeira (RIF), gerados pelo Conselho de Controle de Atividades Financeiras (COAF). Essa análise verifica a existência de algum indício de crime de lavagem de dinheiro e, se for o caso, inicia uma investigação. Essa análise é realizada de forma manual, o agente deve ler o RIF e catalogar em uma planilha todos os envolvidos e operações financeiras realizadas. Esse processo é custoso, pois o RIF pode ter dezenas de páginas. Além disso, vários relatórios são gerados mensalmente, o que agrava a demora no processamento dos RIFs. Esse projeto vem propor uma avaliação das tecnologias atuais de Mineração de Texto, mais especificamente o Reconhecimento de Entidades Nomeadas (REN) em português. A aplicação de Reconhecimento de Entidades Nomeadas ao RIF visa automatizar o processo de extração de informações do texto, submetendo o documento a um sistema computacional que faça sua leitura detalhada e retorne as informações contidas no relatório, como entidades, operações, valores, datas e vínculos entre as entidades. Dessa forma, pretende-se tornar mais ágil a análise dos RIFs. Após a leitura automatizada do texto contido no RIF, as informações extraídas podem ser armazenadas em uma de base dados e disponibilizadas de forma estruturada. Essa etapa automatizada irá facilitar a etapa seguinte de análise de vínculo, na qual consiste em detectar vínculos por meio de uma ferramenta de diagramação e análise de redes. Além disso, a organização e armazenamento dessas informações, também permitiria o cruzamento dos vínculos de diversos RIFs e manter o seu histórico.

Palavras-chave: Lavagem de dinheiro; Relatório de Inteligência Financeira; Mineração de texto; Processamento em linguagem natural; Aprendizado de máquina.

ABSTRACT

One of the responsibilities of the Federal Police to analyze the Financial Intelligence Reports (RIF), generated by the Council for Financial Activities Control (COAF). This analysis checks for evidence of a money laundering crime and, if applicable, initiates an investigation. This analysis is performed manually, the agent must read the RIF and catalog in a spreadsheet all those involved and financial operations carried out. This process is costly because the RIF may have dozens of pages. In addition, several reports are generated on a monthly basis, which aggravates the delay in processing RIFs. This project proposes an assessment of current text mining technologies, more specifically the Recognition of Named Entities (REN) in Portuguese. The application of Recognition of Named Entities to the RIF aims to automate the process of extracting information from the text, submitting the document to a computer system that reads it in detail and returns the information contained in the report, such as entities, operations, values, dates and links between entities. Thus, it is intended to make the analysis of RIFs more agile. After the automated reading of the text contained in the RIF, the extracted information can be stored in a database and made available in a structured way. This automated step will facilitate the next step of link analysis, which consists of detecting links using a diagramming and network analysis tool. In addition, the organization and storage of this information would also allow the crossing of the links of several RIFs and maintain their history.

Keywords: Money laundry; Financial Intelligence Report; Text mining; Natural language processing; Machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A lavagem de dinheiro remonta ao início do século XX com o surgimento das primeiras organizações criminosas, as chamadas máfias, que despontaram principalmente nos Estados Unidos (BRAGA, 2010). Com a evolução, tanto do montante de receitas advindas de atividades ilícitas, quanto das formas de transacionar essas quantias na rede bancária mundial, surgem os paraísos fiscais, países que permitem manter o dinheiro protegido das autoridades dos países de origem. Neste contexto, a Suíça tornou-se o principal destino deste dinheiro (PINTO, 2007).

O surgimento desses paraísos fiscais, o narcotráfico, o crime organizado, a criação de bancos transnacionais, o avanço da Tecnologia da Informação e o próprio mercado financeiro globalizado são fatores que potencializaram e facilitaram o crescimento da lavagem de dinheiro (LEFORT, 1997).

Lustosa (2009, p. 1) define que lavagem de dinheiro "é uma forma genérica de referir-se ao processo ou conjunto de operações de ocultar a origem de dinheiro ou dos bens resultantes das atividades delitivas e integrá-los no sistema econômico ou financeiro, em operações capazes de converter o dinheiro sujo em dinheiro limpo".

A lavagem de dinheiro é um processo complexo no qual podem ser identificadas quatro fases: pré-lavagem, ocultação, dissimulação e integração. Na fase inicial, chamada de pré-lavagem, é onde ocorre a captação e concentração de recursos ilícitos, resultado de um ato criminoso qualquer. Nessa fase, é o momento ideal para se detectar um crime de lavagem de dinheiro, pois o autor do crime ainda não recorreu às medidas que possam esconder ou disfarçar a origem do ativo (ARAS, 2007).

A ocultação consiste basicamente em esconder e afastar o ativo da origem ilícita para que se possa evitar que seja rastreado e uma das maneiras é realizar diversas transações com valores fracionados, que desobriga a sua comunicação às autoridades financeiras (MENDRONI, 2015).

A próxima etapa é a dissimulação, que tem por fim disfarçar a origem criminosa dos valores, camuflando evidências através de uma série de complexas transações financeiras internacionais em países que não cooperam com o combate à lavagem de dinheiro, conhecidos por "paraísos fiscais" (BRAGA, 2010).

Por fim, a fase de integração é onde se tem os benefícios dos ativos como se fossem lícitos, seja através da compra de bens ou no investimento em empresas comerciais criadas e operando de forma legal. Nessa fase o Ministério Público terá extrema dificuldade de provar que esses ativos ou bens tiveram origem em recursos criminosos, pois tudo foi realizado de forma a não existir testemunhas ou documentos que comprovem a lavagem de dinheiro (ARAS, 2007).

O cenário estabelecido nos anos 80, com a facilidade criada pela globalização comercial e comunicação entre os diversos mercados criou uma série de novas oportunidades para que organizações criminosas pudessem executar a lavagem de dinheiro:

> Com o incremento da globalização econômica, era de se esperar que os fatores positivos que favoreceram a interação dos mercados globais fossem apropriados por organizações criminosas nacionais (as tradicionais "máfias") e já então por grupos criminosos transnacionais, especializados em contrabando, tráfico de drogas, tráfico de pessoas, tráfico de armas, tráfico de animais silvestres, entre outros delitos. De fato, a maior facilidade de interação à distância com a difusão das telecomunicações e da internet, a maior facilidade de transporte de bens por todo o globo e a eliminação de barreiras domésticas à livre circulação de pessoas e valores são fatores que não foram ignorados pelos operadores de atividades ilícitas desencadeadas em vários pontos do Brasil e de outros países, com o fim de adquirir, transportar e distribuir drogas, mercadorias contrafeitas, armas e munições. O quadro logístico montado para atender a legítimos negócios internacionais passou a ser utilizado por organizações criminosas de todo o mundo. E as vantagens econômicas advindas desses negócios ilícitos passaram a transitar pela economia global, contando com as mesmas facilidades dos capitais legítimos. (ARAS, 2007, p. 1).

Com a intenção de impedir a utilização de valores ou bens que tiveram origem nas atividades das organizações criminosas, por volta da década de 1980, diversas nações iniciaram alterações em suas leis criando dispositivos que, inicialmente, tipificavam o uso, a movimentação, a disponibilização ou a ocultação de recursos que tiveram origem no narcotráfico.

Somente no fim da década de 1980 foram tomadas medidas mais concretas para o combate da lavagem de dinheiro em âmbito internacional. Capitaneada pela ONU, foi realizada a Convenção de Viena em 1988, com o objetivo de combater o tráfico internacional de entorpecentes (ONU, 1988). E em 1989, a cúpula do G-7, que foi realizada em Paris, criou o Grupo de Ação Financeira contra a Lavagem de Dinheiro e o Financiamento do Terrorismo - GAFI/FATF, com o objetivo de elaborar políticas de proteção ao sistema bancário e às instituições financeiras e identificar suas vulnerabilidades. O GAFI publicou em 1990 uma série

de 40 recomendações (GAFI, 1990). As Recomendações do GAFI (1990) definem as medidas essenciais que os países devem adotar para:

a) Identificar os riscos e desenvolver políticas e coordenação doméstica;
b) Combater a lavagem de dinheiro, o financiamento do terrorismo e sua proliferação;
c) Aplicar medidas preventivas para o setor financeiro e outros setores designados;
d) Estabelecer poderes e responsabilidades para as autoridades competentes (por exemplo: autoridades investigativas, policiais e fiscalizadoras) e outras medidas institucionais;
e) Aumentar a transparência e disponibilidade das informações sobre propriedade de pessoas jurídicas e de outras estruturas jurídicas;
f) Facilitar a cooperação internacional.

Com a intenção de se padronizar as legislações dos diversos países de combate à lavagem de dinheiro oriundos do tráfico de drogas, do crime organizado e da corrupção, foram formulados diversos tratados internacionais multilaterais além do GAFI. A intenção era manter uma economia mundial saudável e proteger a economia de mercado para funcionar de forma justa. Essa preocupação despertou o interesse de inúmeros governos nacionais (ARAS, 2007).

Mas os ataques terroristas aos Estados Unidos em 11 de setembro de 2001, feitos pela Al Qaeda, foram determinantes para um avanço significativo de um sistema global antilavagem de ativos. Os Estados Unidos fizeram diversas alterações em suas leis com o objetivo de combater o terrorismo e suas fontes de recursos. Entre elas a criação da *USA Patriot Act* (Lei Patriótica), extremamente rigorosa e também polêmica, que permitia, por exemplo, a interceptação sem necessidade de ordem judicial, de e-mails e ligações telefônicas de qualquer pessoa ou organização suspeita de envolvimento com terrorismo. Além disso, para combater o financiamento do terrorismo de forma global, passaram a pressionar fortemente a criação de legislações internacionais e domésticas através dos diversos foros internacionais da ONU, OEA, OCDE e GAFI. Juntamente com os Estados Unidos, outros países que também foram ou poderiam ser alvos de ataques terroristas, caso de diversos países europeus, também criaram novas legislações de cooperação internacional com o intuito de coibir os crimes realizados por organizações criminosas e de lavagem de dinheiro. Desde então, mesmo os países que não são

potenciais alvos de terrorismo, como o Brasil, têm adotado o mesmo padrão de legislação utilizando a cooperação internacional (ARAS, 2007).

O crime organizado é uma ameaça aos regimes democráticos bem como à segurança nacional e mundial. E utiliza a lavagem de dinheiro como um eficiente meio de garantir e aumentar cada vez mais suas atividades ilícitas. Esse mecanismo prejudica o desenvolvimento econômico nacional e afeta negativamente as relações com outros países. A guerra civil na Colômbia é um clássico exemplo de como o desenvolvimento social e econômico dessa nação foram afetadas pelo crime organizado do narcotráfico (ROMANTINI, 2003).

Quanto menor o índice de desenvolvimento de um país, maior o impacto negativo dos efeitos do crime organizado e da lavagem de dinheiro. Sua prática constante é um retrato de diversos crimes que a precedem e são uma ameaça para sociedade, como narcotráfico, corrupção, tráfico de armas, terrorismo, entre outros. Por isso, os governos dos países tiveram que atentar-se a esse problema mundial chamado lavagem de dinheiro. Os danos que essa ilicitude pode trazer à sociedade são diversos e extremamente complexos. Segundo Aras (2007, p. 2), os efeitos diretos e indiretos que o dinheiro sujo circulando numa nação podem causar são:

a) Extinguir empreendimentos honestos que não possuem a facilidade de contar com recursos ilícitos;
b) Provocar variações artificiais nas bolsas de valores;
c) Levar à falência empresas concorrentes de outras que utilizam recursos de lavagem de dinheiro, causando desemprego e outros problemas sociais;
d) Favorecer o surgimento de monopólios e oligopólios, cobrando preços abusivos e fornecendo serviços e produtos de pior qualidade, prejudicando assim a população;
e) Diminuir a arrecadação de impostos pelo governo;
f) Causar distorções no mercado financeiro e torna o ambiente econômico instável;
g) Diminuir o índice de desenvolvimento humano;
h) Aumentar a corrupção;
i) Aumentar a insegurança pública;
j) Reduzir os investimentos em educação e saúde;
k) Expandir a desigualdade social através do enriquecimento ilícito.

1.1 COMBATE À LAVAGEM DE DINHEIRO NO BRASIL

Em 1998, foi promulgada a lei 9.613 chamada de “Lei de Lavagem de Capitais” ou também de “Lei de Lavagem de Dinheiro” com o intuito de disponibilizar uma lei específica para esse tipo de crime e também adequar o país à Convenção de Viena e às recomendações do GAFI (OBREGON, 2001). Essa lei instituiu o sistema brasileiro de prevenção e combate ao crime de lavagem de dinheiro através da criação do cadastro nacional de clientes do sistema financeiro nacional (BRASIL, 1998, art. 10A), da instituição da responsabilidade administrativa de sujeitos obrigados (BRASIL, 1998, arts. 9º e 12), estabelecimento de regras de *compliance* (adequação) para certos sujeitos obrigados que são integrantes de setores econômicos relevantes (BRASIL, 1998, arts. 9º a 11) e, por fim, a criação do Conselho de Controle de Atividades Financeiras (COAF) que tem como função a prevenção e combate à lavagem de dinheiro e ao financiamento do terrorismo e possui as seguintes competências (BRASIL, 1998, arts. 14 a 17):

a) Receber, examinar e identificar as ocorrências suspeitas de atividades ilícitas;
b) Comunicar às autoridades competentes para a instauração dos procedimentos cabíveis (quando concluir pela existência de crimes previstos na referida lei, de fundados indícios de sua prática, ou de qualquer outro ilícito);
c) Coordenar e propor mecanismos de cooperação e de troca de informações que viabilizem o combate à ocultação ou dissimulação de bens, direitos e valores;
d) Disciplinar e aplicar penas administrativas;
e) Regular os setores econômicos para os quais não haja órgão regulador ou fiscalizador próprio.

As atribuições do COAF são imprescindíveis para coibir as ilicitudes do crime organizado que utilizam a lavagem de dinheiro. Além disso, as informações disponibilizadas pelo COAF são essenciais para investigação criminal e contribuem para a localização dos ativos oriundos da lavagem de dinheiro e seus autores e outros envolvidos, permitindo assim a

recuperação judicial dos produtos resultantes do crime e a devida condenação dos culpados (ARAS, 2007).

O COAF atua em conjunto com diversos órgãos e entidades financeiras com o objetivo de evitar que a utilização desses diversos setores nas práticas de lavagem de dinheiro. São os principais órgãos e entidades integrados ao COAF:

a) Comissão de Valores Mobiliários – CVM;
b) Secretaria de Previdência Complementar – SPC;
c) Superintendência de seguros privados – SUSEP;
d) Conselho Federal de Corretores de Imóveis – COFECI;
e) Associação Brasileira das Empresas de Cartões de Crédito e Serviços – ABECS;
f) Associação Brasileira das Entidades Fechadas de Previdência Privada – ABRAPP;
g) Federação Brasileira de Bancos – FEBRABAN.

Estes, quando detectam operações atípicas realizadas pelos usuários de suas empresas representadas e supervisionadas ou mesmo se há incompatibilidade financeira entre as transações efetuadas e a capacidade econômica desses usuários, têm a obrigação de informar ao COAF. O controle deve ser realizado não apenas sobre as pessoas físicas, mas também em qualquer empresa, grupo ou conglomerado (MARQUES, 2014).

Essas entidades, também conhecidas por sujeitos obrigados, têm a obrigação de conhecerem seus clientes. Eles são os que possuem as melhores condições de avaliar se compras, vendas, movimentações ou aplicações específicas têm alguma característica suspeita, pois conhecem os rendimentos e atividades dos clientes pessoas físicas e os lucros e estrutura dos clientes pessoas jurídicas. Esses sujeitos obrigados têm o poder de coibir negociações fraudulentas e que ativos de origem criminosa tenham livre circulação, funcionando como uma torre de vigia. Eles devem possuir em sua estrutura áreas de *compliance* (adequação) para manter o histórico das operações financeiras por 5 anos, além de possuir a completa identificação dos seus clientes. Podemos aqui observar a grande importância desses sujeitos obrigados como uma engrenagem essencial para o combate à lavagem de dinheiro (ARAS, 2007).

O Banco Central, através da Carta Circular nº 2.826/01, normatiza vários eventos que são definidos como *red flags*, bandeiras vermelhas, que alertam as instituições financeiras a acompanharem e informarem o Banco Central:

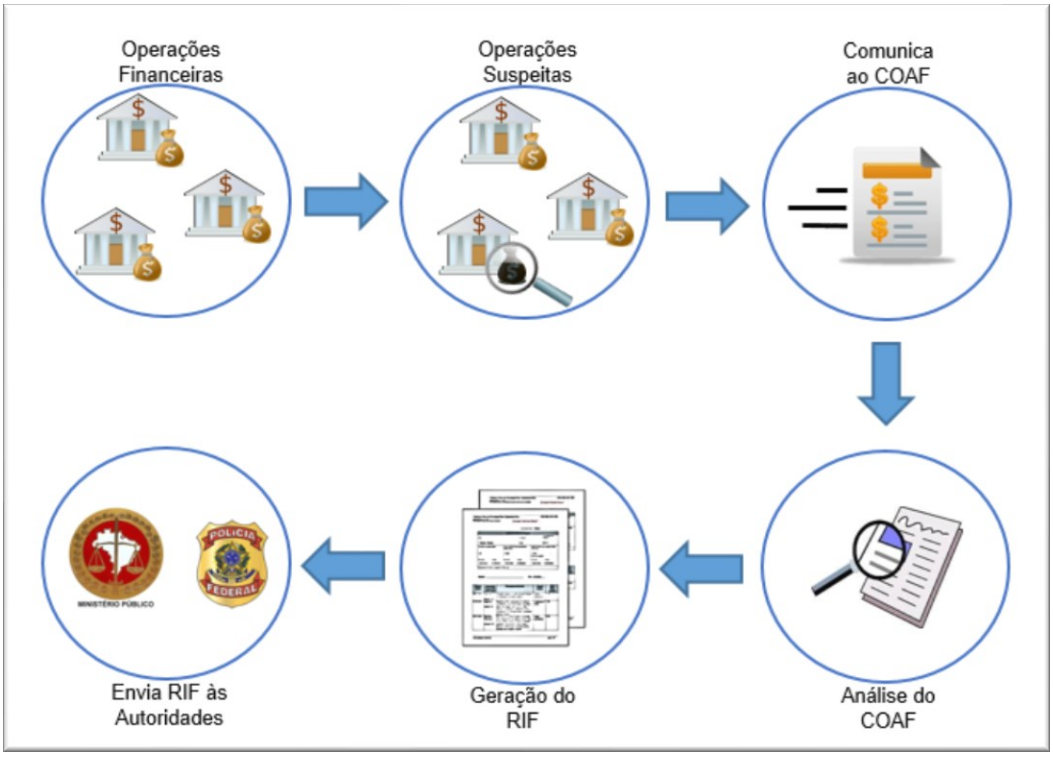
a) Alterações substanciais na rotina da conta bancária;
b) Grande atividade por *wire transfer* (transferência bancária);
c) Operações sem sentido econômico;
d) Uso de várias contas simultaneamente;
e) Movimentação incompatível com o negócio ou a profissão;
f) Relações com paraísos fiscais;
g) Estruturação de operações com fracionamento de depósitos ou remessas;
h) Recusa em informar origem de recursos ou a própria identidade;
i) Inconsistência documental.

Apesar de não haver tipificação do crime de não comunicação de operações suspeitas, o descumprimento da lei de lavagem de dinheiro, lei nº 9.613/98, os sujeitos obrigados respondem administrativamente ao COAF e órgãos reguladores e seus dirigentes podem ser criminalmente processados como participantes ou coautores de crime de lavagem de dinheiro por omissão (ARAS, 2007).

Recebida a comunicação de operações suspeitas, o COAF realiza a análise para detectar se existem evidências de crime de lavagem de dinheiro e, se realmente forem detectadas, se será feito um intercâmbio de informações com as autoridades competentes (MARQUES, 2014). Conforme é disciplinado pelo artigo 15 da Lei nº 9.613/1998: "O COAF comunicará às autoridades competentes para a instauração dos procedimentos cabíveis, quando concluir pela existência de crimes previstos nesta Lei, de fundados indícios de sua prática, ou de qualquer outro ilícito" (BRASIL, 1998, art. 15).

São consideradas autoridades competentes para receberem os informes do COAF, o Ministério Público Federal, a Polícia Federal e o Ministério Público do respectivo Estado, os quais poderão proceder com o bloqueio da operação financeira suspeita, iniciar uma investigação criminal e, se for o caso, propor a ação penal (ARAS, 2007). O informe disponibilizado pelo COAF às autoridades competentes é chamado de Relatório de Inteligência Financeira (RIF). Podemos observar, na Figura 1, as etapas percorridas por esse processo, que também é conhecido como "Inteligência Financeira".

Figura 1: Processo de Inteligência Financeira.



Fonte: Elaborado pelo autor.

Segundo o site do COAF, no ano de 2018, a produtividade de RIFs e as ações decorrentes são:

> De janeiro a novembro de 2018, o Coaf produziu 6.786 Relatórios de Inteligência Financeira (RIF), os quais relacionaram 348.984 pessoas físicas ou jurídicas, e consolidaram 302.648 comunicações de operações financeiras. Atualmente, a base de dados reúne mais de 16,7 milhões de comunicações de operações financeiras. Desse total, aproximadamente 2,8 milhão de comunicações foram recebidas somente em 2018, provenientes dos setores econômicos obrigados a comunicar.
> A atuação do Coaf, juntamente com o Ministério Público e autoridades policiais, possibilitou, de janeiro a novembro de 2018, o bloqueio judicial de R$ 125 milhões no Brasil e no exterior, relacionados a investigações sobre lavagem de dinheiro e crimes relacionados. (COAF, 2015, http://www.fazenda.gov.br/orgaos/coaf).

Durante o desenvolvimento desse trabalho, em 19/08/2019, foi publicada a Medida Provisória 893/2019[1], alterando o nome de COAF para UIF – Unidade de Inteligência Financeira e transferindo-o da estrutura do Ministério da Fazenda para o Banco Central do

[1] http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2019/Mpv/mpv893.htm

Brasil. Em 07/01/2020 foi sancionada a Lei 13.974/2020[2] confirmando a transferência definitiva para o Banco Central do Brasil, mas o nome retornou a ser COAF.

1.2 DELIMITAÇÃO DO PROBLEMA

A Polícia Federal é uma das principais autoridades competentes responsáveis pela investigação de crimes de lavagem de dinheiro. Recebe do COAF o RIF, que relata as transações suspeitas. Atualmente as áreas de inteligência da PF analisam o RIF de modo manual e identificam entidades, valores e operações realizadas (ZAINA, 2020).

Essas informações são tabuladas, permitindo a sua leitura por ferramentas de Análise de Vínculos e a geração de diagramas de relacionamentos entre as entidades, facilitando a detecção de organizações criminosas.

Figura 2: Detecção de Vínculos.

Fonte: Elaborado pelo autor.

No âmbito da Polícia Federal, o RIF é recebido pela área de repressão a crimes financeiro. Esta área realiza uma avaliação preliminar do documento para indicar qual será a delegacia responsável por sua análise. Estando na delegacia específica, o RIF é analisado minuciosamente por um policial para tentar detectar indícios necessários que possam retratar uma prática criminosa e, consequentemente, a instauração de um procedimento investigativo formal (POLÍCIA FEDERAL, 2013).

[2] http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2020/Lei/L13974.htm

O RIF é um documento em formato PDF (*Portable Document Format*), e é escrito em linguagem natural, não estruturado, com relatos descritos pelos analistas do COAF de onde deverão ser extraídos as referências aos possíveis vínculos. Abaixo podemos ver um exemplo fictício de um relato contido no RIF:

> 1. A empresa X1 Ltda., sediada em Brasília/DF, atuante no ramo de administração de consórcios, com capital social de R$ 100.000,00, foi objeto de comunicações de operações financeiras de que trata a Lei 9.613/98. Desde 02/12/2013, figuram como sócios Fulano de Tal e Beltrano da Silva, que até março de 2014, eram funcionários da empresa X2 Ltda., com rendas mensais de R$ 1.836,00 e R$ 1.522,00, respectivamente.
> 1.1. Conforme o comunicante, a X1 Ltda. movimentou, entre 01/09/2014 e 15/10/2014, o total de R$ 35.052.870,00, na conta corrente nº 12345, agência nº 0001, do Banco X3 S.A., da cidade de Taguatinga/DF.

Dependendo do nível de complexidade do RIF, a análise poderá ser realizada apenas com a sua leitura e, nos casos mais complexos, as entidades e vínculos detectados deverão ser digitados e armazenados em planilhas eletrônicas ou em banco de dados e, posteriormente, a utilização de softwares analíticos para a análise desses vínculos.

Na fase de detecção e tabulação de vínculos, é essencial que seja realizada com o maior grau de precisão possível pois, caso contrário, irá comprometer uma correta análise na fase posterior. A Figura 3 é um exemplo de informações tabuladas após análise do RIF.

Figura 3 - Informações do RIF tabuladas.

ORDEM	RIF	ITEM RIF	ORIGEM ou DEPOSITANTE CPF/CNPJ	ORIGEM ou DEPOSITANTE NOME	DESTINO ou SACADOR CPF/CNPJ	DESTINO ou SACADOR NOME	RESPONSÁVEL pelo DEPÓSITO ou SAQUE CPF/CNPJ	RESPONSÁVEL pelo DEPÓSITO ou SAQUE NOME	TIPO OPERAÇÃO Selecione	VALOR (EM R$) somente números	DATA/PERIODO
01	12345	R-1.1.1	01.222.888/0001-70	Auto Posto 1 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda			Transferência	R$ 4.411.827,00	01/09/2014 a 15/10/2014
02	12345	R-1.1.1	01.555.999/0001-30	Auto Posto 2 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda			Transferência	R$ 3.655.000,00	01/09/2014 a 15/10/2014
03	12345	R-1.1.1	01.777.222/0001-52	Auto Posto 3 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda			Transferência	R$ 1.963.150,00	01/09/2014 a 15/10/2014
04	12345	R-1.1.1	01.888.444/0001-88	Auto Posto 4 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda			Transferência	R$ 3.124.624,00	01/09/2014 a 15/10/2014
05	12345	R-1.1.1	01.222.777/0001-01	Auto Posto 5 Ltda	19.555.999/0001-01	X1 Fomento Mercantil Ltda			Transferência	R$ 4.300.049,00	01/09/2014 a 15/10/2014
06	12345	A-1.1.	19.555.999/0001-01	X1 Fomento Mercantil Ltda	858.756.543-53	Fabiana Siqueira	087.654.321-00	Ricardo Gil	Saque	R$ 100.000,00	06/10/2014
07	12345	R-1.1.2	19.555.999/0001-01	X1 Fomento Mercantil Ltda	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda			Transferência	R$ 8.572.000,00	01/09/2014 a 15/10/2014
08	12345	R-1.1.2	19.555.999/0001-01	X1 Fomento Mercantil Ltda	06.999.555/0001-06	Logs Logística e Serviços Ltda			Transferência	R$ 8.331.317,00	01/09/2014 a 15/10/2014
09	12345	R-2.1	858.756.543-53	Fabiana Siqueira	02.333.777/0001-13	Zel Petróleo Ltda	02.333.777/0001-13	Zel Petróleo Ltda	Depósito	R$ 130.000,00	05/11/2012
10	12345	R-3.	06.999.555/0001-06	Logs Logística e Serviços Ltda	02.333.777/0001-13	Zel Petróleo Ltda			Transferência	R$ 4.734.930,00	01/09/2014 a 15/10/2014
11	12345	R-3.	06.999.555/0001-06	Logs Logística e Serviços Ltda	15.151.151/0001-09	Contan Logística Ltda			Transferência	R$ 496.614,00	01/09/2014 a 15/10/2014
12	12345	R-4.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	02.333.777/0001-13	Zel Petróleo Ltda			Transferência	R$ 4.637.200,00	01/09/2014 a 15/10/2014
13	12345	R-5.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	011.444.333-40	Cicrano Ferreira	Saque	R$ 200.000,00	01/10/2014
14	12345	R-5.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	011.444.333-40	Cicrano Ferreira	Saque	R$ 100.000,00	03/10/2014
15	12345	R-5.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	011.444.333-40	Cicrano Ferreira	Saque	R$ 250.000,00	07/10/2014
16	12345	R-5.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	011.444.333-40	Cicrano Ferreira	011.444.333-40	Cicrano Ferreira	Saque	R$ 150.000,00	10/10/2014
17	12345	R-6.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	099.222.555-87	Beltrano de Sousa	099.222.555-87	Beltrano de Sousa	Saque	R$ 200.000,00	08/10/2014
18	12345	R-6.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	099.222.555-87	Beltrano de Sousa	099.222.555-87	Beltrano de Sousa	Saque	R$ 100.000,00	09/10/2014
19	12345	R-6.	07.555.222/0001-40	Logs Distribuidora de Petróleo Ltda	099.222.555-87	Beltrano de Sousa	099.222.555-87	Beltrano de Sousa	Saque	R$ 100.000,00	10/10/2014

Fonte: Elaborado pelo autor.

Devido à grande quantidade de RIFs enviados à Polícia Federal e baixa quantidade de recursos humanos disponibilizados para a tarefa de análise, seria bom ter uma solução automatizada para leitura do RIF e detecção das entidades e eventos. Além do ganho com a agilidade de leitura dos RIFs, também haveria uma melhor padronização do trabalho de análise.

Intenta-se, neste trabalho, apresentar discussões empreendidas no âmbito da CI e as abordagens dos problemas por ela estudados, bem como teorias voltadas à recuperação da informação. Além disso, levantamos aspectos quanto ao reconhecimento de entidades nomeadas em textos em português. Finaliza-se o mesmo, observando os problemas enfrentados pelo profissional da informação, especificamente, das ciências policiais, quando este exerce a função de analista de informações de documentos de investigação, além de buscar estabelecer as relações entre a CI e a Ciência Policial no campo de REN.

1.3 OBJETIVO

Desenvolver um corpus a partir de RIFs para o Reconhecimento de Entidades Nomeadas que permitam a leitura automatizada do RIF e a consequente detecção das entidades que serão tabuladas e utilizadas na análise do agente.

1.3.1 Objetivos Específicos

a) Analisar as tecnologias de Mineração de Texto que permitem a leitura e o reconhecimento de entidades em português.
b) Pesquisar trabalhos encontrados na literatura sobre reconhecimento de entidades em textos.
c) Extrair informações dos relatórios de inteligência financeira usando técnicas e ferramentas de mineração de textos para o reconhecimento de entidades nomeadas.
d) Analisar a aplicação do corpus spaCy em reconhecimento de entidades nomeadas nos RIFs.
e) Desenvolver um corpus baseado nas informações dos RIFS.

1.4 CONTRIBUIÇÕES DO TRABALHO

Nesse trabalho foram desenvolvidos e analisados corpora em português para o reconhecimento de entidades nomeadas em documentos RIFs. Nesse contexto, foi desenvolvido um corpus específico baseado nas informações extraídas dos RIFs e comparado com o corpus de domínio público, o spaCy.

Ao final, foi construído um segundo corpus unindo o corpus específico baseado em RIFs, com o corpus de domínio público spaCy, o que demonstrou uma melhora de 5% na acurácia para o REN.

1.5 ESTRUTURA DO TEXTO

No próximo capítulo será tratada a revisão bibliográfica, inicialmente discutindo os conceitos e tecnologias tratados neste trabalho. Em seguida, é definido o protocolo para Revisão Sistemática da Literatura (RSL) para identificar trabalhos que tratam de Reconhecimento de Entidades Nomeadas (REN) em português e outros idiomas, e demonstrando os resultados encontrados.

No capítulo seguinte será apresentada a revisão da literatura com os conceitos e tecnologias abordados nesse trabalho. O capítulo 3 apresentará a metodologia proposta do trabalho, detalhando a descrição dos experimentos desenvolvidos e discutida a melhor forma de alcançar os objetivos do trabalho. Já no capítulo 4, os resultados serão apresentados e discutidos, enquanto o capítulo 5 tratará as conclusões e considerações finais do trabalho.

2 REVISÃO DA LITERATURA

2.1 CONCEITOS E TECNOLOGIAS

2.1.1 Ciência da Informação e Ciência Policial

Os estudos na ciência da informação podem ter o enfoque em área de conhecimento específica como medicina, engenharia civil, computação, dentre outras, ou em um contexto mais abrangente, como contexto artístico, industrial, tecnológico, etc. No enfoque mais abrangente, a informação pode ser utilizada para tomada de decisão, quando relacionada ao conhecimento gerencial ou administrativo, e também pode ser econômico, adquirindo valor agregado e sendo utilizada para ações políticas e sociais (PINHEIRO, 2004).

A disseminação e vasta utilização do termo informação, tem inclusive contribuído para que a sociedade tenha outra visão da função da biblioteca e da documentação (CAPURRO; HJORLAND, 2007). Mas até mesmo a definição e utilização da palavra informação pode causar confusão nessa nova ciência e, por isso, Capurro e Hjorland alertam:

> Quando usamos o termo informação em CI, deveríamos ter sempre em mente que informação é o que é informativo para uma determinada pessoa. O que é informativo depende das necessidades interpretativas e habilidades do indivíduo (embora estas sejam frequentemente compartilhadas com membros de uma mesma comunidade de discurso) (CAPURRO; HJORLAND, 2007, p. 155).

É importante conceituar e diferenciar informação, conhecimento e dado, pois são termos que possuem uma amplitude semântica e diversas perspectivas de análise. Dado pode ser considerado como a menor partícula de informação e não permite extrair algum significado à compreensão humana. Informação permite extrair algum significado quando se analisa um conjunto de dados. E conhecimento é o que se pode obter da informação disponibilizada dependendo de seu contexto. O processo de obtenção ou não de conhecimento a partir das informações observadas, dependerá da percepção de cada observador, tornando uma reação estritamente individual (LIMA; ALVARES, 2012).

A informação é o que induz a produção de conhecimento, assim que chega ao cérebro e impacta os neurônios. Então começam a acontecer, de forma sucessiva ou simultânea, processos de percepção, apreensão, análise, classificação, arquivo em memória, avaliação que constituem o conhecimento pessoal, subjetivo e condicionado pelo substrato individual e cultural de cada indivíduo. Numa elaboração mental posterior, mais complexa, o conhecimento passa a constituir as ideias, linhas de pensamento. Essas são as que voltam a se converter em informação útil, quando surge a ocasião (CURRÁS, 2010).

O conhecimento então é formado pela informação, e a informação é um conhecimento possível de ser registrado em algum suporte físico, seja impresso, digital, oral ou audiovisual. E que transmite algum significado (LE COADIC, 2004). A Ciência da Informação tem como objeto de estudo a própria informação, permeando seus conceitos e definições. E tem como principal fenômeno analisado em seus estudos a geração, transferência e utilização da informação (PINHEIRO, 2004).

Dentre as definições de Ciência da Informação, podemos destacar a de Capurro (2003) de que "essa ciência tem como objeto a produção, seleção, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso da informação" (GRIFFITH, 1980 apud CAPURRO, 2003). E a de Borko (1968) que "Em essência, a Ciência da Informação investiga as propriedades e o comportamento da informação, o uso e a transmissão da informação, e o processamento da informação, visando uma armazenagem e uma recuperação ideal" (BORKO, 1968, p. 4).

A informação armazenada, processada ou gerada pelos sistemas de informação deve refletir a sua função social. Pois os sistemas de informação existem necessariamente para auxiliar alguma atividade humana, por isso, devem sempre levar em consideração os pressupostos sociais que motivaram a necessidade de seu desenvolvimento (CAPURRO; HJORLAND, 2007).

Mas a globalização da informação influencia em transformações sociais, culturais e políticas. E a sociedade demanda que a segurança pública também evolua e utilize os avanços tecnológicos na gestão da informação. Por outro lado, a utilização da tecnologia da informação também fomenta novas formas de crimes, gerando novas ameaças à segurança da sociedade. (MARTIN; AGUILAR, 2011).

Segundo Bauman (2007), a "Sociedade Líquida", que é o sentimento de complexidade das novas dinâmicas informacionais globais, que surgiram ao fim da Segunda Guerra em meio à infinidade de informações disponíveis, fez com que novos padrões surgissem. Entre eles, a

Ciência Policial pois, com a descoberta de novos tipos de crime e que não se limitam às fronteiras de um país, tem o intuito de aumentar o sentimento de segurança da sociedade.

Por se tratar de uma ciência recente e ainda estar sendo formulada, a Ciência Policial possui pouca literatura. Mas, para ir ao encontro das necessidades da sociedade por segurança, tem evoluído na medida em que novas soluções são apresentadas baseadas em metodologia científica. A Ciência Policial surge no contexto sociocultural, sendo retrato da pós modernidade (ALMEIDA et al., 2007).

Morales e Cândido (2018) analisando a Ciência Policial, concluem que:

> O novo paradigma da ciência imposto pela pós-modernidade e a sua visão sobre a realidade passa obrigatoriamente pela aceitação da interdisciplinaridade do conhecimento, tendo como consequência imediata a aceitação de novas ciências, como a Ciência da Informação e a Ciência Policial. A Ciência Policial é uma ciência de natureza empírica pertencente ao ramo de estudos das ciências sociais aplicadas, a qual supre seus vazios epistemológicos utilizando-se dos conceitos de outras ciências sedimentadas, como a Teoria Geral Jurídica, a Sociologia, a Antropologia, a Psicologia, a História, a Ciência da Informação. Seu campo de estudo não se limita à atividade policial finalística, qual seja, a segurança pública, mas a organização e estrutura da polícia como sistema de conhecimento, em busca de inovação através da gestão do conhecimento, o seu insumo basilar (MORALES; CÂNDIDO, 2018, p. 5).

Devido a esse cenário, podemos ver que o avanço do crime organizado na utilização das mais modernas tecnologias, para cometer e ocultar crimes, tem sido cada vez maior. Principalmente nas últimas décadas, com o desenvolvimento da internet, transações eletrônicas e a integração dos mercados financeiros nacionais. E isto reforça ainda mais a necessidade de evolução da Ciência Policial nessas áreas, de forma a evitar ou esclarecer até mesmo os crimes mais avançados. Ela necessita enfrentar, de maneira concreta, os diversos fenômenos que ocorrem em nossa sociedade que tem contínuas mudanças. Para isso é necessário que as forças policiais sejam treinadas de forma a conhecerem e utilizarem as mais avançadas técnicas de investigação.

Analisando a realidade da Ciência Policial, Fentanes (1972 apud PEREIRA, 2015, p. 60) diz que “a denominação Ciência implica em afirmar que o estudo da polícia assume a qualidade de conhecimento científico, considerado como um sistema de conhecimentos”. E também podemos entender que a Ciência Policial é uma combinação de intenções e procedimentos de várias áreas relacionadas no contexto atividade policial, que não é apenas a

estrita atividade da Polícia, mas também influências externas que possam impactar a ordem pública e a atividade policial. (JASCHKE, 2005).

2.1.2 Mineração de Textos

As evoluções científicas e tecnológicas têm sido marcantes em nossa era, tornando acessíveis, à grande parte da população mundial, avançados equipamentos que possibilitam o acesso à internet nos mais diversos locais do mundo e de maneira rápida e estável. Essa facilidade produz uma infinidade de informações que são armazenadas e disponibilizadas. Mas esse grande volume de informação torna o trabalho de extração de informação muito difícil. (SOUZA; CLARO, 2014).

Segundo (SANTOS et al., 2014), grande parte dessas informações encontra-se armazenada em documentos textuais em forma de linguagem natural. Pois a popularização alcançada pela internet nos anos recentes aumentou, de maneira significativa, o volume de informações. Sem dúvida, isso facilitou o processo de disseminação do conhecimento, mas como consequência desse mesmo fato, cada vez mais temos dificuldades para encontrar as informações realmente relevantes. Isso é devido à grande quantidade de dados disponibilizados na web. A previsão é que em 2020 seja atingido o volume de 44 trilhões de gigabytes (GONÇALVES, 2018).

Esse constante desenvolvimento da Tecnologia da Informação, principalmente das redes de computadores e pela internet, fez surgir uma infinidade de documentos digitais com textos que representam a linguagem com a qual as pessoas se comunicam no dia a dia, seja de maneira formal ou informal. Mas as tradicionais linguagens de desenvolvimento de software não possuem a capacidade de detectar a ideia que o texto deseja expressar, pois geralmente possuem ambiguidades e contextos específicos que apenas a cognição do ser humano consegue compreender. (MACHADO et al., 2010).

Na intenção de solucionar essas dificuldades na descoberta de conhecimento em textos não estruturados, foi desenvolvida a mineração de textos (*text mining*), que disponibiliza uma série de técnicas que possibilitam navegar, organizar e descobrir informação nesses textos, de forma inteligente. A Mineração de Textos consiste na extração de informações úteis em textos não estruturadas escritos em linguagem natural e envolve várias áreas da informática como mineração de dados, aprendizado de máquina, recuperação de informação e linguagem

computacional, para poder transformar os textos analisados em algo compreensível para o computador.

Essa tecnologia auxilia na descoberta de informações desconhecidas anteriormente. Por isso, essa tecnologia não pode ser confundida com a de um de mecanismo de busca, pois neste caso o usuário já conhece o que deseja pesquisar. (ARANHA; VELLASCO; PASSOS, 2007). Encontrar termos relevantes em um grande volume de documentos em linguagem natural e definir suas relações e padrões, pode ser descrito como o principal objetivo da tecnologia de Mineração de Texto. (SERAPIÃO; SUZUKI; MARQUES, 2010).

Uma definição sucinta para a Mineração de Textos é descrita por Pezzini (2017):

> A mineração de textos é uma extensão da mineração de dados, e pode ser definida como um processo de extração de informações desconhecidas e úteis de documentos textuais escritos em linguagem natural. Como a maioria das informações são armazenadas em forma de texto, a mineração de textos possui alto valor comercial, e pode ser aplicada em áreas como medicina e atendimento ao cliente. (PEZZINI, 2017, p. 1).

Ao se realizar o processamento da Mineração de Textos, deverá ser utilizada a Análise Semântica ou a Análise Estatística. A primeira tem como objetivo a análise da funcionalidade dos termos encontrados no texto, conforme uma pessoa que executa a leitura, contextualiza e entende o significado semântico[3], sintático[4], morfológico[5] e pragmático[6] da palavra. Na estatística, o objetivo é detectar a frequência de cada termo no texto, não tendo relevância seu contexto ou significado. As duas abordagens podem ser utilizadas de maneira individual ou combinadas entre si. (CARRILHO JUNIOR, 2007).

A análise semântica leva em consideração a relevância de cada palavras e avalia a ordem em que os termos do texto estão colocados. Utiliza as técnicas de Processamento de Linguagem Natural como fundamento. Analisa a forma, inflexões das palavras, o contexto e como a sua interpretação pode demonstrar resultados diferentes. Nessa análise são considerados o conhecimento, a interpretação do texto e o que ele representa. (CORDEIRO, 2005).

---

[3] Semântico: Conhecimento do significado das palavras, independente do contexto. Também designa outros significados mais complexos, podem ser obtidos pela combinação destas palavras.

[4] Sintático: É o conhecimento estrutural das listas de palavras e de como elas podem ser combinadas para produzir sentenças.

[5] Morfológico: É o conhecimento da estrutura, da forma e das inflexões das palavras.

[6] Pragmático: É o conhecimento do uso da língua em diferentes contextos e como estes afetam seu significado e a interpretação.

Por outro lado, na análise estatística o texto é representado em um formato de blocos de informação e não importa com a forma com a qual está disposto, mas sim apenas a raridade dos termos e sua constância. (SOARES, 2016).

A literatura que trata sobre mineração de textos normalmente a divide em fases ou etapas, mas não há um consenso definido sobre a quantidade e quais são essas fases. Como podemos observar na Tabela 1, essas etapas propostas em cada modelo são semelhantes. Em alguns casos as fases são agrupadas de maneira mais genérica e, em outros casos, as fases são expandidas de forma mais detalhada. Foram escolhidos os modelos de: Ebecken, Lopes e Costa (2003), que propõem duas definições de etapas de mineração de textos, um simplificado com três etapas e outro detalhado com oito; Corrêa, Marcacini e Rezende (2012) propõem cinco etapas; Carvalho (2017) detalha em oito fases; e Aranha, Vellasco e Passos (2007) apresentam cinco etapas.

Tabela 1 - Modelos de fases da Mineração de Textos.

<table>
<tr><th>Ebecken, Lopes e Costa (2003) Simplificado</th><th>Aranha, Vellasco e Passos (2007)</th><th>Corrêa, Marcacini e Rezende (2012)</th><th>Ebecken, Lopes e Costa (2003) Detalhado</th><th>Carvalho (2017)</th></tr>
<tr><td rowspan="2"></td><td rowspan="2">Extração</td><td rowspan="2">Identificação do problema</td><td rowspan="2">Seleção dos textos</td><td>Base de dados não estruturada</td></tr>
<tr><td>Seleção dos termos da anamnese</td></tr>
<tr><td rowspan="2">Pré-processamento dos documentos</td><td rowspan="2">Pré-processamento</td><td rowspan="2">Pré-processamento</td><td rowspan="2">Preparação dos dados</td><td>Tokenization</td></tr>
<tr><td>Remoção de stopwords</td></tr>
<tr><td rowspan="3">Extração de padrões com agrupamento de textos</td><td>Transformação</td><td rowspan="3">Extração de padrões</td><td>Indexação e normalização</td><td>Normalização</td></tr>
<tr><td rowspan="2">Mineração</td><td>Cálculo da relevância dos termos</td><td>Relevância dos termos</td></tr>
<tr><td>Seleção dos termos</td><td>Seleção dos termos</td></tr>
<tr><td>Avaliação dos resultados</td><td>Análise</td><td>Pós-processamento</td><td>Pós-processamento ou análise de resultados</td><td>Apresentação dos resultados</td></tr>
<tr><td></td><td></td><td>Uso do conhecimento</td><td></td><td></td></tr>
</table>

Fonte: Elaborado pelo autor.

Como apresentado na Tabela 1, apesar de dividirem ou agruparem de diversas maneiras e se referirem de maneira distinta as suas várias etapas, essencialmente, o conjunto de etapas têm funções bastante semelhantes. Mas, nessa dissertação, utilizamos a proposta de Aranha, Vellasco e Passos (2007), que entendemos ser a que exemplifica de maneira mais equilibrada as etapas da Mineração de Texto. Conforme a Figura 4, este modelo proposto é dividido nessas cinco etapas:

Figura 4 - Etapas da mineração de textos.

Fonte: Elaborado pelo autor baseado em Aranha, Velasco e Passos (2007).

Em suma, essas cinco etapas propostas podem ser assim definidas como:

a) Extração (Coleta) – É a coleta de dados, que irá criar uma base de dados de documentos, também chamada de Corpus ou Corpora. Essa fase exige um esforço considerável para que seja possível ter uma boa qualidade nos documentos que irão compor a base de forma que permita a obtenção de conhecimento ao fim do processo. Maiores detalhes na Seção 2.1.2.1.
b) Pré-processamento – Tem como objetivo definir a forma como a massa de textos será apresentada. A preparação dos dados, criando um primeiro nível de estruturação e utiliza a aplicação de algoritmos com técnicas de Processamento de Linguagem Natural (PLN). A Seção 2.1.2.2 explica melhor essa etapa.
c) Transformação – Criação de índices para acesso mais rápido na recuperação de dados e utiliza técnicas de Recuperação de Informação (RI). Esse processo organiza os diversos termos obtidos na base de dados de documentos e, dessa forma, facilita o acesso e recuperação dos mesmos. Da mesma forma que índice de livro, quanto melhor estruturado estiver o índice, o processo de recuperação será mais ágil. Podemos estender essa etapa de uma melhor maneira na Seção 2.1.2.3.
d) Mineração – Aquisição do conhecimento através de cálculos, inferências e extração de conhecimento utilizando técnicas de Descoberta do Conhecimento (DC). Essa fase utilizará inferências, cálculos e algoritmos para conseguir extrair conhecimento e descoberta de comportamentos e padrões. Esse processo só será possível após disponibilizar a estrutura de dados textuais e um acesso ágil de recuperação. Na Seção 2.1.2.4 explora detalhadamente essa etapa.

e) Análise – Essa última etapa é realizada por pessoas, que têm interesse no conhecimento extraído, através da leitura e interpretação dos resultados obtidos e permite a tomada de decisões. Finalmente, a Seção 2.1.2.5 demonstra essa etapa especificamente.

2.1.2.1 Extração

A primeira fase, extração, também chamada de coleta, tem como objetivo buscar e recuperar de dados para construir a base de textos, na qual será realizada a extração de conhecimento (CARRILHO JUNIOR, 2007). Essa coleta de documentos tem como objetivo a obtenção de documentos que tenham alguma relevância com o conhecimento que se pretende conseguir. Dentre os diversos tipos de textos que podem ser utilizados, é possível utilizar livros, e-mails, páginas de sites, fóruns de internet, blogs, arquivos de texto em diversos formatos, etc. Técnicas como o Processamento de Linguagem Natural e Recuperação de Informação são utilizadas nessa fase. (MARTINS et al., 2003).

A extração de informação de texto é feita em dados dispostos em textos semiestruturados ou não estruturados. Os dados semiestruturados possuem algum nível de estrutura, como um modelo de currículo pré-definido e alguns tipos de formulários. Os não estruturados, geralmente estão em um formato de texto livre, não possuindo um padrão de formatação. Apesar da estrutura que o texto possui, essa se trata somente do padrão que é definido nas regras linguísticas e que têm por objetivo tornar o texto compreensível ao ser humano, mas não aos computadores. Enfim, os dados estruturados que já possuem regras para seu armazenamento e disponibilização, como os bancos de dados relacionais. (FORTE, 2015).

Um dos maiores desafios da mineração de textos é a descoberta da localização de armazenamento dos dados para realizar sua coleta. Após serem localizados, os documentos que realmente são importantes para a busca de conhecimento devem ser recuperados. Essa descoberta é realizada basicamente nos seguintes ambientes diversos, na estrutura de arquivos de um disco rígido, em um banco de dados com sua estrutura de tabelas ou no ambiente da Internet. Nos dois primeiros casos, a base textual está disponibilizada de maneira mais simples, tendo sua estrutura estática. Já no caso da Internet, sua disponibilidade é dinâmica, devendo-se utilizar programas chamados de crawler ou webcrawler, que são robôs que navegam automaticamente explorando a Internet com seus infindáveis sites, sejam de páginas

institucionais, redes sociais, repositórios acadêmicos, jornais e revistas, entre outros (ARANHA; VELLASCO; PASSOS, 2007).

2.1.2.2 Pré-processamento

O pré-processamento de textos baseia-se em realizar, sobre a base textual coletada, uma série de processos de transformações, com a finalidade de que essa base não estruturada, ao final dessa etapa possua uma estrutura definida com uma representação atributo-valor. Essa etapa também tem a função de organizar essa base textual e dar-lhe uma melhor qualidade, com o objetivo de deixá-la pronta para seja possível aplicar algoritmos nas próximas etapas, de indexação e mineração. As transformações aplicadas consistem em individualizar as palavras da base textual, identifica-las, realizar stemização, excluir as stop-words, classificar de acordo com a classe gramatical, compactar e tratar os dados com informações corrompidas, desconhecidas e irrelevantes. (ARANHA; VELLASCO; PASSOS, 2007).

Nesta fase, o resultado obtido na coleta ou extração, disponibilizados em linguagem natural, têm um tratamento onde é aplicada uma formatação com o objetivo de padronizar o texto com uma estrutura. Isto é feito de forma com que não perca suas características naturais. Após essa fase, o resultado será uma estrutura representativa dos textos processados, que geralmente é apresentado como uma tabela atributo-valor (MARTINS et al., 2003).

Corrêa, Marcacini e Rezende (2012) fazem essas relevantes observações sobre o pré-processamento:

> Na etapa de pré-processamento se encontra a principal diferença entre os processos de mineração de textos e processos de mineração de dados: a estruturação dos textos em um formato adequado para a extração de conhecimento. Muitos autores consideram essa etapa a que mais tempo consome durante todo o ciclo do processo de mineração de textos. O objetivo do pré-processamento é extrair de textos escritos em língua natural, inerentemente não estruturados, uma representação estruturada, concisa e manipulável por algoritmos de agrupamento de documentos.
> Para tal, são executadas atividades de tratamento e padronização da coleção de textos, seleção dos termos (palavras) mais significativos e, por fim, representação da coleção textual em um formato estruturado que preserve as características necessárias aos objetivos definidos na etapa de identificação do problema.

> Os documentos da coleção podem estar em diferentes formatos, uma vez que existem diversos aplicativos para apoiar a geração e publicação de textos eletrônicos.
> Dependendo de como os documentos foram armazenados ou gerados, há a necessidade de padronizar as formas em que se encontram. Na padronização dos textos, geralmente, os documentos são convertidos para a forma de texto plano sem formatação.
> Um dos maiores desafios do processo de mineração de textos é a alta dimensionalidade dos dados. Uma pequena coleção de textos pode facilmente conter milhares de termos, muitos deles redundantes e desnecessários, que tornam lento o processo de extração de conhecimento e prejudicam a qualidade dos resultados. (CORRÊA; MARCACINI; REZENDE, 2012, p. 5–6).

Para enfrentar o desafio de se conseguir um subconjunto pequeno, mas representativo, dos termos que constam na coleção textual, utiliza-se a seleção de termos. Primeiro, são excluídas as stopwords, que se tratam de termos sem relevância e que não acrescentam significado à pesquisa, pronomes, artigos e advérbios, por exemplo. Essa redução da quantidade de termos diminui o custo computacional nas fases restantes. Após isso, variações morfológicas e termos sinônimos são identificados, através de técnicas como stemização e tesauro, possibilitando a diminuição do conjunto pesquisado ainda mais. Há também a possibilidade de pesquisar por termos compostos, também chamados de n-gramas, termos que possuem mais de um elemento, mas têm apenas um significado semântico. (MANNING; RAGHAVAN; SCHUTZE, 2008).

2.1.2.3 Transformação

Nesta fase, também chamada de indexação, o resultado obtido na coleta ou extração deverá ser indexado ou catalogado. Os documentos adicionados à base textual devem ter suas informações analisadas, identificadas e catalogadas, permitindo sua recuperação. Esse índice deve existir para que possa o sistema possa processar a busca de maneira rápida (WIVES, 2002).

Também podemos entender um índice como um tipo filtro, que permita a seleção apenas dos documentos que realmente são importantes, enquanto os que forem irrelevantes não serão incluídos no resultado (LANCASTER, 1968).

Algoritmos de mineração de textos usam técnicas eficientes de indexação para buscas em bases textuais. Essas técnicas possibilitam buscas extremamente rápidas por palavra-chave, mesmo quando se trata de grandes volumes de textos. Esses algoritmos realizam sofisticados cálculos estatísticos trazendo um evidente ganho de performance para a busca. Mas quando

tratamos de mineração de textos, estamos tratando de um conceito mais abrangente, que permita uma busca mais eficiente por palavras-chave. A busca por palavra-chave na Internet, por exemplo, resulta numa série de páginas, nas quais os termos pesquisados existem, entretanto, não considera as características semânticas, retornando, entre o resultado, ocorrências que são irrelevantes para a pesquisa desejada. Já as técnicas de mineração de textos realizam uma análise mais profunda dos conteúdos contidos na base textual, possibilitando que fatos, relações e padrões sejam identificados da mesma forma (ou o mais próximo possível) que um ser humano perceberia ao ler esses documentos. Com a utilização dessas técnicas, em geral, a informação obtida possui maior relevância, permitindo que esta seja utilizada em diversos outros objetivos, e não apenas na simples localização do termo. Isso permite, por exemplo, que um documento seja categorizado pelo seu contexto e que o grupo semântico dos termos de documento tenha seu significado identificado (ARANHA; VELLASCO; PASSOS, 2007).

Os textos que compõem uma base de dados textual têm os seus termos distribuídos em formato de linguagem natural, não tabulados. Neste caso, não há uma definição anterior do termo e, para o computador, esse termo não passa de uma sequência de caracteres sem significado. Só será possível saber o seu real significado utilizando técnicas de análise de Linguagem Natural (ARANHA; VELLASCO; PASSOS, 2007).

Também é nesta fase que se tenta encontrar a similaridade entre as palavras, através da sua morfologia e significado. Para não comprometer a eficiência desejada na busca, os termos nas consultas devem ter uma boa definição e estrutura, fazendo com que os termos mais adequados sejam localizados pela indexação. A identificação da relação entre os termos de consulta e os que realmente existem no texto é possível utilizando-se técnicas de Análise de Relevância, bem como a função de similaridade. Esta comparação é realizada de forma direta, sendo assim, ambiguidade, sinônimos e polissemia[7] não são levados em consideração, comprometendo o resultado da busca, quando esses casos ocorrerem (SOARES, 2016).

A similaridade dos termos a serem buscados na base textual é calculada através de técnicas de Recuperação da Informação que utilizam diversos modelos conceituais de recuperação: Modelo Booleano, Modelo de Espaço-Vetorial, Modelo Probabilístico, Modelo

[7] Polissemia - Representa a multiplicidade de significados de uma palavra. Do grego polis, significa "muitos", enquanto sema refere-se ao "significado". Portanto, um termo polissêmico é aquele que pode apresentar significados distintos de acordo com o contexto. Apesar de terem a mesma etimologia e se relacionarem em termos de ideia.

Difuso, Modelo de Busca Direta, Modelo de Aglomerados, Modelo Lógico, Modelo Contextual (WIVES, 2002). Dependendo do resultado desejado, deve-se utilizar um ou mais modelos. Segue uma breve definição de cada modelo conceitual de recuperação:

a) Modelo Conceitual ou Textual – Julga que os termos estão presentes no documento e realiza a associação de termos para verificar como estão contextualizados no texto, bem como o contexto da busca realizada. Mas, mesmo sendo eficiente, a busca por similaridade nos documentos associando termos tem o risco de se obter respostas equivocadas ao fugir do contexto;
b) Modelo Booleano – Os documentos são vistos como conjuntos de palavras que são manipulados e descritos utilizando conectivos booleanos (and, or e not). Que permitem unir, retirar ou fazer a intersecção de partes. Numa consulta, os termos informados comparados com os documentos pesquisados e vistos como dois conjuntos. Os documentos que possuem intersecção com o conjunto de termos serão apresentados;
c) Modelo de Espaço-Vetorial – Representa os documentos como sendo um vetor de termos e define um grau de importância para cada termo. O cálculo do grau de importância ou peso pode ser obtido de diferentes maneiras. Entretanto, o que mais se utiliza é ter pesos maiores para os termos ocorrem mais vezes no documento;
d) Modelo Probabilístico – Através de conceitos de probabilidade e estatística, identifica a probabilidade de um documento ser relevante diante dos termos definidos na consulta. Neste modelo é utilizado o Método Bayesiano, por isso também é conhecido como Modelo Bayesiano;
e) Modelo Difuso – Como no Modelo de Espaço-Vetorial, representa os documentos como vetores de termos com graus. Aqui, no entanto, se trabalha com a teoria de conjuntos difusos, onde a presença de um termo pode não ser exata. Dessa forma, não existe conjunto vazio, mas conjuntos com termos com relevância muito baixa;
f) Modelo de Busca Direta - Esse modelo localiza strings no documento e apresenta a localização de todas as ocorrências do que foi pesquisado. Mas é recomendado que seja usado em uma quantidade pequena de documentos. Também é denominado de Modelo de Busca de Padrões;
g) Modelo de Aglomerados – Com técnicas de agrupamento de documentos, tem como objetivo a identificação de documentos com conteúdo similar e proceder seu

armazenamento ou indexação. A quantidade de palavras similares e frequentes contidas no documento, é que define a similaridade. Ao ser definida uma consulta pelo usuário, um documento relevante é identificado e todos os outros documentos do mesmo grupo são retornados juntos. Também é chamado de Clustering Model.

h) Modelo Lógico – Utiliza a lógica matemática para modelar o processo de recuperação de documentos. Os documentos são modelados através de lógica predicativa, incorpora semântica ao processo de recuperação, e seu trabalho para modelar exige um grande esforço. Dessa forma, o conteúdo dos documentos é conhecido pelo sistema, o que permite um melhor julgamento da sua relevância. (SOARES, 2016).

2.1.2.4 Mineração

A quarta etapa chamada de mineração de dados, ou simplesmente mineração, trata da escolha dos algoritmos que serão utilizados na busca sobre a massa de dados trabalhada até então. Diversas áreas de conhecimento como Banco de Dados, Estatística, Aprendizado de Máquina e Redes Neurais fornecem os algoritmos. Conhecer o funcionamento de cada algoritmo é essencial para que a escolha seja a mais correta para se obter o resultado desejado, pois nenhum algoritmo funcionará de maneira ótima em todas as aplicações de mineração de textos (ARANHA; VELLASCO; PASSOS, 2007).

Carrilho Junior (2007) discorre e exemplifica essa etapa considerando a importância da correta escolha do algoritmo a ser utilizado:

> A fase de Mineração envolve decidir quais algoritmos deverão ser aplicados sobre a massa de dados desenvolvida até o momento. Para tanto, deve se optar por uma ou mais Tarefas de Mineração, que nada mais é do que decidir o que se quer obter de informação. Por exemplo, se a necessidade de informação do usuário é obter o relacionamento entre documentos, verificando o grau de similaridade e a formação de grupos naturais, então a tarefa a ser escolhida é a clusterização. Em contrapartida, se estes grupos de documentos já existem, seja pela execução de algoritmos ou pelo conhecimento prévio de especialistas, então a indicação de onde um novo documento deve ser "encaixado" é conseguida através de algoritmos de classificação.
> Embora as tarefas de clusterização e classificação sejam compartilhadas entre Mineração de Textos e Mineração de Dados, outras são específicas da primeira, como a sumarização e extração de características. No próximo

> capítulo, são exploradas todas as Tarefas de Mineração de Textos, assim como a relação dos principais algoritmos. (CARRILHO JUNIOR, 2007, p. 56).

2.1.2.5 Análise

Esta última fase, que acontece após a mineração de dados, é realizada pelo analista de dados, que vai avaliar e interpretar os resultados obtidos, verificando taxa de erro, tempo de processamento e a complexidade do processo. Depois, o resultado é avaliado pelo especialista do domínio para determinar se é compatível com o que se conhece sobre o domínio. Finalmente, o usuário a quem se destinou a mineração, avalia a utilidade e aplicação desses resultados. (ARANHA; VELLASCO; PASSOS, 2007).

2.1.3 Processamento de Linguagem Natural

A diferença crucial entre a mineração de dados e a mineração de textos é que esta última trata de dados em linguagem natural, não estruturados, enquanto a primeira vai tratar, de maneira exclusiva, de dados estruturados. (REZENDE; MARCACINI; MOURA, 2011; SOARES, 2008). Não é possível aos sistemas de mineração de textos, processar textos não estruturados simplesmente utilizando os algoritmos tradicionais de descoberta de conhecimento (ARANHA; VELLASCO; PASSOS, 2007; GOMES, 2009). Por isso, é necessário empregar técnicas de Processamento de Linguagem Natural (PLN) de maneira ampla, de forma que os dados textuais sejam preparados, permitindo a busca de conhecimento de algum tipo (SANTOS et al., 2014).

Mesmo quando são utilizadas as principais ferramentas para extração de informação, ainda existe uma limitação na automatização da busca por informações. A estrutura textual constituída por uma gramática (conjunto de regras que regem o uso da língua) e por um léxico (conjunto de palavras existente em um idioma para as pessoas expressarem-se, oralmente ou por escrito) constitui uma alta complexidade e um importante desafio para a evolução dessa tecnologia (ALLES, 2018).

O Processamento de Linguagem Natural (PLN) é a tecnologia utilizada para que o computador possa, através da compreensão da estrutura gramatical, entender qual o significado contido em um texto. PLN é uma área da Inteligência Artificial (IA) e um campo de estudo da automatização computacional. Através do entendimento e organização gramatical de linguagem não estruturada, é utilizada em diversas aplicações, entre as quais se destacam

reconhecimento de fala, processamento e sintetização de textos em linguagem natural, tradução automática e extração de significado (FINATTO; LOPES; SILVA, 2015).

Através do PLN, é possível obter informação através do processamento e interpretação de conteúdo em linguagem natural. Pois o PLN também é definido como o estudo de algoritmos e métodos que permitem que modelos computacionais sejam construídos e tenham a capacidade de analisar e compreender textos expressos na língua natural em variados idiomas (MANNING et al., 2014).

Liddy (2001) diz que PLN é um conjunto de técnicas computacionais para analisar e representar textos que ocorrem naturalmente em um ou mais níveis de análise linguística com a finalidade de obter processamento de linguagem semelhante ao humano para uma variedade de tarefas ou aplicações. E Gonzalez e Lima (2003) definem que PLN, em sentido amplo, tem a função de realizar a comunicação entre o computador e a linguagem humana tratando, no nível computacional, os seus diversos aspectos, como sons, palavras, sentenças e discursos, sendo considerados os seus formatos e referências, estruturas e significados, contextos e usos. Essa comunicação poderá ocorre nos diversos níveis de entendimento ou geração, fonético, morfológico, sintático, semântico e pragmático.

Abaixo podemos entender melhor as funções das análises morfológica, sintática e semântica, definidas por Santos et al. (2014):

> A análise morfológica é responsável por definir artigos, substantivos, verbos e adjetivos, armazenados em um tipo de dicionário. Depois de construído o dicionário, a análise sintática faz uso dele procurando mostrar relacionamento entre as palavras e, num segundo momento, verifica sujeito, predicado, complementos nominais e verbais, adjuntos e apostos. Na análise semântica, ocorre o encontro de termos ambíguos, de sufixos e afixos, ou seja, questões de significado associados aos morfemas componentes de uma palavra, o sentido real da frase ou palavra. (SANTOS et al., 2014)

2.1.4 Aprendizado de Máquina

O Aprendizado de Máquina é uma área da Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais capazes de adquirir conhecimento de forma automática. Um sistema de aprendizado é um programa de computador que toma decisões baseado em experiências acumuladas por meio de solução bem-sucedida de problemas anteriores (MICHALSKI; CARBONELL; MITCHELL, 2013).

Seu uso é bastante difundido nos processos para classificação automática de textos e é uma área que estuda a criação de modelos probabilísticos que possuem a capacidade de aprender após ser experimentado em situações similares. Utiliza métodos dedutivos para obter esse aprendizado, extraindo padrões de grandes massas de dados (CHAKRABARTI, 2002).

Os algoritmos desenvolvidos na Aprendizado de Máquina são baseados em estatística, probabilidade e métodos de otimização estocástica. Tais algoritmos aplicados à mineração de texto, utiliza o corpus como base de dados para poder aprender os padrões (MITCHELL, 1997).

Müller e Guido (2016), em seu livro Introduction to Machine Learning with Python, nos diz que:

> Aprendizado de Máquina é sobre extrair conhecimento de dados. É um campo de pesquisa na interseção entre estatística, inteligência artificial e ciência da computação, também conhecida como análise preditiva ou aprendizagem estatística. A aplicação de métodos de aprendizado de máquina nos últimos anos tornou-se onipresente na vida cotidiana. Desde as recomendações automáticas de quais filmes assistir, até qual comida pedir ou quais produtos comprar, até a rádio on-line personalizada e o reconhecimento de seus amigos em suas fotos, muitos sites e dispositivos modernos têm algoritmos de aprendizado de máquina em seu núcleo.
> Quando você olha para sites complexos como Facebook, Amazon ou Netflix, é muito provável que cada parte do site que você está vendo contenha vários modelos de aprendizado de máquina.
> Fora das aplicações comerciais, o aprendizado de máquina teve uma tremenda influência na forma como a pesquisa conduzida por dados é feita hoje. As ferramentas apresentadas neste livro foram aplicadas a diversas questões científicas, como a compreensão de estrelas, a descoberta de planetas distantes, a análise de sequências de DNA e o fornecimento de tratamentos personalizados para o câncer (MÜLLER; GUIDO, 2016, p. 9, tradução nossa).

O Aprendizado de Máquina tem mostrado que pode ser aplicado em diversos trabalhos interdisciplinares, sendo bem-sucedido em muitos casos. E poderemos ver que a Mineração de Textos, especificamente a classificação automática de textos, possui uma forte relação com essa área de estudo.

2.1.5 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas (REN) é um dos principais elementos do PLN. O REN é essencial para várias etapas do PLN, dentre elas a classificação de uma frase ou a checagem de vínculos entre as entidades. O uso de REN no PLN é a extração de informação para identificar e classificar os termos relevantes em uma determinada categoria semântica, analisando um conjunto de palavras. Esses termos podem ser considerados entidades nomeadas

e essas podem variar conforme o domínio de interesse. Nomes de pessoas, locais, organizações, data, hora, valores monetários são os exemplos mais comumente utilizados para identificar os termos relevantes como entidades nomeadas (ALLES, 2018; MARQUES, 2017; NADEAU; SEKINE, 2007).

As aplicações dos estudos em REN podem ser feitas em diversas áreas como sites de busca na internet, indexação de documentos, ferramentas e sites de tradução ou extrações de informação mais complexas (MARQUES, 2017). E é possível ser categorizado em três tipos (CHITICARIU et al., 2010):

a) REN utilizando aprendizado de máquina - Que utiliza técnicas de aprendizagem automática, onde são gerados modelos que, conforme o domínio de interesse, permitem realizar a previsão de ocorrências;
b) REN baseado em regras - É realizado na definição de heurísticas na forma de expressões regulares, conforme os termos estão organizados no texto analisado e as diversas características desses termos;
c) REN baseada em abordagem híbrida - Que utiliza os dois tipos anteriores, aprendizagem de máquina e regras.

Os tipos de entidades nomeadas definidas para serem detectadas neste trabalho são as demonstradas na Tabela 2:

Tabela 2 – Tipos de Entidades Nomeadas

Tipo	Sigla	Conteúdo
Pessoa	PER - Person	Nome de pessoas
Organização	ORG - Organization	Empresas e organizações
Local	LOC - Location	Lugar, cidade, estado, país, bairro
Transações Financeiras	MISC - Miscellaneous	Transferência, cheque, DOC, TED

Fonte: Elaborado pelo Autor

Ao aplicar a detecção de entidades nomeadas no texto de exemplo abaixo, trecho de um RIF fictício, vamos obter as entidades listadas na Tabela 3.

> Segundo informado, Raimundo Loudeiro Ribeiro Filho atuaria como representante comercial (autônomo), com renda mensal de R$13.000,00. É sócio das empresas Confecções R Loudeiro Ltda (ramo de confecção de peças

> de vestuário, exceto roupas íntimas e as confeccionadas sob medida), e Sambasul Engenharia e Construções Ltda (ramo de construção de edifícios). A conta seria utilizada para movimentar recursos oriundos das atividades da sua loja de confecção de roupas. Os créditos somaram R$ 1.293.275,38, sendo R$ 603.487,92 por meio de depósitos de cheques realizados em Blumenau/SC, Gaspar/SC, Londrina/PR, Ribeirão Preto/SP e São Paulo, dos quais R$ 30.000,00 efetuados em espécie, e R$ 40.906,18 efetuados em terminais de autoatendimento, e R$ 689.787,46 provenientes de TEDs e transferências entre contas.

Tabela 3 – Entidades Nomeadas encontradas.

Tipo	Conteúdo
PER	Raimundo Loudeiro Ribeiro Filho
ORG	Confecções R Loudeiro Ltda
ORG	Sambasul Engenharia e Construções Ltda
MISC	cheques
LOC	Blumenau/SC
LOC	Gaspar/SC
LOC	Londrina/PR
LOC	Ribeirão Preto/SP
LOC	São Paulo
MISC	TEDs
MISC	transferências

Fonte: Elaborado pelo Autor

O tipo de REN que será abordado por esse trabalho é o que utiliza aprendizado de máquina. Portanto, para esse passo importante no REN são utilizados algoritmos de treinamento que se baseiam num conjunto de textos com marcações de entidades, chamado de corpus, para possibilitar a identificação automática dessas entidades. Será construído um corpus próprio a partir diversos RIFs anotados, com o qual serão realizados os testes. Mas também serão feitos testes com um corpus público, para compararmos os resultados.

Neste contexto, este trabalho fez uma análise da literatura sobre ferramentas e técnicas utilizadas para o REN que poderiam ser utilizadas em textos de inteligência financeira (RIF) escritos em português e compartilhados pelo COAF. A primeira etapa desse processo é a identificação das ferramentas e técnicas utilizadas, através de uma pesquisa na literatura, sobre a aplicação do REN em língua portuguesa. A segunda etapa visa a escolha e utilização das técnicas e ferramentas, identificadas na primeira etapa, para a aplicação do REN em relatórios de inteligência financeira (RIF). É fundamental, também encontrarmos na literatura, os principais conjuntos de textos anotados (corpora) em português.

2.2 TRABALHOS RELACIONADOS

2.2.1 Trabalhos Encontrados em Pesquisas Fora da RSL

O foco principal deste trabalho é o estudo e aplicação do reconhecimento de entidades nomeadas (REN) em textos em língua portuguesa para aplicação em Relatórios de Inteligência Financeira. Nesse sentido, essa seção apresenta os resultados da Revisão Sistemática da Literatura em busca de trabalhos sobre a aplicação de REN em texto de língua inglesa e Portuguesa.

Alles, Giozza e Albuquerque (2018) formularam o trabalho: "Processamento de Linguagem Natural para Classificação de Entidades Nomeadas no Diário Oficial da União Brasileiro", que tem como objetivo deste trabalho avaliar qual a melhor ferramenta de REN no contexto das informações não estruturadas do Diário Oficial da União - DOU. As entidades pesquisadas foram Artigo, Cargo, Data, Evento, Lei, Lugar, Número, Organização, Pessoa, Processo e Valor Monetário. E as ferramentas avaliadas foram OpenNLP, Stanford CoreNLP, NLTK e Tensorflow Syntaxnet. No trabalho foi utilizado o corpus Amazônia[8] para o treinamento e teste, mas também construiu um corpus próprio a partir de textos do DOU. Verificou que as ferramentas OpenNLP e CoreNLP realizaram uma extração de entidades nomeadas, enquanto NLTK e Syntaxnet fizeram extração de classificações morfossintáticas.

Avaliando que a ferramenta OpenNLP teve o melhor resultado para o objetivo do trabalho, foi feita então uma comparação entre o corpus Amazônia e o corpus DOU utilizando a OpenNLP. O resultado da comparação não é muito claro, pois como corpus Amazônia extraiu-se 1.852.036 tokens como categoria "Pessoa" com uma acurácia de 75,87%. Já com o corpus DOU, extraiu-se apenas 19.298 da mesma categoria, com uma acurácia de 90,67%. Entretanto, os autores afirmam que na utilização da ferramenta OpenNLP em conjunto com o corpus DOU observou-se o melhor resultado na extração de entidades tanto em relação à quantidade quanto à qualidade. Este trabalho foi encontrado e selecionado na RSL.

[8] https://www.linguateca.pt/Floresta/ - Contém 4.6 milhões de palavras (cerca de 275 mil frases) retiradas do sítio colaborativo Overmundo, um coletivo virtual que tem como objetivo expressar a produção cultural brasileira.

O trabalho “Comparative Analysis of Portuguese Named Entities Recognition Tools”, de Amaral et al. (2014), realiza um comparativo entre quatro ferramentas para REN em português, FreeLing, LanguageTasks, Palavras e NERP-CRF. O experimento foi realizado sobre o corpus HAREM[9]. As ferramentas experimentadas são baseadas em técnicas de processamento de linguagem natural e também em aprendizado de máquina. A ferramenta NERP-CRF é baseada em campos aleatórios condicionais, um modelo de aprendizado de máquina não supervisionado que está sendo usado para reconhecimento de entidades nomeadas em vários idiomas, enquanto as outras ferramentas seguem abordagens de linguagem natural mais tradicionais.

Os testes consideraram as entidades pessoa, local e organização. Os resultados dos testes das quatro ferramentas foram bastante semelhantes. Tiveram um valor de medida F respectivamente de 53% para NERP-CRF, 55% para LanguageTasks, 54% para FreeLing e 57% para PALAVRAS. Também foi encontrado e selecionado na RSL.

Silva e Caseli (2015), no trabalho “Reconhecimento de Entidades Nomeadas em Textos em Português do Brasil no Domínio do e-Commerce”, têm como foco as informações não estruturadas disponibilizadas nos sites de e-commerce. Utiliza basicamente como conteúdo do seu corpus as descrições dos produtos disponibilizados para serem comercializados. E tem como objetivo reconhecer as entidades: Modelo, Marca, Dimensão, Grandeza, Cor, Utilidade, Material, Parte, Objeto e Produto. Utiliza técnicas de Aprendizado de Máquina e CRF - Conditional Random Fields, modelo probabilístico que etiqueta e segmenta dados sequenciais implementado pela ferramenta CRFSharp, que avalia cada palavra do corpus e gera o modelo de treinamento. Também foi utilizada a ferramenta BRAT para anotação das entidades que constam no corpus.

Como resultado, a revocação e precisão variaram bastante dependendo da entidade, sendo o pior resultado para a entidade “Objeto” com 17,86% e 31,25% respectivamente, e o melhor da entidade “Marca” com 100% em ambos. A média de revocação e precisão, considerando todas as entidades, foi de 87,59% nos dois casos. Não foi encontrado pela pesquisa da RSL, pois não estava indexado nos repositórios pesquisados.

Corbett e Boyle (2018) propõem o trabalho “Chemlistem: Chemical Named entity Recognition Using Recurrent Neural Networks” que faz uma avaliação da utilização de técnicas

[9] https://www.linguateca.pt/HAREM/ - O HAREM é uma avaliação conjunta na área do reconhecimento de entidades mencionadas em português. Muito simplificadamente, é uma iniciativa que pretende avaliar o sucesso na identificação e consequente classificação automática dos nomes próprios na língua portuguesa.

de redes neurais artificiais conhecida como “deep learning”, aprendizagem profunda, como alternativa ao CRF - Conditional Random Fields. Essa técnica não faz a tokenização das palavras e sim rotula a sequência de caracteres. Isso se mostrou mais adequado a textos da área química. O estudo apresenta vários sistemas de REN na área química. O primeiro sistema traduz os tradicionais idiomas baseados em CRF em uma estrutura de aprendizagem profunda, usando recursos ricos em token e incorporações de palavras neurais, e produzindo uma sequência de tags usando redes bidirecionais de memória de curto prazo (LSTM) - um tipo de rede neural recorrente.

O segundo sistema evita o conjunto de recursos - e até mesmo a tokenização - em favor da rotulagem de caracteres, usando encartes de caracteres neurais e várias camadas LSTM. O terceiro sistema é um conjunto que combina os resultados dos dois primeiros sistemas implementado com o nome de ChemListem e utiliza a ferramenta Oscar4 Tokeniser. Os resultados obtidos são de precisão de 91,47%, revocação de 89,21% e medida-F de 90,33%. Analisa textos em inglês e possui um escopo muito específico da área química. Também não estava indexado nos repositórios pesquisados, mas, ainda que estivesse, seria rejeitado por não trabalhar com textos em português.

O trabalho “Identificação de Termos Relevantes em Relatórios Usando Text Mining”, de Bastos (2017), apresenta um sistema capaz de extrair automaticamente condições clínicas, descrições de condições clínicas e zonas de incidência das condições clínicas em relatórios da tiróide. Identifica ocorrências relevantes, entidades e extrai automaticamente os termos relevantes. O trabalho possui 8 fases: pré-processamento de texto clínico, identificação de possíveis ocorrências relevantes, identificação de frases relevantes, part-of-speech-tagging e lematização de frases relevantes, identificação de entidades, construção de sequências das entidades identificadas, a associação das entidades às ocorrências relevantes e a extração de termos relevantes. 1.690 relatórios da tiróide foram analisados, sendo 200 escolhidos aleatoriamente e anotados para avaliar a eficácia do sistema. Os resultados obtidos apontam para uma acurácia de 98.9% na extração de condições clínicas, 98.5% para a extração das zonas de incidência das condições clínicas e uma medida-F de 97.8% para a extração de descrições de condições clínicas. O sistema indica que 95.4% dos termos relevantes são corretamente extraídos. Foi encontrado e selecionado na RSL.

A proposta de Zhu et al. (2017), “GRAM-CNN: A Deep Learning Approach with Local Context for Named Entity Recognition in Biomedical Text” propõe uma nova abordagem de aprendizado profundo de ponta a ponta para tarefas de REN biomédicas que aproveitam os contextos dos termos com base em n-grama e incorporação de palavras via Convolutional Neural Network (CNN). Chamam essa abordagem de GRAM-CNN. Para rotular automaticamente uma palavra, esse método usa as informações locais em torno de uma palavra. Portanto, o método GRAM-CNN não requer nenhum conhecimento específico ou engenharia de recursos e pode ser aplicado teoricamente a uma ampla gama de problemas de REN existentes.

Utilizou técnicas de CRF - Conditional Random Fields, camada que foi implementada com a ferramenta TensorFlow. E, para fazer a tokenização, utilizou a ferramenta NLTK. Foram avaliados três corpora da área de biomedicina, Biocreative II, NCBI e JNLPBA. Em comparação com outros métodos de REN da área biomédica, o GRAM-CNN obteve os melhores resultados tendo precisão de 90,41%, revocação de 81,32% e acurácia de 87,26%. Além de trabalhar com textos em inglês, o escopo deste trabalho é extremamente específico para os dados de textos da área de biomedicina, a qual possui muitas entidades com nomes compostos longos, com mais de três palavras. Não estava indexado nos repositórios pesquisados, mas, ainda que estivesse, seria rejeitado por não trabalhar com textos em português.

2.3 MÉTODO UTILIZADO

A estrutura deste trabalho se baseia na proposta de revisão sistemática da literatura de Kitchenham (2004). O objetivo deste estudo é localizar evidências empíricas sobre o reconhecimento de entidades nomeadas em textos da língua portuguesa. A questão de pesquisa é definida da seguinte forma:

> RQ1. Quais estudos apresentam implementações práticas de reconhecimento de entidades nomeadas em textos em português com resultados objetivos, como precisão, revocação, medida-F ou acurácia no REN?

A partir desta questão como fundamento da pesquisa, outras questões podem ser formuladas, às quais podemos definir como questões norteadoras, que terão o objetivo de

conduzir e auxiliar na extração e compilação dos dados das publicações encontradas nesta RSL, que são as seguintes:

RQ2. Quais as principais ferramentas de REN utilizadas nos estudos selecionados?
RQ3. Quais os principais corpora de dados utilizados nos estudos selecionados?

2.3.1 Mecanismo e String de Busca

Nessa RSL, foram analisados os artigos publicados no ACM Digital Library, Capes, El Compendex, IEEE Digital Library, Repositório UFSC, Repositório UP e Scopus, com a intenção de se obter a resposta à questão de pesquisa definida nesta seção. Foram utilizados os próprios mecanismos de busca desses repositórios.

Foi criada uma *string* formada por palavras-chave e por operadores lógicos E (AND) e OU (OR) para executar a busca, aqui apresentada:

> (("named entity" OR "named entities") AND ("extract" OR "classification" OR "classify" OR "classifying" OR "extraction" OR "recognition" OR "recognize") AND ("accuracy" OR "f-measure" OR "precision" OR "f1-measure" OR "recall" OR "supervised learning" OR "natural language processing") AND ("Portuguese" OR "Brazil" OR "Brasil" OR "Portugal")) OR (("entidade nomeada" OR "entidades nomeadas") AND ("extrair" OR "classificação" OR "classificar" OR "extração" OR "reconhecimento" OR "reconhecer") AND ("acurácia" OR "precisão" OR "f-measure" OR "f1-measure" OR "recall" OR "aprendizagem supervisionada" OR "processamento de linguagem natural") AND ("Portuguese" OR "Portugues" OR "Português" OR "Brazil" OR "Brasil" OR "Portugal"))

2.3.2 Critérios de Inclusão e Exclusão

Foi adotada uma abordagem de quatro etapas para a busca. Na primeira etapa, foi feita a busca usando a *string* nos mecanismos de buscas dos repositórios, obtendo os seguintes resultados iniciais:

Tabela 4 - Repositórios pesquisados

Repositório	**Quantidade**
ACM Digital Library	23
Capes	84
El Compendex	157
IEEE Digital Library	14
Repositório UFSC	18
Repositório UP	67
Scopus	481
Total	**844**

Fonte: Elaborado pelo Autor

Na segunda etapa, foram estabelecidos os critérios de inclusão (CI), Tabela 5, e os critérios de exclusão (CE), Tabela 6. Foram lidos os títulos, resumos e palavras-chave de cada uma das publicações. Após a leitura, 139 publicações foram identificadas como duplicadas, 679 rejeitadas pela aplicação dos critérios CI e CE, restando um total de 26 publicações.

Tabela 5 - Critérios de Inclusão

Critérios de Inclusão	
CI1	Os estudos devem ser relacionados ao reconhecimento de entidades nomeadas em português.
CI2	Os estudos devem apresentar resultados como acurácia, precisão, revocação (recall) ou medida-F (f-measure).
CI3	Os estudos devem apresentar seu corpus de texto.

Fonte: Elaborado pelo Autor

Tabela 6 - Critérios de Exclusão

Critérios de Exclusão	
CE1	Os estudos não são relacionados ao reconhecimento de entidades nomeadas.
CE2	Os estudos não relacionados ao REN em português.
CE3	Estudos sem corpus de texto.
CE4	Estudos que não apresentam resultados.
CE5	Estudos em andamento ou indisponíveis.
CE6	Estudos duplicados.

Fonte: Elaborado pelo Autor

Na terceira etapa, foram lidas as 26 publicações restantes. A seguir, a RSL prossegue na extração dos dados de qualidade dos artigos.

2.4 ANÁLISE DA REVISÃO SISTEMÁTICA DA LITERATURA

Aplicados devidamente os critérios de inclusão e exclusão, Kitchenham (2004) indica, como próximo passo, a importância de verificação da qualidade dos artigos selecionados e reduzir a amplitude do resultado da pesquisa. Apresentaremos, nesta seção, os critérios de qualidade necessários para este estudo, avaliando se estes critérios estão presentes ou não nas publicações avaliadas.

2.4.1 Qualidade dos Artigos Selecionados

É importante realizar uma avaliação da qualidade dos artigos selecionados inicialmente, para que se tenha mais clareza da intenção dos critérios de inclusão e exclusão adotados. Assim, é possível explicar se as diferenças na qualidade dos estudos refletem em diferenças nos seus resultados, além de definir a importância de cada trabalho através de pesos, após estarem consolidados. E poderá servir como um direcionamento para uma melhor interpretação dos resultados. Apesar de não existir um consenso, quando tratamos de qualidade, com ela a validade interna e externa da RSL aumenta e o viés da pesquisa diminui. (KITCHENHAM, 2004).

A definição dos critérios de qualidade é dividida considerando-se seis tópicos que, de acordo com as orientações de Kitchenham et al. (2002), devem estar presentes em pesquisas empíricas:

a) Contexto do experimento;
b) Planejamento do experimento;
c) Condução do experimento e coleta dos dados;
d) Análise;
e) Apresentação dos resultados;
f) Interpretação dos resultados.

Os critérios de qualidade que foram observados nas 26 publicações restantes desta RSL, são apresentados na Tabela 7 (KITCHENHAM, 2004):

Tabela 7 - Critérios de Qualidade

Critérios de Qualidade	
CQ1	O estudo está baseado em pesquisas empíricas ou em relatos de experiência com base em relatórios ou na opinião de especialistas?
CQ2	Existe uma definição clara dos objetivos da pesquisa?
CQ3	Existe uma descrição adequada do contexto em que a pesquisa foi realizada?
CQ4	O planejamento da pesquisa foi adequado para abordar os objetivos da pesquisa?
CQ5	A estratégia de extração de dados foi adequada aos objetivos da pesquisa?
CQ6	A análise dos dados foi suficientemente rigorosa estatisticamente?
CQ7	Existe uma indicação clara dos resultados e métricas usadas (AUC, F1-measuare, recall, accuracy, precision...)?

Fonte: Elaborado pelo Autor

Para cada critério de qualidade foi definida uma pontuação: se atende, 1 ponto; se atende parcialmente, 0,5 ponto; e se não atende, 0 ponto. A seguir, podemos observar na Tabela 8 a consolidação dos critérios de qualidade pelos artigos selecionados.

Tabela 8 - Percentuais dos Critérios de Qualidade Encontrados nas Publicações

CQ1	CQ2	CQ3	CQ4	CQ5	CQ6	CQ7
24,0	22,5	23,5	20,5	21,5	18,0	14,5
80,0%	75,0%	78,3%	68,3%	71,7%	60,0%	48,3%

Fonte: Elaborado pelo Autor

Podemos verificar que temos altos percentuais referentes aos critérios de qualidade CQ1, CQ2, CQ3, CQ4 e CQ5 (80,0%, 75,0%, 78,3%, 68,3% e 71,7%, respectivamente), demonstrando que a maior parte das pesquisas foi baseada em dados empíricos, com base em relatórios ou na opinião de especialistas, possui objetivos bem definidos, procura definir claramente o contexto onde a pesquisa foi realizada, teve um planejamento adequado para abordar os objetivos da pesquisa e também uma estratégia de extração de dados adequada aos objetivos da pesquisa.

Por outro lado, vemos claramente que não houve uma preocupação com a análise e apresentação dos resultados dos experimentos realizados nas publicações. Realmente, o CQ7, que trata por demonstrar os resultados e métricas alcançados chegou a apenas 48,3% dos artigos analisados, ou seja, menos de 15 trabalhos, obtendo o pior valor dentre todos os critérios. Utilizamos como base para o entendimento de “análise com rigor” as 5 diretrizes apresentadas por Kitchenham et al. (2002):

a) Especificação de todo e qualquer procedimento usado para controlar múltiplos testes;

b) Considerar o uso de “análise às cegas”;
c) Realizar análises minuciosas (identificação e tratamento de valores atípicos ou de observações que influenciam a análise);
d) Assegurar que os dados não violam as restrições aplicadas pelos testes a serem utilizados;
e) Aplicar procedimentos de controle de qualidade apropriados para a verificação dos resultados.

Das 26 publicações avaliadas, somando-se a pontuação dos 7 critérios, apenas as que obtiveram uma pontuação superior a 3,5 pontos foram selecionadas para a próxima fase da RSL. Sendo que 5 publicações não alcançaram a pontuação mínima, restando assim 21.

2.4.2 Trabalhos Relacionados

Com as 21 publicações restantes após a aplicação dos critérios de qualidade, realizamos o trabalho de extração de dados, preenchendo um formulário com informações detalhadas obtidas após a leitura dos trabalhos. Os campos definidos são: Data Publicação, Autores, Título, Objetivo do Estudo, Tipo de Estudo/Ferramenta, Método/Técnica, Corpus de Texto, Medidas (recall, acurácia, precisão e medida-F), Resultados e Conclusões.

Na Tabela 9, é apresentada a lista dos 21 trabalhos restantes selecionados:

Tabela 9 - Trabalhos Relacionados

	Artigo	Ano	Autor(es)
1	A bootstrapping approach for training a NER with conditional random fields	2011	Teixeira, Jorge; Sarmento, Luís e Oliveira, Eugênio
2	A Deep Learning Approach to Named Entity Recognition in Portuguese Texts	2018	Fernandes, Ivo André Domingues
3	A metadata geoparsing system for place name recognition and resolution in metadata records	2011	Freire, Nuno; Borbinha, José; Calado, Pável e Martins, Bruno
4	Adapting an Entity Centric Model for Portuguese coreference resolution	2016	Fonseca, Evandro B.; Vieira, Renata e Vanin, Aline
5	Bidirectional LSTM with a context input window for named entity recognition in tweets	2017	Peres, Rafael; Esteves, Diego e Maheshwari, Gaurav
6	Comparative analysis of Portuguese named entities recognition tools	2014	Amaral, Daniela; Fonseca, Evandro; Lopes, Lucelene e Vieira, Renata
7	Entity extraction within plain-text collections WISE 2013 challenge - T1: Entity linking track	2013	Abreu, Carolina; Costa, Flávio; Santos, Laécio; Monteiro, Lucas;

			Peres, Luiz; Lustosa, Patrícia e Weigang, Li
8	Exploiting named entity taggers in a second language	2005	Solorio, Thamar
9	Extracting and structuring open relations from Portuguese text	2016	Collovini, Sandra; Machado, Gabriel e Vieira, Renata
10	FS-NER: A lightweight filter-stream approach to named entity recognition on twitter data	2013	Oliveira, Diego Marinho de; Laender, Alberto; Veloso, Adriano e Silva, Altigran da
11	Identificação de termos relevantes em relatórios usando text mining	2017	Bastos, Pedro da Silva
12	Learning named entity recognition in Portuguese from Spanish	2005	Solorio, Thamar e Lopez-Lopez, Aurelio
13	Machine learning algorithms for Portuguese named entity recognition	2007	Duarte, Julio Cesar e Milidiú, Ruy Luiz
14	Multi-level NER for Portuguese in a CG framework	2013	Bick, Eckhard
15	Named entities for hot topics ranking and ontology navigation aid	2009	Bruckschen, Mírian; Vieira, Renata e Rigo, Sandro
16	Named entity extraction from Portuguese web text	2017	Pires, André Ricardo Oliveira
17	Named Entity Recognition in Twitter Using Images and Text	2019	Esteves, Diego; Peres, Rafael; Lehmann, Jens e Napolitano, Giulio
18	Portuguese corpus-based learning using ETL	2008	Milidiú, Ruy Luiz; Santos, Cícero Nogueira dos e Duarte, Julio Cesar
19	Portuguese part-of-speech tagging using entropy guided transformation learning	2008	Santos, Cícero Nogueira dos; Milidiú, Ruy Luiz e Rentería, Raul
20	Processamento de Linguagem Natural para classificação de entidades nomeadas no Diário Oficial da União Brasileiro	2018	Alles, Vanderlei; Giozza, William e Alburquerque, Robson de Oliveira
21	Second HAREM: Advancing the state of the art of named entity recognition in Portuguese	2010	Freitas, Cláudia; Mota, Cristina; Santos, Diana; Oliveira, Hugo Gonçalo e Carvalho, Paula

Fonte: Elaborado pelo Autor

Os trabalhos “Identificação de termos relevantes em relatórios usando text mining” de Bastos (2017), “Comparative analysis of Portuguese named entities recognition tools”, de Amaral et al. (2014) e “Processamento de Linguagem Natural para classificação de entidades nomeadas no Diário Oficial da União Brasileiro” de Alles et al. (2018), também foram encontrados na pesquisa fora da RSL e analisados na seção 3.2.1. Dos demais trabalhos selecionados ao final da RSL, iremos apresentar uma breve descrição do que trata cada um deles.

2.4.2.1 Descrição dos Trabalhos Selecionados

O trabalho “A bootstrapping approach for training a NER with conditional random fields” de Teixeira et al. (2011) apresenta uma abordagem de bootstrapping para o treinamento de um sistema de reconhecimento de entidades nomeadas (NER). Realiza a anotação de nomes

das pessoas em um conjunto de dados de 50.000 itens de notícias. Usando esse conjunto de treinamento, foi criado um modelo de classificação baseado em Campos Aleatórios Condicionais (CRF). Com o modelo pronto, são feitas anotações adicionais ao corpus inicial e é treinado um novo modelo de classificação. O ciclo é repetido até o modelo NER estabilizar. Cada iterações de bootstrapping foi avaliada calculando a precisão e a revocação do modelo NER na anotação de uma pequena coleção padrão-ouro (HAREM), a precisão e a revocação do método de anotação de inicialização CRF em uma pequena amostra de notícias e a correção e o número de novos nomes identificados. Os resultados obtidos se estabilizam após 7 iterações, atingindo valores de precisão de 83% e revocação de 68%.

O trabalho “A Deep Learning Approach to Named Entity Recognition in Portuguese Texts” de Fernandes (2018) analisa a viabilidade de utilizar arquiteturas de deep learning no reconhecimento de entidades em português. O trabalho analisa e prepara os dados textuais, define um método de avaliação que será usado para testar e comparar os modelos criados, e implementa e testa múltiplas arquiteturas de deep learning. Trabalha com dados textuais anotados e não anotados. Utiliza os datasets anotados HAREM I GC, HAREM II GC, MiniHAREM GC e WikiNER. Os dados não anotados são utilizados no processo de bootstrapping. Conclui que os resultados não são muito bons, mas destaca a importância do trabalho para entender as dificuldades e o potencial da arquitetura de deep learning. Os resultados obtidos com notícias jornalísticas foram precisão de 71,23%, revocação de 15,38% e medida-F de 25,3%.

Freire et al. (2011), no artigo “ A metadata geoparsing system for place name recognition and resolution in metadata records”, descrevem uma abordagem para realizar o reconhecimento e a resolução de nomes de locais mencionados nos registros descritivos de metadados de bibliotecas digitais típicas. Implementam uma solução utilizando uma técnica baseada em dicionário para reconhecimento de nomes de lugares geográficos, independente de qual a linguagem, e aprendizado de máquina para escolher um possível candidato à resolução. Dois métodos de avaliação foram utilizados. Enfim, foi utilizada validação cruzada, que mostrou resultados de precisão de 99% com revocação de 55%, ou uma revocação de 79% com precisão de 86%.

O artigo de Fonseca et al. (2016) apresenta a adaptação de um Modelo Centrado em Entidade para Português para a Coreference Resolution, considerando 10 categorias de

entidades nomeadas. O modelo foi avaliado utilizando o corpus português do HAREM e os resultados foram de 81,0% de precisão e 58,3% de revocação.

Peres et al. (2017) no artigo “Bidirectional LSTM with a context input window for named entity recognition in tweets” propõe um NER para tweets em português. Através de um novo corpus padrão de tweets anotado para Pessoa, Local e Organização (PLO). Além disso, demonstra várias experiências NER usando modelos baseados em Long Short Term Term Memory (LSTM). Obteve resultado de 52,78% de medida-F.

O trabalho de Abreu et al. (2013) “Entity extraction within plain-text collections WISE 2013 challenge - T1: Entity linking track”, é sobre a participação na conferência do WISE 2013, que propôs um desafio no qual as equipes devem rotular entidades em textos simples, com base em um determinado conjunto de entidades, o conjunto de dados do Wikilinks, que compreende 40 milhões de menções sobre 3 milhões de entidades. O artigo descreve uma estratégia direta e não supervisionada, dupla, para extrair e marcar entidades, com o objetivo de obter resultados precisos na identificação de nomes próprios e conceitos concretos, independentemente do domínio. A solução proposta é baseada em um pipeline de módulos de processamento de texto que inclui um analisador léxico. Para validar a solução proposta, foi feita a avaliação estatística dos resultados, através de várias medidas no estudo de caso fornecido. Os resultados obtidos foram de uma precisão média de 80,4% e de revocação média de 65,5%.

“Exploiting named entity taggers in a second language”, trabalho de Solorio (2005), apresenta um método para reconhecimento de entidades nomeadas (NER) que não dependa de recursos linguísticos complexos e não utiliza ferramentas dependentes de idioma. As únicas informações que utilizadas são extraídas automaticamente dos documentos, sem intervenção humana. Obteve bons resultados em espanhol, superando um sistema NER nessa língua. Em português, utilizando o corpus HAREM, teve resultados que podem ser considerados bons, pela proposta apresentada. Obteve médias de 56,1% de precisão, 46,8% de revocação e 50,3% de medida-F.

O artigo “Extracting and structuring open relations from Portuguese text” de Collovini et al. (2016) apresenta a extração e estruturação de relações abertas entre entidades nomeadas a partir de textos em português. Aplicou o modelo Conditional Random Fields (CRF) para a extração de descritores de relações entre entidades nomeadas pertencentes às categorias Pessoa, Local e Organização. Utilizando o corpus HAREM em conjunto com as ferramentas NLTK e Mallet, obteve resultados de 71% de precisão, 58% de revocação e 64% de medida-F.

O trabalho de Oliveira et al. (2013) “FS-NER : A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data” propõe uma nova abordagem para o reconhecimento de entidades nomeadas nos dados do Twitter, denominada FS-NER (Filter-Stream Named Entity Recognition). Que se caracteriza pelo uso de filtros que processam mensagens não rotuladas do Twitter, mais prático do que as abordagens supervisionadas existentes baseadas em CRF. Esses filtros não dependem da linguagem, o FS-NER pode ser aplicado a diferentes línguas sem exigir uma adaptação trabalhosa. Por meio de uma avaliação sistemática usando três coleções do Twitter e considerando sete tipos de entidade, o FS-NER apresenta um desempenho 3% melhor do que uma linha de base baseada em CRF, com uma medida-F média de 67%.

O trabalho “Learning named entity recognition in Portuguese from Spanish”, de Solorio e Lopez-Lopez (2005), apresenta um método prático para adaptar um sistema NER do espanhol para o português, utilizando o corpus HAREM. Baseia-se no treinamento de um algoritmo de aprendizado de máquina que utiliza recursos internos e externos. Os recursos externos são fornecidos por um sistema NER em espanhol, enquanto os recursos internos são extraídos automaticamente dos documentos. Os resultados para o idioma português foram precisão de 87,7%, revocação de 94% e medida-F de 90,8%.

Duarte e Milidiú (2007), no trabalho “Machine learning algorithms for Portuguese named entity recognition”, apresenta sete abordagens de aprendizado de máquina que utilizam HMM – Hidden Markov Modeling, TBL – Transformation Based error-driven Learning e SVM - Support Vector Machines, para resolver o NER em português, com o corpus SNR-CLIC. O desempenho de cada abordagem é avaliado empiricamente. O extrator baseado em SVM obteve o melhor resultado, com medida-F de 88.11%.

“Multi-level NER for Portuguese in a CG framework” é o trabalho apresentado por Bick (2003). O artigo descreve e avalia um sistema NER de base linguística para o português, o PALAVRAS, baseado em informações léxico-semânticas, correspondência de padrões e regras morfossintáticas, e regras de gramática orientadas ao contexto. Utiliza o corpus CETEM Público. Os resultados iniciais para textos de notícias em vários domínios, ao distinguir seis tipos de nomes diferentes, teve uma média de medida-F de 91,85% e 93,6% para subtipo de substantivos próprios.

No trabalho “Named entities for hot topics ranking and ontology navigation aid”, Bruckschen et al. (2009) propõe o aplicativo SeRELeP, que utiliza técnicas de PLN para a identificação de tópicos importantes em um portal de notícias. Utiliza o corpus HAREM e através da utilização do reconhecimento de entidades nomeadas, da marcação semântica e da identificação de identidade, é possível adquirir conhecimento dos textos analisados e, de modo automático, gerar ontologias e classificar tópicos importantes. Obteve como resultados precisão de 77%, revocação de 69% e medida-F de 73%.

Pires (2017), em sua dissertação “Named entity extraction from Portuguese web text” realiza uma avaliação de ferramentas NER, com o objetivo de identificar a melhor abordagem e configuração para o idioma português, utilizando como domínio as notícias do SIGARRA. Utilizou a coleção HAREM e um subconjunto anotado manualmente das notícias do SIGARRA para avaliar as ferramentas. Foi realizada inicialmente uma análise de desempenho das ferramentas Stanford CoreNLP, OpenNLP, spaCy e NLTK com o conjunto de dados HAREM, sendo que a ferramenta Stanford CoreNLP teve o melhor resultado, com 56,10%. Depois foi realizado um estudo de hiperparâmetros para verificar qual a melhor configuração para cada ferramenta, sendo que todas as ferramentas tiveram melhores resultados. Por fim preparou um novo corpus chamado SIGARRA News Corpus, contendo 905 notícias anotadas e 12.644 entidades anotadas. E, utilizando a melhor configuração, esse novo corpus foi treinado, tendo resultado de 86,86% de medida-F para o Stanford CoreNLP.

“Named Entity Recognition in Twitter Using Images and Text”, trabalho de Esteves et al. (2018), propõe vários níveis numa nova arquitetura sem dependência de recurso linguístico específico ou regra codificada. De um modo diferente das abordagens tradicionais, utiliza a extração de imagens e texto para classificar entidades nomeadas. Testes experimentais com o conjunto de dados Ritter obtiveram resultado de 59% de medida-F.

Milidiú et al. (2008), com o trabalho “Portuguese corpus-based learning using ETL”, Propõe modelos de Aprendizado de Transformação Guiada por Entropia (ETL) para a marcação de parte do discurso, separação de frases substantivas e reconhecimento de entidades nomeadas. Utiliza os corpora Mac-Morpho e Tycho Brahe para a marcação de parte do discurso. Para a separação de frases substantivas, utiliza o corpus SNR-CLIC. E para o reconhecimento de entidades nomeadas, os corpora HAREM, MiniHAREM e LearnNEC06. O ETL necessita apenas do conjunto de treinamento. Ele também simplifica a incorporação de novos recursos de entrada que são usadas com sucesso nos sistemas baseados em ETL. O trabalho apresenta

diversos resultados, um para cada corpus. Para o corpus HAREM, como exemplo, teve resultados médios de 71,35% de precisão, 68,74% de revocação e 70,02% de medida-F.

"Portuguese part-of-speech tagging using entropy guided transformation learning", trabalho de Santos et al. (2008), apresenta o Aprendizado de Transformação Guiada por Entropia (ETL) como uma nova estratégia de aprendizado de máquina, utilizando árvores de decisão e Aprendizado Baseado em Transformação (TBL). Aplica a estrutura ETL para a marcação de trechos em português e utiliza os corpora Mac-Morpho e Tycho Brahae. Essa proposta com o ETL obteve acurácias de 96,75% e 96,64% para Mac-Morpho e Tycho Brahae, respectivamente.

E, finalmente, o artigo de Freitas et al. (2010), "Second HAREM: Advancing the state of the art of named entity recognition in Portuguese", apresenta o Segundo HAREM abordando o reconhecimento de entidades nomeadas (NER). Nesta edição foram incluídos o reconhecimento, a normalização de entidades temporais e o ReRelEM, a detecção de relações semânticas entre entidades nomeadas. O trabalho mostra, além da configuração resumida do Segundo HAREM, os recursos e ferramentas disponíveis desenvolvidos: as coleções de ouro, um conjunto de documentos cujas entidades nomeadas e relações semânticas entre essas entidades foram anotadas manualmente, as ferramentas de pontuação, o SAHARA, aplicativo da Web que permite avaliação interativa, e a segunda coleção HAREM (que contém a versão não anotada da coleção dourada), bem como os resultados dos sistemas participantes. O melhor resultado foi o do sistema Priberam, com medida-F de 58%.

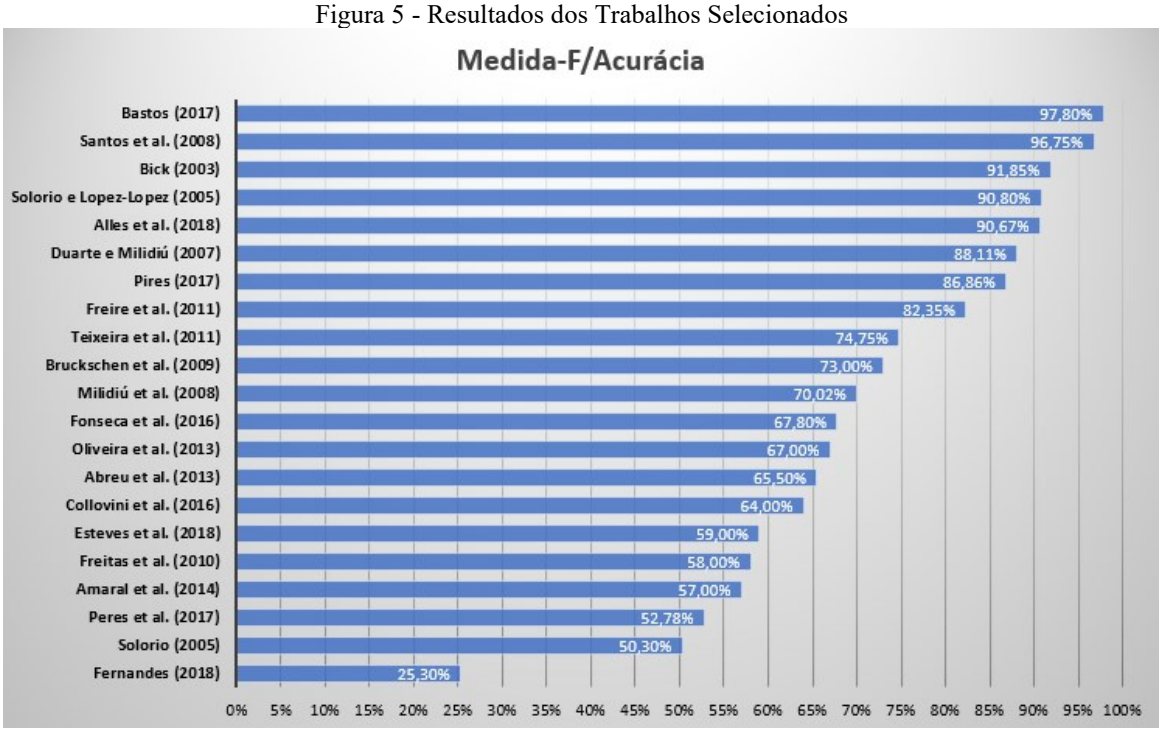
2.5 RESULTADOS DA RSL

Nesta seção, apresentaremos os principais resultados encontrados a partir da RSL que pesquisa métodos e técnicas para reconhecimento de entidades nomeadas em português. Em busca de evidências empíricas dos artefatos que tratam desse tema, na tentativa de termos as respostas à nossa questão de pesquisa e às questões norteadoras que foram definidas na Seção 2.3.

RQ1. Quais estudos apresentam implementações práticas de reconhecimento de entidades nomeadas em textos em português com resultados objetivos, como precisão, revocação, medida-F ou acurácia no REN?

- A resposta à pergunta RQ1 pode ser verificada na Tabela 9, que lista os trabalhos selecionados nas diversas fases da RSL e que atendem à questão, sendo estudos que apresentam resultados objetivos de implementações práticas de REN em textos em português. Um total de 21 trabalhos dos 844 inicialmente selecionados. Na Figura 5 podemos observar o gráfico comparativo dos resultados dos 21 trabalhos selecionados (Alguns trabalhos apresentam acurácia como resultado). Mas a conclusão de quais as melhores ferramentas utilizadas pode se seguir apenas por essa comparação, é importante considerar que os testes foram realizados com corpora e textos diferentes, além de entidades diversas. Para que houvesse uma definição real da melhor ferramenta, ou até mesmo o melhor corpus, só seria possível com testes na mesma condição.

Figura 5 - Resultados dos Trabalhos Selecionados



Fonte: Elaborado pelo Autor

RQ2. Quais as principais ferramentas de REN utilizadas nos estudos selecionados?

- Quando se trata das ferramentas de REN que foram abordadas pelos trabalhos, alguns utilizaram bibliotecas prontas e disponibilizadas publicamente para implementar seus testes, como Stanford CoreNLP, OpenNLP, NLTK, spaCy, entre outras. Enquanto outros trabalhos implementaram seu próprio algoritmo,

tendo, alguns desses trabalhos, realizado testes comparativos entre os dois tipos de ferramentas. Na Tabela 10 podemos verificar quais as ferramentas que foram utilizadas:

Tabela 10 – Ferramentas utilizadas nos trabalhos selecionados.

Ferramenta	Quant. Utilizado
OpenNLP	3
NLTK	3
Stanford CoreNLP	2
Tensorflow Syntaxnet	2
FreeLing	1
LanguageTasks	1
Palavras	1
NERP-CRF	1
TreeTagger	1
FS-NER	1
spaCy	1
Priberam	1
SeRELeP	1
Mallet	1

Fonte: Elaborado pelo Autor

RQ3. Quais os principais corpora de dados utilizados nos estudos selecionados?

- Em relação aos corpora de texto utilizados, alguns utilizaram um construído especificamente para sua pesquisa, que também é chamado de corpus próprio. Outros utilizaram corpus disponibilizado para o público, como o HAREM. E ainda existem alguns trabalhos que realizaram testes comparativos entre o corpus público e o próprio. Na Tabela 11 podemos verificar quais os corpora que foram utilizados:

Tabela 11 – Corpora utilizados nos trabalhos selecionados.

Corpus	Quant. Utilizado
Corpus Próprio	3
HAREM	10
Mac-Morpho	2
MiniHAREM	2
Tycho Brahae	2
Amazônia	1
CETEMPúblico	1
LearnNEC06	1
Ritter	1
SIGARRA	1
SNR-CLIC	1
Wikilinks	1
WikiNER	1

Fonte: Elaborado pelo Autor

3 PROCEDIMENTOS METODOLÓGICOS

O objetivo deste trabalho é construir um corpus que permita realizar a detecção de entidades nomeadas em relatórios de inteligência financeira. Para atingir esse objetivo, propõe-se seguir os seguintes passos apoiados por materiais e métodos conforme mostrado na Figura 6.

Figura 6 - Passos da Metodologia

Fonte: Elaborado pelo Autor

Para tanto, descrevemos a coleta de dados na Seção 3.1, transformação e limpeza dos dados na Seção 3.2, criação do corpus na Seção 3.4, análise de ferramentas REN 3.3, treinamento do modelo na Seção 3.5, validação do modelo na Seção 3.6, testes e resultados na Seção 3.7 e, por fim, discussão dos resultados na Seção 4.

a) Coleta de dados: apresentar as características físicas do RIF, confecção de RIFs fictícios e o universo de RIFs utilizados neste trabalho.
b) Transformação e limpeza dos dados: demonstra as necessidades de uma biblioteca para leitura dos arquivos em PDF, limpeza de determinados caracteres e geração de RIFs em formato TXT.
c) Criação e análise de corpus: mostra como será criado o corpus próprio baseado em RIFs e analisa um corpus público para realizar uma comparação.
d) Ferramenta REN: identificação das ferramentas/técnicas/algoritmos a serem usados em REN e quais entidades serão reconhecidas.
e) Treinamento do modelo: demonstra como será o treinamento baseado no corpus anotado do RIF e qual a configuração de treinamento para esse corpus.

f) Validação do modelo: com o conjunto de RIFs utilizados na construção do corpus, realiza treinamentos específicos para aplicar a validação cruzada e apresenta seus resultados.
g) Testes dos modelos: com o modelo treinado no item “e”, realiza testes e a apresentação dos resultados.
h) Discussão: analisa os resultados obtidos.
i) Apresentação e análise dos resultados dos testes com o modelo do corpus RIF e comparação com os resultados validação;
j) Apresentação e análise dos resultados dos testes com o modelo em português do spaCy;
k) Apresentação e análise dos resultados dos testes com o modelo treinado do corpus RIF em conjunto com o modelo em português do spaCy e comparação entre os resultados dos três testes.

3.1 RELATÓRIO DE INTELIGÊNCIA FINANCEIRA

O Relatório de Inteligência Financeira (RIF) é produzido pelo Conselho de Controle de Atividades Financeiras (COAF), como foi detalhado na Seção 1.1. Ele é distribuído às autoridades competentes para realizar a análise e possível investigação das informações relatadas (COAF, 2015). O COAF também determina que esses relatórios são sigilosos por força de lei, conforme podemos verificar no próprio site do COAF (2015):

> O resultado das análises de inteligência financeira decorrentes de comunicações recebidas, de intercâmbio de informações ou de denúncias é registrado em documento denominado Relatório de Inteligência Financeira – RIF.
> Quando o resultado das análises indicar a existência de fundados indícios de lavagem de dinheiro, ou qualquer outro ilícito, os Relatórios de Inteligência Financeira são encaminhados às autoridades competentes para instauração dos procedimentos cabíveis.
> O conteúdo do RIF é protegido por sigilo constitucional, inclusive nos termos da Lei Complementar 105, de 2001, não estando, portanto, sujeito às classificações da Lei 12.527, de 2011. O órgão destinatário do RIF é responsável pela preservação do sigilo.
> Existem dois tipos de relatório: espontâneo (de ofício): elaborado por iniciativa do Coaf a partir da análise de comunicações ou denúncias; e de

> intercâmbio: elaborado para atendimento a solicitação de intercâmbio de informações por autoridades nacionais ou por Unidades de Inteligência Financeira.

Por motivo de sigilo e a fim de evitar questionamentos dos órgãos responsáveis pela confecção e análise dos RIFs, além de pessoas físicas e jurídicas relacionadas, não foram utilizados RIFs reais. A hipótese de trabalhar com RIFs reais de maneira sigilosa, sem expor suas informações, foi tema de discussão durante o trabalho e entende-se que: 1) esta pode ser considerada uma limitação do trabalho e discutimos este assunto na Seção 5, de conclusão; 2) a anonimização dos RIF pode levar a especulações de órgãos responsáveis e pessoas (física e jurídicas) envolvidas nos RIF selecionados, podendo, caso sejam conhecidos alguns dos valores ou conteúdo de um RIF real, levar ao processo inverso de desanonimização; e 3) a criação de RIF a partir de exemplos de RIF reais é viável, desde que busque não apenas manter o formato de escrita e apresentação dos dados, mas, também de possíveis crimes a serem investigados. Escolhe-se, assim, a terceira opção, pois acredita-se que o essencial para este trabalho é a identificação de entidades nomeadas específicas e não a relação entre estas entidades ou o significado semântico do texto, apesar de este último também ser importante. Dessa forma, construímos um grupo de RIFs com informações fictícias, mas com estrutura parecida e que representasse, mesmo que em universo menor, a diversidade dos RIFs reais.

Foi construído um conjunto de 35 RIFs de variados tamanhos, entre 2 e 15 páginas, de poucas a muitas entidades, com os arquivos em formato PDF e com as mesmas fontes e outros padrões utilizados nos RIFs reais. Apresentamos um exemplo RIF na Figura 7:

Figura 7 - Trecho de RIF

Comunicações de Operações de que trata a Lei 9.613/98

Comunicações recebidas dos setores obrigados nos termos das normas emanadas das autoridades supervisoras.

1 - JOSE BARROSO DA SILVA FILHO

1.1

Relacionados	CPF/CNPJ	Tipo do Envolvimento
RODAUTO DISTRIBUIDORA DE AUTOMÓVEIS LTDA.	08.888.867/0003-03	Outros
JOSE LUCIANO FERREIRA	018.888.888-18	Outros
LEANDRO NIVALDO ADM E COR DE SEGS LTDA	08.888.848/0001-52	Outros
OMEGA SUPERTROCA COMERCIO DE LUBRIFICANTES E SERVICOS LTDA	08.888.850/0001-90	Outros
JOSE BARROSO DA SILVA FILHO	158.888.818-53	Titular
LOURENÇO & LOVATO LTDA - ME	28.888.808/0001-70	Outros
IMPEMAX BRASIL LTDA - ME	28.888.860/0001-05	Outros
TEREZA PACHECO HEITOR	28.888.846/0001-83	Outros
CHARGER ADMINISTRACAO COMERCIO E SERVICOS LTDA.	28.888.864/0001-93	Outros
RICARDO MALTA LOMBA VIEIRA	338.888.868-74	Outros
ZOROASTRO ANTÔNIO PEREIRA	518.888.846-04	Outros
ALBERGSON JOSE DA SILVA	598.888.836-04	Outros

Segmento: Banco Central - Atípicas

Instituição Financeira	Local	Agência - Sufixo CNPJ	Conta	Período	Valor em R$
Banco do Brasil S.A.	OLIMPIA-SP	MARISTA - 6560	8783	2/10/2017 até 2/7/2018	2.208.677,00

Créditos R$: 1.087.480,00 **Débitos R$:** 1.121.197,00

Informações Adicionais: Período de análise: 02.10.2017 a 02.07.2018 Cliente cadastrado como servidor público estadual aposentado com renda mensal de R$ 6.944,09. Principais créditos: TEDs no valor de R$ 1.036.983,66 Proventos no valor de R$ 18.470,69 Depósitos em Cheques (compensado) no valor de R$ 17.960,00 Depósitos Online no valor de R$ 9.383,00 Imposto de Renda no valor de R$ 3.840,22 Algumas origens dos recursos: OMEGA SUPERTROCA COMERCIO DE LUBRIFICANTES E SERV - 08888850000190 - R$ 647.250,00 CHARGER ADMINISTRACAO COMERCIO E SERVICOS LTDA. - 24888884000193 - R$ 225.000,00 IMPEMAX BRASIL LTDA - ME - 28888860000105 - R$ 64.275,26 MARLENE PEREIRA LOURENÇO TRAVEL - 28888808000170 - R$ 36.382,00 Titular em análise – R$ 27.000,00 RICARDO MALTA LOMBA VIEIRA - 33888886874 - R$ 17.500,00 LEANDRO NIVALDO ADMINISTRADORA E CORRETORA DE SEG - 08888848000152 - R$ 10.110,00 ARGUEL COMERCIO IMPORTACAO E EXPORTACAO EIRELI - 28888846000183 - R$ 7.421,40 Débitos mais expressivos: Saques no valor de R$ 753.895,98 TEDs emitidas no valor de R$ 311.405,00 Compras com Cartão no valor de R$ 14.933,23 Banco 24 Horas no valor de R$ 9.530,00 Cheques Emitidos no valor de R$ 9.279,77 Pagamentos de Títulos no valor de R$ 6.272,26 Transferências no valor de R$ 6.100,00 Os saques, ocorreram por meio de 265 transações, em sua maioria, de valores iguais ou inferiores a R$ 9.752,33, impossibilitando o conhecimento dos destinatários dos recursos em razão de quantias abaixo do mínimo exigido para a identificação. Alguns destinos dos valores: MARLENE PEREIRA LOURENÇO TRAVEL - 28888808000170 - R$ 180.265,00 JOSE LUCIANO FERREIRA - 01888888818 - R$ 100.000,00 ZOROASTRO ANTONIO PEREIRA - 51888884604 - R$ 19.140,00 RODAUTO DISTRIBUIDORA DE AUTOMOVEIS LTDA - 08888867000303 - R$ 12.000,00 ALBERGSON JOSE DA SILVA - 59888883604 - R$ 6.000,00 INFORMAÇÕES

Fonte: Elaborado pelo Autor

3.2 TRANSFORMAÇÃO, LIMPEZA E LEITURA DOS RIFS

Para a leitura dos dados dos RIFs em PDF pelo Python, foram testadas algumas bibliotecas de leitura de arquivos PDF, como PDFMiner[10] e PyPDF2[11]. Essas ferramentas não leram corretamente os arquivos. Os problemas detectados foram: não conseguir ler dados das tabelas com exatidão; não ler fontes muito pequenas; e inserir espaços entre as letras de algumas palavras.

Enfim, foi testada a biblioteca Tika[12], que funcionou quase sem erros, necessitando apenas algumas correções após carregar o arquivo em memória. Os erros que ainda persistiram e as respectivas soluções foram:

a) Apresentar no texto uma string “cid:160”, que significa que um caractere de código 160, que é um tipo de espaço, não pode ser interpretado. Esta string foi substituída por um espaço em branco;
b) Em alguns casos, o caractere “i” é carregado como “iiii”. Esta string foi substituída por um “i”;
c) Apresentou espaços duplicados em alguns casos. Estes foram substituídos por um espaço simples;
d) As quebras de linha foram retiradas para não separarem um nome de entidade.

O script de limpeza e tratamento foi útil tanto para o processo de REN, como para a geração dos arquivos dos RIFs em formato TXT para realizar a marcação das entidades, tornando o texto mais limpo ao retirar caracteres desnecessários que causavam eventuais erros.

3.3 FERRAMENTA DE REN

Uma das ferramentas bastante utilizada para processamento de linguagem natural é o spaCy[13], uma biblioteca de código aberto gratuita em Python. Essa biblioteca permite identificar sobre o que o texto trata, o que as palavras significam no contexto, quais seus relacionamentos,

[10] https://pdfminer-docs.readthedocs.io/pdfminer_index.html
[11] https://pythonhosted.org/PyPDF2/
[12] Apache Tika - a content analysis toolkit - https://tika.apache.org/
[13] https://spacy.io/

quais empresas e produtos são mencionados ou quais textos são semelhantes entre si, entre outras informações.

O spaCy foi projetado especificamente para ajudar na criação de aplicativos que processam e "compreendem" grandes volumes de texto. Ele pode ser usado para criar sistemas de extração de informações ou de compreensão de linguagem natural ou para pré-processar o texto para aprendizado profundo (SPACY, 2020).

Um recurso importante do spaCy é calcular os resultados de precisão, revocação e medida F, através da classe Scorer[14]. Ao treinar os modelos usando o comando spacy train, incluindo as categorias de texto nos dados de treinamento no formato JSON, o Scorer e o nlp.evaluate relatam as pontuações de classificação de texto.

Os testes de implementação foram realizados com o Python 3.7 e o spaCy 2.2.3.

3.3.1 Modelo em Português

O spaCy disponibiliza diversos modelos pré-treinados em vários idiomas. Esses modelos são divididos em dois tipos, modelos principais, que são modelos pré-treinados de uso geral para prever entidades nomeadas, *tags* de parte do texto e dependências sintáticas. Pode ser usado imediatamente e ajustado em dados mais específicos. E modelos iniciais, que são modelos em branco para receberem os pacotes iniciais de aprendizagem com pesos pré-treinados com os quais você pode inicializar seus modelos para obter melhor precisão. Eles podem incluir vetores de palavras, que serão utilizados como recursos durante o treinamento, ou outros modelos pré-treinados (SPACY, 2020).

Os modelos pré-treinados disponibilizados pelo spaCy são nos idiomas chinês, dinamarquês, holandês, inglês, francês, alemão, grego, italiano, japonês, lituano, norueguês, polonês, português, romeno e espanhol. Neste trabalho foi utilizado o modelo em português pt_core_news_sm[15] versão 2.2.5. O modelo para o idioma português não possui tantos recursos como o modelo em inglês. Enquanto o modelo em inglês permite o reconhecimento de 18 tipos de entidade, o português utiliza um esquema de anotação de REN menos refinado e reconhecem as seguintes entidades (SPACY, 2020):

[14] https://spacy.io/api/scorer
[15] https://spacy.io/models/pt

e) PER – Pessoa ou família nomeada;
f) LOC – Nome do local definido politicamente ou geograficamente (cidades, províncias, países, regiões internacionais, massas de água, montanhas);
g) ORG – Nome da entidade corporativa, governamental ou outra entidade organizacional;
h) MISC – Entidades diversas, eventos, nacionalidades, produtos ou obras de arte.

O modelo português foi construído a partir dos corpora Bosque[16] e WikiNer[17]. O corpus Bosque tem sua origem no projeto Floresta Sintá(c)tica[16] que é um conjunto de frases (corpus) analisadas morfossintaticamente. Esse projeto é uma colaboração entre a Linguateca[18] e o projeto VISL[19]. Contém textos em português (do Brasil e de Portugal) anotados (analisados) automaticamente pelo analisador sintático PALAVRAS e revistos por linguistas. Tendo por objetivo o treino e avaliação dos analisadores morfossintáticos, estudos que se baseiam em corpus para investigação da língua e seus aspectos, além dos sintáticos, também semânticos e discursivos. A Floresta Sintá(c)tica possui quatro partes, diferindo quanto ao gênero textual, quanto ao modo, escrito ou falado e quanto ao grau de revisão linguística (FLORESTA, 2010):

a) Bosque[16] – Totalmente revisto por linguistas;
b) Selva[16] – Parcialmente revisto;
c) Floresta Virgem[16] – Não revisto;
d) Amazônia[16] – Não revisto.

O corpus Bosque, que é uma "floresta" integralmente revista por linguistas, é composto por 9.368 frases, retiradas do CETENFolha[20] e do CETEMPúblico[21]. Trata-se da

[16] https://www.linguateca.pt/Floresta/
[17] https://hackage.haskell.org/package/chatter-0.9.1.0/docs/NLP-Corpora-WikiNer.html
[18] https://www.linguateca.pt/
[19] https://visl.sdu.dk/
[20] CETENFolha – Corpus de Extratos de Textos Electrónicos NILC/Folha de S. Paulo. Inclui o texto da Folha de S. Paulo do ano de 1994 (as 365 edições), incluindo cadernos não-diários, num total ligeiramente inferior a 24 milhões de palavras (versão 1.0).
[21] CETEMPúblico – Corpus de Extratos de Textos Electrónicos MCT/Público. Composto de aproximadamente 180 milhões de palavras em português europeu, criado pelo projeto Processamento Computacional do Português, que deu origem à Linguateca, em protocolo entre o Ministério da Ciência e da Tecnologia (MCT) português e o jornal PÚBLICO em Abril de 2000.

parte mais correta da Floresta, sendo assim o mais aconselhado para pesquisas em que se priorize a precisão dos resultados e não a quantidade (FLORESTA, 2010).

3.4 CORPUS BASEADO NOS RIFS

Para a divisão do conjunto de RIFs para treinamento e teste, foi utilizado o método Hold-out na proporção 85%/15% (AMARAL, 2013). Assim, dos 35 RIFs confeccionados para este trabalho, 30 foram utilizados na construção do corpus, e consequente treinamento, e 5 reservados para a realização dos testes. Também definimos que seriam marcadas as entidades PER – pessoa, LOC – local, ORG – organização e MISC – que foi utilizado na marcação das transações financeiras informadas.

Para realizar a marcação das entidades, foi utilizada a ferramenta Dataturks[22], que permite anotação em diversos tipos de documentos.

Apesar dos RIFs estarem no formato PDF e o Dataturks realizar a marcação em PDF também, nos testes iniciais foi observado que ocorriam diversos erros, principalmente ao executar a classe Scorer. Por isso, optou-se por transformar os documentos PDF em TXT, utilizando um script conforme foi exposto na Seção 3.2. Isto foi útil para realizar a marcação das entidades, tornando o texto mais limpo retirando caracteres desnecessários que causavam os erros.

[22] https://dataturks.com/

Figura 8 - Anotação de Entidades

Fonte: Gerado no https://dataturks.com/

Após a anotação das entidades dos 30 RIFs, com um total de 3.132 entidades anotadas, o Dataturks permite exportar nos formatos JSON[23] e Stanford NER[24]. A saída NER dos dados no Dataturks está muito próxima do formato usado pelo spaCy, a diferença é que o spaCy utiliza tuplas do Python que não são suportadas pelo padrão JSON, portanto, é necessário usar um script para converter dados em formato JSON em dados spaCy. Abaixo podemos observar as diferenças entre os dois formatos:

a) Formato spaCy:

```
[('RICARDO MALTA LOMBA VIEIRA trabalha na IMPEMAX BRASIL LTDA na cidade de Recife - Brasil. Juntamente saques com sua esposa TEREZA PACHECO HEITOR', [(0, 26, 'PER'), (39, 58, 'ORG'), (72, 78, 'LOC'), (81, 87, 'LOC'), (122, 143, 'PER')])]
```

[23] JSON, um acrônimo de JavaScript Object Notation, é um formato compacto, de padrão aberto independente, de troca de dados simples e rápida entre sistemas, especificado por Douglas Crockford em 2000, que utiliza texto legível a humanos, no formato atributo-valor.

[24] https://nlp.stanford.edu/ner/

b) Formato JSON gerado pelo Dataturks:


```
{"content" : "RICARDO MALTA LOMBA VIEIRA trabalha na IMPEMAX BRASIL LTDA na cidade
de Recife - Brasil. Juntamente saques com sua esposa TEREZA PACHECO HEITOR",
"annotation" : [{"label" : ["PER"], "points" : [{"start" : 0, "end" : 25, "text" :
"RICARDO MALTA LOMBA VIEIRA"}]}, {"label" : ["ORG"], "points" : [{"start" : 39,
"end" : 57, "text" : "IMPEMAX BRASIL LTDA"}]}, {"label" : ["LOC"], "points" :
[{"start" : 72, "end" : 77, "text" : "Recife"}]}, {"label" : ["LOC"], "points" :
[{"start" : 81, "end" : 86, "text" : "Brasil"}]}, {"label" : ["PER"], "points" :
[{"start" : 122, "end" : 142, "text" : "TEREZA PACHECO HEITOR"}]}], "extras" : null,
"metadata" : {"first_done_at" : 1591670090000, "last_updated_at" : 1591670090000,
"sec_taken" : 0, "last_updated_by" : "0Ic1iIcJxfU9bHb4H9tCcZI2Kx33", "status" :
"done", "evaluation" : "NONE"}}
```


Então, baseado nesses 30 RIFs, temos um corpus RIF, que será treinado (Seção 3.5), validado (Seção 3.6) e testado (Seção 3.7).

Conforme pôde ser avaliado nos trabalhos selecionados na RSL, Seção 2.4.2, alguns trabalhos fizeram a opção de criar um corpus próprio. Alles (2018, p. 25) fez essa opção e justifica “utilizar um corpus já existente é em geral inviável, uma vez que pode gerar resultados diversos do pretendido, além de ter o desempenho reduzido, pois o processo de seleção de textos leva em conta o objetivo pretendido para um contexto específico”. Por isso, para comparar com os resultados do corpus RIF, também será utilizado nos testes o modelo em português do spaCy, pt_core_news_sm. Além disso, esse modelo do spaCy será treinado em conjunto com o corpus RIF.

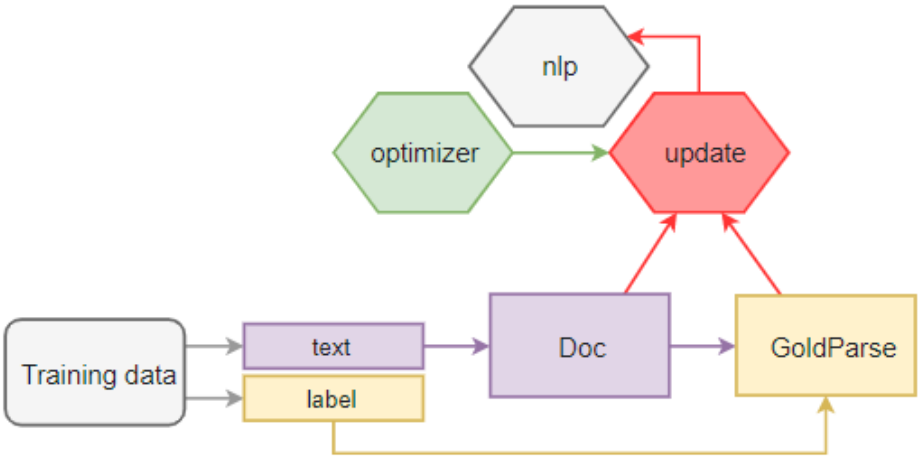
3.5 TREINAMENTO DO MODELO

Para treinar um modelo é necessário que se tenha dados de treinamento, que são exemplos de texto anotados com os rótulos que se deseja que o modelo preveja. É o que chamamos de corpus. O treinamento faz a leitura do corpus e utiliza o aprendizado supervisionado que se baseia em técnicas ou algoritmos específicos de Aprendizado de Máquina. O algoritmo aprende vários padrões para cada tipo de rótulo a partir dos dados de treinamento. Uma vez concluído esse processo, temos um modelo treinado. Esse modelo é estatístico e pode ser usado para prever os tipos de palavras em futuras amostras de dados de teste. Assim, a máquina realmente aprendeu, com base em amostras de dados de treinamento anteriores, como prever o tipo para novas amostras de dados não vistas (SARKAR, 2016).

Após ter o realizado treinamento, outro texto, não anotado, é apresentado ao modelo e este fará uma previsão. Sabendo quais são as entidades ou outras informações que devem ser reconhecidas corretamente, podemos fornecer feedback ao modelo com intuito de melhorar o modelo construído. A intenção de treinar um modelo, não é apenas para que ele memorize os exemplos, mas que possua uma teoria que possa ser generalizada em outros exemplos (ALLES, 2018). Por isso, os dados de treinamento devem ser os mais representativos possível dos dados que queremos processar. Por exemplo, ao treinar um modelo na Wikipédia, onde há pouquíssimas sentenças na primeira pessoa, é bem provável que terá um desempenho ruim no Twitter. Da mesma forma, um modelo treinado em textos jurídicos provavelmente terá um desempenho ruim em textos jornalísticos (SPACY, 2020).

Em geral, o conjunto de dados que temos é dividido em duas ou três divisões chamadas de conjuntos de dados de treinamento, validação (opcional) e teste, respectivamente. Para entender como o modelo está funcionando, além dos dados de treinamento, são necessários dados de avaliação (SARKAR, 2016). Testar o modelo apenas com os dados com os quais ele foi treinado resultará em resultados com praticamente 100% de exatidão, ou seja, overfitting. O algoritmo espera treinar os modelos providos a partir de documentos inteiros, não apenas frases simples. Se o corpus contiver apenas frases únicas, os modelos não aprenderão a processar documentos com várias frases, resultando em baixo desempenho em texto real.

Figura 9 - Reconhecimento no spaCy



Fonte: (SPACY, 2020)

Nesse esquema demonstrado na Figura 9, podemos entender melhor como funciona o fluxo do processo de reconhecimento do spaCy. Abaixo a explicação detalhada de cada fase:

> O objeto GoldParse coleta os exemplos de treinamento anotados, que são chamados de padrão-ouro. É inicializado com o objeto Doc a que se refere e argumentos de palavra-chave que especificam as anotações, como tags ou entidades. Seu trabalho é codificar as anotações, mantê-las alinhadas e criar as estruturas de dados em nível C necessárias para um acesso eficiente. Aqui está um exemplo de um GoldParse simples para tags de parte do discurso:
>
> ```
> vocab = Vocab(tag_map={"N": {"pos": "NOUN"}, "V": {"pos": "VERB"}})
> doc = Doc(vocab, words=["I", "like", "stuff"])
> gold = GoldParse(doc, tags=["N", "V", "N"])
> ```
>
> Usando o Doc e suas anotações de padrão ouro, o modelo pode ser atualizado para aprender uma frase de três palavras com as tags de parte do discurso designadas. O mapa de tags faz parte do vocabulário e define o esquema de anotação. Se você estiver treinando um novo modelo de idioma, isso permitirá que você mapeie as tags presentes no banco de árvores em que você treina para o esquema de tags do spaCy.
>
> ```
> doc = Doc(Vocab(), words=["Facebook", "released", "React", "in", "2014"])
> gold = GoldParse(doc, entities=["U-ORG", "O", "U-TECHNOLOGY", "O", "U-DATE"])
> ```
>
> O mesmo vale para entidades nomeadas. As letras adicionadas antes dos rótulos referem-se às tags do esquema BILUO - O é um token fora de uma entidade, U uma única unidade de entidade, B o início de uma entidade, I um token dentro de uma entidade e L o último token de uma entidade.
> Obviamente, não basta mostrar um modelo apenas um exemplo uma vez. Especialmente se você tiver apenas alguns exemplos, convém treinar para várias iterações. A cada iteração, os dados de treinamento são embaralhados para garantir que o modelo não faça generalizações com base na ordem dos exemplos. Outra técnica para melhorar os resultados do aprendizado é definir uma taxa de abandono, uma taxa na qual aleatoriamente "abandona" características e representações individuais. Isso torna mais difícil para o modelo memorizar os dados de treinamento. Por exemplo, um abandono de 0,25 significa que cada recurso ou representação interna tem uma probabilidade de 1/4 de ser descartado (SPACY, 2020, spacy.io/usage/training, tradução nossa).

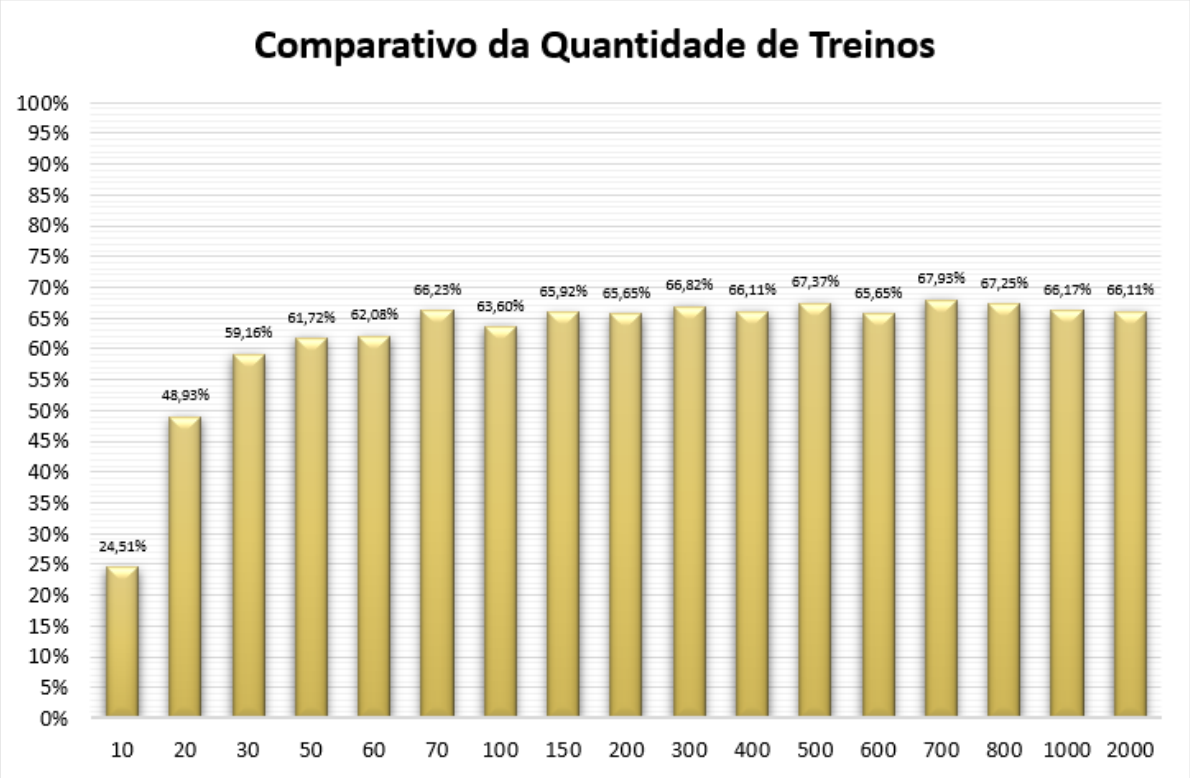
É necessário que se busque não repetir os mesmos exemplos várias vezes, pois o modelo deverá ficar "viciado" e terá dificuldades para reconhecer outros exemplos. Se isto acontecer, a função de perda será alterada, na prática. O otimizador vai tentar uma alternativa para minimizar a perda, mas isso pode gerar outras consequências, quando estiver trabalhando em exemplos que não está mais prestando atenção. Para tentar de evitar esse efeito nocivo do modelo estar "viciado" ou de "esquecimento generalizado", uma alternativa é adicionar outros exemplos que possam ser "lembrados" pelo modelo. E assim, acrescentar mais anotações com frases anotadas com entidades reconhecidas automaticamente pelo modelo original. Esse é um

processo de ajustes específicos para cada caso, de maneira empírica, deverá testar e melhorar seu modelo até conseguir a melhor solução (SARKAR, 2016; SPACY, 2020).

Para definir a configuração de treinamento ideal do corpus RIF, foram realizados testes variando a quantidade de iterações. Esses testes foram feitos treinando os 30 RIFs anotados e definidos para compor o corpus. E os resultados foram obtidos submetendo os 5 RIFs em conjunto ao modelo treinado. Apesar de serem apenas 5 RIFs para teste, eles possuem uma quantidade considerável de entidades, 926 (Seção 3.7), em relação ao corpus RIF, 3.132 (Seção 3.4).

Foram realizados treinamentos variando a quantidade de iterações (treinos) de 10 a 2.000. Na Figura 9 podemos entender que os treinamentos, a partir de 70 iterações, tiveram resultados muito próximos, demonstrando que o modelo se estabilizou. Assim, definimos adotar a utilização dos modelos com 200 treinos para os testes e para a validação.

Figura 10 - Definição da melhor Quantidade de Treinos



Fonte: Elaborado pelo Autor

3.5.1 Taxa de Abandono

A Taxa de Abandono – Dropout, aplicada ao REN, é uma técnica utilizada pelo algoritmo de treinamento que consiste em abandonar alguns trechos do corpus, baseada em

probabilidade, fazendo com que, a cada iteração do treinamento, o conjunto de termos do corpus esteja diferente. Isso evita o overfitting (superajuste), problema que é comum ocorrer num treinamento, e é caracterizado pelo aprendizado do modelo apenas para reconhecer os próprios termos do corpus mas tem dificuldade de identificar outros termos. A técnica do dropout faz com que o aprendizado do treinamento seja mais robusto e permita um melhor reconhecimento de novos termos (FONSECA, 2018; VIEIRA, 2018).

O dropout é definido num valor entre 0 e 1. Uma taxa de 0,3, por exemplo, significa que até 30% dos termos serão abandonados durante o treinamento. Segundo Srivastava et al. (2014), a taxa de 0,5 é a próxima do ideal para a maior parte dos tipos de treinamentos e, por este motivo, define-se a taxa a ser utilizada nos testes deste trabalho como 0,5.

3.6 VALIDAÇÃO CRUZADA – K-FOLD

Para realizar a validação do modelo, escolhemos a técnica de validação de dados conhecida como Validação Cruzada K-fold, que é uma técnica que utiliza todos os exemplos tanto para os treinamentos, quanto para os testes. Essa técnica consegue apresentar resultados mais precisos que outras técnicas de validação cruzada como Hold-out e Leave-One-Out (SCHREIBER et al., 2017). No caso deste último, o K-fold é considerado muitas vezes superior e não requer um desempenho de processamento de recursos computacionais tão alto quanto o Leave-One-Out exige (SCHREIBER et al., 2017). Abaixo poderemos ver uma explicação detalhada do funcionamento do K-fold:

> Dado uma base de dados hipotética em que conste 100 registros, e definindo o k=10 a base de dados será dividido em 10 subconjuntos onde cada subconjunto terá 10 registros cada. Após a divisão em subconjuntos, será utilizado um subconjunto, para ser utilizado na validação do modelo e os conjuntos restantes são utilizados como treinamento. O processo de validação cruzada é então repetido K (10) vezes, de modo que cada um dos K subconjuntos sejam utilizados exatamente uma vez como teste para validação do modelo.
> Por exemplo, dados 10 subconjuntos B1, B2... B10 o primeiro passo do K-Fold é utilizar B1 para teste e de B2 a B10 para treino. No segundo passo, B2 é utilizado para teste e todo o restante para treino, incluindo B1 que foi usado para teste no primeiro passo, no terceiro passo até o décimo será aplicada a mesma lógica sucessivamente. O resultado final da validação K-Fold é o desempenho médio do classificador nos K testes. O objetivo de repetir os

> testes diversas vezes é com o intuito de aumentar a confiabilidade da estimativa da precisão do classificador (SCHREIBER et al., 2017, p. 4).

Neste trabalho, a validação é feita usando K=5, conforme apresentado na Figura 10. São 30 RIFs definidos para o treinamento, dividindo o conjunto de RIFs em grupos de 6 RIFs cada. Após essa divisão, é realizado o treinamento com 24 RIFs distintos e a validação com 6 RIFs do conjunto. O treinamento utiliza a mesma configuração definida anteriormente de 200 treinos (iterações) e dropout de 0.5, como apresentado nas Seções 3.5 e 3.5.1, respectivamente.

Para a verificação dos resultados das validações, utiliza-se a classe Scorer do spaCy. Esta classe calcula os resultados de precisão, revocação e medida-F com base nas entidades reconhecidas em comparação com as entidades anotadas do arquivo submetido ao REN. São apresentados os resultados para cada tipo de entidade e um geral, que considera todos os tipos.

Na Figura 11, os conjuntos da cor azul são dos RIFs treinados, enquanto o conjunto da cor laranja é o dos RIFs reservados para a validação. A cada iteração, a mesma operação é feita, retirando o próximo grupo do treinamento para ser validado e retornando para o treinamento o grupo validado na iteração anterior, até que todos os RIFs tenham participado tanto do treinamento quanto da validação.

Figura 11 - Validação K-fold

Fonte: Elaborado pelo Autor

3.7 TESTES

Após a validação, os 30 RIFs são treinados em conjunto, criando o modelo do corpus RIF. Os 5 RIFs reservados para a fase de testes têm as suas entidades anotadas no Dataturks, para possibilitar o cálculo dos resultados pela classe Scorer.

É importante que os RIFs utilizados para o teste tenham uma quantidade alta de entidades anotadas, de forma que o resultado não seja fora de um padrão, que desejamos identificar, pelo motivo de ter poucas entidades ou até não possuir de algum tipo de entidade

(PER, ORG, LOC e MISC) (SARKAR, 2016; SPACY, 2020). Assim, foram escolhidos RIFs que possuem uma quantidade alta de entidades. Os 5 RIFs selecionados possuem entre 97 e 432 entidades anotadas, totalizando 926 entidades marcadas. Na Tabela 12 podemos verificar a distribuição de entidades por tipo de entidade e por arquivo de teste:

Tabela 12 – Entidades anotadas nos RIFs de teste.

Testes	PER	ORG	LOC	MISC	Total
Teste RIF 1	36	22	31	9	**98**
Teste RIF 2	91	208	98	35	**432**
Teste RIF 3	49	78	24	11	**162**
Teste RIF 4	55	29	8	9	**101**
Teste RIF 5	61	64	2	6	**133**
Total	**292**	**401**	**163**	**70**	**926**

Fonte: Elaborado pelo Autor

Além dos testes realizados com o modelo treinado do corpus RIF, também são realizados testes com o modelo em português do spaCy chamado pt_core_news_sm e com os dois modelos treinados em conjunto, para comparação dos resultados dos três cenários.

O teste apresenta as métricas de precisão, revocação e medida-F para cada RIF e, também, são detalhados para cada tipo de entidade (PER, ORG, LOC e MISC). Para o conjunto de RIFs do teste é apresentado o desvio padrão para todas as métricas.

3.8 MÉTRICAS DE AVALIAÇÃO

Em diversas áreas do conhecimento, para a avaliação da qualidade dos resultados são utilizadas de forma ampla as medidas de desempenho de precisão (*precision*), revocação (*recall*) e medida-F (*F-Measure*). Enquanto a precisão é uma medida de fidelidade, a revocação é uma medida de completude. Já medida-F é a média harmônica ponderada entre a precisão e a revocação (MATOS et al., 2009).

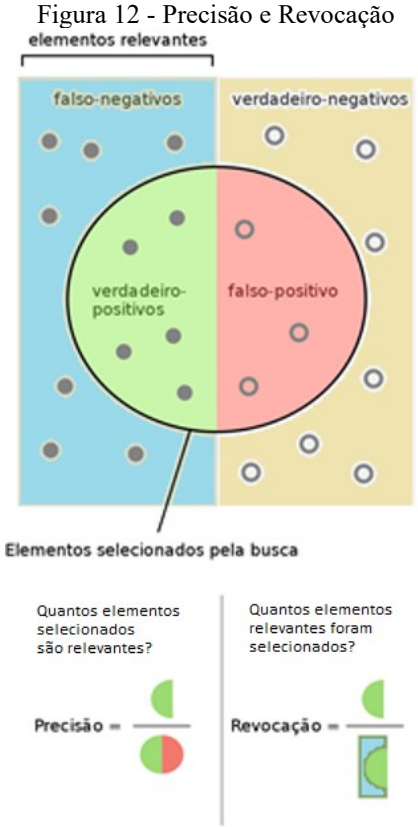
Precisão e revocação contribuem na avaliação de sistemas na área de Recuperação de Informação (RI) como medidas padrão. Mas as áreas de Extração da Informação, Inteligência Artificial, Aprendizado de Máquina, Processamento de Linguagem Natural (PLN), Reconhecimento de Entidades Nomeadas (REN), entre outras, também se utilizam bastante desses métricas.

Observando a Figura 12, no contexto do Reconhecimento de Entidades Nomeadas, podemos ter um conjunto de entidades nomeadas que esperamos que sejam reconhecidas

conforme o tipo com o qual foram definidas. Essas entidades nomeadas compõem o conjunto dos elementos relevantes, conjuntos azul e verde. Não fazem parte de conjunto os demais termos/elementos que não foram nomeados, conjuntos amarelo e vermelho. Os elementos dentro do círculo (verde e vermelho) são os termos que foram reconhecidos como as entidades nomeadas, e os que não estão dentro da área do círculo são os termos que não foram reconhecidos. Para entendermos melhor a forma de cálculo da precisão e da revocação, vamos detalhar as seguintes definições:

a) Verdadeiro-positivo – São os termos/elementos que **deveriam** ser reconhecidos como as entidades nomeadas esperadas e efetivamente **foram** reconhecidos;
b) Falso-positivo – São os termos/elementos que **não deveriam** ser reconhecidos como as entidades nomeadas esperadas e efetivamente **foram** reconhecidos;
c) Falso-negativo – São os termos/elementos que **deveriam** ser reconhecidos como as entidades nomeadas esperadas e efetivamente **não foram** reconhecidos;
d) Verdadeiro-negativo – São os termos/elementos que **não deveriam** ser reconhecidos como as entidades nomeadas esperadas e efetivamente **não foram** reconhecidos.

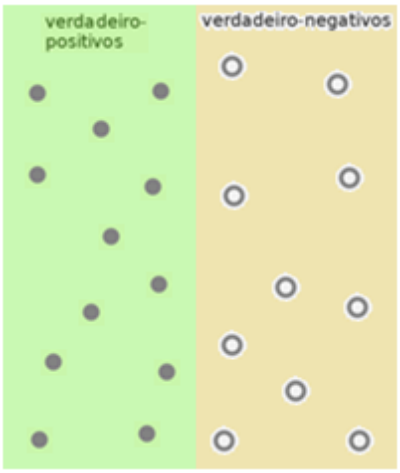
Figura 12 - Precisão e Revocação



Fonte: (OBSERVATÓRIO DE DADOS, 2020)

O reconhecimento ideal e desejado é que todas e apenas as entidades/elementos da parte esquerda (azul e verde) do conjunto total fossem reconhecidas, o que daria um resultado de 100% para a precisão e para a revocação (MATOS et al., 2009). Nesse caso teríamos apenas verdadeiro-positivos e verdadeiro-negativos, e a representação gráfica seria assim:

Figura 13 - Matriz Ideal



Fonte: Elaborado pelo Autor

Precisão é a definição de quanto da informação reconhecida é correta. Conforme vemos na Figura 11, a precisão é definida pela quantidade de elementos relevantes recuperados dividido por todos os elementos recuperados. No contexto de REN, todos os termos reconhecidos corretamente dividido pelo total de termos reconhecidos correta e incorretamente. A precisão é representada pela seguinte fórmula, onde VP é a quantidade de verdadeiro-positivos e FP é a quantidade de falso-positivos:

$$Precisão = \frac{VP}{VP + FP} \quad (1)$$

Já a revocação, que também é conhecida como cobertura ou sensibilidade, é a definição de quanta informação foi reconhecida. Verificando novamente a Figura 11, a revocação é definida pela quantidade de elementos relevantes recuperados dividido por todos os elementos relevantes. No contexto de REN, todos os termos reconhecidos corretamente dividido pelo total de termos que deveriam ter sido reconhecidos. A revocação é representada

pela seguinte fórmula, onde VP é a quantidade de verdadeiro-positivos e FN é a quantidade de falso-negativos:

$$Revocação = \frac{VP}{VP + FN} \tag{2}$$

Por fim, a medida-F, como já foi dito, é o cálculo da média harmônica ponderada da precisão e da revocação. A fórmula da medida-F permite que se atribua um peso, representado por β, para distinguir a importância da precisão ou da revocação. Se $\beta < 1$, a precisão terá um peso maior, mas, se $\beta > 1$, a revocação terá mais peso. Enfim, se $\beta = 1$, precisão e revocação terão o mesmo peso, essa é forma comumente utilizada nos sistemas da área de RI (MÔRO, 2018). Assim, utilizaremos a fórmula simplificada e tradicional da medida-F, aonde a precisão será representada por P e a revocação por R:

$$Medida\ F = \frac{2\ x\ P\ x\ R}{P + R} \tag{3}$$

4 RESULTADOS

Nesta seção, apresentam-se os resultados para a validação e os testes do modelo. Estes são divididos entre os arquivos RIFs de teste, depois o modelo treinado pelo spaCy e, por fim, o modelo em conjunto dos dois modelos anteriores.

São apresentadas as métricas de precisão, revocação e medida-F de cada iteração da validação e de cada arquivo de teste. Também, são detalhados os resultados separados para cada tipo de entidade (PER, ORG, LOC e MISC), além de um resultado geral, contendo todos os tipos. E, para cada grupo de resultados, apresenta-se também a média e o desvio-padrão.

4.1 RESULTADOS DA VALIDAÇÃO

De acordo com a metodologia definida na Seção 3.6, realizou-se a validação cruzada K-fold com cinco iterações. Os 30 RIFs anotados e definidos para o treinamento são divididos em 5 conjuntos de 6 RIFs e, a cada uma das 5 iterações, separa-se um dos conjuntos e faz-se o treinamento com os outros 4 conjuntos, num total de 24 RIFs. Em seguida, o conjunto separado é submetido ao modelo treinado para o reconhecimento de suas entidades e tem seus resultados calculados pela classe Scorer. Isso é repetido para os 5 conjuntos. O script de treinamento está descrito no Apêndice A e o de reconhecimento no Apêndice B.

Na Tabela 13 e na Figura 14 vemos o resultado geral para todos os tipos de entidades anotadas e os resultados detalhados por cada tipo de entidade.

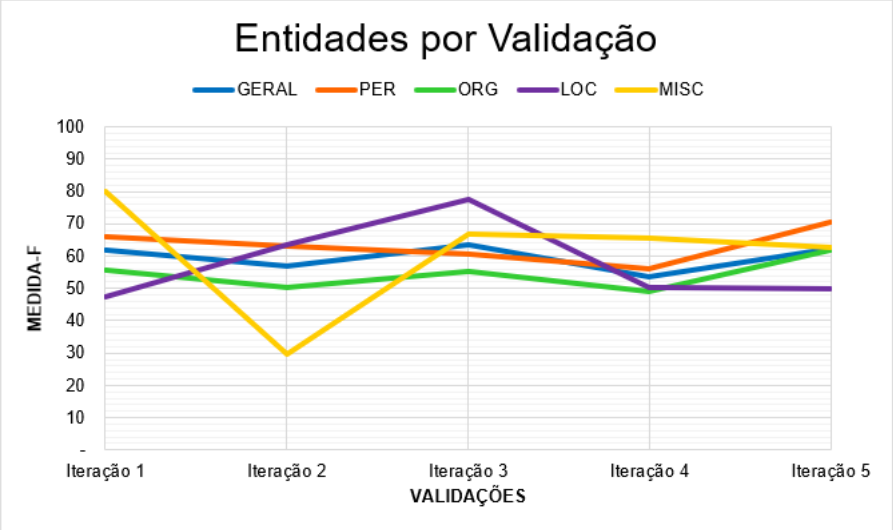
Tabela 13 - Resultados da Validação.

Validações	**Precisão**	**Revocação**	**Medida-F**
Iteração 1	67,57	57,07	61,88
Iteração 2	62,12	52,73	57,04
Iteração 3	68,35	59,01	63,33
Iteração 4	49,92	58,30	53,79
Iteração 5	63,19	61,79	62,48
Média	**62,23**	**57,78**	**59,71**
Desvio Padrão	**7,39**	**3,31**	**4,12**
Tipo de Entidade PER – Pessoa			
Iteração 1	64,44	67,28	65,83
Iteração 2	72,53	55,93	63,16
Iteração 3	63,89	57,50	60,53
Iteração 4	50,52	63,09	56,11
Iteração 5	63,76	78,92	70,53

Média	**63,03**	**64,54**	**63,23**
Desvio Padrão	**7,90**	**9,22**	**5,43**
Tipo de Entidade ORG – Organização			
Iteração 1	68,36	46,90	55,63
Iteração 2	57,14	44,83	50,24
Iteração 3	62,07	50,00	55,38
Iteração 4	46,17	52,57	49,16
Iteração 5	62,84	60,72	61,76
Média	**59,32**	**51,00**	**54,44**
Desvio Padrão	**8,36**	**6,18**	**5,04**
Tipo de Entidade LOC – Local			
Iteração 1	54,37	41,79	47,26
Iteração 2	64,62	62,69	63,64
Iteração 3	84,62	71,74	77,65
Iteração 4	39,42	69,23	50,23
Iteração 5	82,09	35,95	50,00
Média	**65,02**	**56,28**	**57,75**
Desvio Padrão	**19,00**	**16,36**	**12,82**
Tipo de Entidade MISC – Transações Bancárias			
Iteração 1	88,89	72,73	80,00
Iteração 2	23,53	40,00	29,63
Iteração 3	50,00	100,00	66,67
Iteração 4	76,52	57,52	65,67
Iteração 5	50,51	81,97	62,50
Média	**57,89**	**70,44**	**60,89**
Desvio Padrão	**25,52**	**22,94**	**18,72**

Fonte: Elaborado pelo Autor

Figura 14 – Entidades por validação.



Fonte: Elaborado pelo Autor

É possível observar que o desvio padrão das entidades dos tipos LOC e MISC tem valores bem mais altos que as entidades PER e ORG. Isto acontece por haver uma quantidade menor de entidades LOC e MISC, tanto nos RIFs da validação quanto nos RIFs do treinamento.

Essa baixa quantidade pode causar distorções na precisão e na revocação, apresentando ora resultados muito altos, ora muito baixos. Alguns RIFs nem mesmo possuem esse tipo de entidade.

Das 3.132 entidades anotadas nos 30 RIFs treinados, 1.307 são do tipo LOC, 997 são do tipo PER, 478 são do tipo LOC e 350 são do tipo MISC. Quanto maior a quantidade de entidades anotadas de um tipo, melhor a qualidade do corpus e maior a precisão das informações reconhecidas (ALLES, 2018).

4.2 RESULTADOS DOS TESTES

4.2.1 Resultados dos testes com o corpus RIF

De acordo com a metodologia definida na Seção 3.7, realizou-se os testes com os 5 RIFs reservados e anotados com o modelo treinado baseado no corpus RIF. Na Tabela 14 e na Figura 15 vemos o resultado geral para todos os tipos de entidades anotadas e os resultados detalhados por cada tipo de entidade.

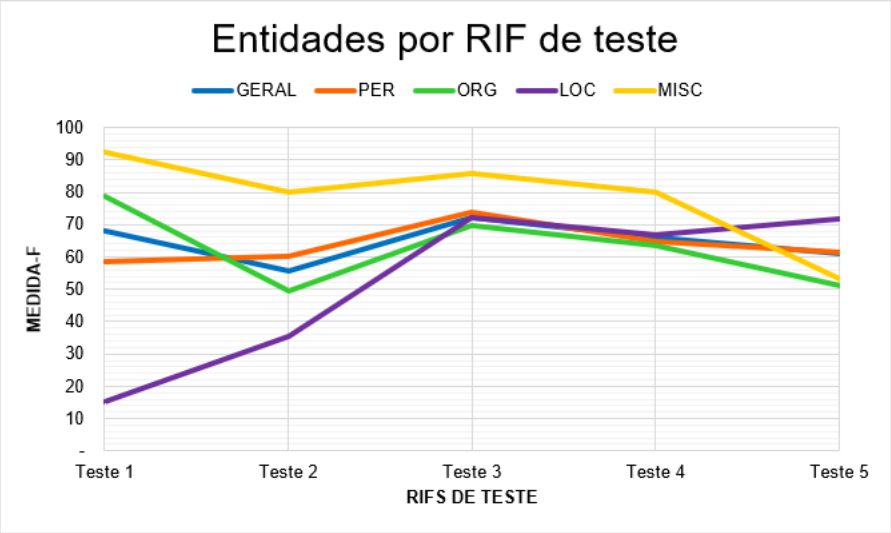
Tabela 14 - Resultados dos testes com o corpus RIF.

Testes	**Precisão**	**Revocação**	**Medida-F**
Teste RIF 1	69,53	66,53	67,94
Teste RIF 2	54,72	56,86	55,77
Teste RIF 3	73,89	70,73	72,27
Teste RIF 4	66,12	65,51	65,81
Teste RIF 5	66,27	56,70	61,11
Média	**66,11**	**63,27**	**64,58**
Desvio Padrão	**7,11**	**6,24**	**6,36**
Tipo de Entidade PER – Pessoa			
Teste RIF 1	66,67	52,46	58,72
Teste RIF 2	65,96	55,36	60,19
Teste RIF 3	76,09	71,43	73,68
Teste RIF 4	62,89	67,03	64,89
Teste RIF 5	68,97	55,56	61,54
Média	**68,12**	**60,37**	**63,80**
Desvio Padrão	**4,96**	**8,33**	**5,97**
Tipo de Entidade ORG – Organização			
Teste RIF 1	79,37	78,13	78,74
Teste RIF 2	40,91	62,07	49,32
Teste RIF 3	70,51	68,75	69,62
Teste RIF 4	63,77	63,46	63,61
Teste RIF 5	48,00	54,55	51,06

Média	**60,51**	**65,39**	**62,47**
Desvio Padrão	**15,87**	**8,74**	**12,45**
Tipo de Entidade LOC – Local			
Teste RIF 1	10,00	66,67	15,38
Teste RIF 2	33,33	37,50	35,29
Teste RIF 3	73,91	70,83	72,34
Teste RIF 4	68,09	65,31	66,67
Teste RIF 5	86,36	61,29	71,70
Média	**54,34**	**60,32**	**52,28**
Desvio Padrão	**31,64**	**13,20**	**25,68**
Tipo de Entidade MISC – Transações Bancárias			
Teste RIF 1	85,72	100,00	92,31
Teste RIF 2	100,00	66,67	80,00
Teste RIF 3	90,00	81,82	85,71
Teste RIF 4	86,67	74,29	80,00
Teste RIF 5	57,14	50,00	53,33
Média	**83,91**	**74,56**	**78,27**
Desvio Padrão	**16,00**	**18,48**	**14,84**

Fonte: Elaborado pelo Autor

Figura 15 – Entidades por RIF.



Fonte: Elaborado pelo Autor

Nos testes podemos observar que o desvio padrão é até pior que das validações. Isto pode ser explicado porque os testes são feitos com um RIF de cada vez, enquanto a validação foi realizada com um conjunto de 6 RIFs. Ou seja, menor quantidade de entidades anotadas quando se faz o reconhecimento com apenas um RIF, causando as distorções de se obter resultados muito altos ou muitos baixos.

Apesar disto, a média dos resultados nos testes é melhor do que a média dos resultados nas avaliações, conforme vemos na comparação feita na Tabela 15.

Tabela 15 - Comparação dos Resultados da Validação e dos Testes.

		Precisão	Revocação	Medida-F
Validação	Média	62,23	57,78	59,71
	Desvio Padrão	7,39	3,31	4,12
Testes	Média	66,11	63,27	64,58
	Desvio Padrão	7,11	6,24	6,36

Fonte: Elaborado pelo Autor

A melhora da média nos testes pode ser explicada porque na validação os modelos treinados continham 24 RIFs, para cada iteração. Já nos testes, o modelo foi treinado com os 30 RIFs anotados. Mais RIFs contidos no modelo com, consequentemente, mais entidades, ajudam a melhorar o reconhecimento. Conforme afirma Alles (2018, p. 24) que "A qualidade de um corpus depende do seu tamanho, ou seja, quanto mais treinamento, maior é a quantidade de textos anotados o que implica em informações extraídas com mais precisão".

4.2.2 Resultados dos testes com o modelo spaCy

Conforme as definições na Seção 3.7, também foram realizados os testes com os 5 RIFs reservados e anotados com o modelo em português do spaCy chamado pt_core_news_sm. A Tabela 16 e a Figura 16 apresentam o resultado geral para todos os tipos de entidades anotadas e, a seguir, detalhado por cada tipo de entidade.

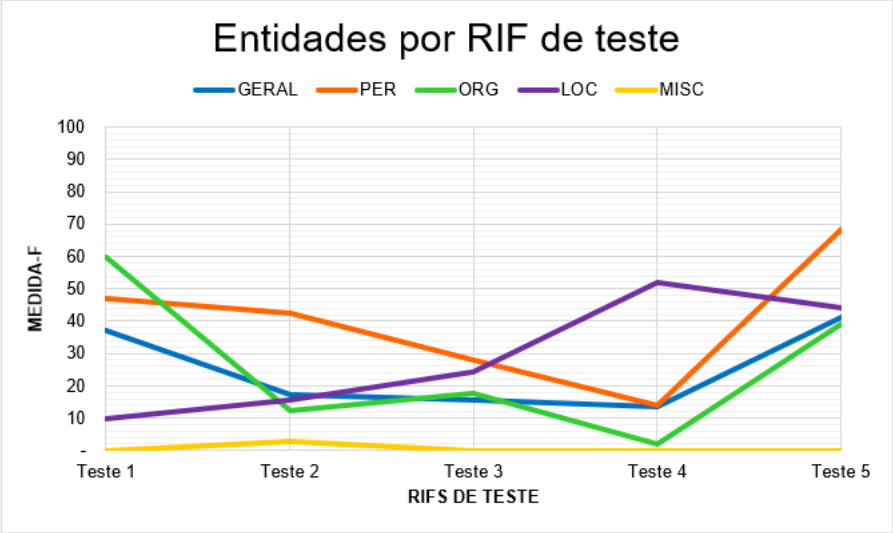
Tabela 16 - Resultados dos testes com o modelo spaCy português.

Testes	Precisão	Revocação	Medida-F
Teste RIF 1	29,44	50,75	37,26
Teste RIF 2	11,40	34,31	17,11
Teste RIF 3	11,55	23,17	15,42
Teste RIF 4	9,56	23,38	13,57
Teste RIF 5	31,35	59,79	41,13
Média	**18,66**	**38,28**	**24,90**
Desvio Padrão	**10,76**	**16,46**	**13,18**
Tipo de Entidade PER – Pessoa			
Teste RIF 1	48,28	45,90	47,06
Teste RIF 2	44,23	41,07	42,59
Teste RIF 3	27,45	28,57	28,00
Teste RIF 4	12,84	15,38	14,00
Teste RIF 5	67,57	69,44	68,49
Média	**40,07**	**40,07**	**40,03**
Desvio Padrão	**20,87**	**20,24**	**20,54**
Tipo de Entidade ORG – Organização			
Teste RIF 1	60,32	59,38	59,84

Teste RIF 2	8,70	20,69	12,24
Teste RIF 3	19,12	16,25	17,57
Teste RIF 4	1,69	1,92	1,80
Teste RIF 5	37,50	40,91	39,13
Média	**25,46**	**27,83**	**26,12**
Desvio Padrão	**23,70**	**22,49**	**23,26**
Tipo de Entidade LOC – Local			
Teste RIF 1	5,41	66,67	10,00
Teste RIF 2	9,30	50,00	15,69
Teste RIF 3	16,67	45,83	24,44
Teste RIF 4	37,39	84,69	51,88
Teste RIF 5	30,77	77,42	44,04
Média	**19,91**	**64,92**	**29,21**
Desvio Padrão	**13,76**	**16,86**	**18,09**
Tipo de Entidade MISC – Transações Bancárias			
Teste RIF 1	0,00	0,00	0,00
Teste RIF 2	1,40	22,22	2,63
Teste RIF 3	0,00	0,00	0,00
Teste RIF 4	0,00	0,00	0,00
Teste RIF 5	0,00	0,00	0,00
Média	**0,28**	**4,44**	**0,53**
Desvio Padrão	**0,63**	**9,94**	**1,18**

Fonte: Elaborado pelo Autor

Figura 16 – Entidades por RIF com o modelo spaCy português.



Fonte: Elaborado pelo Autor

Podemos verificar nestes testes que os resultados pioram consideravelmente. A média da medida-F dos 5 RIFs neste teste é de 24,90, contra 64,58 no teste com o modelo treinado com o corpus RIF. Temos que levar em consideração que esse modelo do spaCy se baseia num corpus muito mais amplo (SPACY, 2020), conforme explicado na Seção 3.3.1, e desconhece a estrutura de textos do RIF. Por isso a vantagem de se construir um corpus próprio para o RIF (ALLES, 2018), conforme foi destacado na Seção 3.4.

4.2.3 Resultados dos testes com os modelos RIF e spaCy juntos

Por fim, foram realizados os testes, da mesma forma que os anteriores, mas com o modelo treinado com a junção do corpus RIF e o modelo em português do spaCy, pt_core_news_sm. A Tabela 17 apresenta o resultado geral para todos os tipos de entidades anotadas e os resultados detalhados por cada tipo de entidade:

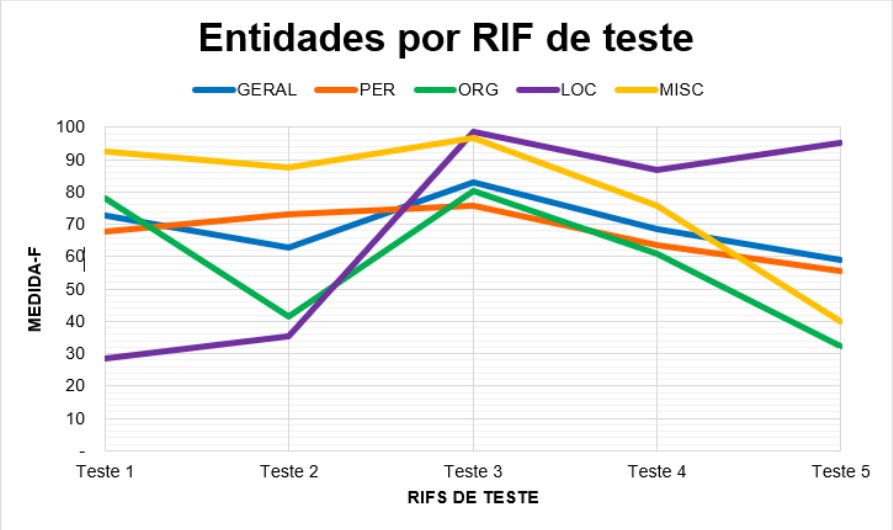
Tabela 17 - Resultados dos testes com o corpus RIF e modelo spaCy.

Testes	Precisão	Revocação	Medida-F
Teste RIF 1	73,85	71,64	72,73
Teste RIF 2	64,58	60,78	62,63
Teste RIF 3	83,03	83,00	83,01
Teste RIF 4	70,52	66,44	68,41
Teste RIF 5	56,60	61,86	59,11
Média	**69,72**	**68,74**	**69,18**
Desvio Padrão	**9,92**	**9,05**	**9,34**
Tipo de Entidade PER – Pessoa			
Teste RIF 1	68,33	67,21	67,77
Teste RIF 2	73,21	73,21	73,21
Teste RIF 3	73,53	78,03	75,72
Teste RIF 4	65,88	61,54	63,64
Teste RIF 5	62,07	50,00	55,38
Média	**68,60**	**66,00**	**67,14**
Desvio Padrão	**4,89**	**10,89**	**8,08**
Tipo de Entidade ORG – Organização			
Teste RIF 1	81,36	75,00	78,05
Teste RIF 2	45,83	37,93	41,51
Teste RIF 3	82,87	77,64	80,17
Teste RIF 4	61,88	60,10	60,98
Teste RIF 5	25,00	45,45	32,26
Média	**59,39**	**59,22**	**58,59**
Desvio Padrão	**24,53**	**17,55**	**21,42**
Tipo de Entidade LOC – Local			
Teste RIF 1	25,00	33,33	28,57
Teste RIF 2	33,33	37,50	35,29
Teste RIF 3	98,08	99,03	98,55
Teste RIF 4	91,01	82,65	86,63
Teste RIF 5	96,67	93,55	95,08
Média	**68,82**	**69,21**	**68,82**
Desvio Padrão	**36,41**	**31,45**	**34,04**
Tipo de Entidade MISC – Transações Bancárias			
Teste RIF 1	85,71	100,00	92,31
Teste RIF 2	100,00	77,77	87,50
Teste RIF 3	96,20	97,44	96,82
Teste RIF 4	80,65	71,43	75,76

Teste RIF 5	42,86	37,50	40,00
Média	**81,08**	**76,83**	**78,48**
Desvio Padrão	**22,74**	**25,19**	**22,90**

Fonte: Elaborado pelo Autor

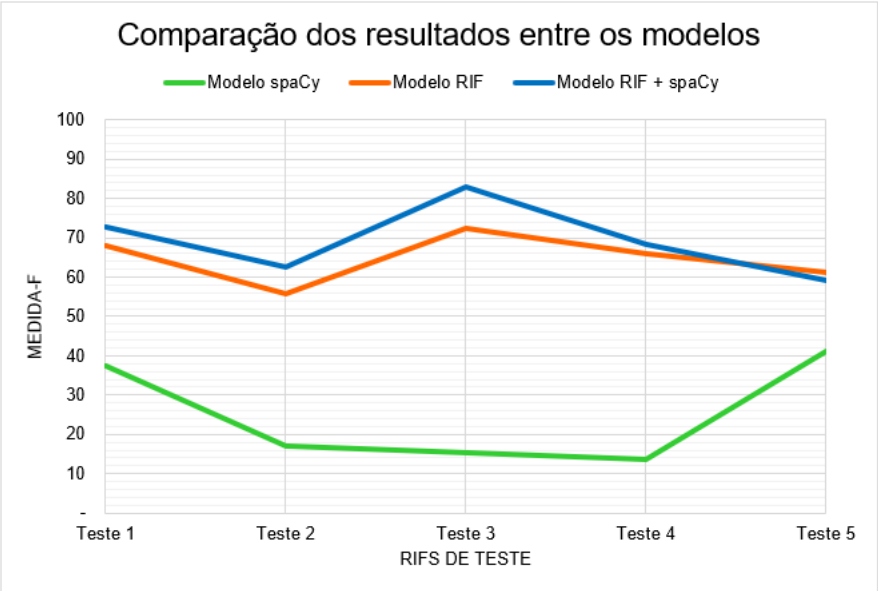
Figura 17 – Entidades por RIF com o corpus RIF e modelo spaCy.



Fonte: Elaborado pelo Autor

Com essa configuração de teste, treinando os dois modelos em conjunto, obtivemos os melhores resultados, em geral. Comparando com os resultados dos primeiros testes, com o modelo treinado com o corpus RIF, apenas um dos 5 RIFs teve piora da medida-F, o RIF do teste 5, que obteve 59,11 e anteriormente teve 61,11. Os demais RIFs tiveram melhora nos resultados, como podemos observar na Figura 18.

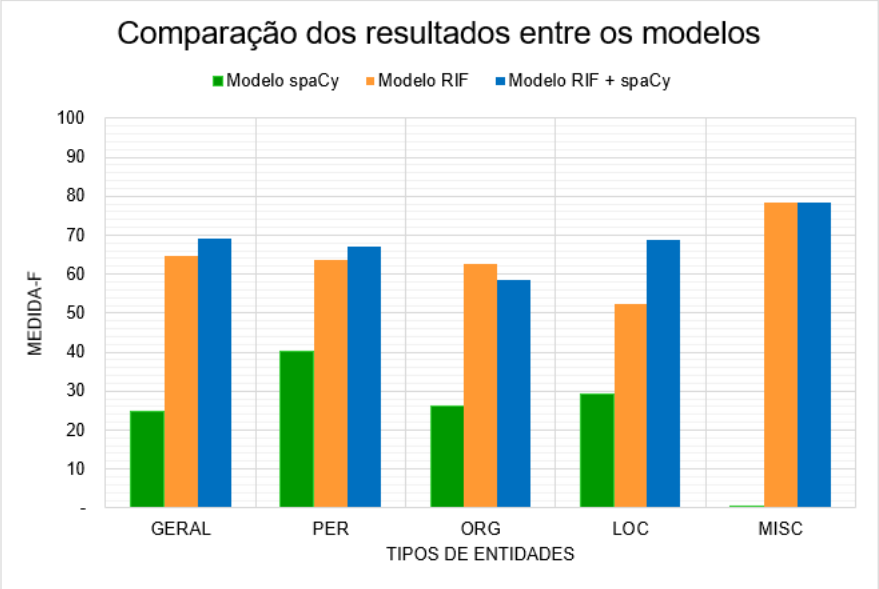
Figura 18 - Resultados por testes



Fonte: Elaborado pelo Autor

Analisando os tipos de entidades, podemos verificar que a única que teve pior resultado na sua média de medida-F foi o tipo ORG, que teve inicialmente 62,47 e nesses últimos testes conseguiu apenas 58,59. Se verificarmos de maneira analítica os demais resultados, de cada tipo, para cada RIF, observamos outros casos em que o resultado precisão, revocação ou medida-F pioram. Depende das especificidades de cada RIF, conforme as entidades anotadas e da estrutura do texto de cada um, conforme verificamos na Figura 19.

Figura 19 - Resultados por tipos



Fonte: Elaborado pelo Autor

Mas, em geral, os resultados demonstram que esses últimos testes tiveram um melhor desempenho, conforme comparação demonstrada na Tabela 18, onde vemos que que os modelos Corpus RIF, spaCy e RIF + spaCy tiveram, respectivamente, resultados de medida-F de 64,58%, 24,90% e 69,18%. A melhora dos resultados com este último modelo pode ser entendida que a união do corpus RIF, por ter termos mais semelhantes aos dos RIFs de teste, com o modelo do spaCy, mais amplo e diverso, foi uma melhor opção, aumentando em cerca de 5% a média da medida-F.

Na Tabela 18, podemos visualizar melhor a comparação entre os três conjuntos de testes:

Tabela 18 - Comparação dos resultados dos testes com os três modelos

Testes		**Precisão**	**Revocação**	**Medida-F**
Corpus RIF	Média	66,11	63,27	64,58
	Desvio Padrão	7,11	6,24	6,36
Modelo spaCy	Média	18,66	38,28	24,90
	Desvio Padrão	10,76	16,46	13,18
RIF + spaCy	Média	69,72	68,74	69,18
	Desvio Padrão	9,92	9,05	9,34

Fonte: Elaborado pelo Autor

5 CONCLUSÃO

A análise do Relatório de Inteligência Financeira – RIF, feita através de sua leitura por policiais federais especializados nessa tarefa, teria uma grande ajuda se esses documentos pudessem ser inicialmente analisados de forma automatizada e suas informações mais relevantes fossem detectadas. Isso traria uma maior agilidade e, reconhecendo corretamente essas informações, diminuiria os riscos de erro na análise subjetiva que o ser humano realiza, podendo este supervisionar e revisar o trabalho automatizado.

Por se tratar de um texto não estruturado, a extração de informações relevantes pode ser realizada através de técnicas de Mineração de Texto, Processamento de Linguagem Natural e Reconhecimento de Entidades Nomeadas.

Atendendo ao objetivo específico "Pesquisar trabalhos encontrados na literatura sobre reconhecimento de entidades em textos", foi realizada uma Revisão Sistemática da Literatura para encontrar os trabalhos que abordam o REN em português e que apresentavam resultados objetivos. Os trabalhos selecionados apresentaram resultados bastante diversos, indo de 25% a 97% de medida-F. É claro que se trata de uma comparação superficial, pois cada trabalho utilizou ferramentas de REN, corpora de dados e massas de testes diferentes.

Para esse trabalho foi utilizada a ferramenta spaCy para o REN e como corpus foi anotado um próprio, baseado em um conjunto de 30 RIFs fictícios, contendo 3.132 entidades anotadas, dos tipos PER, ORG, LOC e MISC. Também foram realizados testes com o modelo em português do spaCy, que é baseado nos corpora Bosque e WikiNer. A construção do corpus RIF cumpriu o objetivo específico de "Desenvolver um corpus baseado nas informações dos RIFS".

O treinamento do corpus RIF foi realizado com 200 iterações e, para a validação do modelo, foi realizada a Validação Cruzada K-fold, em 5 conjuntos de 6 RIFs, apresentando um resultado médio de medida-F de 59,71%. Já para os testes, foram submetidos outros 5 RIFs, contendo 926 entidades.

Inicialmente, os RIFs foram testados sendo submetidos ao modelo treinado do corpus RIF. Tendo um resultado médio de medida-F de 64,58%. Em seguida, foram submetidos ao modelo do spaCy, obtendo um resultado médio de 24,90%. Finalmente, foi feito um treinamento do corpus RIF em conjunto com o modelo em português do spaCy, com esse novo

modelo os resultados obtidos tiveram uma média de 69,18%. Diante dos resultados, pode-se entender que o objetivo específico, “Extrair informações dos relatórios de inteligência financeira usando técnicas e ferramentas de mineração de textos para o reconhecimento de entidades nomeadas”, foi alcançado.

Analisando esses resultados, podemos observar a melhora do resultado do primeiro teste, 64,58%, comparado ao da validação, 59,71%. Isso pode ser explicado pelo universo dos corpora utilizados na validação, com 24 RIFs a cada iteração. Já o corpus completo, com os 30 RIFs, utilizado no teste, possui uma maior quantidade de entidades anotadas, o que aumenta o universo de termos possíveis de serem reconhecidos.

Já o resultado com o teste com o corpus em português do spaCy teve um resultado bem pior, 24,90%. Isso pode ser explicado por se tratar um modelo que foi treinado a partir de outros tipos de textos, com uma estrutura que difere da estrutura dos RIFs. Mas, ao treinar esse modelo em conjunto com o corpus RIF, foi possível conseguir resultados bem mais promissores, com média de 69,18%, superior inclusive ao resultado do teste com o modelo do corpus RIF. E, dessa forma, também foi realizado o objetivo específico “Analisar a aplicação do corpus spaCy em reconhecimento de entidades nomeadas nos RIFs”.

Todos esses passos demonstram o objetivo específico de “Analisar as tecnologias de Mineração de Texto que permitem a leitura e o reconhecimento de entidades em português” e o objetivo geral “Desenvolver um corpus a partir de RIFs para o Reconhecimento de Entidades Nomeadas que permitam a leitura automatizada do RIF e a consequente detecção das entidades que serão tabuladas e utilizadas na análise do agente”, também foram alcançados.

Comparando com os resultados verificados na RSL, o melhor resultado pode ser considerado como mediano, mas, como já foi dito, essa comparação só poderia ser considerada válida se os testes fossem realizados nas mesmas condições, tanto dos termos da massa de teste, como utilizar os mesmos modelos de corpora aos quais foram submetidos. Para isso, seria necessário um trabalho mais amplo para testar as diversas ferramentas de REN e os diversos corpora.

Em relação à sua utilidade para o reconhecimento de entidades no trabalho de análise dos RIFs que é feito pela Polícia Federal, ao questionarmos se seria produtivo e confiável uma automatização para detecção de informações sensíveis e que podem desencadear ou não em investigações, inquéritos e até mesmo processos judiciais, entendemos que uma assertividade na faixa de 70% não é aceitável, mesmo havendo uma posterior revisão pelo analista. Mas também entendemos que a quantidade de RIFs utilizados na construção do corpus foi

relativamente pequena, sendo um limitador para obtermos melhores resultados pois, conforme pode ser observado nos testes, uma maior quantidade de RIFs e termos anotados no corpus aumentou o índice de assertividade.

Como explicado no trabalho, os RIFs utilizados para compor o corpus e a massa de testes, não foram RIFs reais, mas sim fictícios, por se tratar de informações sigilosas. Assim, podemos depreender que havendo uma construção de um corpus com uma quantidade bem mais significativa de RIFs, com treinamento em conjunto um corpus público, que daria mais amplitude ao modelo, com a inclusão periódica de novos RIFs e com a avaliação, a cada novo treinamento, para realizar os ajustes necessários, os resultados seriam melhores e teriam um crescimento contínuo até alcançar índices mais aceitáveis.

A partir deste trabalho, é possível desenvolver outros trabalhos futuros, com o objetivo de se obter um aumento no desempenho do REN. As opções para esses trabalhos são, construir um corpus maior baseado em RIFs, treinar um modelo a partir do corpus RIF em conjunto com outros corpora públicos, utilizar outras ferramentas de REN e implementar uma solução automatizada para leitura do RIF para utilização na Polícia Federal.

REFERÊNCIAS

ABREU, C. et al. Entity extraction within plain-text collections WISE 2013 challenge - T1: Entity Linking Track. **14th edition of the International Conference on Web Information System Engineering (WISE 2013)**, n. May 2014, 2013.

ALLES, V. J. **Construção de um corpus para extrair entidades nomeadas do Diário Oficial da União utilizando aprendizado supervisionado**. [s.l.] Brasília, DF, 2018.

ALLES, V. J.; GIOZZA, W. F.; ALBURQUERQUE, R. DE O. Processamento de Linguagem Natural para classificação de entidades nomeadas no Diário Oficial da União Brasileiro Natural Language Processing to classify named entities of the Brazilian Union Official Diary. **2018 13th Iberian Conference on Information Systems and Technologies (CISTI)**, p. 1–6, 2018.

ALMEIDA, D. P. DOS R. DE et al. Paradigmas Contemporâneos da Ciência da Informação: a recuperação da informação como ponto focal. **Revista Eletrônica Informação e Cognição (Cessada)**, v. 6, nº 1, 2007.

AMARAL, D. O. F. et al. Comparative Analysis of Portuguese Named Entities Recognition Tools. **Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014**, p. 2554–2558, 2014.

AMARAL, D. O. F. DO. **O Reconhecimento De Entidades Nomeadas Por Meio De Conditional Random Fields Para a Língua Portuguesa**. [s.l.] Pontifícia Universidade Católica do Rio Grande do Sul, 2013.

ARANHA, C. N.; VELLASCO, M. M. B. R.; PASSOS, E. P. L. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional**. [s.l.] Universidade Católica do Rio de Janeiro, 2007.

ARAS, V. Sistema nacional de combate à lavagem de dinheiro e de recuperação de ativos. **Revista Jus Navigandi, nº 1411, Teresina**, 2007.

BASTOS, P. DA S. Identificação de termos relevantes em relatórios usando text mining. **Repositório Aberto da Universidade do Porto**, 2017.

BAUMAN, Z. **Tempos líquidos**. Rio de Janeiro: Zahar, 2007.

BICK, E. Multi-level NER for Portuguese in a CG framework. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 2721, n. 1998, p. 118–125, 2003.

BORKO, H. Ciência da Informação: O que é isto? **American Documentation**, v. 19, n. 1, 1968.

BRAGA, J. T. DOS S. **Lavagem de dinheiro – Origem histórica, conceito e fases**. Disponível em: http://www.ambitojuridico.com.br/site/index.php?n_link=revista_artigos_leitura&artigo_id=8425. Acesso em: 12 jun. 2018.

BRASIL. **Lei da Lavagem de Dinheiro**, 1998. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/L9613.htm. Acesso em jun 2018.

BRUCKSCHEN, M.; VIEIRA, R.; RIGO, S. Named entities for hot topics ranking and ontology navigation aid. **Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, HT'09**, p. 373–374, 2009.

CAPURRO, R. Epistemologia e Ciência da Informação. **V Encontro Nacional de Pesquisa em Ciência da Informação**, 2003.

CAPURRO, R.; HJORLAND, B. O conceito de Informação. **Perspectivas em Ciências da Informação**, v. 12, n. 1, p. 148–207, 2007.

CARRILHO JUNIOR, J. R. **Desenvolvimento de uma Metodologia para Mineração de Textos**. [s.l.] Pontifícia Universidade Católica do Rio e Janeiro, 2007.

CHITICARIU, L. et al. **Domain adaptation of rule-based annotators for named-entity recognition tasks**. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. **Anais**...Cambridge, MA: Association for Computational Linguistics, 2010

COAF. **Relatório de Inteligência Financeira - RIF**. Disponível em: http://coaf.fazenda.gov.br/menu/a-inteligencia-financeira/relatorio-de-inteligencia-financeira-rif. Acesso em: 1 jul. 2020.

COLLOVINI, S.; MACHADO, G.; VIEIRA, R. **Extracting and Structuring Open Relations from Portuguese Text**. (J. Silva et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Cham: Springer International Publishing, 2016

CORBETT, P.; BOYLE, J. Chemlistem: chemical named entity recognition using recurrent neural networks. **Journal of Cheminformatics**, v. 10, n. 1, 2018.

CORDEIRO, A. D. **Gerador Inteligente de Sistemas com Auto-aprendizagem para Gestão de Informações e Conhecimento**. [s.l.] Universidade Federal de Santa Catarina, 2005.

CORRÊA, G. N.; MARCACINI, R. M.; REZENDE, S. O. Uso da mineração de textos na análise exploratória de artigos científicos. p. 36, 2012.

CURRÁS, E. Ontologias, taxonomia e tesauros em teoria de sistemas e sistemática. **Thesaurus Editora**, p. 24–25, 2010.

DUARTE, J. C.; MILIDIÚ, R. L. Machine learning algorithms for Portuguese named entity recognition. **Inteligencia Artificial**, v. 11, n. 36, p. 67–75, 2007.

ESTEVES, D. et al. Named Entity Recognition in Twitter Using Images and Text. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 10544 LNCS, p. 191–199, 2018.

FERNANDES, I. A. D. A Deep Learning Approach to Named Entity Recognition in Portuguese Texts. **Repositório Aberto da Universidade do Porto**, 2018.

FINATTO, M. J. B.; LOPES, L.; SILVA, A. C. Processamento de linguagem natural, linguística de corpus e estudos linguísticos: uma parceria bem-sucedida. **Domínios de Lingu@gem**, v. 9, n. 5, p. 41–59, 2015.

FLORESTA. **Projeto Floresta Sintá(c)tica**. Disponível em: https://www.linguateca.pt/Floresta/. Acesso em: 1 jul. 2020.

FONSECA, E. B.; VIEIRA, R.; VANIN, A. Adapting an Entity Centric Model for Portuguese coreference resolution. **Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016**, p. 150–154, 2016.

FONSECA, E. R. **Reconhecimento de implicação textual em português**. [s.l.] Universidade de São Paulo, 2018.

FORTE, A. C. B. Análise de comentários de clientes com o auxílio a técnicas de Text Mining para determinar o nível de (in)satisfação. 2015.

FREIRE, N. et al. A metadata geoparsing system for place name recognition and resolution in metadata records. **Proceedings of the ACM/IEEE Joint Conference on Digital Libraries**, p. 339–348, 2011.

FREITAS, C. et al. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. **Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010**, p. 3630–3637, 2010.

GAFI. **Padrões Internacionais de Combate à Lavagem de Dinheiro e ao Financiamento do Terrorismo e da Proliferação**, 1990. Disponível em: http://www.fatf-gafi.org/about/historyofthefatf/. Acesso em: 1 jul. 2020.

GOMES, R. M. **Mineração de Textos na Desambiguação de Sentido de Palavras Dirigida por Técnicas de Agrupamento sob o Enfoque da Mineração de Textos**. [s.l.] Rio de Janeiro, RJ, 2009.

GONÇALVES, L. **2019: o ano do ecossistema digital orientado por dados**, 2018.

GONZALEZ, M.; LIMA, V. L. S. DE. Recuperação de Informação e Processamento da Linguagem Natural. **XXIII Congresso da Sociedade Brasileira de Computação**, p. 3, 2003.

JASCHKE, H.-G. La ciencia policial – Enfoque europeo. **CEPOL - European Police College**, p. 10, 2005.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK, Keele**

University, v. 33, n. 2004, p. 1–26, 2004.

KITCHENHAM, B. A. et al. Preliminary Guidelines for Empirical Research in Software Engineering. **IEEE Transactions Software Engineering**, v. 28, n. 8, p. 721–734, 2002.

LANCASTER, F. W. **Information Retrieval systems: Characteristics, Testing and Evaluation**. New York, NY, USA: Wiley, 1968.

LE COADIC, Y.-F. A ciência da informação. **Brasília: Briquet de Lemos**, 2004.

LEFORT, V. M. N. **El Lavado de Dinero: Nuevo Problema para el Campo Jurídico**. [s.l.] Editorial Trillas, 1997.

LIDDY, E. D. **Natural Language Processing**. 2nd. ed. New York: Marcel Decker, Inc, 2001.

LIMA, J. L. O.; ALVARES, L. Organização da Informação e do conhecimento: conceitos, subsídios interdisciplinares e aplicações. **Revista Brasileira de Biblioteconomia e Documentação, São Paulo**, v. 8, n. 2, C, p. 1–21, 2012.

LUSTOSA, D. S. DE M. Aspectos gerais do crime de lavagem de dinheiro (Lei 9.613/98). **Âmbito Jurídico, Rio Grande, XII, n. 70**, 2009.

MACHADO, A. P. et al. Mineração de Texto em Redes Sociais Aplicada à Educação a Distância. **Colabor@ - A Revista Digital da CVA-RICESU, nº 23**, v. 6, 2010.

MANNING, C. et al. **The {S}tanford {C}ore{NLP} Natural Language Processing Toolkit**. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. **Anais**...Baltimore, Maryland: Association for Computational Linguistics, 2014

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008.

MARQUES, N. J. F. O papel do COAF no combate ao crime de lavagem de dinheiro. **Conteudo Juridico, Brasilia-DF**, 2014.

MARQUES, P. E. C. P. **Padrão De Financiamento À Pesquisa Em Dengue a Partir Do Diário Oficial Da União**. [s.l.] Instituto de Comunicação e Informação Científica e

Tecnológica em Saúde - ICICT, 2017.

MARTIN, M. Á. B.; AGUILAR, L. J. Los efectos de la globalización en el ámbito de la seguridad y defensa. **Inteligencia y Seguridad: Revista de análisis y prospectiva**, v. 10, p. 11–28, 2011.

MARTINS, C. A. et al. **Uma experiência em mineração de textos utilizando clustering probabilístico clustering hierárquico**. Disponível em: https://docplayer.com.br/120210562-Uma-experiencia-em-mineracao-de-textos-utilizando-clustering-probabilistico-clustering-hierarquico.html. Acesso em: 1 jun. 2019.

MATOS, P. F. et al. Relatório Técnico "Métricas de Avaliação". p. 16, 2009.

MENDRONI, M. B. As três fases do crime de Lavagem de Dinheiro. **Crime de Lavagem de Dinheiro. 3ª ed. São Paulo: Atlas**, 2015.

MICHALSKI, S. R.; CARBONELL, G. J.; MITCHELL, M. T. **Machine Learning an Artificial Intelligence Approach**. [s.l.] Springer Publishing Company, Incorporated, 2013.

MILIDIÚ, R.; DOS SANTOS, C.; DUARTE, J. Portuguese Corpus-Based Learning Using ETL. **J. Braz. Comp. Soc.**, v. 14, p. 17–27, 2008.

MORALES, P. D. A.; CÂNDIDO, A. C. Contribuições da ciência da informação para a afirmação da ciência policial: evidências na gestão do conhecimento. **Encontro Nacional de Pesquisa em Ciência da Informação**, v. n. XIX ENA, 2018.

MÔRO, D. K. **Reconhecimento de Entidades Nomeadas em Documentos de Língua Portuguesa**. [s.l.] Universidade Federal de Santa Catarina, 2018.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **LingvisticæInvestigationes**, v. 30, n. 1, p. 3–26, 2007.

OBREGON, S. R. D. G. P. Lavagem de Dinheiro. **Argumentum (UNIMAR)**, v. 1, p. 75–82, 2001.

OBSERVATÓRIO DE DADOS. **Observatório de dados/Precisão e revogação -**

Wikiversidade. Disponível em: https://pt.wikiversity.org/wiki/Observatório_de_dados/Precisão_e_revogação. Acesso em: 22 jun. 2020.

OLIVEIRA, D. M. DE et al. FS-NER : A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data. **WWW '13: 22nd International World Wide Web Conference**, p. 597–604, 2013.

ONU. **CONVENÇÃO DE 20 DE DEZEMBRO DE 1988 Convenção das Nações Unidas contra o Tráfico Ilícito de Estupefacientes e Substâncias Psicotrópicas, Viena - Áustria**, 1988. Disponível em: https://www.iberred.org/pt/convenios-penal/convencao-de-20-de-dezembro-de-1988-convencao-das-nacoes-unidas-contra-o-trafico. Acesso em: 1 jul. 2020.

PEREIRA, E. DA S. **Introdução às Ciências Policiais: a Polícia entre Ciência e Política**. São Paulo: Almedina, 2015.

PERES, R.; ESTEVES, D.; MAHESHWARI, G. Bidirectional LSTM with a context input window for named entity recognition in tweets. **Proceedings of the Knowledge Capture Conference, K-CAP 2017**, n. May, p. 2–5, 2017.

PEZZINI, A. Mineração De Textos: Conceito, Processo E Aplicações. **Revista Eletrônica do Alto Vale do itajaí**, v. 5, n. 8, p. 058–061, 2017.

PINHEIRO, L. V. R. Informação - Esse Obscuro Objeto da Ciência da Informação. **MORPHEUS: Revista eletrônica em Ciências Humanas Informação e Sociedade, Rio de Janeiro, UNIRIO**, v. 02, n. 04, 2004.

PINTO, E. Lavagem de Capitais e Paraísos Fiscais. São Paulo: Atlas. p. 50, 2007.

PIRES, A. R. O. Named entity extraction from Portuguese web text. **Repositório Aberto da Universidade do Porto**, 2017.

POLÍCIA FEDERAL. **Manual Prático de Combate à Lavagem de Dinheiro e aos Crimes Financeiros (reservado)** Brasília - DF, 2013.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. **Revista de Sistemas de**

Informação da FSMA, n. 7, p. 7–21, 2011.

ROMANTINI, G. L. O Desenvolvimento Institucional Do Combate À Lavagem De Dinheiro No Brasil Desde a Lei 9.613/98. p. 1–234, 2003.

SANTOS, C. N. DOS; MILIDIÚ, R. L.; RENTERÍA, R. P. **Portuguese Part-of-Speech Tagging Using Entropy Guided Transformation Learning**. (A. Teixeira et al., Eds.)Computational Processing of the Portuguese Language. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2008

SANTOS, R. E. S. et al. Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. **Revista de Sistemas e Computação**, v. 4, n. 2, p. 116–125, 2014.

SARKAR, D. **Text Analytics with Python**. Bangalore: Allite Books, 2016.

SCHREIBER, J. N. C. et al. Técnicas de Validação de Dados para Sistemas Inteligentes: Uma Abordagem do Software SDBayes. **XVII Colóquio Internacional de Gestão Universitária**, p. 18, 2017.

SERAPIÃO, P. R. B.; SUZUKI, K. M. F.; MARQUES, P. M. DE A. Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia. **Radiologia Brasileira**, v. 43, n. 2, p. 103–107, 2010.

SILVA, L. H.; CASELI, H. M. **Reconhecimento de entidades nomeadas em textos em português do Brasil no domínio do e-commerce**. Anais do IV Student Workshop on Information and Human Language Technology. **Anais**...2015Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/tilic/2015/010.pdf. Acesso em: 1 jul. 2020.

SOARES, F. A. **Mineração de Textos na Coleta Inteligente de Dados na Web**. [s.l.] Rio de Janeiro, RJ, 2008.

SOARES, V. DA S. Mineração de Textos para Identificar Perfis de Satisfação de Clientes. p. 56, 2016.

SOLORIO, T. Exploiting Named Entity Taggers in a Second Language. n. June, p. 6, 2005.

SOLORIO, T.; LOPEZ-LOPEZ, A. **Learning Named Entity Recognition in Portuguese from Spanish**. 2005

SOUZA, E. N. P.; CLARO, D. B. Extração de relações utilizando features diferenciadas para português. **Linguamática**, v. 6, n. 2, p. 57–65, 2014.

SPACY. **spaCy - Everything you need to know**. Disponível em: https://spacy.io/usage/spacy-101. Acesso em: 17 jun. 2020.

TEIXEIRA, J.; SARMENTO, L.; OLIVEIRA, E. A Bootstrapping Approach for Training a NER with Conditional Random Fields. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [s.l: s.n.]. v. 7026 LNAIp. 664–678.

VIEIRA, J. P. A. **Análise de Métodos de Extração de Aspectos em Opiniões Regulares**. [s.l.] Universidade Federal do Piauí, 2018.

WIVES, L. K. **Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva**. [s.l.] Universidade Federal do Rio Grande do Sul, 2002.

ZAINA, R. **Identifição de Entidades Destaque na Análise de Relatórios de Inteligência Financeira**. [s.l.] Universidade Federal de Santa Catarina, 2020.

ZHU, Q. et al. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. **Bioinformatics**, v. 34, n. 9, p. 1547–1554, 2017.

APÊNDICE A – Script de Treinamento

```
import random
import spacy

################### Train Spacy NER.############
from pandas._libs import json
from spacy.gold import GoldParse
from spacy.scorer import Scorer

def convert_dataturks_to_spacy(dataturks_JSON_FilePath):
    try:
        training_data = []
        lines = []
        with open(dataturks_JSON_FilePath, 'r', encoding = "utf-8") as f:
            lines = f.readlines()
        for line in lines:
            data = json.loads(line)
            text = data['content']
            entities = []
            for annotation in data['annotation']:
                point = annotation['points'][0]
                labels = annotation['label']
                if not isinstance(labels, list):
                    labels = [labels]
                for label in labels:
                    entities.append((point['start'], point['end'] + 1, label))
            training_data.append((text, {"entities": entities}))
        return training_data
    except Exception as e:
        logging.exception("Unable to process " + dataturks_JSON_FilePath + "\n" + "error =
" + str(e))
        return None

def evaluate(ner_model, examples):
    scorer = Scorer()
    for input_, annot in examples:
        doc_gold_text = ner_model.make_doc(input_)
        gold = GoldParse(doc_gold_text, entities=annot['entities'])
        pred_value = ner_model(input_)
        scorer.score(pred_value, gold)
    return scorer.scores

global ner

TRAIN_DATA = convert_dataturks_to_spacy("RIFS_todos.json")

nlp = spacy.blank('pt')  # create blank Language class
if 'ner' not in nlp.pipe_names:
    ner = nlp.create_pipe('ner')
    nlp.add_pipe(ner, last=True)
else:
        ner = nlp.get_pipe("ner")

for _, annotations in TRAIN_DATA:
    for ent in annotations.get('entities'):
        ner.add_label(ent[2])
```

```
other_pipes = [pipe for pipe in nlp.pipe_names if pipe != 'ner']
with nlp.disable_pipes(*other_pipes):
    optimizer = nlp.begin_training()
    for itn in range(150):
        print("Statring iteration " + str(itn))
        random.shuffle(TRAIN_DATA)
        losses = {}
        for text, annotations in TRAIN_DATA:
            nlp.update(
                [text],
                [annotations],
                drop=0.5,
                sgd=optimizer,
                losses=losses)
        print(losses)

nlp.to_disk("modelo_rif_todos")
```

APÊNDICE B – Script de REN

```
import spacy

from pandas._libs import json
from spacy.gold import GoldParse
from spacy.scorer import Scorer
from tika import parser

def evaluate(ner_model, examples):
    scorer = Scorer()
    for input_, annot in examples:
        doc_gold_text = ner_model.make_doc(input_)
        gold = GoldParse(doc_gold_text, entities=annot)
        pred_value = ner_model(input_)
        scorer.score(pred_value, gold)
    return scorer.scores

def convert_dataturks_to_spacy(dataturks_JSON_FilePath):
    try:
        training_data = []
        lines = []
        with open(dataturks_JSON_FilePath, 'r', encoding = "utf-8") as f:
            lines = f.readlines()

        for line in lines:
            data = json.loads(line)
            text = data['content']
            entities = []
            for annotation in data['annotation']:
                point = annotation['points'][0]
                labels = annotation['label']
                if not isinstance(labels, list):
                    labels = [labels]

                for label in labels:
                    entities.append((point['start'], point['end'] + 1, label))

            training_data.append((text, {"entities": entities}))
        return training_data
    except Exception as e:
        logging.exception("Unable to process " + dataturks_JSON_FilePath + "\n" + "error =
" + str(e))
        return None

def lerPDF(arquivoPDF):
    raw = parser.from_file(arquivoPDF)
    conteudo=(raw['content'])
    return conteudo

listaArquivos=['Rif_93849.pdf']
stringSaida=''

arquivos=[]
newline=chr(10)
for arq in range(len(listaArquivos)):
    arquivoPDF = open(listaArquivos[arq], 'rb')
    string_temp=lerPDF(arquivoPDF)
    string_temp = string_temp.replace('(cid:160)', ' ')
```

```
    string_temp = string_temp.replace(' iiii', 'i')
    string_temp = string_temp.replace('iiii ', 'i')
    string_temp = string_temp.replace('iiii', 'i')
    string_temp = string_temp.replace(chr(160), chr(32))
    string_temp = string_temp.replace(newline, chr(32))
    string_temp = string_temp.replace(chr(12), chr(32))
    string_temp = string_temp.replace('  ', ' ')
    string_temp = string_temp.replace('  ', ' ')
    string_temp = string_temp.replace('  ', ' ')
    stringSaida = stringSaida + string_temp
    arquivos.append(string_temp)
    arquivoPDF.close()

text = stringSaida

nlp = spacy.load('modelo_rif_todos_pt')

doc = nlp(text)

i = 0
d = set()
for entity in doc.ents:
    print(entity.text, entity.label_)

entidades = convert_dataturks_to_spacy("Rif_93849.json")

for item in range(tamanho):
    ent_input = entidades[item][0]
    dictlist = entidades[item][1]['entities']
    print(dictlist)
    entidades2.append((ent_input, dictlist))

results = evaluate(nlp, entidades2)

print(results)
```