UNIVERSIDADE FUMEC
FACULDADE DE CIÊNCIAS EMPRESARIAIS
MESTRADO PROFISSIONAL EM SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO

Hector Matheus Soares Vieira

O USO DA MINERAÇÃO DE TEXTOS PARA O INCREMENTO DA SEGURANÇA DENTRO DE SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO DA ÁREA FINANCEIRA

Belo Horizonte
2020

Hector Matheus Soares Vieira

O USO DA MINERAÇÃO DE TEXTOS PARA O INCREMENTO DA SEGURANÇA DENTRO DE SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO DA ÁREA FINANCEIRA

Dissertação apresentada ao Curso de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC, como parte dos requisitos para obtenção do título de mestre.

Área de Concentração: Sistemas de Informação.

Linha de pesquisa: Tecnologia e Sistemas de Informação.

Orientador: Prof. Dr. Luiz Cláudio Gomes Maia.

Belo Horizonte
2020

Dados Internacionais de Catalogação na Publicação (CIP)

V658u Vieira, Hector Matheus Soares, 1990-

O uso da mineração de textos para o incremento da segurança dentro de sistemas de recuperação da informação da área financeira / Hector Matheus Soares Vieira. - Belo Horizonte, 2020.

84 f. : il.

Orientador: Luiz Cláudio Gomes Maia

Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento), Universidade FUMEC, Faculdade de Ciências Empresariais, Belo Horizonte, 2020.

1. Aprendizagem. 2. Processamento de linguagem natural (Computação). 3. Recuperação da informação. I. Título. II. Maia, Luiz Cláudio Gomes. III. Universidade FUMEC, Faculdade de Ciências Empresariais.

CDU: 65.011:681.3.6

Ficha catalográfica elaborada pela Biblioteca Universitária-FUMEC

Dissertação intitulada "**O uso da mineração de textos para o incremento da segurança dentro de sistemas de recuperação da informação da área financeira**" de autoria de Hector Matheus Soares Vieira, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Luiz Cláudio Gomes Maia – Universidade FUMEC
(Orientador)

Prof. Dr. Fernando Silva Parreiras – Universidade FUMEC
(Examinador Interno)

Prof. Dr. Rodrigo Moreno Marques – UFMG
(Examinador Externo)

Prof. Dr. Fernando Silva Parreiras
Coordenador do Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC

Belo Horizonte, 26 de outubro de 2020.

Luiz Maia.

Fernando Silva Parreiras

Rodrigo Moreno Marques

ABARIS

REQUESTED

TITLE	Assinatura de ata e contra-capas Universidade
FILE NAME	6da6b905-58bd-4a63-a607-76acc1799492.pdf
REQUEST ID	signature_request_ac5afacc-1144-4008-938e-0010a
REQUESTED BY	Karem Estefani Oliveira De Paula
STATUS	Completed

Professor (luiz.maia@fumec.br)

SENDED
12/02/2021
01:34:22UTC±0

SIGNED
19/02/2021
14:41:19UTC±0
191.185.140.62

Professor (fernando.parreiras@fumec.br)

SENDED
19/02/2021
14:41:19UTC±0

SIGNED
23/02/2021
20:58:57UTC±0
187.111.30.10

Professor (rodrigomorenomarques@yahoo.com.br)

SENDED
23/02/2021
20:58:57UTC±0

SIGNED
23/02/2021
21:08:24UTC±0
186.195.101.117

COMPLETED
23/02/2021
21:08:24 UTC±0
The document has been completed.

Assinado Por:
EVELYN FERNANDA DE LELIS
MOREIRA DE
FREITAS:03475835630
Validade: 15/06/2022
Emissor: AC LINK RFB v2
Data: 24/02/2021 09:03

Assinado Por:
FERNANDO SILVA
PARREIRAS:03073186646
Validade: 09/08/2021
Emissor: AC PRODEMGE RFB G4
Data: 25/02/2021 16:22

AGRADECIMENTOS

A todos os professores da universidade FUMEC, que me proveram, e continuam provendo, um conhecimento imensurável, em especial ao meu orientador Dr. Luiz Cláudio Gomes Maia, por toda a sua ajuda no desenvolvimento deste trabalho;

Aos meus pais, que tornaram este caminho possível ao acreditarem na educação como um caminho essencial;

À minha esposa Luciana, por todo apoio e compreensão durante esta jornada;

Aos meus irmãos, por todo seu companheirismo e incentivo;

A todos os colegas de estudo que estiveram presentes nesta jornada;

meus sinceros agradecimentos.

“O pior naufrágio é não partir.”

(Amyr Klink)

RESUMO

Processos de mineração de textos e processamento de linguagem natural são áreas de estudo difundidas dentro da ciência atual e com aplicações reais no cenário corporativo, sendo assim, o objetivo deste trabalho foi identificar qual dessas técnicas pode ser utilizada para incrementar a segurança documental em sistemas de recuperação da informação e também quais as melhores formas de isto ser feito de forma aplicada. Apresentando as condições dos sistemas em estudo, a pesquisa se propõe a avaliar como a combinação de algoritmos de inteligência artificial pode auxiliar no processo de controle de acesso aos documentos, avaliando qual a acurácia dos algoritmos selecionados de aprendizagem de máquina para identificar se o documento pertence ao usuário que está tentando acessá-lo. Resultados satisfatórios trouxeram a possibilidade de ser utilizado um módulo extra de segurança, a fim de se evitar o acesso a um documento restrito devido a um possível erro dentro de seus atributos e/ou metadados. A fim de ter uma base de comparação de resultados e métodos utilizados, foram levantados estudos anteriores que possuem afinidade com o tema, auxiliando, assim, a escolha de passos adotados nos futuros métodos e encontrado o estado da arte atual dentro de mineração de textos em documentos. Foram escolhidos dois tipos de algoritmos que pudessem fazer o processo de recuperar os atributos do documento para servir de base para o controle de acesso. Os algoritmos escolhidos foram os de *Support Vector Machine* (SVM) e *Bidirectional Long Short-Term Memory (*BiLSTM*)*. Foi realizada a aprendizagem de máquina dentro do *dataset*, levantado anteriormente, de documentos de um sistema de recuperação da informação pertencente à área financeira. Durante os testes, ficou evidente que o processo de extração *bag-of-words* se torna ineficaz, mas modelos utilizando conjuntos mais extensos de palavras foram capazes de resultados acima de 90%. Foi testada também a utilização de um segundo conjunto de documentos utilizado por pessoas distintas, porém, a mudança nos indicadores se mostrou muito tímida. Por fim, os testes realizados com a utilização do modelo BiLSTM obtiveram uma melhor acurácia, próxima dos 99%. Com estes resultados foi possível sugerir formas nas quais possa ser incrementada a segurança dos documentos com os usos dos métodos apresentados.

Palavras-chave: Processamento de linguagem natural. Aprendizagem de máquina. Aprendizagem Profunda. Segurança da Informação.

ABSTRACT

The processes of text mining and natural language processing are increasingly widespread within different uses on our current world, so the objective of work identifies which of these techniques can be used to increase the document security in information retrieval systems and also which is the best way that this can be achieved. This project brings a theoretical base that did intend to bring information on important topics on the topics of Artificial Intelligence, Metadata, Information Retrieval Systems, Information Security, Natural language processing. To have a basis to compare results and methods used were raised some previous studies that have some similarities with the topic, helping on this on choosing the steps adopted in future methods and thus finding the current state-of-the-art within document classifications. At this point 2 types of algorithms were chosen that could classify a document to serve as a basis for access control, the chosen algorithms were the Support Vector Machine SVM and Bidirectional Long Short-Term Memory BiLSTM. The learning process was executed on the dataset that were been previously collected from documents of a financial information retrieval system. During the tests. It was evident that the bag-of-words model becomes ineffective, but models using a larger number of words were able to obtain results above 90% of accuracy, the use of a different set of documents used by different people was also tested, but the change in the indicators proved to be small on results. After all, tests were made using the BiLSTM model obtained excellent accuracy above the mark of 99 %. With these results, it was possible to suggest ways where the security of documents can be increased with the uses of the presented methods.

Keywords: Natural language processing. Machine learning. Deep Learning. Information Security.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

ABAC	*Attribute-based access control*
BiLSTM	*Bidirecional LSTM*
CNN	*Convolutional Neural Network*
CSV	*Comma-separated values*
DL	*Deep Learning*
GPU	*Graphics Processing Unit*
HCP	*The Health Care Personnel*
IA	Inteligência Artificial
KDT	*Knowledge Discovered in Text*
LSTM	*Long-Story Short Memory*
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
NB	*Naïve Bayes*
NIST	*National Institute of Standards and Technology*
NLACP	*Natural Language Access Control Policy*
PPSIGC	Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento
RBAC	*Role-based access control*
RN	Redes neurais
RNE	Redes Neurais Convulsionais
SI	Sistema de Informação
SRI	Sistema de Recuperação da Informação
SVM	*Support Vector Machine*
XACML	*Extensible Access Control Markup Language*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto

Grandes volumes de documentos demandam sistemas próprios para que seja possível chegarem ao usuário de forma eficiente. Neste contexto, possuímos os Sistemas de Recuperação da Informação (SRI), que, entre outras funções, têm o objetivo de dar acesso à informação potencialmente contida nos documentos neles registrados (ARAUJO, 1994).

É possível enxergar que a sociedade utiliza o documento como uma forma de comunicação confiável. Nota-se, principalmente no ambiente formal, que há muita comunicação sob a forma de documentos, e, assim, dentro de sistemas de informações da área financeira, os documentos são amplamente utilizados dentro do processo de comunicação.

Tendo em vista que esse grande volume de documentos pode conter informações sensíveis e/ou pessoais, a segurança é um ponto crucial para a confiança do usuário dentro dos sistemas de recuperação da informação. Nesse sentido, não se deve medir esforços para aumentar a segurança provida, sobretudo, em sistemas da área financeira, cujos documentos podem conter informações muito sensíveis ao negócio.

Sistemas de informações têm origem que remonta as bibliotecas de Terracota e de Alexandria (ARAUJO, 1994), e, caso bibliotecas físicas tivessem a necessidade de restringir o acesso a alguma informação, era necessário criar áreas reservadas para tal. Dentro do meio digital, esse controle é feito de uma forma mais eficiente, podendo o sistema possuir diversos grupos com acesso a documentos específicos, ou mesmo um controle de acesso baseado em atributos, em que o atributo dos usuários e do documento irá determinar seu acesso.

Esta informação contida em documentos possui um valor definido por Cronin (1990) como sendo o valor da informação mensurado tomando como base alguns pontos, conforme explicitado abaixo:

a) Valor de uso: baseia-se na utilização final que se fará com a informação;
b) Valor de troca: é aquele que o usuário está preparado para pagar; esse valor varia de acordo com as leis de oferta e demanda, podendo também ser denominado valor de mercado;
c) Valor de propriedade: reflete o custo substitutivo de um bem;
d) Valor de restrição: surge no caso de informação secreta ou de interesse comercial, quando o uso fica restrito apenas a algumas pessoas.

Com isto explicitado, percebe-se que trabalhar na segurança do documento pode evitar que este valor se perca como ocorre no valor da restrição, ou mesmo que esse valor seja transferido de forma não autorizada para outra pessoa.

Com a quantidade de documentos gerados em expansão, o estudo em SRI tem se ampliado, juntamente com estudos em novas frentes. O presente trabalho busca explorar estudos com aprendizagem de máquina aplicados em SRI.

A capacidade computacional tem se expandido com o passar do tempo. A Lei de Moore diz que essa capacidade dobra a cada 2 anos. Em paralelo a este crescimento, a aprendizagem de máquina se torna comum e disponível para ser utilizada dentro de novos cenários, podendo melhorar diversos processos. Um exemplo é a Netflix, que consegue prever quais filmes as pessoas gostariam de assistir e fazendo recomendações, ou o Google, que é capaz de identificar o que a pessoa deseja saber com base em seu histórico de busca (BEAM; KOHANE, 2018). Dessa forma é possível atualmente encontrar pesquisas sobre diversos sistemas de recomendação além dos filmes, como as músicas (CHEN *et al.*, 2019) e os livros (PARK *et al.*, 2018).

Aplicações de redes neurais têm ajudado médicos em diagnósticos com o uso de segmentação de imagens, separando partes com tumores de imagens de cérebros obtidas através de ressonância magnética (HAVAEI *et al.*, 2017; MYRONENKO, 2018; ZHOU 2020), assim como também a segmentação de órgãos e lesões (GU *et al.*, 2018; JHA *et al.*, 2020; FAN *et al.*, 2020).

Dentro da área de processamento de linguagem natural isso também ocorre, através do uso de algoritmos de aprendizagem profundos treinados que são utilizados no processo de tradução de sentenças de um determinado idioma para outro (SENNRICH; HADDOW; BIRCH, 2016; RIKTERS; PINNIS; KRIŠLAUKS, 2018; EDUNOV *et al.*, 2018).

1.2 Lacuna a ser explorada

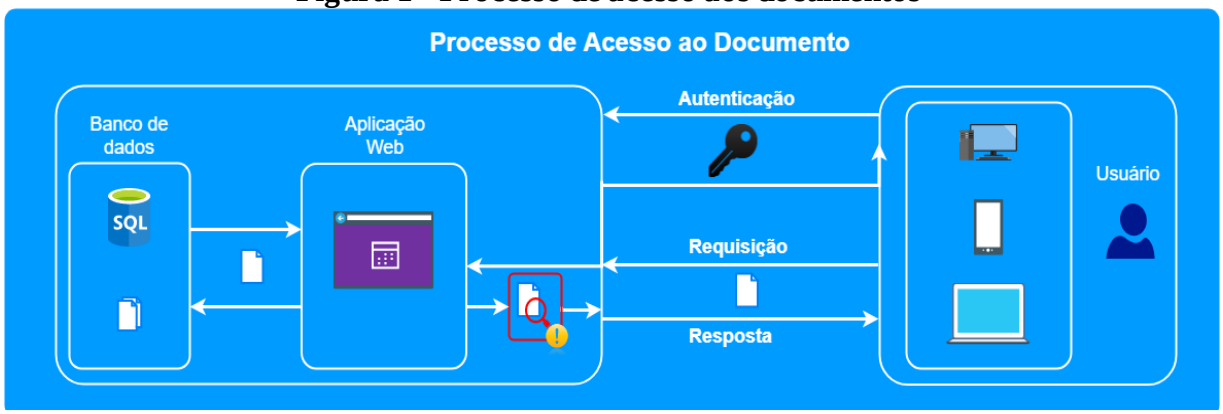
A pesquisa busca explorar uma lacuna de uso dos sistemas de classificação de documento, com o uso de inteligência artificial de forma aplicada, como forma de incremento da segurança em sistema de recuperação da informação. Propõe-se a avaliar a capacidade desses métodos de classificação, assim como possíveis limitações para seu uso.

Assim, esta pesquisa tem o foco em aumentar a segurança de documentos, alterando o processo padrão de acesso a eles, e procurando descobrir se é possível utilizar algoritmos de classificação de documentos para recuperar informações utilizadas no processo de controle de

acesso, fornecendo, assim, uma visão semântica. Desta forma será possível definir uma camada de segurança que não se baseará em uma fonte diferente de dado. Sendo assim possível restringir o acesso a um documento mesmo que classificado de forma errônea pelo usuário.

Durante o processo padrão de autenticação, o usuário faz a autenticação do sistema com o envio de seu usuário e senha, e, assim que o sistema o identifica, irá retornar as informações que seu dispositivo irá utilizar para se autenticar. Durante a navegação no sistema, o usuário poderá fazer a requisição de um arquivo que esteja disponível; os documentos disponíveis são baseados em atributos que o usuário e o documento possuem em comum, como fundo e investidor. No processo normal, esse documento é enviado diretamente ao usuário x, e na proposta desta pesquisa há uma segunda camada de validação antes do documento ser enviado ao usuário, porém esta verificação levará em conta os atributos do usuário, mas de maneira automática irá classificar com base em seu conteúdo, com o uso de inteligência artificial. Dessa forma, será criada esta segunda forma de validação, que não depende somente da classificação feita pelo usuário previamente, conforme processo que pode ser observado na Figura 1.

Figura 1 - Processo de acesso aos documentos



Fonte: Elaborada pelo autor

1.3 Problema

Dentro do cenário em que sistemas de recuperação da informação possuem sob seu controle acesso a documentos com base em atributos, e estes atributos são definidos por usuários, uma má classificação poderia colocar em risco a confidencialidade do documento. Camadas de segurança e criptografias podem ser anuladas, caso os dados informados para os sistemas tenham sido inseridos de forma errônea. Como exemplo, um documento de um investidor mal categorizado e exibido para outro poderá causar danos tanto ao investidor original, por ter dados pessoais expostos a outrem, quanto para o sistema – que se tornará menos confiável.

O caso exemplificado pode parecer distante, porém, a empresa de consultoria Delloite (2012) classificou erros e omissões de funcionários como uma das três principais ameaças para o setor financeiro em 2011. Isto deixa claro que erros por parte do usuário ocorrem, e isto é uma ameaça possível à segurança dos dados.

Para Valentim e Ançanelo (2018), a informação é, em si, uma mercadoria, considerando o fato de que ela pode ser negociada como qualquer outro produto. Sendo assim, qualquer vazamento de informação implica diretamente em perda de capital. É preciso trabalhar, então, para que se diminuam ao máximo os riscos envolvidos no processo de recuperação da informação.

Com as informações apresentadas, a presente pesquisa busca avaliar o uso de métodos de inteligência artificial para identificar atributos utilizados como controle de acesso, de forma automática, e usar esta informação para melhorar o controle de acesso aos documentos, possibilitando, desta forma, a criação de um sistema mais seguro em que os erros do usuário poderão ter menores impactos, trazendo mais confiabilidade.

1.3.1 Pergunta

Quais os ganhos e as limitações do uso das técnicas atuais da mineração de texto com aprendizado de máquina visando incrementar a segurança dos documentos dentro de sistemas de recuperação da informação da área financeira?

1.4 Objetivo Geral

Esta pesquisa tem o objetivo de avaliar a viabilidade do uso de mineração de textos e aprendizado de máquina na otimização do processo de controle de acesso de documentos dentro de sistemas de recuperação da informação pertencentes à área financeira.

1.4.1 Objetivos específicos

A fim de alcançar o objetivo principal deste trabalho, foram definidos os seguintes objetivos específicos:

a) Identificar métodos capazes de avaliar o conteúdo do documento para a análise de permissão;
b) Evidenciar o desempenho dos métodos selecionados, incluindo performance e acurácia;

c) Comparar resultados obtidos para identificação de possível uso e possíveis formas de implementação.

1.5 Justificativa

Observa-se que a segurança é essencial em *softwares* financeiros, e o controle de acesso é parte fundamental da segurança da informação, levando-nos a supor que sistemas mais seguros tendem a ser mais atrativos aos olhos de possíveis compradores, assim como garantem a privacidade dos envolvidos.

Documentos dentro desses sistemas podem conter informações sensíveis, que causariam grandes danos financeiros e à privacidade diante de um possível vazamento ou acesso não autorizado. Sendo assim, a proteção dos documentos contidos em sistemas financeiros representa um investimento com relevante potencial de retorno.

O documento digital oferece inúmeros benefícios para sua manutenção, como a redução de espaço físico, o acesso à informação, a segurança do documento contra a ação de agentes externos como cupins e traças, além de situações imprevistas como inundações, incêndios, entre outros.

Porém, proporcional ao volume de documentos, há um aumento de problemas a serem resolvidos, como a dificuldade de achar a informação requerida, a organização e o tratamento da informação e o controle de acesso. Diante dessas dificuldades, os sistemas de recuperação de informação devem continuamente aumentar sua maturidade e funcionalidade.

Para isso, contamos com processos de classificação, estes são mais antigos do que o próprio sistema de recuperação da informação, que vêm evoluindo há anos em conjunto com a classificação de livros em biblioteca. Para Barbosa, a classificação é definida como "Um processo mental pelo qual coisas, seres ou pensamento, são reunidos segundo as semelhanças ou diferenças que apresentam." (BARBOSA, 1969, p. 13).

Sendo assim, a classificação é fundamental para qualquer sistema de recuperação da informação, desde bibliotecas até sistemas atuais. Com o uso da classificação é possível ao usuário acessar e recuperar, com agilidade, o documento desejado. Nesse sentido, um Sistema de Recuperação da Informação (SRI) que não é capaz de entregar o documento que o usuário deseja é um SRI ineficiente.

Dentro da classificação dos documentos, os metadados são dados sobre o documento que não necessariamente estão contidos em seu conteúdo. Estes são bases da classificação em certos sistemas de recuperação da informação, auxiliando na interoperabilidade dos sistemas e

fornecendo ao usuário detalhes sobre a informação, ou criando uma listagem filtrada de documentos, o que contribui para a rápida recuperação da informação e também para o controle de acesso deste documento.

Os metadados dos arquivos podem ser utilizados pelos sistemas de recuperação da informação para o controle da permissão do usuário aos documentos, principalmente em questão de visualização, em que se pode observar no *dataset* a ser utilizado, por exemplo, a divisão de fundos de investimento.

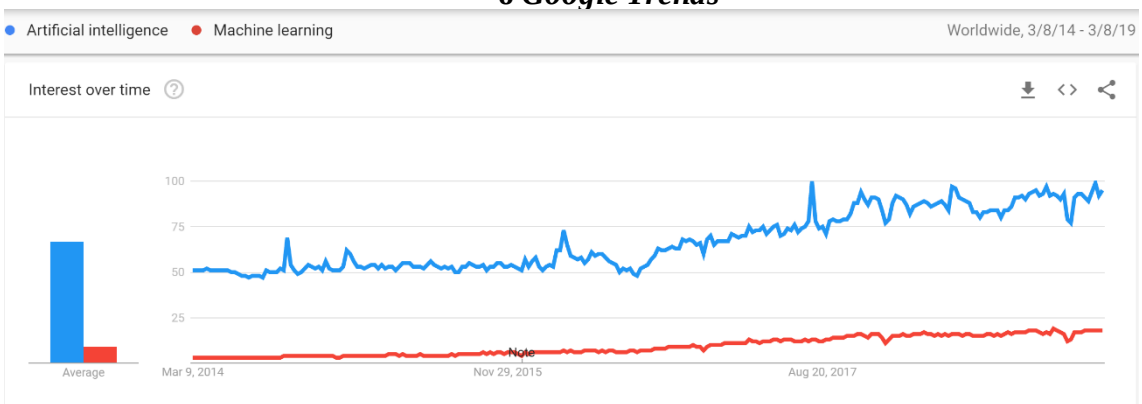
Esse método de controle de permissão, no entanto, é definido pelo usuário quando ele preenche as informações de metadados, deixando uma possível vulnerabilidade a erro durante o preenchimento feito pelo usuário, que poderia conceder acesso indevido a documentos de caráter confidencial.

Porém, paralelamente ao crescimento, em volume, dos documentos digitais, tem-se o aumento da capacidade computacional, trazendo a capacidade de se trabalhar com os conceitos de inteligência artificial – conceito antigo que está em amplo crescimento.

Com algoritmos baseados em inteligência artificial e a evolução da capacidade computacional, tem-se conseguido soluções que antes não eram possíveis, além de melhorias nas já existentes e a promoção de outras opções para o uso.

O interesse por inteligência artificial e aprendizado de máquina vem crescendo nos últimos anos, como é possível observar na Figura 2, que demonstra o interesse em pesquisas realizadas no Google em termos como *Artificial Inteligence* e *Machine Learning*. Com o passar do tempo, o termo, criado por John McCarthy em 1956, foi difundido no mundo da computação.

Figura 2 - Interesse nos tópicos *Artificial Intelligence* e *Machine Learning* de acordo com o *Google Trends*



Fonte: Google Trends.

Apresentando as condições dos sistemas em estudo, a pesquisa se propõe a avaliar como a combinação de algoritmos de inteligência artificial pode auxiliar no processo de controle de acesso aos documentos, avaliando qual a acurácia dos algoritmos posteriormente selecionados de aprendizagem de máquina para identificar se o documento pertence ao usuário que está tentando acessá-lo. Resultados satisfatórios podem trazer a possibilidade de ser utilizado como um módulo extra de segurança, a fim de se evitar o acesso a um documento restrito devido a um erro dentro de seus metadados.

1.6 Aderência ao Objetivo de Pesquisa do Programa

A atual dissertação foi apresentada ao Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento (PPSIGC), o qual possui o objetivo, entre outros, de disseminar o conhecimento científico e tecnológico de natureza interdisciplinar por meio de suas linhas de pesquisas "Gestão da Informação e do Conhecimento" e "Tecnologia e Sistemas de Informação", estando a atual dissertação relacionada à segunda linha.

Referente à trilha, relaciona-se com a trilha TI Recuperação da Informação, na qual se explora o uso de ferramentas para a melhor eficiência de sistemas com o foco em recuperação da informação, dentre elas técnicas de inteligência artificial e processamento de linguagem natural.

Esta pesquisa busca ampliar o conhecimento acadêmico existente sobre o tema através dos experimentos realizados, assim como trazer uma visão adquirida de forma empírica – de possíveis melhorias na área profissional –, com o vasto conhecimento científico gerado através de pesquisas anteriores.

A pesquisa apresenta possíveis melhorias para sistemas de recuperação da informação da área financeira mediante os resultados encontrados durante a sua execução e analisando-os a partir da visão empírica em conjunto com o conhecimento acadêmico apresentado na fundamentação teórica.

1.7 Estrutura da Dissertação

O presente documento divide-se em quatro principais capítulos: Introdução, Fundamentação Teórica e Metodologia, Execução e Considerações Finais.

A introdução busca trazer o contexto que será explorado adiante. Nela, poderá ser encontrada uma breve introdução ao tema explorado, assim como a lacuna a qual este trabalho

se propôs a preencher, o problema de pesquisa, que contém a pergunta que foi respondida no desenvolvimento do trabalho e nas considerações finais. Também indica a justificativa e expõe os motivos da realização da pesquisa e, por fim, como ela se adere ao PPSIGC.

A Fundamentação Teórica apresenta tópicos a serem abordados durante a pesquisa, informando ao leitor os conceitos necessários para a compreensão dos itens da pesquisa. Serão apresentados os seguintes itens: Inteligência Artificial; Aprendizado de Máquina; Metadados; Sistemas de Recuperação da Informação; e Mineração de Texto.

A Metodologia apresenta o aspecto metodológico necessário para a realização desta pesquisa a partir do método científico. No capítulo é apresentado o método escolhido para a realização da pesquisa, contando com a preparação dos arquivos para experimento, os algoritmos utilizados, assim como as fórmulas para a mensuração dos resultados obtidos.

No capítulo de execução, os passos apontados dentro da metodologia são seguidos e, com isso, são obtidos os resultados que são base para o alcance dos objetivos propostos anteriormente. Após os dados apresentados é feita uma discussão sobre os dados obtidos, e eles são colocados juntos para serem comparados.

Durante as considerações finais é respondida a pergunta de pesquisa e evidenciado o resultado dos objetivos apontados durante a introdução. Também são indicados possíveis usos da proposta de mineração de texto juntamente com a sugestão de pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Inteligência Artificial

A inteligência artificial começou a ser estudada logo após a Segunda Guerra Mundial e teve seu conceito definido em 1956. É considerada um dos campos mais recentes das áreas de ciências e engenharias (RUSSEL; NORVIG, 2014). No Quadro 1, pode-se observar a apresentação de diversas visões de cientistas sobre a inteligência artificial desde 1978 até 1998.

Quadro 1 - Inteligência artificial para alguns cientistas

Pensando como um humano	**Pensando racionalmente**
"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal." (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)
Agindo como seres humanos	**Agindo racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole *et al*., 1998) "AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Fonte: (RUSSEL; NORVIG, 2014, p. 24)

> Em linhas gerais, as que estão na parte superior da tabela se relacionam a processos de pensamento e raciocínio, enquanto as definições da parte inferior se referem ao comportamento. As definições do lado esquerdo medem o sucesso em termos de fidelidade ao desempenho humano, enquanto as definições do lado direito medem o sucesso comparando-o a um conceito ideal de inteligência, chamado de racionalidade. Um sistema é racional se "faz a coisa certa", dado o que ele sabe. (RUSSEL; NORVIG, 2014, p. 24).

2.1.1 Aprendizado de Máquina

O aprendizado de máquina traz a proposta de um aprendizado automático para as máquinas, por meio do uso de reconhecimento de padrões. Esta proposta traz algoritmos que são capazes de analisar dados e aprender a partir de seus erros.

Dentro da área de inteligência artificial, temos o aprendizado de máquina, que é definido como o estudo e a modelagem computacional em suas múltiplas manifestações (CAMASTRA; VINCIARELLI, 2008). Este estudo divide-se em três principais linhas:

a) **Estudo orientado a tarefas**: desenvolvimento de sistemas de aprendizagem para melhora da performance de tarefas determinadas;
b) **Simulação cognitiva**: simulação computacional e investigativa sobre o processo humano de aprendizagem;
c) **Análise teórica**: estudo teórico sobre as possibilidades de aprendizagem, métodos e algoritmos, independente do domínio de aplicação.

Algoritmos de aprendizagem precisam ser treinados para realizarem o aprendizado, o que consiste em ajustar os pesos e coeficientes para a execução da tarefa em específico. Este treinamento pode ser feito das seguintes formas apresentadas na sequência.

2.1.1.1 Aprendizado Supervisionado

O aprendizado supervisionado é uma das formas que a máquina tem para aprender. É semelhante à aprendizagem com um professor; o dado de amostra é colocado com padrões de entrada e de saída. No aprendizado supervisionado, os objetivos já possuem um valor definido com classes ou valores reais (RUSSEL; NORVIG, 2014).

Como exemplo de uma tarefa para a aprendizagem supervisionada, temos o reconhecimento de fontes manuscritas (RUSSEL; NORVIG, 2014) em que o programa receberá, como entrada, diversas escritas e sua respectiva correspondência; em seguida, tentará descobrir qual o valor equivalente a uma letra não apresentada anteriormente.

Esta forma de aprendizagem requer uma massa de dados rotulada para sua execução, ou seja, como pré-requisito, a massa de dados que será utilizada já deve ter sido previamente rotulada para que, através de métodos como os apresentados abaixo, o algoritmo seja capaz de trazer resultados esperados.

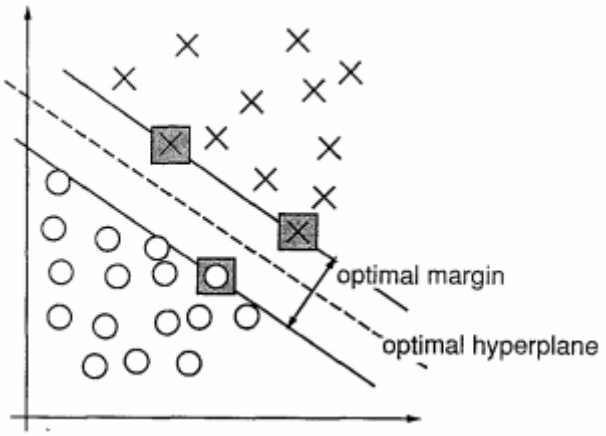
2.1.1.2 Aprendizagem de Classificação

Similar à apresentada anteriormente, com um número finito de saída, cada saída é chamada de classe; então, o programa avalia a entrada para determinar uma classe ou categoria de saída. É um método mais comum no uso de reconhecimento de padrões segundo Russel e Norvig (2014).

2.1.2 Support Vector Machine (SVM)

O método de máquinas de Vetores de Suporte é um método de aprendizado de máquina; é um dos mais importantes dentro da área de aprendizado para resolver problemas de regressão e classificação (RANA; SINGH, 2016). Este método recebe como entrada um conjunto de itens e forma um novo espaço de características em uma nova dimensão cujos padrões poderão ser linearmente separados, conforme apresentado na **Error! Reference source not found.**, tornando o hiperplano de separação ótimo (CORTES; VAPINIK, 1995).

Figura 3 - Exemplo de divisão de um problema em um ambiente bidimensional



Fonte: (CORTES; VAPINIK, 1995, p. 275)

2.1.3 Deep Learning

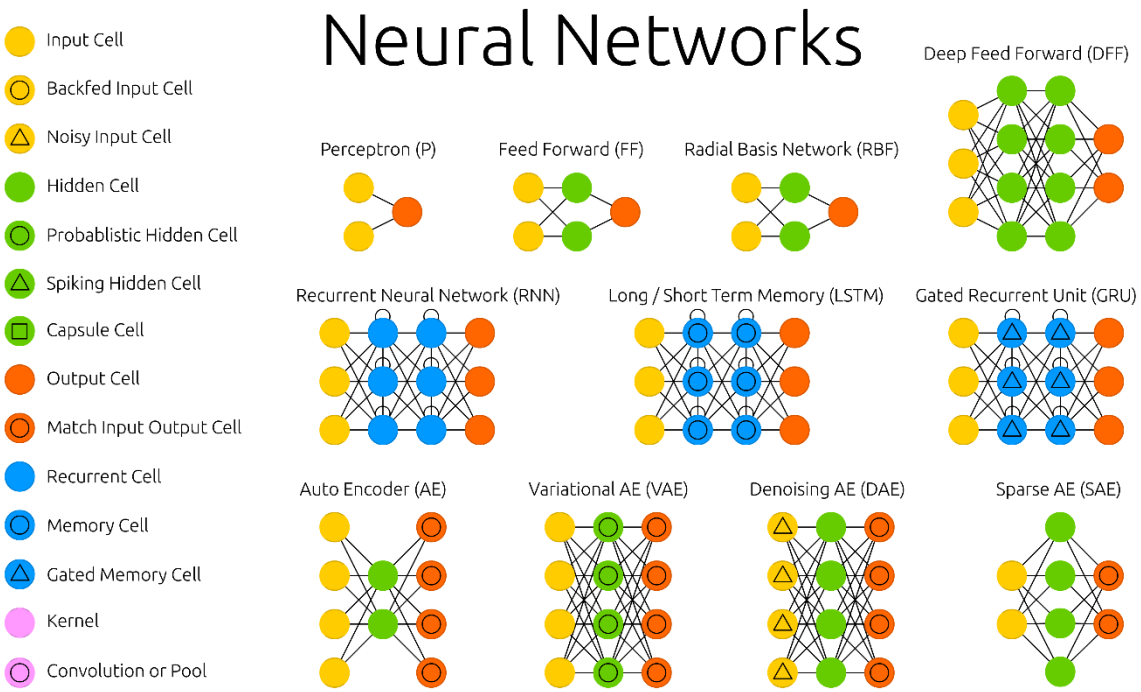
Dentro da área de *Machine Learning* existem os estudos ligados a *Deep Learning* (DL) em que Redes Neurais são construídas com base em uma analogia entre a estrutura presente e a utilizada por cérebro humano por meio dos neurônios (BENGIO, 2009).

> O *Deep Learning* permite que modelos computacionais baseados em processamentos de multicamadas de processamento aprendam representações de dados com múltiplos leveis de abstração. Estes métodos melhoraram drasticamente o estado da arte em reconhecimento de fala, reconhecimento visual de objetos, detecção de objetos e muitos outros domínios, como descoberta de medicamentos e genômica. (LECUN; BENGIO; HINTON, 2015, p. 436, tradução nossa).[1]

[1] Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics

Na Figura 4, são demonstrados diferentes tipos de redes neurais. Estas variam entre si pela quantidade de camada e tipos de células, e, desta maneira, cada tipo de rede pode ser customizado para a necessidade do cenário. Dentro desta pesquisa, por exemplo, foi utilizada a rede LSTM, que possui as células de entradas, células de memória, que serão utilizadas como memória de cada palavra, e as células de saída, que irão apresentar as respostas.

Figura 4 - Exemplos de Redes Neurais



Fonte: Adaptado de (VAN VEEN, 2016)

As Redes Neurais (RN) artificiais são modelos computacionais baseados no sistema nervoso central, ou seja, no cérebro, sendo geralmente apresentadas como um sistema de neurônios interconectados, que, segundo Caudill (1987), recebem diversos valores de entrada e apresentam na saída o resultado esperado.

Sua estrutura base são nós interconectados que funcionam como simples unidades de processamento conhecidos como neurôns, que atuam como pesos e conexões ponderadas que se juntam aos neurônios. Dessa forma, os neurônios de uma rede neural artificial devem estar conectados entre si e são dispostos em camadas, sendo que normalmente cada camada possui o seu comportamento, e quanto mais camadas, mais complexo o processamento e maior a aprendizagem, e também maior o tempo de processamento.

Alguns procedimentos que podem ser realizados em células internas são: probabilística, recorrente, memória e eliminação de ruídos, conforme demonstrado na Figura 4. Cada um dos usos possui aplicações diferentes e varia conforme o resultado esperado. Por exemplo, as redes

neurais recorrentes, segundo Graves, Fernandez e Schmidhuber (2007), são redes que foram originalmente desenvolvidas como uma extensão das redes neurais para dados sequenciais e são um conjunto de algoritmos artificiais de redes neurais especialmente úteis para o processamento de dados sequenciais, tais como: som, dados de séries temporais ou vídeos. Essas redes também permitem operar sobre sequências de vetores na entrada, saída, ou em ambos.

O Perceptron de Múltiplas Camadas (*Multilayer Perceptron* - MLP), conforme abordado por Carlos Junior (2011), é uma rede de múltiplas camadas, composta por um conjunto de unidades sensoriais que formam a camada de entrada, uma, ou mais camadas ocultas e uma camada de saída, sendo que a função dessas camadas intermediárias é transformar o problema em linearmente separável, de forma que possa ser resolvido pela rede.

Uma variação da arquitetura de redes neurais recorrentes é a *Long Short Term Memory* (LSTM), capaz de aprender dependências de longo prazo, funcionando bem para uma variedade de problemas, porém seu melhor uso é para o áudio, onde o algoritmo pode reconhecer padrões.

Pode-se observar que o *Deep Learning* possui uma grande variedade de aplicação, tendo sido aplicado em diversas áreas, como classificação, recuperação da informação, como será explorado neste estudo, e também em regressão, redução de dimensionalidade, modelagem de texturas, modelagem de movimentos, segmentação, robótica e filtragem colaborativa (BENGIO, 2009).

O *Deep Learning* pode ser usado na mineração de textos, para que se possa identificar informações consideradas importantes, e também no contexto da segurança; neste último, através dos algoritmos utilizados por ele, é possível fazer um sistema de arquitetura profunda, que vise processar os dados que ali estão sendo colocados, para assim identificar se alguma ameaça poderá estar tentando obter aqueles dados, na medida em que o avanço da tecnologia e, consequentemente, das informações aumenta.

DL pode ser utilizado em diversos momentos, a fim de se assegurar um processamento de dados eficaz, mas, para que seja possível utilizar essa metodologia, se faz necessário entender e definir bem alguns objetivos, a instrumentação de sistemas, estabelecer estimativas do trabalho e modificar o sistema inúmeras vezes para que acompanhe o avanço da tecnologia em detrimento a um maior fluxo de informações que circulam nas mais variadas plataformas (SATO *et al.*, 2018).

O *Deep Learning* foi otimizado devido ao avanço em conjunto dos computadores e da tecnologia em si, portanto, a mineração de texto foi possível devido a estes dois acontecimentos, fazendo com que as *Graphics Processing Unit* (GPU) atuassem justamente para a realização

dessa atividade, em consenso com todo o aparato tecnológico necessário para a realização de tal ato (MARUMO, 2018). O DL atua fazendo a coleta e a extração desses dados que são apresentados nesses textos, para que apenas as informações importantes sejam extraídas e assim possam ser levadas em consideração frente a todas as outras informações presentes naquele documento; e descritores de texto é o nome dado corretamente ao que se é feito (MARUMO, 2018).

Apesar da sua otimização, o DL conta com uma gama de algoritmos que podem vir a impedir o uso por parte de pessoas que não possuem o conhecimento suficiente para lidar com ele, apesar de atualmente ser o melhor recurso para que análises sejam feitas, não só documentais, mas também em vídeos, sendo aplicado em fala, áudio, imagens, e em uma série de outros componentes que podem ser analisados. O DP é o programa escolhido para a execução de tal função, por oferecer uma resposta e um resultado difícil de ser encontrado em qualquer outro (PONTI; COSTA, 2017).

O *DL* é como um guia para que tal informação seja obtida em um curto espaço de tempo. Isto pode ser visualizado na recuperação de dados, que é algo bastante utilizado na mineração de textos, servindo para a coleta de informações em textos muito densos, ou que as informações desejadas não estejam claras. Desse modo, o DP pode ser inserido para que assim seja possível obter as informações e seja feito o compartilhamento daquela informação (PONTI; COSTA, 2017).

2.1.3.1 Long Short Term Memory

De acordo com Greef *et al.* (2015), *Long Short Term Memory* foi inserido na literatura no início do ano de 1995, para que se pudesse utilizar e ter disponível um novo modelo de redes neurais, servindo para mudar o conceito frente a variantes da arquitetura que era empregada pelas ferramentas até então comumente utilizadas pelas pessoas para o armazenamento de informações.

A LSTM foi desenvolvida por Hochreiter (HOCHREITER; SCHMIDHUBER, 1997) como um tipo de rede neural recorrente para resolver um problema matemático. A ideia da rede é utilizar um sistema de portões que possibilitam escrever, ler e apagar informações da memória, permitindo uma correlação entre informações acima de mil entradas anteriores com a entrada atual.

Dessa forma, a camada LSTM consiste em múltiplas sub redes que se conectam de forma recorrente, conhecidas como blocos de memória, ativadas por três unidades: o portão de

entrada, o portão de esquecimento e o portão de saída. Esses portões permitem que as células armazenem e acessem as informações durante longos períodos de tempo, resolvendo, assim, o problema matemático do gradiente que desaparecia, conforme Graves, Fernandes e Schmidhuber (2007).

O carregamento de informações, algo extremamente importante e indispensável no LSTM, é feito através de Células, que possuem a função de guardar as informações e/ou excluir, dependendo do que seja considerado importante ou não, ou seja, as informações podem ser excluídas no meio do processo, se a célula interpretar que aquela informação não é tão importante, e desse modo pode ser excluída, dando maior espaço para aquelas consideradas importantes e indispensáveis. Estes são os pontos de um LSTM, e essas especificações fazem parte de todo um sistema que está amplamente definido, desde o seu funcionamento, até mesmo os dados que não são tão importantes de serem armazenados (JUNIOR, 2019).

As redes LSTM, em seu módulo, destacam-se das demais por conta de apresentar uma estrutura neural totalmente diferente de outras, no caso das redes neurais não LSTM, apenas uma camada neural é apresentada, enquanto na LSTM são apresentadas quatro, fazendo com que esse tipo de rede alcance um destaque maior diante de outras, e assim seja mais utilizada, tendo em vista que quanto mais células presentes, mais informações podem ser armazenadas por mais tempo, fazendo com que seja possível uma análise em um determinado espaço de tempo, sem a necessidade daqueles dados serem armazenados novamente, tendo em vista que a LSTM executa essa função de forma satisfatória e totalmente evoluída (JUNIOR, 2019).

Com o armazenamento de dados em uma larga escala de tempo, essa rede neural é muito utilizada para que seja possível fazer previsões, por exemplo, o índice de chuva em determinadas cidades, com espaço de sete meses. A rede LSTM, a partir dessas informações passadas armazenadas, poderá ser utilizada para que seja possível fazer o planejamento de qual volume pluvial poderá acontecer em uma escala dos meses selecionados. Isso é importante para que esse tipo de programação seja aplicado e suas especificidades sejam utilizadas ao máximo, em contraponto a outras redes neurais (JUNIOR, 2019).

O LSTM possui a capacidade de fazer com que aconteça uma economia de tempo para que os arquivos presentes na célula sejam excluídos automaticamente, sem a necessidade de uma ação mecânica. Esse ato só é possível devido a uma chave existente chamada de *sigmoide*, que executa justamente essa função, a fim de que seja possível manter armazenados apenas os arquivos realmente necessários para a execução das funções para as quais a LSTM fora designada, que é o armazenamento de dados por um longo tempo (CHRISTOPHER, 2015).

A eficácia das Redes Neurais, por conta da quantidade de informações que podem ser processadas, faz com que esse modelo de rede neural ocupe um importante papel no que se possui de conhecimento sobre o método de análise de dados em células de forma rápida, precisa e eficaz, além de ser possível o armazenamento em grande quantidade de inúmeras informações (CHRISTOPHER, 2015).

A rede LSTM faz a leitura sequencial das memórias em um único sentido, para o incremento de sua performance em cenários específicos, como, por exemplo, trabalhos com processamento de linguagem natural em que a ordem inversa pode auxiliar no processo desejado.

Graves e Schmidhuber (2005) utilizam o conceito de Bidirecional LSTM (BiLSTM) utilizando a concepção de que a memória das células é lida em ambos os sentidos. O processo de BiLSTM superou o processo unidirecional existente em 2005 no reconhecimento de fonemas, processo este utilizado no reconhecimento de fala, alcançando uma acurácia superior com a necessidade de menos épocas de treinamento.

Com o BiLSTM, o algoritmo passa a ser alimentado com os dados originais uma vez do começo ao fim, e uma vez do fim ao começo, fazendo com que, a depender da tarefa, ele possa ter uma maior capacidade de aprendizado.

2.2 Processamento de Linguagem Natural

A comunicação humana é feita de forma diferente da comunicação realizada pelas máquinas. Com o intuito de fazer as máquinas interpretarem a comunicação que nós utilizamos, foi criado o processamento de linguagem natural.

Pode-se afirmar, de forma simplificada, que com o processamento de linguagem natural o computador passou a conversar utilizando linguagem humana por meio do tratamento de diversos aspectos da computação humana, como sons, palavras, sentenças, discursos considerando contexto, estruturas, referências e significados, com o entendimento classificado em alguns níveis. São eles:

a) **Fonológico e fonético**: trata do relacionamento das palavras com os sons que produzem;
b) **Morfológico**: trata da construção das palavras a partir de unidades de significado primitivas e de como classificá-las em categorias morfológicas;

c) **Sintático**: trata do relacionamento das palavras entre si, cada uma assumindo seu papel estrutural nas frases, e de como as frases podem ser partes de outras, constituindo sentenças;
d) **Semântico**: trata do relacionamento das palavras com seus significados e de como eles são combinados para formar os significados das sentenças;
e) **Pragmático**: trata do uso de frases e sentenças em diferentes contextos, afetando o significado (MAIA; SOUZA, 2010).

Exemplos práticos do uso de processamento de linguagem natural usualmente utilizados são os aparelhos de assistentes pessoais: o aparelho deve escutar o comando enviado pelo usuário, através de voz ou por escrito, e interpretá-lo considerando o contexto para executar alguma tarefa determinada.

As referidas técnicas serão utilizadas nesta pesquisa para avaliar a viabilidade de uma análise do conteúdo de um documento e a possível vinculação do acesso ao documento pelos usuários que estão tentando acessá-lo.

2.2.1 Mineração de Texto

A mineração de texto é o processo de extrair informação útil (conhecimento) de dentro de um documento de texto que não está estruturado. Também pode ser conhecida como *Knowledge Discovered In Text* (KDT). Para isso, utiliza-se de diversas outras ferramentas conhecidas como Processamento de Linguagem Natural ou Descoberta de Conhecimento em Banco de Dados (BARION; LAGO, 2018).

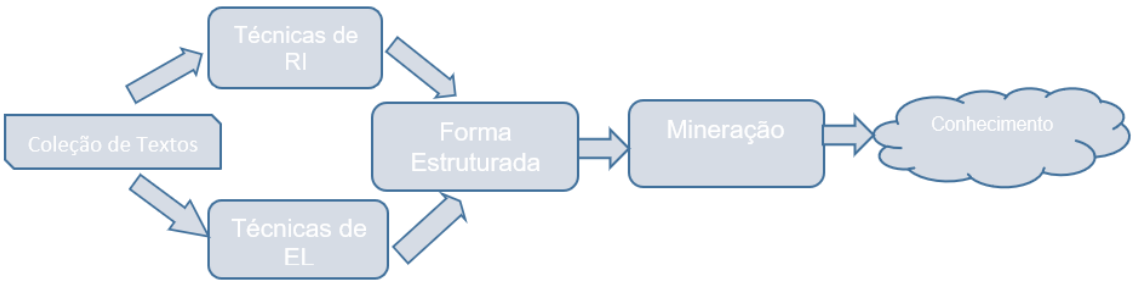
Para Aranha e Passos, a Mineração de textos "é um conjunto de métodos usados para navegar, organizar, achar e descobrir informações em bases textuais. Pode ser vista como uma extensão da área de Data Mining, focada na análise de textos." (ARANHA; PASSOS, 2006, p. 2).

Considera-se que 80% das informações estão armazenadas em formato de texto de forma não estruturada, conferindo à mineração de texto um grande valor comercial (BARION; LAGO, 2018).

Conforme demonstrado na Figura 5, diversas técnicas são aplicadas para que uma coleção de textos possa se tornar conhecimento. Primeiramente são utilizadas técnicas de Recuperação da Informação (RI) ou técnicas de extração da informação, possibilitando a

construção de uma fonte estruturada onde será aplicada a mineração para que todos esses textos possam ser convertidos em conhecimento.

Figura 5 - Técnicas de transformação de coleções de textos em conhecimento



Fonte: Adaptada de (CORRÊA, 2003)

2.2.2 Pré-processamento

O pré-processamento é uma parte essencial para o funcionamento de qualquer sistema de recuperação da informação. Busca facilitar processos de indexação e processamento de linguagem natural.

O pré-processamento possui as seguintes etapas:

a) Análise Léxica: Etapa responsável por converter o texto dos documentos em sequência de palavras em palavras candidatas a serem os termos dos índices. Além de palavras, existem estudos sugerindo a separação em sintagmas nominais que podem possuir um maior valor informacional (BARION; LAGO, 2008);
b) Remoção de *stop-words*: Etapa responsável por realizar a limpeza de palavras que se repetem em demasia durante o texto – geralmente são preposições, artigos, conjunções, alguns verbos, nomes, adjetivos e advérbios (BARION; LAGO, 2008);
c) *Stemming*: Etapa em que as palavras são convertidas em uma forma padrão, com isso são removidos sufixos e prefixos com o objetivo de eliminar o número de palavras a serem armazenadas. Exemplo: a palavra construção pode possuir diversas variações, como construir, construção, construído, construindo (BARION; LAGO, 2008);
d) Seleção de termos índices: Nesta etapa são selecionados quais itens devem ir para o índice; a escolha é feita pela categoria da palavra ou por seu agrupamento. Comumente, os substantivos possuem uma grande capacidade semântica (BARION; LAGO, 2008);

e) Determinação de pesos: Etapa responsável por determinar qual a relevância de cada termo índice selecionado. Diversos fatores são levados em conta, como quantidade de repetição do índice, ou mesmo sua categoria.
f) Criação de Tesauros: É recomendável que após isso se crie tesauros ou vocabulários controlados para criar associações entre termos conhecidos e possibilitar uma busca hierárquica.

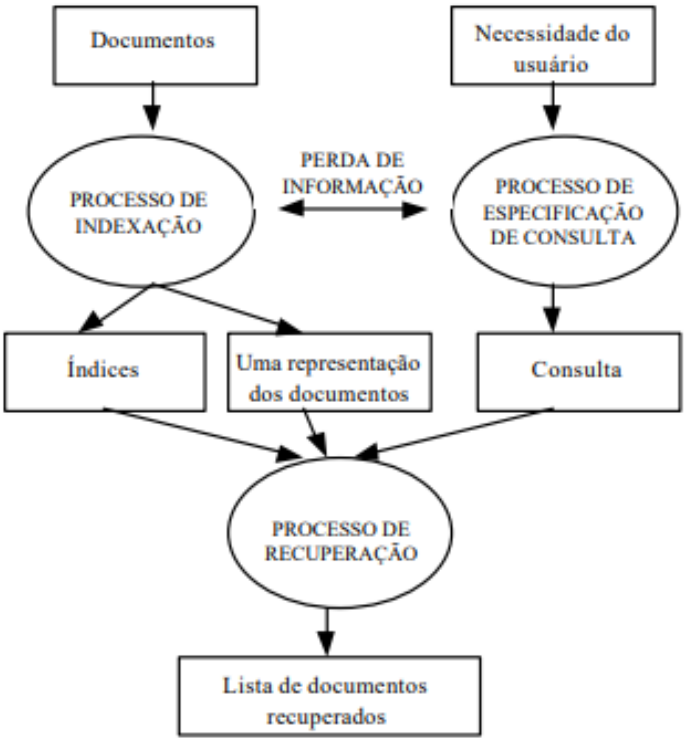
2.3 Sistemas de Recuperação da Informação

Segundo Souza (2005), as funções desempenhadas por um sistema de recuperação da informação são:

a) Representação das informações contidas nos documentos, usualmente por meio de processos de indexação e descrição dos documentos;
b) Armazenamento e gestão física e/ou lógica desses documentos e de suas representações;
c) Recuperação das informações representadas e dos próprios documentos armazenados, de forma a satisfazer as necessidades de informação dos usuários. Para isso é necessário que haja uma interface na qual os usuários possam descrever suas necessidades e questões – e possam analisar os resultados recuperados pelo SRI.

Sistemas de recuperação da informação fazem parte do grupo de sistemas de informação dentro da área de ciência da computação. É uma área de estudos responsável por estudar o armazenamento e a recuperação automática da informação. Conforme Cardoso (2004), um sistema de recuperação da informação geralmente possui a estrutura conforme representado abaixo na Figura 6.

Figura 6 - Processo de recuperação de documentos dentro de um SRI



Fonte: Adaptada de (CARDOSO, 2004)

Conforme a figura acima, temos as duas entradas básicas nos sistemas de recuperação da informação: Documentos e Necessidade do usuário. De um lado, temos a entrada de documentos que é o lado onde será criada a base de dados de documentos dentro do sistema de recuperação da informação, podendo conter tanto arquivos com formato texto quanto diversos outros tipos de arquivos, a depender da necessidade de um sistema de recuperação da informação.

Logo após, existe o processo de indexação, que é imprescindível para uma performance aceitável em um sistema de recuperação da informação. Esta parte, geralmente separada dos outros serviços, faz a leitura dos documentos no sistema para a criação de um índice que é posteriormente apresentado.

A criação do índice acelera o sistema, pois evita que a cada filtragem do usuário seja necessário o acesso a todos os documentos existentes no sistema para se descobrir qual o documento que se encaixa nos filtros fornecidos pelos usuários.

Na outra ponta, temos a entrada feita pelo usuário que deseja recuperar alguma informação que, supostamente, estará dentro do sistema de recuperação da informação.

O usuário, a partir do uso da interface, podendo esta ser escrita ou visual, irá construir uma consulta no sistema. A forma mais comum desse tipo de entrada é por meio de alguma

palavra-chave em algum campo de busca, mas não se limita apenas a isso, pode-se, por exemplo, utilizar um filtro de datas.

A consulta será executada com base nos índices já construídos e irá retornar uma listagem de documentos que se encaixe na consulta criada. Após isso, o usuário poderá acessar a representação do documento, dando-se por completo o ciclo de recuperação da informação.

Segundo Cardoso (2004), grande volume de informação é perdida durante o processo da criação da consulta em função de uma distância semântica entre a real necessidade do usuário e o desejo expresso na consulta. Contanto, quanto mais informações são fornecidas para que os usuários possam criar sua busca, diminui-se a capacidade de perda de informação entre a real necessidade do usuário e o que ele expressou na consulta.

2.3.1 Categorização e Extração de Informação

Para Cardoso, "categorização é o processo de classificar documentos em categorias pré-definidas. Sua maior aplicação tem sido para atribuir categorias a documentos e posteriormente utilizar estas categorias para suportar recuperação e filtragem de informação." (CARDOSO, 2004, p. 37).

É um processo importante, porém, pode tornar-se complexo em sistemas de recuperação da informação; pode, ainda, auxiliar muito no processo de filtro de documentos dentro do sistema.

As categorias podem ser consideradas como um conjunto de características que tendem a ser mais estáticas do que os perfis de filtragem. Sistemas de recuperação da informação apresentam um baixo desempenho no conceito apresentado devido ao alto índice de volume de informação que podem possuir e a uma quantidade limitada de categorias que podem ser utilizadas (CARDOSO, 2004).

Um portal de documentos de investidores é um exemplo de categoria. É possível haver a categorização sobre os fundos de investimentos, ou até mesmo a categoria sobre a qual investidor pertence cada investimento.

2.3.2 Metadados

Metadado é definido por Milstead e Feldman (1999) como dados sobre os dados, descrevendo atributos e conteúdo sobre o documento ou trabalho original. O termo geralmente

é aplicado em recursos digitais, como sistemas de recuperação da informação, mas não se limita a documentos. Pode ser usado também em itens como *dataset*, imagens, músicas e outros.

Inicialmente, bancos de dados de imagens só permitiam buscas com o uso de textos que estavam contidos dentro dos metadados da imagem (MILSTEAD; FELDMAN, 1999). Por exemplo, se alguém fosse pesquisar fotos de cachorros, o *software* não abriria imagens para identificar se existiam cachorros nela; em vez disso, buscaria nos metadados que possuíam a palavra "cachorro".

Já Duval *et al.* (2002) definem metadados como uma ferramenta primária para gerenciar informações de recursos no mundo *web;* para eles, metadados são um ponto-chave para uma infraestrutura de informação, ajudando a criar ordem dentro do caos da *web*. Sobre metadados, Duval *et al.* (2002) ainda afirmam que este continua sendo o padrão de maior sucesso para encontrar livros e periódicos. Porém, é uma criação com custo impraticável quando aplicada ao cenário da *web* com uma grande quantidade de documentos.

> [...] existe uma ampla gama de criação de metadados que podem ser automatizados em algum grau, e que podem crescer em importância à medida que avanços em áreas como processamento de linguagem natural, mineração de dados, perfil e algoritmos de reconhecimento de padrões se tornam mais eficazes. (DUVAL *et al*., 2002, p. 8, tradução nossa).[2]

Como o termo, apesar de muito usado na *web*, possui um longo período de tempo de criação, existem vários conjuntos de metadados já definidos, como demonstrado no Quadro 2, em que podem ser observados diversos padrões de metadados utilizados no passado.

Quadro 2 - Conjuntos de metadados

PADRÃO	RESPONSÁVEL	ANO
MARC - *Machine Readable Cataloging Record*	*Library of Congress (LC)*	1960
Dublin Core	OCLC	1968
GILS - *Government Information Location Service*	*National Archives dos EUA*	1992
EAD - *Encoded Archival Description*	Universidade da Califórnia	1993
RDF - *Resource Description Framework Schema*	W3C - *World Wide Web Consortium*	1998

Fonte: (MAIA; SOUZA, 2010)

[2] Between these two extremes lies a broad range of metadata creation that can be automated to some degree, and which can be expected to grow in importance as advances in such areas as natural language processing, data mining, profile and pattern recognition algorithms become more effective.

2.4 Segurança da Informação

Esse termo pode ser entendido por mecanismos que asseguram o funcionamento de um determinado sistema e exige uma atenção frente aos dados que são solicitados para que o acesso a determinado site seja feito. Para que não aconteça o roubo de informações e as pessoas acabem sendo prejudicadas por esses acontecimentos, os dados precisam estar seguros, em todos os âmbitos de coleta de dados, seja em documentos, eletrônicos e quaisquer outros modelos usados para tal ato, as informações devem ser asseguradas mediante uso de sistemas de informação que de fato façam com que os dados não sejam desviados dos seus respectivos fins (BIANCHI; FONSECA, 2014).

A configuração ao acesso de informações é fundamental para que seja possível verificar quais pessoas estão permitidas e aptas a fazerem isso, para que justamente se evite o compartilhamento dessas informações com pessoas não autorizadas, tais como dados pessoais que são inseridos ao se cadastrar em determinados sites ou quando se faz algum tipo de solicitação, como o login. Assim, as informações inseridas nos formulários devem ser protegidas e utilizadas apenas para aquele tipo de solicitação requisitada. O uso não autorizado de dados faz com que as empresas criem mecanismos para evitar a incidência desses acontecimentos (BIANCHI; FONSECA, 2014).

A segurança da informação é tida como um mecanismo forte o bastante para que se evite o compartilhamento de informações pessoais e/ou corporativas em detrimento de ameaças existentes, que possuem o único objetivo de prejudicar a vítima com o roubo de informações, mediante o uso dos dados obtidos, e diretamente fazer com que as pessoas fiquem seguras e confiantes em colocar seus dados e assim preencher os formulários presentes em sites, por exemplo, para que determinado acesso às informações seja feito, como compras em sites que ocasionalmente só podem ser feitas mediante cadastro, para que seja possível identificar o comprador e possíveis fraudes à informação que possam vir a ocorrer (FERNANDES, 2013).

Existem princípios básicos que norteiam a segurança da informação: Confidencialidade, Disponibilidade, Integridade, Legalidade e Autenticidade, todos esses são aspectos inerentes a uma boa segurança da informação, para que assim os dados que são colocados em destaque possam estar seguros diante da grande existência de ameaças que trabalham para o roubo desses dados, fazendo com que as pessoas sintam-se receosas em fazer determinado tipo de cadastro, principalmente em redes sociais, que constantemente são alvos de ataques por diversos grupos de criminosos cibernéticos (FERNANDES, 2013).

Tendo em vista a necessidade cada vez maior de criação de mecanismos que evitem o uso indevido dos dados, a ISSO/IEC 177/99 criou a CID, que é responsável por protocolar como a segurança da informação deve ser cumprida, seguindo três principais bases que servem indispensavelmente para a segurança de todos os envolvidos. São elas: Confidencialidade, Integridade e Disponibilidade, evidenciando perfeitamente que todos os outros pontos colocados devem ser considerados importantes ao se tratar de segurança da informação.

Quanto à segurança da informação é importante evidenciar que as ameaças estão presentes, sendo assim, todos os cuidados disponíveis devem ser adotados, de forma a proteger dados trafegados, quantidade e tipos de acesso, entre outros (FERNANDES, 2013).

Os hackers possuem conhecimento suficiente para adquirir as informações que estão presentes nos sites, desse modo, os investimentos em Segurança da Informação estão cada vez mais em destaque por conta do avanço exponencial da tecnologia, e das ameaças à segurança crescerem na mesma proporção ou de uma forma infinitamente maior (JUNIOR, 2019).

Uma das formas de fazer isso é através da mineração de textos, capaz de identificar as informações mais importantes em um site e aquelas que podem ser alvo de ataque com maior incidência por parte das pessoas que trabalham para roubar essas informações, sendo assim, são indispensáveis o seu uso ao se tratar de segurança da informação (JUNIOR, 2019).

2.4.1 Controle de Acesso

Dentro de sistemas de recuperação da informação, principalmente relacionados à área financeira, a segurança é um item primordial e torna-se mais complexa com o aumento do número de documentos. Nesse sentido, o controle de acesso tem se tornado um item de importante estudo na área da ciência da informação.

> As organizações modernas acumulam grandes volumes de informações, incluindo informações confidenciais. A violação dessas informações primeiro leva à sua exposição e, em seguida, pode resultar em violação de privacidade e perdas financeiras. A segurança de um sistema de informações baseado em computador deve, por design, proteger a confidencialidade, a integridade e a disponibilidade do sistema que contém informações confidenciais (GORDON, 2002, p. 341, tradução nossa).[3]

[3] Modern organizations accumulate huge volumes of information including sensitive information. The breach of this information first leads to its exposure and then may result in a breach of privacy and financial losses. Security of a computer-based information system should, by design, protect the confidentiality, integrity, and availability of the system that contains sensitive information.

Os métodos de controle de acesso mais utilizados atualmente são classificados como *Role-Based Access Control* (RBAC) e *Attribute-Based Access Control* (ABAC). Um método comum utilizado principalmente dentro do ambiente de redes de controle de acesso é o ACL.

2.4.2 Role-Based Access Control (RBAC)

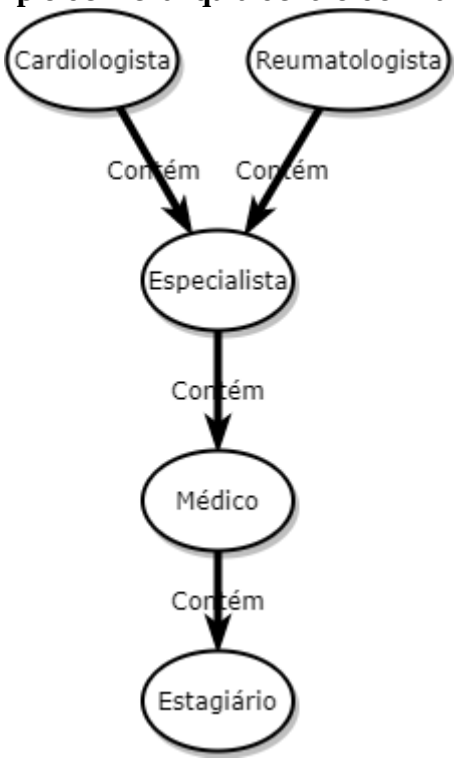
Apresentado em 1992 por D.F. Ferraiolo e D.R. Kuhn, pode ser traduzido por controle de acesso baseado em função. Segundo Gallaher *et al.* (2002), este método tende a ser implementado em empresas com mais de 500 funcionários por tornar mais fácil o gerenciamento em grandes ambientes, atribuindo permissões ao ambiente de forma mais próxima de como é a organização da empresa, concedendo o acesso baseado nas necessidades que a função do funcionário necessita.

O método de controle de acesso RBAC possui três principais itens: o usuário, que é uma pessoa; a função, que é uma coleção de funções de trabalhos; e a operação, que representa o modo particular de acesso a um ou a mais objetos protegidos (Ferraiolo, Cugini, & Kuhn, 1995).

As funções podem assumir papéis hierárquicos definindo que algumas funções podem conter outras funções. É uma maneira natural para fazer sua organização, representando autoridade, responsabilidade e competência (FERRAIOLO; CUGINI; KUHN, 1995).

A Figura 7 apresenta um exemplo de como ocorre a hierarquia em sistemas utilizando esse tipo de framework. Arquivos relacionados a *Cardiologista*, por exemplo, somente poderiam ser acessados por usuários nesta função; porém, documentos da função *Médico* seriam compartilhados com os usuários que estão na parte superior da figura.

Figura 7 - Exemplo de hierarquia dentro do Framework RBAC



Fonte: Adaptada de (FARRAIOLO; CUGINI; KUNH, 1995)

2.4.3 Attribute-Based Access Control (ABAC)

O *National Institute of Standards and Technology* (NIST), órgão estadunidense responsável por promover competitividade industrial, define o ABAC como uma metodologia de controle de acesso lógica, em que a autorização para executar o conjunto de operações é determinada pelo valor do atributo do sujeito, objeto, operação solicitada, e, em alguns casos, condições do ambiente e de políticas, regras ou relacionamentos que descrevem a operação baseada nesta série de atributos (HU *et al.*, 2013).

Segundo Hu *et al.* (2013), o ABAC é um modo de controle de acesso baseado em atributo que foi introduzido em 2003 devido ao crescimento da arquitetura orientada a serviços. Possui um modo de permissão mais individual e, ao mesmo tempo, mais complexo de gerenciar.

Atributos (*Attributes*): são as características que definem os aspectos específicos de um usuário, objeto, condições do ambiente e/ou ações demandadas (HU *et al.*, 2013).

Sujeito (*Subject*): é a entidade ativa que causa o fluxo de informação entre os objetos ou muda o estado do sistema. Existem alguns tipos de sujeitos: o usuário requisitante pode ser um deles, mas o sujeito também pode ser algum mecanismo, como, por exemplo, um indexador funcionando através da autenticação de um desses anteriores. Os sujeitos podem possuir atributos (HU *et al.*, 2013).

Objeto (*Object*): é a informação passiva do sistema (dispositivos, arquivos, tabelas, processos, programas) que contém ou recebe informação. O acesso ao objeto implica no acesso à informação que ele possui (HU *et al.*, 2013).

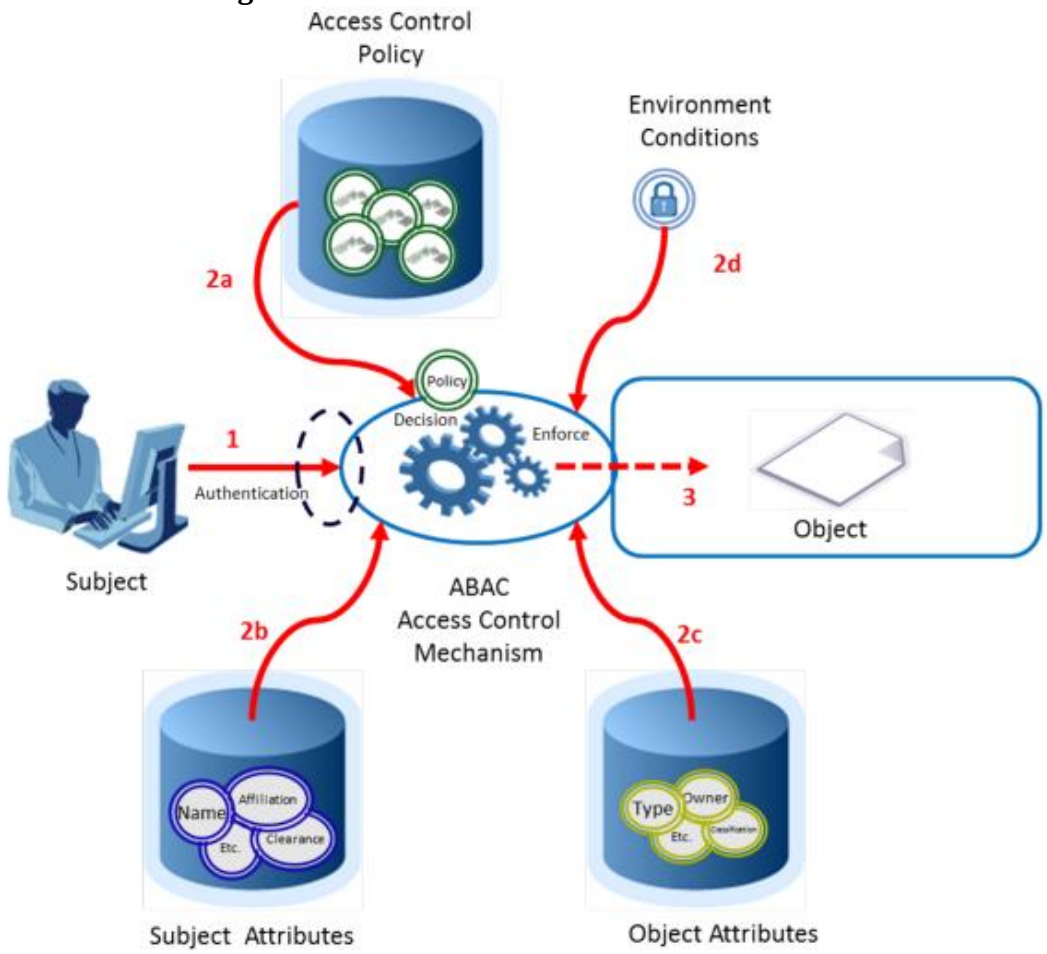
Operação (*Operation*): é a execução da função ou tarefa demandada pelo sujeito. Segundo C. Hu et al. (2013), a operação pode incluir, segundo Hu *et al.* (2013):

a) Ler;
b) Escrever;
c) Editar;
d) Deletar;
e) Copiar;
f) Ser autor;
g) Executar;
h) Modificar.

Política (*Policy*): representa a regra ou o relacionamento que irá definir a autorização das operações que o sujeito poderá executar (HU *et al.*, 2013).

Na Figura 8, pode-se ver o funcionamento básico do controle de acesso baseado em atributos: os atributos do usuário, juntamente com atributos do objeto ao qual se deseja ter acesso, são enviados para o mecanismo de controle e este irá avaliar, com base nos atributos do sujeito, atributos do objeto e políticas de controle de acesso – se o sujeito em questão possui o acesso ao objeto. Essas informações são analisadas de acordo com as políticas de acesso que podem estar descritas no formato *Extensible Access Control Markup Language* (XACML).

Figura 8 - Funcionamento básico ABAC



1. Subject Requests Access to Object
2. Access Control Mechanism Assesses a) Rules, b) Subject Attributes, c) Object Attributes, and d) Environment Conditions to Determine Authorization
3. Subject is Given Access to Object if Authorized and Denied Access if Not authorized

Fonte: (HU *et al.*, 2013).

O ABAC possui algumas vantagens em relação ao RBAC, pois este nem sempre atende às demandas do mundo real. Decisões fora da função do usuário – por exemplo, um usuário da área do departamento legal precisa de um acesso a algum documento da área de direitos humanos – fugiriam do escopo da função (*role*) do usuário, mas poderia ter uma aplicação mais prática em um sistema com base em atributos.

Dentro do ABAC, diversos itens podem ser considerados para o controle de acesso do item. Um documento que possua identificação com um fundo de investimento, por exemplo, só poderia ser acessado por usuários que possuam o mesmo tipo de atributo.

Com o ABAC em sistemas de larga escala, as regras de controle de acesso para sistemas de larga escala podem ser simplificadas. Ao mesmo tempo, ele pode fornecer flexibilidade e escalabilidade, essenciais para grandes sistemas distribuídos (Shen, 2009).

2.5 Trabalhos Relacionados

Este capítulo destina-se a apresentar pesquisas realizadas anteriormente com objetivos similares, pesquisas estas utilizadas para guiar a criação da metodologia. Estas pesquisas se focam nos temas de controle de acesso com uso de Linguagem natural e Classificação de documentos.

O termo *Natural Language Access Control Policy* (NLACP) é usado por alguns autores para denominar o controle de acesso automático utilizando processos de processamento de Linguagem Natural.

Inglesant *et al.* (2008) pesquisaram sobre a dificuldade de implementar um modelo de controle de acesso baseado em processamento de linguagem natural frente ao *Role-based access control* (RBAC), que é o modelo mais utilizado em empresas com mais de 500 funcionários (GALLAHER *et al.*, 2002). Utilizando uma linguagem controlada para designar a permissão, o estudo concluiu que o modelo, sozinho, não é suficiente para gerir as permissões gerais dos sistemas devido a algumas falhas que não puderam ser entendidas pelo analisador de linguagem natural.

Segundo Xiao *et al.* (2012), em 2012, a ferramenta Text2Policy foi publicada com o objetivo de identificar sentenças em linguagem natural sobre políticas de controle de acesso. O objetivo foi extrair permissões de frases como "*The Health Care Personnel (HCP) does not have the ability to edit the patient's security question and password*" (XIAO et al., 2012).

O Text2Policy passou a ser utilizado em outros artigos para explorar a extração de políticas de controle de acesso com base em linguagem natural, tornando-se uma fonte para melhorias no futuro. Sua precisão no artigo supracitado foi de 88,7% na classificação da sentença.

Em *Identification of Access Control Policy Sentences from Natural Language Policy Documents*, os autores Narquei, Khanpour e Takabi (2017) fizeram o trabalho de traduzir sentenças em linguagem natural para transformá-las em um sistema com o controle de acesso do modo ABAC (*Attribute-based access control*), que poderia superar limitações do modelo mais comum utilizando RBAC com uma precisão de 90% na identificação de sentenças referentes a permissões.

Esses estudos focam na usabilidade do usuário, permitindo a este atribuir permissões nos sistemas, e denotam um crescente interesse sobre o uso do processamento de linguagem natural dentro da área de controle de acesso.

Em *One-Class SVMs for Document Classification* (MANEVITZ; YOUSSEF, 2001), diversos algoritmos são comparados a mecanismos de *Support Machine Vector (SVM)*. Nessa publicação é feita a análise de documentos da agência de notícias *Reuters* utilizando-se de aprendizagem supervisionada. O autor compara algoritmos e sua performance nas seguintes classificações:

a) **Representação binária**: representação de 0 ou 1; quando o termo aparece no documento, é classificado como 1; caso não apareça, a classificação é 0 (MANEVITZ; YOUSSEF, 2001).
b) **Representação de frequência**: apresenta a frequência de vezes que o termo aparece no texto no documento específico (MANEVITZ; YOUSSEF, 2001).
c) **Representação *tf-idf*:** esta representação *tf-idf* é a abreviação do termo inglês *term frequency inverse document frequency*, que é o resultado da função tf-idf(termo) = frequência(termo) * [log(n/N(termo)+1] (onde n representa o total de palavras no dicionário, e N, a função dada pelo total de documentos em que o termo aparece) (MANEVITZ; YOUSSEF, 2001).
d) **Representação Hadamard**: o autor classifica essa representação como

> A representação do produto Hadamard foi descoberta experimentalmente; consiste no vetor m dimensional em que a i-ésima entrada é o produto da frequência da i-ésima palavra-chave no documento e sua frequência em todos os documentos (no conjunto de treinamento) (MANEVITZ; YOUSSEF, 2001, p. 141, tradução nossa). [4]

Já Eui-Hong e Karypis (2002) apresentam um algoritmo linear para a classificação de documentos. Sua análise mostra a similaridade, permitindo classificar documentos pela proximidade de comportamento ajustando dinamicamente o distanciamento de termos.

Nas pesquisas citadas acima, foram usados documentos na língua inglesa; isso indica que os mesmos estudos utilizando a língua portuguesa podem apresentar resultados diferentes, com performance do algoritmo diferente.

Na língua portuguesa, Maia e Souza (2010) publicaram sobre o uso de sintagmas nominais na classificação automática de documentos eletrônicos. Os autores, utilizando-se do processamento de linguagem natural, extraíram sintagmas nominais e avaliaram a ocorrência

---

[4] The Hadamard product representation was discovered experimentally; it consists of the m dimensional vector where the i-th entry is the product of the frequency of the i-th keyword in the document and its frequency over all documents (in the training set)"

de uma classificação superior dos documentos àquela que ocorre somente com o uso da classificação com palavras.

Também na língua portuguesa, Araújo *et al.* (2020) publicam um *dataset* contendo documentos digitalizados do Supremo Tribunal Federal contendo cerca de 692 mil documentos com informações rotuladas, possibilitando pesquisas dentro da área de processamento de linguagem natural, entre outros. A base de dados é importante por se tornar uma base de comparação entre diversos algoritmos que buscam realizar tarefas de reconhecimento de linguagem natural na língua portuguesa. Este também apresenta resultados de base com algoritmos selecionados para a classificação do documento se utilizando dos modelos de *Naïve Bayes (NB)*, SVM, BiLSTM, e Convolutional Neural Network (CNN). O *dataset* foi dividido em três versões BVic com todos os dados, incluindo dados não rotulados. O Mvic, o *dataset* de tamanho médio, contendo 44.855 pastas, 628.820 arquivos e mais de dois milhões de páginas, por fim, apresentou também o dataset SVic para facilitar o compartilhamento com uma amostra limitada a 100 amostras de pastas, 94.267 documentos e 339.478 páginas.

Nas pesquisas de Araújo *et al.* (2020), os algoritmos utilizados para estabelecer uma linha de base dos documentos, os algoritmos de CNN e o BiLSTM, superaram a performance dos outros algoritmos dentro do *dataset* médio, sendo eles *Naïve Bayes* e SVM, em todas as categorias de documentos disponíveis, tendo o primeiro apresentado a pior performance. Já com o menor *dataset* SVic, as melhores performances foram de SVM e CNN, respectivamente. Na Figura 9 pode-se observar todos os resultados do score F1 para as categorias pesquisadas, e é interessante observar que as soluções baseadas em redes neurais nem sempre podem obter o melhor resultado, tornando necessária uma análise mais profunda para o caso de cada aplicação.

Figura 9 - Resultados dos modelos aplicados ao dataset VICTOR

Dataset	Model	Acórdão	ARE	Despacho	Others	RE	Sentença	Weighted	Average
MVic	NB	49.20	32.08	39.82	89.38	38.06	37.80	84.77	47.72
	SVM	65.41	52.62	59.34	95.85	64.52	69.75	92.88	67.92
	BiLSTM	**72.84**	57.82	**60.07**	97.11	67.74	69.96	94.33	**70.92**
	CNN	71.06	**58.11**	56.04	**97.37**	**68.71**	**72.35**	**94.64**	70.61
SVic	NB	66.40	36.07	51.15	93.24	55.89	55.99	88.93	59.79
	SVM	81.15	**58.06**	**67.88**	96.85	74.66	**79.30**	94.25	**76.32**
	BiLSTM	85.82	52.12	51.01	97.15	74.06	76.70	94.65	72.81
	CNN	**86.43**	55.92	59.88	**97.30**	**76.23**	79.29	**94.72**	75.84

Fonte: (ARAÚJO *et al.*, 2020)

A pesquisa (OORD; DIELEMAN, 2014) foi escrita por funcionários do site de músicas Spotify, em que eles utilizaram redes convolucionais para criar um sistema de recomendações de música com base nas similaridades musicais. Suas entradas foram fragmentos de 3 segundos

de duração e foi utilizada uma GPU NVIDIA GeForce GTX780Ti que levou entre 18 e 36 horas para o treinamento.

Os autores (CHOI, 2017) apresentam uma arquitetura híbrida, utilizando CNN e RNN, para classificação musical. A CNN foi utilizada para a extração de características e RNN para sumarização temporal das características extraídas. O método CRNN foi comparado com outras três estruturas de CNN, obtendo um resultado melhor a respeito do tempo de treinamento e número de parâmetros.

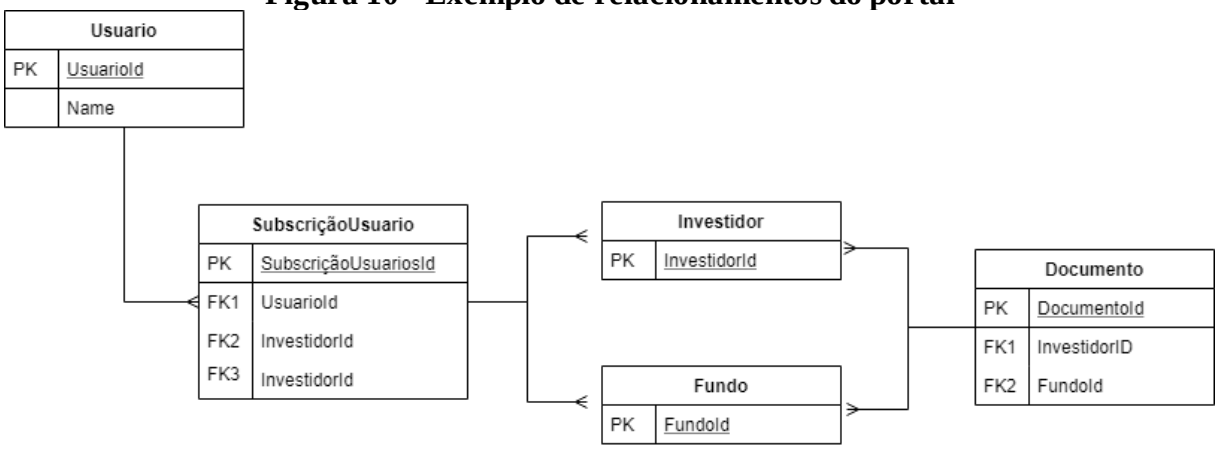
3 METODOLOGIA

3.1 Contextualização do Ambiente

O ambiente escolhido para realizar a pesquisa foi um portal que provê informações financeiras para investidores. Um dos objetivos principais do referido portal é ser um sistema de recuperação da informação. Usuários o acessam com o objetivo de encontrar balanços e documentos financeiros referentes a diversos temas, como informações fiscais, notas de distribuição de capitais e informações sobre portfólio. Um exemplo de documento com dados fictícios está presente no Apêndice A, e um exemplo de um formulário real é apresentado dentro do Apêndice B, que apresenta o endereço que pode conter o nome do fundo como *Parternship* e do Investidor como *Partner*.

A forma de permissão dos documentos é feita por meio dos atributos que o usuário e o documento possuem. Cada documento irá possuir informações de um único investidor e um único fundo, enquanto o usuário pode possuir registros de diversos investidores e diversos fundos. Com o cruzamento desses dados, a permissão do documento é concedida, e a estrutura simplificada de entidades e relacionamentos pode ser observada na Figura 10.

Figura 10 - Exemplo de relacionamentos do portal



Fonte: Elaborada pelo autor

Os dados disponibilizados para este estudo são de dois portais distintos, identificados como portal A e portal B, ambos presentes em um portal de investidores atual. Foram criados durante o período de 2016 até 2018. O portal A conta com um total de 9.459 documentos e o portal B possui um volume consideravelmente menor, contando com 1.142 documentos, tendo sido utilizados 1.102. Todos os documentos estão na língua inglesa e em formato PDF, alguns são nato-digitais, outros são digitalizados após assinatura feita à mão. Possuem um tamanho médio de 218 KB, totalizando, os dois portais, um tamanho total de 2,15 GB.

O controle atual de permissão é feito de modo próximo ao ABAC, em que os atributos do usuário e do documento são capazes de determinar se ao usuário está permitido acessar aquele arquivo ou não.

Os algoritmos serão treinados a fim de encontrar uma classe dentro do documento, a classe que o algoritmo deverá classificar o documento baseado ao investidor e ao fundo ao qual ele pertence. Os tipos destes documentos estão detalhados abaixo, juntamente com a quantidade de investidores e fundos existentes que foram utilizados para os treinamentos dos algoritmos.

Os documentos utilizados são de variados tipos como pode-se observar no Quadro 3, que apresenta do lado esquerdo as categorias de documentos do Portal A e do lado direito as categorias dos documentos presentes no Portal B.

Quadro 3 - Categorias de documentos

Portal A	**Portal B**
Demonstração de Contas	Demonstração de Contas
Distribuição de Lucros	Distribuição de Lucros
Documentos Legais do Investidor	Documentos Legais do Investidor
Documentos Legais do Fundo	Documentos Legais do Fundo
Documentos Fiscais	Documentos Fiscais
Relatórios Trimestrais e Anuais	Relatórios Trimestrais e Anuais
Auditorias de Finanças	Auditorias de Finanças
	Chamada de Capital
	Diversos

Fonte: Elaborado pelo autor

O tipo de classificação que será necessário realizar é a classificação *multi-class*, em que diferente de uma saída binária, o algoritmo, durante execução da classificação de investidores no Portal A, precisará de – do meio de um universo de 408 investidores – classificar o documento em uma classe deste conjunto. No Quadro 4 pode-se observar as categorias e variedades nas quais os algoritmos irão ser colocados para classificação.

Quadro 4 - Descrição de quantidade de documentos

Portal de investidores	**Investidor**	**Fundo**	**Total de documentos**
Portal A	408	15	8.950
Portal B	42	3	1.102

Fonte: Elaborado pelo autor

3.2 Escolha dos Algoritmos

Foram escolhidos algoritmos distintos com o objetivo de avaliar performance e possibilidade de utilização dentro do ambiente e do experimento apresentado. Os dois algoritmos escolhidos são de inteligência artificial, porém nem todos utilizam o conceito de redes neurais para o seu funcionamento.

O primeiro algoritmo escolhido, o algoritmo *Support Vector Machine* (SVM), é um dos mais importantes dentro da área de classificação (RANA; SINGH, 2016). O algoritmo não utiliza redes neurais e isto foi um dos motivos da escolha, podendo, dessa forma, apresentar uma massa de resultados diferentes, ampliando a capacidade de análise e comparação. Este algoritmo permitiu também realizar testes utilizando o conceito de *bag of words* ou um determinado número de n-gramas.

Algoritmos de redes neurais são predominantes dentro do estado da arte, dentro das diversas subtarefas de classificação na área de processamento de linguagem natural, isto pode ser notado ao se analisar o site paperswithcode.com, que possui uma categoria que se propõe a apresentar o estado da arte da subtarefa de classificação de documentos em diversos *dataset* públicos, e nota-se que na data desta pesquisa todas as maiores acurácias eram atingidas com o uso de redes neurais.

Sendo assim, o modelo de LSTM foi escolhido por possuir um comportamento de memória ao qual pode ser útil durante a análise de documento quando consideramos que a ordem das palavras interfere dentro do sentido da frase. Este método irá analisar a sequência de palavras com memória a aprender, e, com base nesta memória de palavras, agregado a isso, foi utilizado uma camada bi-direcional que fará com que a sequência seja analisada tanto no sentido da leitura do texto quanto no sentido inverso, conforme arquitetura BiLSTM. O modelo foi similar ao apresentado por Huang, Xu e Yu (2018), e partes importantes do código utilizado estão presentes no Anexo B.

3.3 Processo de Validação

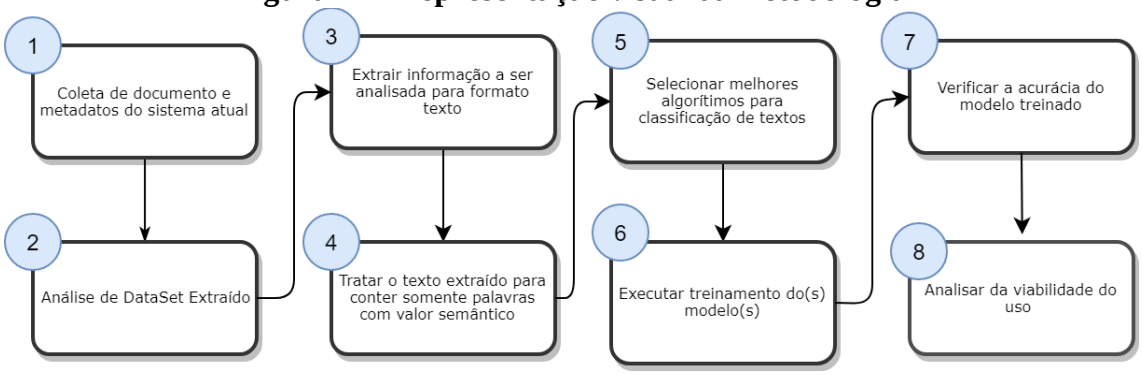
A proposta desta dissertação tem como metodologia um estudo quantitativo sobre a performance percebida dos algoritmos selecionados.

Todos os passos estão detalhados abaixo:

a) O primeiro passo a ser tomado é a coleta dos documentos. A pesquisa tem o objetivo de avaliar documentos da área financeira; logo, serão extraídos documentos financeiros e seus respectivos metadados.
b) Com os documentos já extraídos, será necessário fazer a análise dos documentos recuperados para tentar identificar se todos possuem o conteúdo e se eles estão em formato de texto. (Documentos contendo conteúdo em formato de imagem poderiam não fornecer informações relevantes nesse caso.)
c) Neste passo, toda a informação que está contida nos documentos será extraída para formato de texto, criando o conceito de *bag of words*, sendo assim padronizada e facilitando a leitura dos futuros algoritmos que serão utilizados.
d) Nesse ponto ocorre o tratamento no texto para extrair palavras que não possuam valores semânticos. Essas são extraídas semelhantemente à forma apresentada no parágrafo de Indexação; isso se faz necessário para agilizar o processamento, evitando tempo perdido com itens que não auxiliarão no processo de identificação do documento.
e) Nesta etapa ocorrerá a criação dos algoritmos de aprendizagem de máquina.
f) Etapa em que será executado o treinamento dos modelos referentes aos algoritmos:
    - *Bidirecional LSTM;*
    - *Support Vector Machine* (SVM).
g) Após o treinamento, serão realizados testes de performance para avaliar a capacidade do modelo de identificar os atributos do documento, do usuário que poderá acessá-lo, e o tempo que é necessário para esta análise ocorrer.
h) Com base nos resultados apresentados, será feita a análise para identificar se o método proposto nesta pesquisa possui viabilidade para ser colocado em prática levando-se em conta diversos aspectos, como tempo de processamento, investimentos necessários e taxa de acurácia.

Os passos podem ser observados de forma visual na Figura 11.

Figura 11 - Representação visual da metodologia



Fonte: Elaborada pelo autor

3.4 Validação dos Resultados

Com base nos experimentos realizados, os resultados de todos os algoritmos serão inseridos em diversas matrizes de confusão. São matrizes direcionais com duas classes: uma irá expressar os valores reais, e a outra, os valores preditos pelos algoritmos.

Para avaliar a performance dos algoritmos de uma forma mais ampla será utilizada a matriz de confusão, como observado no

Quadro 5, para determinar não somente os acertos, mas observar o resultado como um todo. O uso da matriz de confusão se torna interessante para o trabalho, por além de apresentar o acerto total do algoritmo em questão chamado de verdadeiro positivo, esta também apresenta outros fatores sendo eles.

- Verdadeiro Positivo
  - O algoritmo previu corretamente que o documento pertence ao usuário
- Verdadeiro Negativo
  - O algoritmo previu corretamente que o documento não pertence ao usuário
- Falso Positivo
  - O algoritmo previu erroneamente que o documento pertence ao usuário, neste caso autorizando o acesso indevido ao documento.
- Falso Negativo
  - O documento previu erroneamente que o documento não pertence ao usuário, neste caso negando o acesso ao usuário ao documento que ele deveria ter o acesso.

O quadro 5 demonstra visualmente a matriz de confusão.

Quadro 5 - Matriz de confusão duas classes

		Valor Previsto	
		Classe A	Classe B
Valor Verdadeiro	Classe A	Verdadeiro Positivo	Falso Negativo
	Classe B	Falso Positivo	Verdadeiro Negativo

Fonte: Elaborado pelo autor

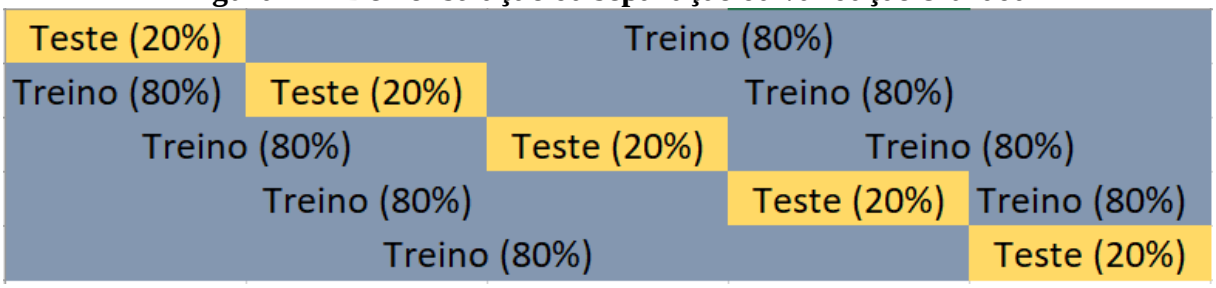
Com a matriz de confusão preenchida, serão extraídos os indicadores abaixo para melhor base de comparação. Os indicadores abaixo serão utilizados como medidas para a comparação de performance dos algoritmos.

- Acurácia
  - (Verdadeiro Positivo + Verdadeiro Negativo) / Total
- Precisão
  - Verdadeiro Positivo / (Verdadeiro Positivo + Falso Positivo)
- Recall
  - Verdadeiro Positivo / Verdadeiro Positivo + Falso Negativo
- F1
  - 2 * Precisão * Recall / Precisão + Recall
- Tempo de classificação
  - Tempo total de testes do algo110ritmo / Número de itens testados

Para certificar que o modelo esteja preparado para lidar com os dados futuros é necessário aplicar um modelo de re-amostragem. Neste caso será utilizado o processo de validação cruzada, o objetivo com este processo é retirarmos o máximo de aleatoriedade de dentro das análises. Enquanto no processo normal iremos escolher 80% do modelo de dados de forma aleatória para o treinamento e 20% para testes, processo que irá produzir os dados de desempenho do modelo, no processo cruzado os dados serão separados e avaliados em cenário de amostragem. A

Figura 12 abaixo ilustra o processo de separação dos valores.

Figura 12 - Demonstração da separação da validação cruzada



Fonte: Elaborada pelo autor

4 EXECUÇÃO

4.1 Preparação

O primeiro passo para a análise foi converter os documentos de PDF para arquivo texto, para isso foi utilizada uma biblioteca capaz de extrair o conteúdo do PDF, com isso, para cada documento foi criado um arquivo de texto com o conteúdo que foi extraído. Isso também resultou na criação de uma grande quantidade de dados que podem ser considerados ruídos, como linhas em branco e diversos espaços.

Com os arquivos de textos extraídos, o próximo passo foi converter esses arquivos de textos para o formato CSV (*Comma-separated values*), formato de arquivo texto onde colunas são separadas por vírgula, com isso, o resultado é um arquivo único em que cada linha irá representar um documento e cada vírgula irá separar os valores que serão utilizados durante o treinamento e teste dos algoritmos. No final desse processo, teremos um novo arquivo contendo todo o conteúdo presente em todo o conjunto de documentos que pretendemos analisar, e logo na frente as categorias rotuladas Investidor e Fundo. Este formato aberto de documento facilita a leitura para outras linguagens e implementação de futuros algoritmos.

4.2 Algoritmos de Aprendizagem

Os algoritmos foram desenvolvidos e executados utilizando a linguagem Python dentro da IDE de desenvolvimento *Jupyter*. Todos os testes foram realizados e mensurados dentro do ambiente de desenvolvimento que conta com um processador Intel Core I9 9900K, 32gb de memória RAM e uma placa de vídeo Nvidia 1070 TI; tais configurações podem afetar diretamente o tempo apresentado para execução e testes.

Como forma de explorar diversos possíveis usos de algoritmos de inteligência artificial, a pesquisa foi executada em diferentes experimentos variando o uso de dados e algoritmo. Espera-se, assim, deixar de maneira mais clara a exposição dos resultados encontrados.

4.2.1 Experimento 1

4.2.1.1 Descrição do experimento

Neste primeiro experimento, o objetivo principal foi ter uma base inicial de acurácia possível de se obter na extração das categorias pretendidas (Investidor e Fundo); foi utilizado o modelo *Support Vector Machine* (SVM) e executado utilizando de um a seis de n-gramas, os

quais foram criados utilizando a biblioteca *TfIdfVectorMachine*, a *API* utilizada para a modelagem foi a SciKit.

Os testes com diversos n-gramas foram feitos a fim de verificar a precisão do modelo *bag of words* (modelo no qual a ordem das palavras não importa). Os testes foram iniciados com a utilização de unigrama e aumentando até o 6-gram, para observar, assim, a diferença e se houve aumento ou não na precisão, e o tempo demandado para predição das categorias.

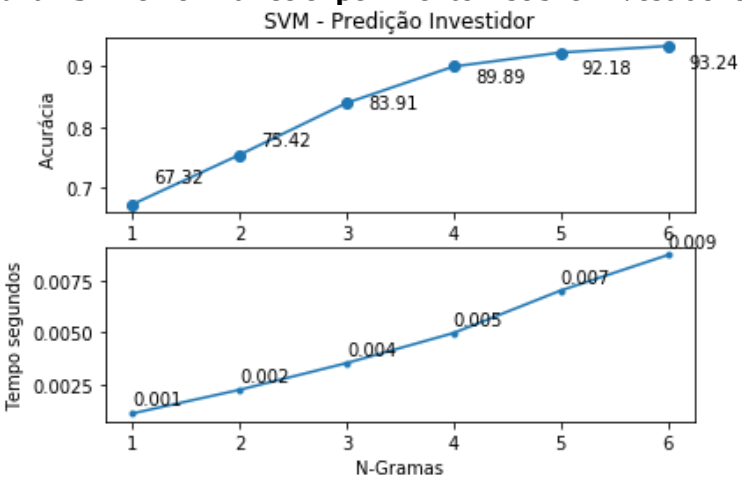
O corpus utilizado foi de documentos do portal de investidor extraídos do Portal A. Este contém 8950 documentos, que foram separados entre documentos para treinamento e documentos para validação, resultando em um total de 7160 documentos para treino/aprendizagem (80%) e de 1790 (20%) para testes de onde o algoritmo será avaliado para medir sua performance.

4.2.1.2 Resultados do experimento

Foi observado um aumento na acurácia do algoritmo acompanhando o aumento do número de n-gramas utilizado, indicando que a ordem das palavras tem bastante importância dentro do corpus utilizado. Entretanto, o tempo gasto para a predição tem um crescimento mais acentuado do que a precisão, indicando também que n-grama utilizado deve ser analisado de acordo com o experimento.

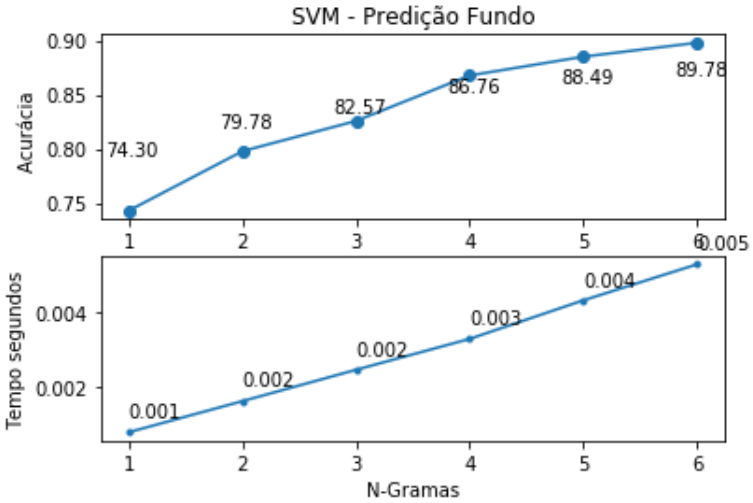
Com o uso de 6-grama o algoritmo foi capaz de ter a melhor acurácia dentro do experimento executado, sendo capaz de predizer 93.24% das vezes o investidor correto quando informado o documento e de acertar em 89.78% a predição de fundo com a mesma entrada de valores, conforme observado nas Figuras 13 e 14.

Figura 13 - Performance experimento 1 sobre Investidores



Fonte: Elaborada pelo autor

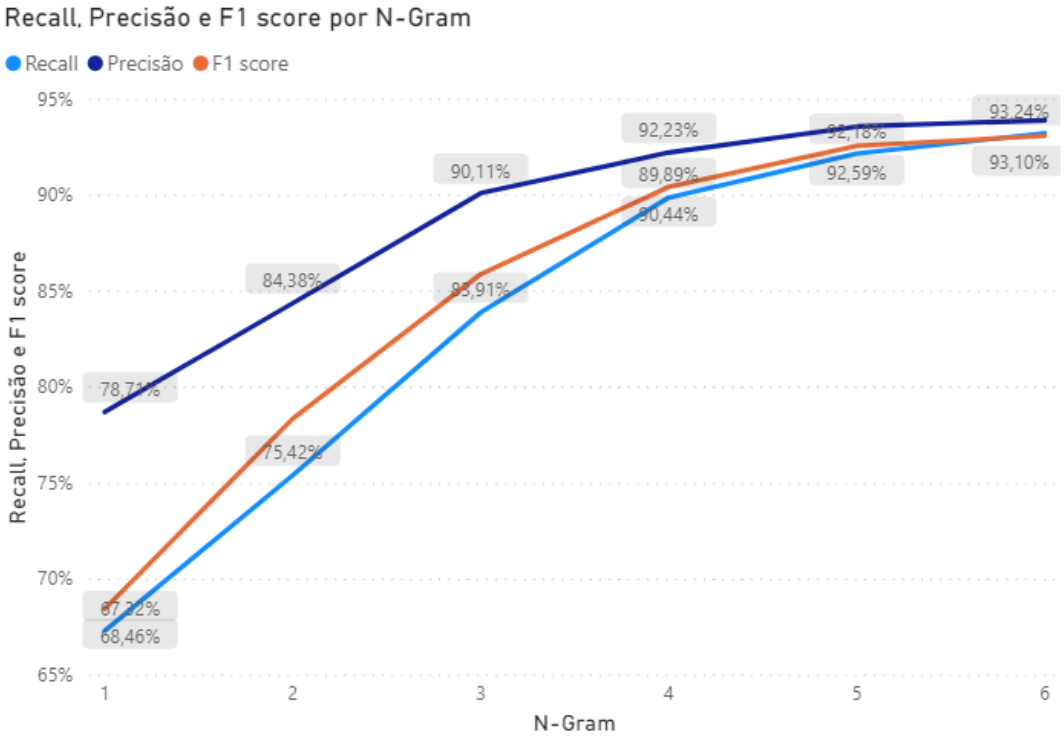
Figura 14 - Performance experimento 1 sobre Fundos



Fonte: Elaborada pelo autor

Na Figura 15 podemos ver a acurácia juntamente com os indicadores de precisão e F1 *score*, durante a classificação dos documentos podemos perceber que, juntamente com a acurácia, todos os indicadores apresentam um aumento dos outros indicadores, o que mantém a tendência observada nos estudos anteriores de que o uso de n-gramas torna o algoritmo em estudo melhor.

Figura 15 - Gráfico de acurácia, precisão e F1 score na classificação de investidores no experimento 1



Fonte: Elaborada pelo autor

4.2.2 Experimento 2

4.2.2.1 Descrição do experimento

Durante este experimento o algoritmo utilizado foi o mesmo do experimento anterior, sem mudanças de códigos, no entanto, a entrada de dados foi alterada, tendo sido inserida, dentro do corpus, a entrada de dois portais distintos: o portal do experimento 1 e um novo portal; isto fornecerá ao software uma entrada de forma mais variada, considerando que foram criados por dois times, com pessoas diferentes, e que estas possuem diferentes escolhas lexicais sobre a forma que é tratado o tema.

Com este experimento procura-se entender como a entrada diversificada irá afetar a aprendizagem do algoritmo. Os portais utilizados são distintos e possuem usuários e volumes de dados diferentes, com isto busca-se descobrir qual é o impacto que causará aumento da massa de dados e como ela irá afetar a acurácia, ou se ela será degradada pois terá uma maior complexidade.

A nova entrada resultou em um total de 10.052 documentos, sendo estes divididos em 8041 documentos para treino/aprendizagem (80%) e 2011 (20%) para testes e apuração da acurácia do algoritmo de aprendizagem. Os resultados apresentados também acompanham o uso da separação de palavras entre 1-grama e 6-grama.

4.2.2.2 Resultados do experimento

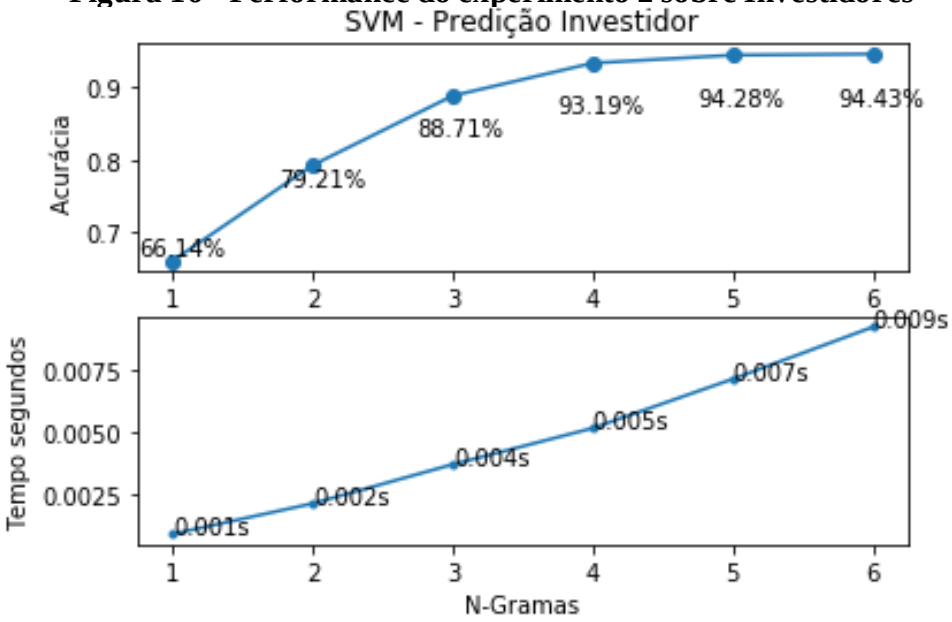
Em relação à predição de investidores, o resultado de maior precisão também foi o de 6-grama, porém, o algoritmo conseguiu acurácias superiores com o uso de n-grama menores; com 4-grama, por exemplo, a diferença da acurácia foi aproximadamente 6,43% superior ao experimento 1, uma diferença superior à diferença final que foi de 1,19%, o que torna possível sinalizar que a massa de dados maior conseguiu melhor fazer com que o algoritmo aprendesse de forma mais eficiente em relação a investidores.

Já em relação à predição de fundos, a diferença teve uma leve queda em 0,62% na acurácia máxima quando comparada ao experimento 1, e apresentando uma curva bem semelhante à curva apresentada dentro do experimento anterior, indicando que dentro da predição de fundos o aprendizado com uma massa mais variada teve uma performance levemente degradada.

A relação de aumento de acurácia com um menor número de n-gramas é interessante por conseguir obter resultados próximos, porém economizando em capacidade computacional durante o treinamento e predição.

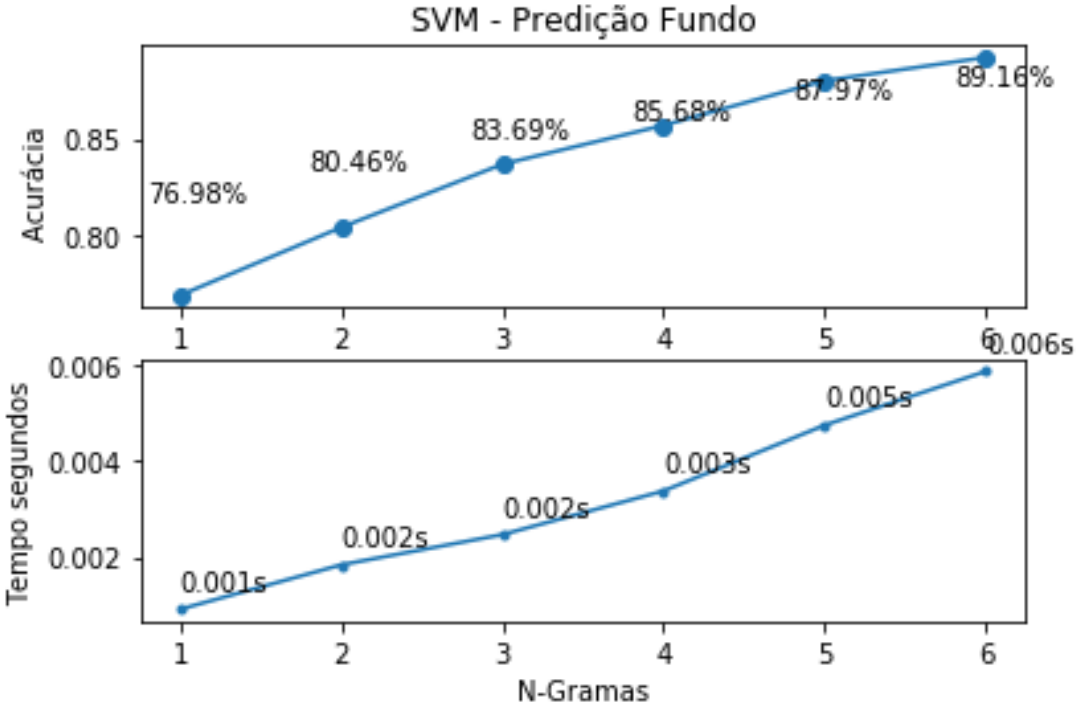
As Figuras 16 e 17 demonstram a evolução dos valores de acurácia e o tempo gasto para a classificação dos documentos.

Figura 16 - Performance do experimento 2 sobre Investidores



Fonte: Elaborada pelo autor

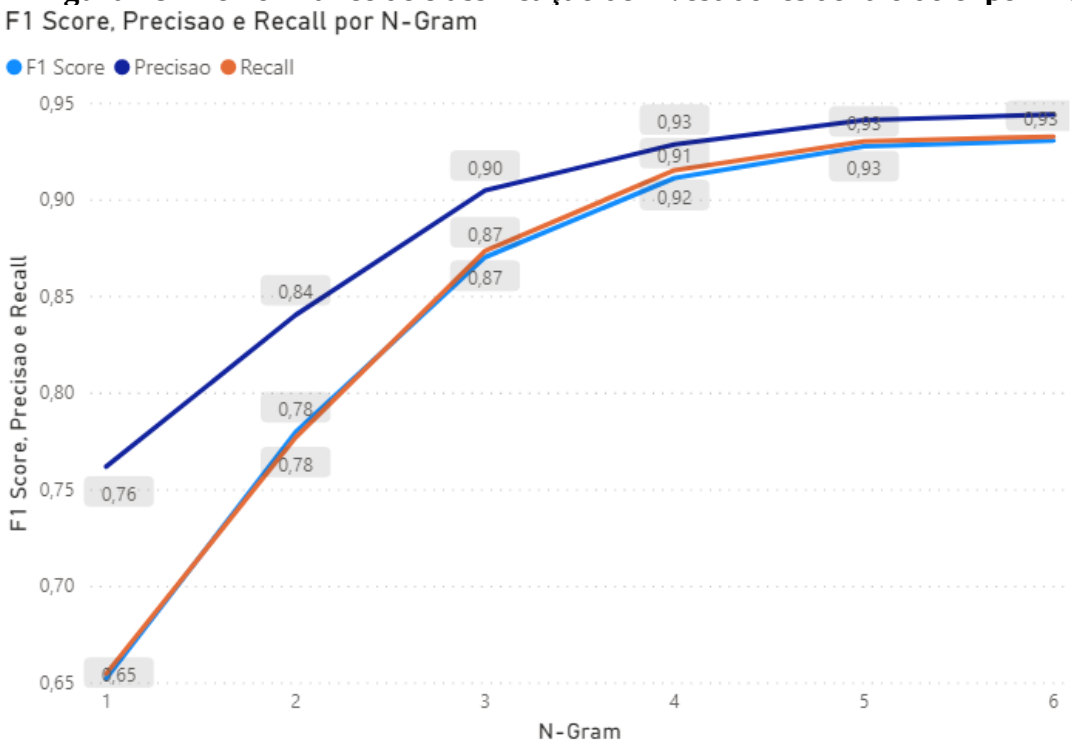
Figura 17 - Performance do experimento 2 sobre Fundos



Fonte: Elaborada pelo autor

A Figura 18 demonstra, assim como no experimento 1, a apresentação dos demais indicadores avaliados já contemplando os dados de classificação de textos presentes dentro dos dados propostos para o experimento 2, em que também podemos observar a tendência de aumento em todos os critérios avaliados ao utilizar n-gram maiores.

Figura 18 - Performance de classificação de investidores dentro do experimento 2



Fonte: Elaborada pelo autor

4.2.3 Experimento 3

4.2.3.1 Descrição do experimento

Neste experimento iremos utilizar a técnica de SVM, mas, no lugar de separarmos nossos dados entre treinamento e teste, iremos fazer a validação cruzada, em que iremos separar os dados em 10 partes e o aprendizado irá ocorrer diversas vezes: 9 partes serão utilizadas para aprendizagem e 1 parte usada para treinamento. Assim, o próximo modelo utilizado para teste será inserido dentro do conjunto de partes para treinamento e outro conjunto será selecionado para os testes, isto será feito até todas as partes serem utilizadas para treinamento e testes.

Os testes com validação cruzada serão realizados com o conjunto de documentos utilizados dentro do experimento 1 e do experimento 2. Espera-se, assim, obter uma resposta

concreta sobre a massa de dados de clientes diferentes, trazendo mais desempenho e eliminando a possibilidade da diferença apresentada ter sido causada devido à aleatoriedade feita durante o momento de escolha de dados para treinamento e testes.

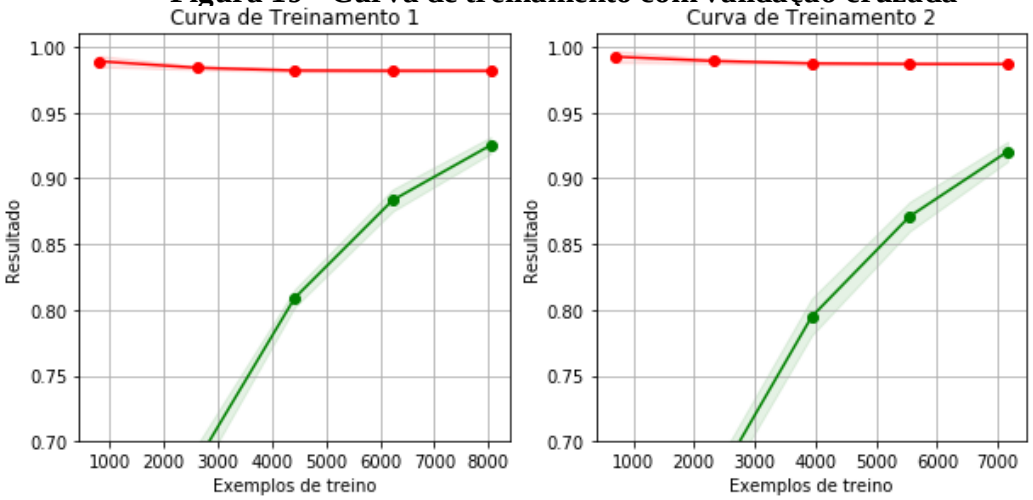
Os testes foram feitos utilizando o modo de 6-gram por ser o resultado de maior valor obtido durante os experimentos anteriores.

4.2.3.2 Resultados do experimento

No gráfico abaixo pode-se observar os resultados com validação cruzada. A linha verde representa o resultado da validação cruzada ao ser treinada com os exemplos e a verde o resultado de cada treinamento.

Comparando as acurácias máximas resultantes da validação cruzada pode-se observar que há o aumento de acurácia com o uso dos dois conjuntos de documentos juntos, porém a acurácia que era de 92,6% passa a ser 93,0%, um aumento desprezível diante do aumento de mais de 2 mil documentos adicionados ao conjunto, como observado na Figura 19.

Figura 19 - Curva de treinamento com validação cruzada



Fonte: Elaborada pelo autor

4.2.4 Experimento 4

4.2.4.1 Descrição do experimento

Para este experimento iremos utilizar o conceito de aprendizagem profunda (*Deep Learning*), utilizando diversas camadas para a aprendizagem. O modelo é utilizado para reconhecimento de entidades nomeadas e foi aplicado neste experimento para extrair o

investidor e o fundo que está contido dentro do texto. Neste experimento, para modelarmos o algoritmo, foi feita a utilização da biblioteca Keras.

Este treinamento foi executado com o uso de uma GPU, pois, devido a sua arquitetura paralelizada, ela possui maior poder de processamento paralelo, que é demandado durante o treinamento do algoritmo criado.

Além das camadas de entrada e de saída, o modelo foi estruturado no total de seis camadas/*features*, que podem ser observadas na Figura 20.

- *Embedding*
  - Esta camada será a responsável por converter a entrada de texto para um formato vetorizado.
- Bidirecional *LSTM*
  - Trata-se de uma camada responsável por gerar os valores de ativação. Estes valores são baseados na entrada de dados fornecidos pela camada anterior (*Embedding*) e irá calcular estes valores baseado nas palavras antecedes e subsequentes.
- *TimeDistributed*
  - Camada responsável por gerar ativações baseadas em valores temporais.

Figura 20 - Camadas do experimento 4

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	(None, 100)	0
embedding_2 (Embedding)	(None, 100, 100)	5387800
bidirectional_2 (Bidirection	(None, 100, 100)	60400
time_distributed_2 (TimeDist	(None, 100, 50)	5050
crf_2 (CRF)	(None, 100, 3)	168

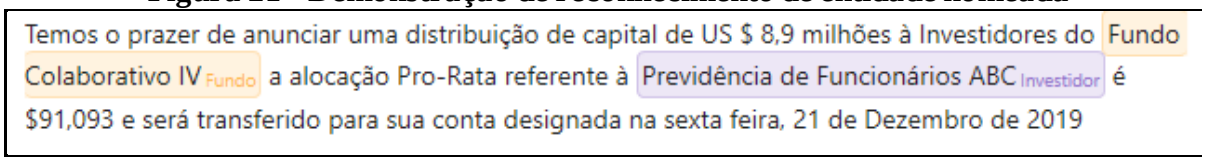
Total params: 5,453,418
Trainable params: 5,453,418
Non-trainable params: 0

Fonte: Elaborada pelo autor

O conceito de entidade nomeada requer que a entrada seja feita por sentenças, logo, o texto foi quebrado em um total de 9375 sentenças nas quais o algoritmo foi treinado para

reconhecer as palavras com a função semântica procurada, conforme demonstrado na Figura 21.

Figura 21 - Demonstração de reconhecimento de entidade nomeada



Fonte: Elaborada pelo autor

O treinamento acontece através de épocas, que são basicamente a quantidade de vezes que o algoritmo irá percorrer o conjunto de dados fornecidos. No experimento utilizado dentro desta pesquisa, foram seis épocas para o treinamento.

As camadas utilizadas são do tipo Bidirecional LSTM (*Long-Short Memory*), em que o algoritmo faz a leitura das sentenças nos dois sentidos, no objetivo de aumentar a capacidade de identificação do algoritmo.

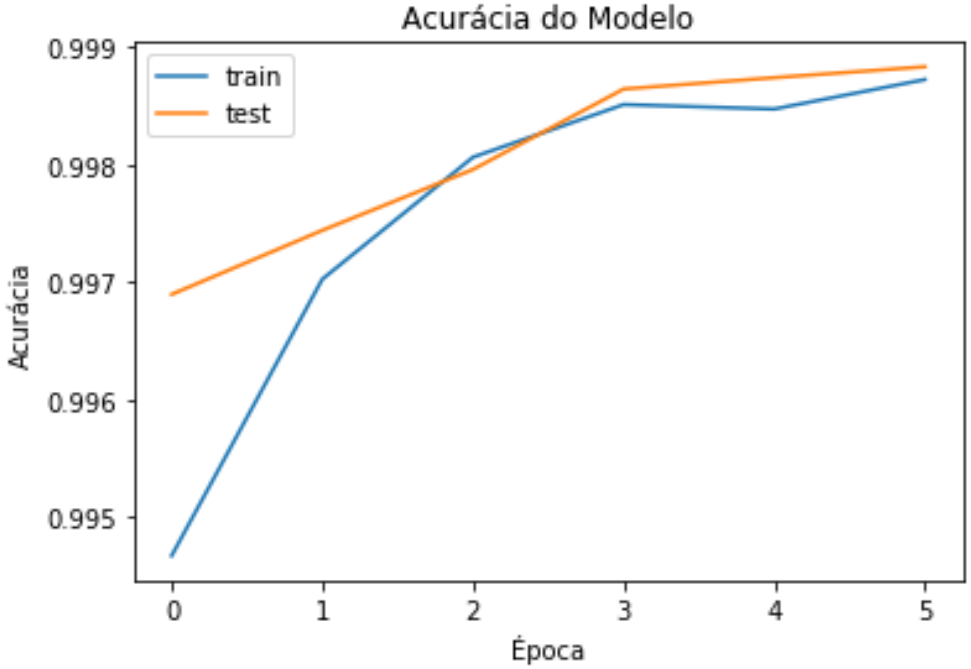
Neste experimento, apesar de todas as sentenças terem sido extraídas dos documentos, somente as sentenças que contêm informação com fundo ou investidor foram utilizadas para treinamento e teste, o que resultou em um total de 9375 sentenças, que foram divididas em 80% para treinamento e 20% para testes.

4.2.4.2 Resultados do experimento

O modelo apresentou uma acurácia final de 99,8%, conforme pode ser observado na

Figura 22, um valor bem superior aos experimentos anteriores, e o tempo médio de predição foi de 0.002s, um tempo semelhante aos testes com 2-grama dos experimentos anteriores, conforme demonstrado na Figura 22.

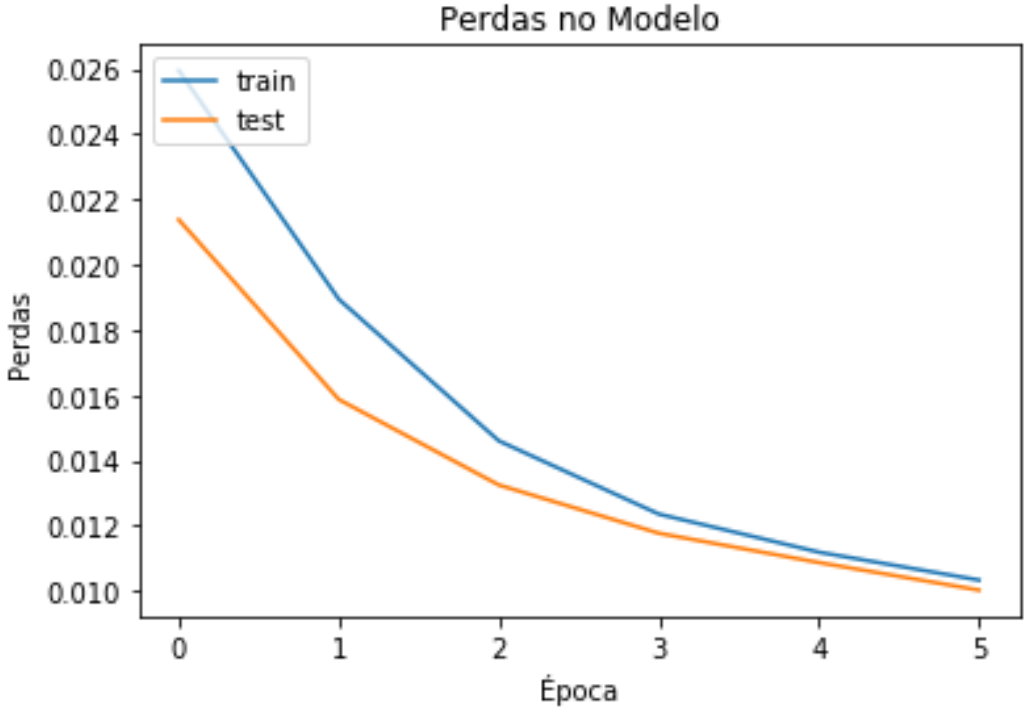
Figura 22 - Acurácia do Modelo no experimento 3



Fonte: Elaborada pelo autor

Quanto à perda, o modelo também vem diminuindo durante as épocas, valor inversamente proporcional com a acurácia, como pode ser observado na Figura 23.

Figura 23 - Perdas no modelo do experimento 3



Fonte: Elaborada pelo autor

Em termos dos outros parâmetros de precisão e *f1 score*, este modelo teve uma média de precisão de 97,1% na classificação das palavras como Investor ou Fundo, e um score F1 de 96,3%, contando com um recall de 95,6%.

4.3 Consolidação dos Resultados

O Quadro 7 demonstra os valores consolidados dos resultados considerados pelo autor como mais importantes dentro do estudo proposto. É interessante notar que o BiLSTM possui a maior acurácia dentre os métodos selecionados e com o tempo menor que a execução do modelo com maior acurácia dentro do SVM. Além da diferença de algoritmos, o uso da GPU para acelerar o processamento também pode ter afetado o tempo necessário para realizar a classificação.

Quadro 6 - Desempenho de Algoritmos para categorização de investidores.

	SVM BAG-OF-WORDS	SVM 6-GRAM	BILSTM
ACURÁCIA	67,3%	93,2%	99,8%
PRECISÃO	79,3%	94,0%	97,1%
RECALL	70,3%	93,0%	95,6%
F1 SCORE	70,4%	92,9%	96,3%
TEMPO	0,0005s	0,004s	0,001s

Fonte: Elaborado pelo autor

Outro cenário observado foi o fato de a classificação do fundo ter menor eficiência mesmo com um conjunto menor de possibilidades do que o observado dentro da classificação de investidores. Possivelmente a causa é a posição de menor destaque para o fundo, como podemos observar no documento Anexo A, o nome do investidor pode ter um destaque maior dentro do texto, até mesmo para a melhor visualização e controle de pessoas que administram o SRI. Isto, com o uso de modelos que não utilizem o modelo *bag-of-words*, pode representar uma vantagem, podendo observar que com aplicação de 1-gram o modelo de fundo apresenta nos dois experimentos acurácia superior de 74% contra 67% no experimento, e 76% contra 66% no experimento 2.

O uso da validação cruzada demonstrou que a diferença do experimento 1 e experimento 2 foi baixa, o que deixa em aberto se o aumento de precisão com o aumento de uma gama mais variada de documento de fato irá ocorrer em futuras implementações com conjunto de dados separados.

5 CONSIDERAÇÕES FINAIS

Através da pesquisa demonstramos que grande parte das falhas nos sistemas Delloite (2012) são advindas de erros humanos e que podem acontecer a qualquer momento durante o período que o usuário está operando o sistema, até mesmo durante uma classificação do documento ou inserção de metadados, criando, assim, um cenário onde para o algoritmo implementado tudo está correto, ou seja, os atributos do usuário compatíveis com os atributos do documento, porém, na prática, o documento contém este atributo informado de forma errônea, e esta má classificação pode ocasionar um mau permissionamento do documento.

Também com a pesquisa de estudos anteriores foi visto que a mineração de texto em documentos é um assunto em constante exploração. Novos métodos têm sido criados, com o avanço da capacidade computacional acessível e cada vez mais atingindo maiores números de precisão e performance, e constante atuação de algoritmos de *Deep Learning*.

Assim, com os experimentos realizados, este trabalho se propôs a analisar a capacidade de algoritmos de mineração de texto em serem capazes de aumentar a segurança de sistemas de recuperação da informação. Com esse propósito, foi traçado o objetivo de avaliar a viabilidade do uso de mineração de textos e aprendizado de máquina na otimização do processo de controle de acesso de documentos dentro de sistemas de recuperação da informação da área financeira.

Para o atingimento do objetivo principal, foram traçados objetivos específicos, primeiramente com o objetivo de identificar métodos capazes de avaliar o conteúdo do documento, com o foco em extrair os atributos que são utilizados para o controle de acesso. Assim, foi feito o levantamento do estado da arte, dentre a classificação de textos, e foi encontrado, nas redes neurais, o modelo de *Long Short Term Memory*, aplicado de forma bi-direcional. Este método foi explorado no trabalho com o objetivo de saber sua performance e seu tempo de execução. Como redes neurais podem ter uma alta carga computacional, foi avaliada também a performance do algoritmo fora da gama de redes neurais, sendo assim, foi escolhido o *Support Vector Machine* devido a sua grande utilização em trabalhos relacionados.

Após a execução dos algoritmos, ficou evidenciado, no trabalho, o seu desempenho, dentro dos experimentos realizados. Cada experimento buscou analisar situações distintas que pudessem ser aplicadas para atingir a melhor performance no processo proposto de extrair os atributos que posteriormente serão utilizados como base do sistema para o controle de acesso ao documento.

No estudo do SVM foi observado um aumento de acurácia com o uso de n-gramas, o que descarta a possibilidade de uso do tipo *bag-of-words*, que obteve uma acurácia de 66.14%

e 76.98% para a classificação de investidor e fundo, respectivamente. Porém, apresentou resultados melhores a cada aumento de n-grama durante o treinamento, chegando à acurácia de 94,43% e 89,16% para a classificação de investidor e fundo, respectivamente, quando utilizado 6-gram. Foi observado no experimento 2 que o uso de documentos de fontes variadas teve um impacto levemente positivo na acurácia na classificação do documento. Assim, para uma melhor comparação, os dois experimentos foram executados com a avaliação feita de modo de validação cruzada, o que demonstra que o incremento de documentos aumentou a performance de forma sutil.

Já no campo das redes neurais, resultados de acurácia foram satisfatórios para os objetivos da pesquisa, a rede neural foi capaz de alcançar uma performance de 99,8% na identificação de investidores e fundos dentro de sentenças e com uma velocidade de cálculo de 0,001s, mais preciso e com uma classificação mais ágil do que outros algoritmos. Um resultado relativamente superior ao apresentado nos outros exemplos foi que este método é baseado na sequência das palavras, logo, o estudo com diferentes n-gramas não foi relevante.

Com os dados apresentados torna-se possível atingir o último objetivo específico, o de analisar resultados para identificação de possível uso e possíveis formas de implementação.

Com uma acurácia superior a 90% e um tempo de processamento inferior a 0,1s, os algoritmos de SVM com n-grama 6 e o algoritmo de BiLSTM são uma poderosa ferramenta para a obtenção de informações sobre o documento de forma automática, que independe da informação inserida pelo usuário do sistema. Considerando cenários de sistemas que utilizam métodos de controle de acesso como o ACL, ABAC, RBAC, os atributos são essenciais durante este controle, e os experimentos realizados demonstram que é possível a extração destas informações de dentro dos documentos, podendo assim ser utilizados das formas descritas abaixo.

Um dos métodos pode agir de maneira preventiva, utilizando-se do código de análise de documento para a verificação dos atributos durante o momento de cadastro do documento, e apresentando para o usuário possíveis erros de classificação e uma base mais confiável, mas documentos legados continuam sem a cobertura desta nova implementação.

Outro método é uma execução agendada em que os algoritmos iriam analisar os documentos armazenados no sistema de recuperação da informação e informar ao usuário sobre possíveis erros durante a inserção ou alteração dos atributos. Este método traz como vantagem poder tratar documentos legados, mas como desvantagem a necessidade de uma atuação tardia do usuário, o que pode deixar uma lacuna de tempo até o problema ser identificado e corrigido.

Como última opção, um módulo poderá analisar todo o tráfego requisitado pelo usuário, e com isso executar o algoritmo no arquivo requisitado, compararando as listas de acesso às quais o usuário tem permissão. Com isso, o sistema pode bloquear a saída do arquivo caso identifique que a classificação deste documento não é a mesma a qual o usuário deveria ter acesso. Este método tem a vantagem de tratar todos os documentos, mas a desvantagem de causar indisponibilidade de documentos em alguma situação específica, em que existe 0,2% de erro, e onde não foi possível identificar o atributo, podendo causar uma indisponibilidade, além de incrementar o tempo para o *download* do arquivo.

Documentos da área financeira podem conter informações altamente sensíveis, com risco de perda de valores financeiros e/ou danos à privacidade de pessoas, sendo assim, devemos nos esforçar cada dia mais para a proteção de dados pessoais dos usuários. Esta necessidade cria gastos em segurança da informação justificáveis e relevantes, pois em casos de vazamentos o dano à imagem da empresa e possíveis multas previstas em leis de proteção a dados – que vem sendo implantadas pelo mundo, como *General Data Protection Regulation* (Lei Europeia que regula os dados pessoais de cidadãos europeus) e Lei Geral de Proteção de Dados (que faz o mesmo papel no Brasil) – podem exceder em muito os valores em investimentos. Por isso, a necessidade de implementação das soluções propostas.

Com o conhecimento gerado é possível afirmar que ferramentas de mineração de texto são uma forma de se otimizar o processo de controle de acesso de documentos dentro de sistemas de informações da área financeira. Através do uso dos algoritmos é possível se obter os atributos através do conteúdo dos documentos, com isso, uma segunda fonte de informação pode ser utilizada dentro de controles de acesso como RBAC, ABAC ou ACL, todos dependentes do atributo inserido e se beneficiando dos algoritmos que irão extrair a informação do texto.

5.1 Sugestão de estudos posteriores

Como sugestão para estudos posteriores, registra-se o estudo utilizando algoritmos de pré-treinamento. Estes algoritmos demandam uma base de dados mais ampla em que a máquina saberá identificar estruturas semânticas, mas ao custo de um maior esforço computacional.

REFERÊNCIAS

ARANHA, Christian; PASSOS, Emmanuel. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v. 5, n. 2, p. 1-8, ago. 2006. https://doi.org/10.21529/RESI.2006.0502001. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171. Acesso em: fev. 2021.

ARAUJO, Pedro Henrique Luz de *et al.* VICTOR: a dataset for Brazilian legal documents classification. In: CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12., 2020, Marseille, FR. **Proceedings of the […].** Marseille: European Language Resources Association, 2020. p. 1449-1458. Disponível em: https://www.aclweb.org/anthology/2020.lrec-1.181.pdf. Acesso em: fev. 2021.

ARAÚJO, Vania Maria Rodrigues Hermes de. Sistemas de informação: nova abordagem teórico-conceitual. **Ciência da Informação**, v. 24, n. 1, 1995. Disponível em: http://revista.ibict.br/ciinf/article/view/610/612. Acesso em: fev. 2021.

BARBOSA, Alice Príncipe. **Teoria e prática dos sistemas de classificação bibliográfica**. Rio de Janeiro: Instituto Brasileiro de Bibliografia e Documentação, 1969.

BARBOSA, Tharlis da Silva; REIS, Flávio Alexandre dos. O uso de ferramentas open source para aplicações de segurança em redes corporativas: um estudo baseado em firewalls. **Revista Saber Digital**, v. 5, n. 1, p. 72-90, dez. 2012. Disponível em: http://revistas.faa.edu.br/index.php/SaberDigital/article/view/650/514. Acesso em: fev. 2021.

BARION, Eliana Cristina Nogueira; LAGO, Décio. Mineração de textos. **Revista de Ciências Exatas e Tecnologia**, v. 3, n. 3, p. 123-140, 2008. https://doi.org/10.17921/1890-1793.2008v3n3p123-140. Disponível em: https://revista.pgsskroton.com/index.php/rcext/article/view/2372/2276. Acesso em: fev. 2021.

BEAM, Andrew L.; KOHANE, Isaac S. Big data and machine learning in health care. **JAMA**, v. 319, n. 13, p. 1317-1318, 2018. https://doi.org/10.1001/jama.2017.18391.

BENGIO, Yoshua. **Learning deep architectures for al:** foundations and trends(r) in machine learning. 2. ed. Hanover: Now Publishers, 2009.

BIANCHI, Luciano Henrique Teixeira; FONSECA, Jacqueline. **Java Athentication e Authorization Service como mecanismo de segurança e controle de acesso em aplicações web.** Instituto de Gestão em Tecnologia da Informação, 2017. Disponível em: https://core.ac.uk/download/pdf/211938542.pdf. Acesso em: fev. 2021.

BRASIL. Lei nº 12.682, de 9 de julho de 2012. Dispõe sobre a elaboração e o arquivamento de documentos em meios eletromagnéticos. **Diário Oficial da União**, Brasília, 10 jul. 2012. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12682.htm. Acesso em: fev. 2021.

CAMASTRA, Francesco; VINCIARELLI, Alessandro. **Machine learning for audio, image and video analysis:** Theory and applications. New York: Springer, 2008.

CARDOSO, Olinda Nogueira Paes. Recuperação de informação. **INFOCOMP: Journal of Computer Science**, v. 2, n. 1, p. 33-38, nov. 2004. Disponível em: http://infocomp.dcc.ufla.br/index.php/infocomp/article/view/46/31. Acesso em: fev. 2021.

CARLOS JUNIOR, Luís Fernando Martins. **Reconhecimento facial utilizando redes neurais.** 2011. Monografia (Conclusão de Curso) Centro Universitário Eurípides de Marília UNIVEM, Marília, SP, 2011. Disponível em: https://aberto.univem.edu.br/bitstream/handle/11077/360/Reconhecimento%20Facial%20Utilizando%20Redes%20Neurais.pdf?sequence=1&isAllowed=y. Acesso em: fev. 2021.

CASTRO, Tiago O.; MACEDO, Daniel F.; SANTOS, Aldri. Controle de acesso IoT escalável e ciente de contexto suportando múltiplos usuários. In: WORKSHOP DE GERÊNCIA E OPERAÇÃO DE REDES E SERVIÇOS, 23., 2018, Campos do Jordão, SP. **Anais [...].** Campos do Jordão, 2018. Disponível em: https://sol.sbc.org.br/index.php/wgrs/article/view/2366/2330. Acesso em: fev. 2021.

CAUDILL, Maureen. Neural networks primer, part I. **AI Expert**, v. 2, n. 12, p. 46-52, Dec. 1987.

CHEN, Chih-Ming et al. Collaborative similarity embedding for recommender systems. In: WWW '19: THE WORLD WIDE WEB CONFERENCE, 2019, San Francisco, USA. **Proceedings of the [...].** New York: Association for Computing Machinery, 2019. p. 2637-2643. Disponível em: https://arxiv.org/pdf/1902.06188.pdf. Acesso em: fev. 2021.

CHOI, Keunwoo *et al.* Convolutional recurrent neural network for music classification. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2017, New Orleans, USA. **Proceedings of the [...].** New Orleans: IEEE, 2017. p. 2392-2396. https://doi.org/10.1109/ICASSP.2017.7952585. Disponível em: https://arxiv.org/pdf/1609.04243.pdf. Acesso em: fev. 2021.

CHRISTOPHER, O. **Understanding LSTM Networks.** 2015. Disponível em: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Acesso em: 15 jun. 2020.

CORREA, Adriana Cristina Giusti. **Recuperação de documentos baseada em informação semântica no ambiente AMMO**. 2003. 102 f. Dissertação (Mestrado em Ciências Exatas e da Terra) - Universidade Federal de São Carlos, São Carlos, 2003. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/522/DissACGC.pdf?sequence=1&isAllowed=y. Acesso em: fev. 2021.

CORTES, Corina; VAPNIK, Vladimir. Support-vector networks. **Machine Learning**, v. 20, p. 237-297, 1995. https://doi.org/10.1007/BF00994018. Disponível em: https://link.springer.com/content/pdf/10.1007/BF00994018.pdf. Acesso em: fev. 2021.

COSTA, Christiano M. *et al.* Aplicação de SDN no gerenciamento de perfis de usuário em dispositivos de rede. In: XXXVI SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS, 36., 2018, Campos do Jordão, SP. **Anais [...].** Porto Alegre: Sociedade Brasileira de Computação, 2018. Disponível em: https://sol.sbc.org.br/index.php/sbrc_estendido/article/view/2506/2468. Acesso em: fev. 2021.

CRONIN, Blaise. Esquemas conceituais e estratégicos para a gerência. **Revista da Escola de Biblioteconomia da UFMG**, v. 19, n. 2, p. 195-220, 1990. Disponível em: http://hdl.handle.net/20.500.11959/brapci/70793. Acesso em: fev. 2021.

DELOITTE. **2012 global financial services industry security study.** 2012. Disponível em: https://www2.deloitte.com/za/en/pages/financial-services/articles/global-fs-industry-security-study-2012.html. Acesso em: 26 jun. 2020.

DUVAL, Erik *et al.* Metadata principles and practicalities. **D-Lib Magazine**, v. 8, n. 4, p. 1-10, Apr. 2002. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.8711&rep=rep1&type=pdf. Acesso em: fev. 2021.

EDUNOV, Segey *et al.* Understanding back-translation at scale. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2018, Brussels. **Proceedings of the [...].** Stroudsburg: Association for Computational Linguistics, 2018. p. 489-500. Disponível em: https://arxiv.org/pdf/1808.09381.pdf. Acesso em: fev. 2021.

EUI-HONG, Ham; KARYPIS, George. Centroid-based document classification: analysis and experimental results. In: In: ZIGHED, D. A.; KOMOROWSKI, J.; ŻYTKOW J. (Ed.). **Principles of data mining and knowledge discovery.** Berlin: Springer, 2002. https://doi.org/10.1007/3-540-45372-5_46.

FAN , D.-P. *et al.* PraNet: parallel reverse attention network for polyp segmentation. In: MARTEL, A. L. *et al.* (Ed.). **Medical image computing and computer assisted intervention**.Berlim: Springer, 2020. https://doi.org/10.1007/978-3-030-59725-2_26.

FERNANDES, Nélia O. Campo. **Segurança da informação**. Cuiabá: Rede e-Tec Brasil, 2013. Disponível em: http://proedu.rnp.br/bitstream/handle/123456789/1538/15.6_versao_Finalizada_com_Logo_IFRO-Seguranca_Informacao_04_04_14.pdf?sequence=1&isAllowed=y. Acesso em: fev. 2021.

FERRAIOLO, David; CUGINI, Janet; KUHN, Richard. Role-based access control (RBAC): features and motivations. ANNUAL COMPUTER SECURITY APPLICATIONS CONFERENCE, 11., 1995, New Orleans, USA. **Proceedings of the [...].**, Maryland: Applied Computer Security Associates, 1995. p. 241-48. Disponível em: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=916537. Acesso em: fev. 2021.

GALLAHER, Michael P. *et al.* **The economic impact of role-based access control**. Washington: National Institute of Standards & Technology, 2002. Disponível em: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=916549. Acesso em: fev. 2021.

GORDON, Lawrence A.; LOEB, Martin P. The economics of information security investment. **ACM Transactions on Information and System Security**, v. 5, n. 4, p. 438-457, Nov. 2002. https://doi.org/10.1145/581271.581274

GRAVES, Alex; FERNANDEZ, Santiago; SCHMIDHUBER, Jürgen. Multi-dimensional recurrent neural networks. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL NEURAL NETWORKS, 17., 2007, Berlin. **Proceedings of the [...].** Berlin: Springer, 2007. p. 549-558. https://doi.org/10.1007/978-3-540-74690-4_56. Disponível em: https://arxiv.org/pdf/0705.2011.pdf. Acesso em: fev. 2021.

GRAVES, Alex; SCHMIDHUBER, Jürgen. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. **Neural Networks**, v. 18, n. 5-6, p. 602-610, July/Aug. 2005. https://doi.org/10.1016/j.neunet.2005.06.042.

GREFF, Klauss *et al.* LSTM: A search space odyssey. **IEEE Transactions on Neural Networks and Learning Systems**, v. 28, n. 10, p. 2222-2232, Mar. 2015. https://doi.org/10.1109/TNNLS.2016.2582924.

GU, Zaiwang *et al*. CE-Net: context encoder network for 2D medical image segmentation. **IEEE Transactions on Medical Imaging**, v. 38, n. 10, p. 2281-2292, Oct. 2018. http://doi.org/10.1109/TMI.2019.2903562. Disponível em: https://arxiv.org/pdf/1903.02740.pdf. Acesso em: fev. 2021.

HARNAD, Stevan. The annotation game: on Turing (1950) on computing, machinery, and intelligence. In: EPSTEIN, Robert; PETERS, Grace (Ed.). **The Turing test sourcebook:** philosophical and methodological issues in the quest for the thinking computer. Boston: Kluwer, 2008. p. 23–66

HAVAEI, Mohammad *et al.* Brain tumor segmentation with deep neural networks. **Medical Image Analysis**, v. 35, p. 18-31, Jan. 2017. https://doi.org/10.1016/j.media.2016.05.004.

HOCHREITER, Seep; SCHMIDHUBER, Jürgen. **Neural Computation,** v. 9, n. 8, p. 1735-1780, 1997.

HU, Vicent C. et al**. Guide to attribute based access control (ABAC) definition and considerations**. Washington: National Institute of Standards and Technology, 2013. (NIST special publication, 800-162). https://doi.org/10.6028/NIST.SP.800-162. Disponível em: https://nvlpubs.nist.gov/nistpubs/specialpublications/NIST.sp.800-162.pdf. Acesso em: fev. 2021.

HUANG, Zhiheng; XU, Wei; YU, Kai. **Bidirectional LSTM-CRF models for sequence tagging.** 2018. Disponível em: https://static.aminer.cn/upload/pdf/604/1492/663/573695fe6e3b12023e511e25_0.pdf. Acesso em: fev. 2021.

INGLESANT, Philip *et al*. Expressions of expertness: the virtuous circle of natural language for access control policy specification. In: SYMPOSIUM ON USABLE PRIVACY AND SECURITY, 4., 2008, Pittsburgh, USA. **Proceedings of the […].** Baltimore: USENIX Association, 2008. p. 77-88. https://doi.org/10.1145/1408664.1408675.

JHA, Debesh *et al*. DoubleU-Net: a deep convolutional neural network for medical image segmentation. In: IEEE INTERNATIONAL SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS, 33., 2020, Minnesota, USA. **Proceedings of the […].** New Orleans: IEEE, 2020. p. 558-564. Disponível em: https://arxiv.org/pdf/2006.04868.pdf. Acesso em: fev. 2021.

JÚNIOR, José R. F. **Redes neurais recorrentes: LSTM**. Medium, 2019. Disponível em: https://medium.com/@web2ajax/redes-neurais-recorrentes-lstm-b90b720dc3f6. Acesso em: fev. 2021.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, v. 521, p. 436-444, 2015. https://doi.org/10.1038/nature14539. Disponível em: https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf. Acesso em: fev. 2021.

MAIA, Luiz Cláudio Gomes; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectiva em Ciência da Informação**, v. 15, n. 1, p. 154-172, jan./abr. 2010. Disponível em: http://portalde periodicos.eci.ufmg.br/index.php/pci/article/download/875/717. Acesso em: fev. 2021.

MANEVITZ, Larry M.; YOUSEF, Malik. One-class SVMs for document classification. **Journal of Machine Learning Research,** v. 2, n. 1, p. 139-154, 2001. https://doi.org/10. 1162/15324430260185574. Disponível em: https://www.jmlr.org/papers/volume2/manevitz 01a/manevitz01a.pdf. Acesso em: fev. 2021.

MARUMO, Fabiano Shiiti. ***Deep learning* para classificação de *fake news* por sumarização de texto.** 2018. Monografia (Conclusão de curso) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Londrina, PR, 2018. Disponível em: http://www.uel.br/ cce/dc/wp-content/uploads/Fabiano_Preliminar.pdf. Acesso em: fev. 2021.

MILSTEAD, Jessica; FELDMAN, Susan. Metadata: cataloging by any other. **Online,** v. 23, n. 1, p. 24-31, Jan./Feb. 1999.

MYRONENKO, Andriy. 3D MRI brain tumor segmentation using autoencoder regularization. In: INTERNATIONAL WORKSHOP, BRAINLES, 4., 2018, Granada, Spain. **Proceedings of the […].** New York: Springer, 2018. p.311-320. https://doi.org/10.1007/978-3-030-11726-9_28. Disponível em: https://arxiv.org/pdf/1810.11654.pdf. Acesso em: fev. 2021.

NAROUEI, Massoud; KHANPOUR, Hamed; TAKABI, Daniel. Identification of access control policy sentences from natural language policy documents. In: IFIP ANNUAL CONFERENCE ON DATA AND APPLICATIONS SECURITY AND PRIVACY, 31., 2017, Philadelphia, USA. **Proceedings of the […].** New York: Springer, 2017. p. 82-100. https:// doi.org/10.1007/978-3-319-61176-1_5.

NIVEDITA, Naidu; DHARASKAR, Rajiv Vasantrao. An effective approach to network intrusion detection. **International Journal of Computer Applications,** v. 1, n. 3, p. 26-32, 2010. https://doi.org/10.5120/89-188.

OORD, Aaron van den; DIELEMAN, Sander. Recommending music on spotify with deep learning. In: NEURAL INFORMATION PROCESSING SYSTEMS CONFERENCE, 28., 2014, Montreal. **Proceedings of the […].** San Diego: Neural Information Processing Systems, 2014.

PARK, Chanyoung *et al.* Collaborative Translational Metric Learning. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 2018, Singapore. **Proceedings of the […].** New Orleans: IEEE, 2018. p. 367-376. https://doi.org/10.1109/ICDM.2018.00052.

PEDREGOSA, Fabian et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011. Disponível em: https://www.jmlr.org/ papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: fev. 2021.

PETERS, Matthew E. *et al*. Deep contextualized word representations. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2018, New Orleans. **Proceedings of the […].** New Orleans: Association for Computational Linguistics, 2018. Disponível em: http://arxiv.org/ abs/1802.05365. Acesso em: fev. 2021.

PONTI, Moacir Antonelli; COSTA, Gabriel B. Paranhos da. Como funciona o Deep Learning. In: VIEIRA, Vaninha; RAZENTE, Humberto L.; BARIONI, Maria Camila N. (Org.). **Tópicos em gerenciamento de dados e informações.** Uberlândia: Sociedade Brasileira de Computação, 2017. p. 63-93.

RANA, Shweta; SINGH, Archana. Comparative analysis of sentiment orientation using SVM and Naïve Bayes techniques. In: INTERNATIONAL CONFERENCE ON NEXT GENERATION COMPUTING TECHNOLOGIES, 2., Dehradun. **Proceedings of the […].** New Orleans: IEEE, 2016. https://doi.org/10.1109/NGCT.2016.7877399.

RIKTERS, Matīss; PINNIS, Mārcis; KRIŠLAUKS Rihards. Training and adapting multilingual NMT for less-resourced and morphologically rich languages. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 11., 2018, Miyazaki, Japan. **Proceedings of the […].** Luxemburg: European Language Resources Association, 2018.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência artificial**. Rio de Janeiro: Elsevier Brasil, 2014.

SATO, Mazakazu *et al.* Application of deep learning to the classification of images from colposcopy. **Oncology Letters**, v. 15, n. 3, p. 3518-3523, Mar. 2018. https://doi.org/10.3892/ol.2018.7762. Disponível em: https://www.spandidos-publications.com/10.3892/ol.2018.7762/download. Acesso em: fev. 2021.

SENNRICH, Rico; HADDOW, Barry; BIRCH Alexandra. Neural machine translation of rare words with subword units. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 54., 2016, Berlin. **Proceedings of the […].** Berlin: Association for Computational Linguistics, 2016. p. 1715-1725. Disponível em: https://arxiv.org/pdf/1508.07909.pdf. Acesso em: fev. 2021.

SHEN, Haibo. A Semantic-aware attribute-based access control model for web services. In: In: HUA, A.; CHANG, S. L. (Ed.). **Algorithms and architectures for parallel processing.** Berlin: Springer, 2009. p. 693-703.https://doi.org/10.1007/978-3-642-03095-6_65.

SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, v. 11, n. 2, p. 161-173, maio/ago. 2006. https://doi.org/10.1590/S1413-99362006000200002. Disponível em: https://www.scielo.br/pdf/pci/v11n2/v11n2a02.pdf. Acesso em: fev. 2021.

TECHTARGET CONTRIBUTOR. **Access control list (ACL).** 2006. Disponível em: https://searchsoftwarequality.techtarget.com/definition/access-control-list. Acesso em: fev. 2021.

VALENTIM, Marta Ligia Pomim; ANÇANELO, Juliana Venâncio. Análise de conceitos sobre valor da informação no âmbito da ciência da informação. **Convergências em Ciência da Informação,** v. 1, n. 1, p. 26-43, 2018. https://doi.org/10.33467/conci.v1i1.9343. Disponível em: https://seer.ufs.br/index.php/conci/article/view/9343. Acesso em: fev. 2021.

VAN VEEN, Fjodor. **Neural network zoo prequel**: cells and layers. The Asimov Institute, 2016. Disponível em: https://www.asimovinstitute.org/author/fjodorvanveen/. Acesso em: 01 jun. 2020.

XIAO, Xusheng *et al.* Automated extraction of security policies from natural-language software documents. INTERNATIONAL SYMPOSIUM ON THE FOUNDATIONS OF SOFTWARE ENGINEERING, 20., 2012, New York. **Proceedings of the […].** New York: Association for Computing Machinery, 2012. https://doi.org/10.1145/2393596.2393608.

ZHANG, Harry. The optimality of naive bayes. In: INTERNATIONAL FLORIDA ARTIFICIAL INTELLIGENCE RESEARCH SOCIETY CONFERENCE, 17., 2004, Miami Beach. **Proceedings of the […].** Florida: AI Research Society, 2004. Disponível em: https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf. Acesso em: 01 jun. 2020.

ZHOU, C. et al. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. **IEEE Transactions on Image Processing**, v. 29, p. 4516-4529, Feb. 2020. https://doi.org/10.1109/TIP.2020.2973510.

APÊNDICES

APÊNDICE A – Exemplo demonstrativo de formato de Documento

21 de dezembro de 2018

Atenção: Hector Vieira

Re: $X.X Milhões de distribuição de capital de Fundo de Investimento ABC

Caro Investidor:

O Fundo de Investimento ABC está fazendo uma distribuição de caixa de US $ X, X milhões para seus associados. Esta distribuição é da empresa de portfólio da Empresa XYZ Ltda, na qual a Empresa investiu US $ X, X milhões em dezembro de 2017. Esses recursos incluem US $ X, X milhões em distribuições do veículo de investimento estabelecido para esse investimento.

Sua alocação proporcional dessa distribuição referente ao Investidor ABC, US $ XXX.XXX, será transferida para sua conta designada na sexta-feira, 21 de dezembro de 2018. Estimamos que sua renda tributável proporcional gerada é de US $ XXX.XXX para o ano encerrado em 31 de dezembro de 2018.

Obrigado, como sempre, por seu apoio contínuo. Se você tiver alguma dúvida, não hesite em entrar em contato conosco.

Atenciosamente,

APÊNDICE B – Formulário 1065 para informações fiscais

651119

Final K-1 ☐ Amended K-1 ☐ OMB No. 1545-0123

Schedule K-1 (Form 1065) 2019

Department of the Treasury
Internal Revenue Service

For calendar year 2019, or tax year beginning / / 2019 ending / /

Partner's Share of Income, Deductions, Credits, etc. ▶ See back of form and separate instructions.

Part I Information About the Partnership

A Partnership's employer identification number

B Partnership's name, address, city, state, and ZIP code

C IRS Center where partnership filed return ▶

D ☐ Check if this is a publicly traded partnership (PTP)

Part II Information About the Partner

E Partner's SSN or TIN (Do not use TIN of a disregarded entity. See inst.)

F Name, address, city, state, and ZIP code for partner entered in E. See instructions.

G ☐ General partner or LLC member-manager ☐ Limited partner or other LLC member

H1 ☐ Domestic partner ☐ Foreign partner

H2 ☐ If the partner is a disregarded entity (DE), enter the partner's:
TIN _____ Name _____

I1 What type of entity is this partner? _____

I2 If this partner is a retirement plan (IRA/SEP/Keogh/etc.), check here ☐

J Partner's share of profit, loss, and capital (see instructions):

	Beginning	Ending
Profit	%	%
Loss	%	%
Capital	%	%

Check if decrease is due to sale or exchange of partnership interest . . ☐

K Partner's share of liabilities:

	Beginning	Ending
Nonrecourse	$	$
Qualified nonrecourse financing	$	$
Recourse	$	$

☐ Check this box if Item K includes liability amounts from lower tier partnerships.

L **Partner's Capital Account Analysis**

Beginning capital account	$
Capital contributed during the year	$
Current year net income (loss)	$
Other increase (decrease) (attach explanation)	$
Withdrawals & distributions	$ ( )
Ending capital account	$

M Did the partner contribute property with a built-in gain or loss?
☐ **Yes** ☐ **No** If "Yes," attach statement. See instructions.

N **Partner's Share of Net Unrecognized Section 704(c) Gain or (Loss)**
Beginning . . . . . $
Ending . . . . . $

Part III Partner's Share of Current Year Income, Deductions, Credits, and Other Items

1	Ordinary business income (loss)	15	Credits
2	Net rental real estate income (loss)		
3	Other net rental income (loss)	16	Foreign transactions
4a	Guaranteed payments for services		
4b	Guaranteed payments for capital		
4c	Total guaranteed payments		
5	Interest income		
6a	Ordinary dividends		
6b	Qualified dividends		
6c	Dividend equivalents	17	Alternative minimum tax (AMT) items
7	Royalties		
8	Net short-term capital gain (loss)		
9a	Net long-term capital gain (loss)	18	Tax-exempt income and nondeductible expenses
9b	Collectibles (28%) gain (loss)		
9c	Unrecaptured section 1250 gain		
10	Net section 1231 gain (loss)		
		19	Distributions
11	Other income (loss)		
		20	Other information
12	Section 179 deduction		
13	Other deductions		
14	Self-employment earnings (loss)		

21 ☐ More than one activity for at-risk purposes*
22 ☐ More than one activity for passive activity purposes*

*See attached statement for additional information.

For IRS Use Only

APÊNDICE C – Códigos utilizados para criação do modelo utilizado no treinamento LSTM

```
from keras.models import Model, Input
from keras.layers import LSTM, Embedding, Dense, TimeDistributed, Dropout, Bidirectional
from keras_contrib.layers import CRF

# Model definition
input = Input(shape=(MAX_LEN,))
model = Embedding(input_dim=n_words+2, output_dim=EMBEDDING, # Entrada
                  input_length=MAX_LEN, mask_zero=False)(input)  # Embedding recebendo vetores da palavras
model = Bidirectional(LSTM(units=50, return_sequences=True,
                           recurrent_dropout=0.1))(model)  # BiDirecional LSTM
model = TimeDistributed(Dense(50, activation="relu"))(model)  # Camada Dense
crf = CRF(n_tags+1)  # CRF layer, Reconhcimento padrões
out = crf(model)  # saida

model = Model(input, out)
model.compile(optimizer="rmsprop", loss=crf.loss_function, metrics=[crf.accuracy])

model.summary()

history = model.fit(X_tr, np.array(y_tr), batch_size=BATCH_SIZE, epochs=EPOCHS,
                    validation_split=0.2, verbose=1)


history = model.fit(X_tr, np.array(y_tr), batch_size=BATCH_SIZE, epochs=EPOCHS,
                    validation_split=0.2, verbose=1)
```

```
# Eval
import time
time_start_predict = time.time()
pred_cat = model.predict(X_te)
time_stop_predict = time.time()

time_predict = (time_stop_predict - time_start_predict)/len(X_te)

print(f'|Predict AVG Time:{time_predict:2f}|')

pred = np.argmax(pred_cat, axis=-1)
y_te_true = np.argmax(y_te, -1)
```

```
from sklearn_crfsuite.metrics  import flat_classification_report,flat_accuracy_score,flat_precision_score,flat_recall_score,flat_f1_score

# Convert the index to tag
pred_tag = [[idx2tag[i] for i in row] for row in pred]
y_te_true_tag = [[idx2tag[i] for i in row] for row in y_te_true]

report = flat_classification_report(y_pred=pred_tag, y_true=y_te_true_tag)

print("acuracia",flat_accuracy_score(y_pred=pred_tag, y_true=y_te_true_tag))
print("precision",flat_precision_score(y_pred=pred_tag, y_true=y_te_true_tag,average='macro'))
print("recall",flat_recall_score(y_pred=pred_tag, y_true=y_te_true_tag,average='macro'))
print("f1",flat_f1_score(y_pred=pred_tag, y_true=y_te_true_tag,average='macro'))

print(report)
```

APÊNDICE D - Códigos utilizados para criação do modelo utilizado no treinamento SVM

```
predicted_percent = []
predicted_time = []
#Aumento de N-Gramas
for i in range(1,7):

    #Criação do modelo
    model_investor = make_pipeline(TfidfVectorizer(stop_words=stopset,ngram_range=(1, i)),
                                   SGDClassifier(loss='hinge', penalty='l2', alpha=1e-2,
                                                 random_state=42, max_iter=25,
                                                 tol=None,n_jobs=-1,verbose=False))

    #Treinamento do modelo
    history = model_investor.fit(train_document_investor, train_investor)

    #Criando teste com 20% e computando tempo
    time_start_predict = time.time()
    predicted_labels_investor = model_investor.predict(test_document_investor)
    time_stop_predict = time.time()
    time_predict = (time_stop_predict - time_start_predict)/len(test_document_investor)

    #cTeste de acurácia
    test_accuracy_investor = accuracy_score(test_investor,predicted_labels_investor)
    test_loss_investor = classification_report(test_investor,predicted_labels_investor,output_dict=True)
```

APÊNDICE E - Códigos utilizados para criação do modelo utilizado no treinamento SVM com validação cruzada

```
predicted_percent_fund = []
predicted_time_fund = []
for i in range(1,7):
    model_fund = make_pipeline(TfidfVectorizer(stop_words=stopset,ngram_range=(1, i)),
                               SGDClassifier(loss='hinge', penalty='l2',
                                             alpha=1e-2, random_state=42,
                                             max_iter=25, tol=None,n_jobs=-1))

    #Treinamento
    model_fund.fit(train_document_fund, train_fund)

    time_start_predict = time.time()
    predicted_labels_fund = model_fund.predict(test_document_fund)
    time_stop_predict = time.time()
    time_predict = (time_stop_predict - time_start_predict)/len(test_document_investor)


    #Calculo acuracia
    test_accuracy_fund = accuracy_score(test_fund,predicted_labels_fund)

    predicted_percent_fund.append([i,test_accuracy_fund])
    predicted_time_fund.append([i,time_predict])
    #Imrpimir acuracia
    test_loss_fund = classification_report(test_fund,predicted_labels_fund,output_dict=True)


    print("Weigthed AVG",test_loss_fund["weighted avg"])
    print("Marco AVG",test_loss_fund["macro avg"])

    print(f'|Test Acc:{test_accuracy_fund*100:.2f}%|')
    print(f'|Predict AVG Time:{time_predict:.2f}|')

    fig, axes = plt.subplots(3, 2, figsize=(10, 15))
    cv = ShuffleSplit(n_splits=10, test_size=0.2, random_state=0)

    plot_learning_curve(model_fund, "Curva de Treinamento", train_document_fund, train_fund, axes=axes[:, 0], ylim=(0.7, 1.01),
                        cv=cv, n_jobs=4)
    plt.show()
```