UNIVERSIDADE NOVE DE JULHO – UNINOVE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO

CINTIA MARIA DE ARAÚJO PINHO

ANÁLISE DE TEXTOS COM APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL: ESTUDO COMPARATIVO PARA CLASSIFICAÇÃO DE FUGA AO TEMA EM REDAÇÕES

São Paulo
2021

CINTIA MARIA DE ARAÚJO PINHO

ANÁLISE DE TEXTOS COM APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL: ESTUDO COMPARATIVO PARA CLASSIFICAÇÃO DE FUGA AO TEMA EM REDAÇÕES

Dissertação apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE, como requisito para obtenção do título de Mestre.

Prof. orientador: Dr. Marcos Antonio Gaspar

Prof. coorientador: Dr. Renato José Sassi

Linha de Pesquisa: Gestão da Tecnologia da Informação e Conhecimento

São Paulo
2021

Pinho, Cintia Maria de Araújo.

Análise de textos com aplicação de técnicas de inteligência artificial: estudo comparativo para classificação de fuga ao tema em redações. / Cintia Maria de Araújo Pinho. 2021.

151 f.

Dissertação (Mestrado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2021.

Orientador (a): Prof. Dr. Marcos Antonio Gaspar.

1. Redações. 2. Avaliação automática. 3. Fuga ao tema. 4. Inteligência artificial.

I. Gaspar, Marcos Antonio. II. Título.

CDU 004

UNINOVE
Universidade Nove de Julho

PARECER – EXAME DE DEFESA

Parecer da Comissão Examinadora designada para o exame de defesa do Programa de Pós-Graduação em Informática e Gestão do Conhecimento, a qual se submeteu a aluna regularmente matriculada Cintia Maria de Araújo Pinho.

Tendo examinado o trabalho apresentado para obtenção do título de "Mestre em Informática e Gestão do Conhecimento", com Dissertação intitulada "ANÁLISE DE TEXTOS COM APLICAÇÃO DE TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL: ESTUDO COMPARATIVO PARA CLASSIFICAÇÃO DE FUGA AO TEMA EM REDAÇÕES", a Comissão Examinadora considerou o trabalho:

(X) Aprovado **( ) Aprovado condicionalmente**

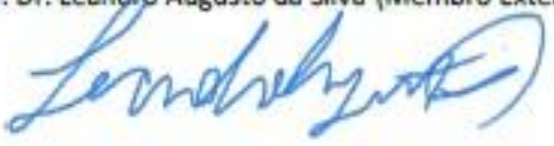
( ) Reprovado com direito a novo exame **( ) Reprovado**

Parecer:
Trabalho aprovado. Providenciar a inserção dos apontamentos realizados pela banca na versão definitiva a ser depositada na Biblioteca Uninove.

EXAMINADORES

Prof. Dr. Marcos Antonio Gaspar (Orientador – UNINOVE)

Prof. Dr. Leandro Augusto da Silva (Membro Externo – MACKENZIE)



Prof. Dr. Cleber Gustavo Dias (Membro Interno – UNINOVE)

Prof. Dr. Renato José Sassi (Coorientador – UNINOVE)

São Paulo, 30 de agosto de 2021.

DEDICATÓRIA

Dedico esse trabalho primeiramente a Deus e todos os espíritos de luz que acompanharam e orientaram nesta e em todas as jornadas de minha vida. Dedico em especial aos meus pais Benício e Fátima que sempre foram um exemplo de vida, sempre me inspirei em suas lutas e procurei seguir seus exemplos de honestidade e trabalho. Sempre me dediquei para ser um orgulho, eles nunca deixaram de acreditar em mim e me admiram incondicionalmente.

Dedico em especial ao meu noivo Anderson Vanin e filho Arthur Gael, esses sim tiveram que ter muita paciência comigo, aguentar todas as minhas crises de stress, nunca deixaram de me incentivar e sempre entenderam que não estaria disponível em todos os momentos em família. Meu filho me inspira em ser uma boa mãe e ser um exemplo a seguir.

Ao Anderson Vanin, meu amor, amigo e parceiro não só agradeço como dedico, me ajudou tanto que é difícil elencar, ter uma pessoa que te ajuda no apoio moral e inclusive nos códigos não é todos que tem, sou privilegiada, ter alguém que sabe e entende os desafios que enfrenta é motivador.

AGRADECIMENTOS

Este trabalho é um resultado de muita colaboração, desde a minha família que sempre foi minha base e inspiração e a cada pessoa que a sua forma colaborou com incentivo, ideias, conhecimento, apoio e principalmente paciência. Agradecer a todos os envolvidos é praticamente impossível, mas tive grandes instituições e pessoas que me acompanharam de forma única para a realização desse sonho.

Primeiramente a Universidade Nove de Julho (UNINOVE) pelo apoio, pela oportunidade de crescimento, aprimoramento acadêmico, pessoal e profissional, assim como pela bolsa de estudos. Nesta instituição encontrei profissionais maravilhosos que me guiaram e orientaram nessa jornada, a começar pelo meu primeiro orientador Prof. Dr. Domingos Napolitano que me apresentou a área de PLN e Inteligência artificial, pela qual nunca tinha trabalhado e me instigou a encarar esse novo aprendizado, além das orientações me proporcionou aulas preciosas que direcionaram esta dissertação, além da participação essencial na publicação de 3 artigos, foi incansável em seus ensinamentos, sou muito grata. Outro agradecimento muito especial vai ao Prof. Dr. Gaspar, meu último orientador pelo qual desde as primeiras aulas me despertou total admiração, não só pelo conhecimento que me trouxe, mas especialmente como um exemplo de ser humano, educação e dedicação, profissional exemplar como orientador sempre trazendo feedbacks rápidos e coerentes, um dia quando eu crescer, vou ser assim. Ao prof. Sassi meu coorientador devo tantos agradecimentos, me salvou de vários equívocos neste trabalho, trouxe ideias muito valiosas, ao assistir suas aulas ficamos admirados ao ver o quanto o entendimento a fundo do que se produz é essencial, me trouxe provas de que apenas a prática não traz bons resultados se você não entende a teoria, vou levar isso para sempre comigo. Agradeço imensamente a todos os professores do PPGI, em especial ao prof. Dr. Fellipe pelas orientações e ensino de Metodologia e sugestões valiosas na banca de qualificação, sempre muito educado e preciso, além de ter colaborado muito para a publicação do artigo que me trouxe grandes conhecimentos na área de classificação. Ao prof. Dr. Cleber que me acompanhou na banca de qualificação com indicações pontuais e de grande pertinência a esta dissertação. Ao Prof. Dr. Ivanir pelo auxílio na publicação do artigo na área de Bibliometria e que sempre apoiou essa ideia desde o primeiro semestre. Agradeço ainda a toda direção e setor administrativo da Uninove que sempre deu o suporte necessário.

Ainda falando de instituição agradeço imensamente a Tecnologia Única, empresa que abraçou essa ideia e encubou esse projeto, é um exemplo a ser seguido, entendendo que o apoio a pesquisa científica é um investimento essencial ao crescimento do país. Em especial ao Tássio, que acreditou na ideia e apresentou para aos diretores da empresa.

Agradeço aos colegas de Universidade que me auxiliaram, acompanharam e incentivaram de diferentes formas, em especial a Amanda super companheira de trabalhos e artigos, outros colegas fantásticos como Paola, Roger, João Barbosa, Célia, Yuri, Pamela e Anderson Vanin.

Também sou muito agradecida aos companheiros de trabalho da ETEC onde atuo como professora, em especial Amanda, Keli, Agda, Joarez, Bárbara e Oséias, professores de português que deram total suporte a este trabalho, utilizando o projeto beta e dando dicas de melhoria e aperfeiçoamento deste trabalho.

RESUMO

O processo de correção manual de redações acarreta algumas dificuldades, dentre as quais aponta-se o tempo dispendido para a correção e para a devolutiva de resposta ao aluno. Para instituições como as universidades e o Exame Nacional do Ensino Médio (ENEM), que se utilizam de redação como avaliação para o ingresso no ensino superior, além das escolas de ensino básico, tal atividade demanda tempo e custo para a avaliação dos textos produzidos. A fuga ao tema é um dos itens avaliados na redação do ENEM e, quando o estudante comete tal falha, sua redação é anulada por não ter desenvolvido os conceitos solicitados na proposta do tema estipulado para a redação. Neste contexto, a análise automática de redações com a aplicação de Processamento de Linguagem Natural (PLN), Mineração de Textos (MT) e outras técnicas de Inteligência Artificial (IA) tem se revelado promissora no processo de avaliação automatizada da linguagem escrita. Face ao contexto exposto, o objetivo desta pesquisa é comparar diferentes técnicas de IA para classificação de fuga ao tema em textos e identificar aquelas que trouxeram melhores resultados. Esta é uma pesquisa aplicada e experimental executada por meio da aplicação de algoritmos e mensuração dos resultados obtidos. Os experimentos delinearam em especial a classificação de 1320 redações de língua portuguesa com 119 temas diferentes. Além da PLN e MT, a pesquisa se utilizou das seguintes técnicas inteligentes de classificação: Redes Neurais Convolucionais (RNC), *Multilayer Perceptron* (MPL), Árvores de Decisão, Florestas Aleatórias, *Gradiente Boosting*, *Ada Boost*, *Stochastic Gradiente Descent*, *Support Vector Machines* e outras técnicas para identificar padrões na base de dados por meio de algoritmos não supervisionados como a clusterização. Os experimentos trouxeram os melhores resultados para o classificador RNC, que obteve acurácia de até 89%, com taxa de Falso Positivo (FP) de 5,7% e Verdadeiro Positivo (VP) de 49%. Outros classificadores com resultados satisfatórios foram MLP e *Gradiente Boosting*, com acurácia de 90% e 74%, VP de 33% e 51% e média de FP de 4% e 20%, respectivamente. Espera-se que a solução desenvolvida nesta pesquisa contribua para impactar positivamente o trabalho de professores e instituições de ensino, por meio da redução de tempo e custos associados ao processo de avaliação de redações.

Palavras-chave: Redações. Avaliação automática. Fuga ao tema. Inteligência artificial.

ABSTRACT

The process of manual correction of essays brings some difficulties, among which is the time spent for the correction and for returning the answer to the student. For institutions such as universities and the Exame Nacional do Ensino Médio (ENEM), which use essays as an evaluation for admission to higher education, as well as elementary schools, this activity demands time and cost to evaluate the texts produced. Changing the subject is one of the items evaluated on ENEM, when a student commits such mistake, his or her text is annulled because the concepts were not developed as requested in the stipulated theme for the essay. In this context, the automatic analysis of essays with the application of Natural Language Processing (NLP), Text Mining (TM) and other Artificial Intelligence (AI) techniques has shown promise in the process of automated assessment of written language. Given the referred context, the aim of this research is to compare different AI techniques for classification of topic avoidance in texts and identify those that brought better results. This is an applied and experimental research executed by applying algorithms and measuring the results obtained. The experiments outlined in particular the classification of 1320 Portuguese language essays with 119 different themes. Besides PLN and MT, the research used the following intelligent classification techniques: Convolutional Neural Networks (CNN), Multilayer Perceptron (MPL), Decision Trees, Random Forests, Gradient Boosting, Ada Boost, Stochastic Gradient Descent, Support Vector Machines, and other techniques to identify patterns in the database through unsupervised algorithms such as clustering. The experiments brought the best results for the RNC classifier, which obtained accuracy up to 89%, with False Positive (FP) rate of 5.7% and True Positive (VP) rate of 49%. Other classifiers with satisfactory results were MLP and Gradient Boosting, with accuracy of 90% and 74%, PV of 33% and 51%, and average FP of 4% and 20%, respectively. It is hoped that the solution developed in this research will contribute to impact the work of teachers and educational institutions by reducing the time and costs associated with the essay evaluation process.

Key words: Essays. Automatic evaluation. Escape from the topic. Artificial intelligence.

LISTA DE ABREVIAÇÕES

AM	**Aprendizado de Máquina**
AP	**Aprendizado Profundo**
C1	**Consulta 1**
C2	**Consulta 2**
Conabe	**Conferência Nacional de Alfabetização Baseada em Evidências**
ENEM	**Exame Nacional do Ensino Médio**
FUVEST	**Fundação Universitária para o Vestibular**
IA	**Inteligência Artificial**
INEP	**Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**
MEC	**Ministério da Educação**
MIT	***Massachusetts Institute of Technology***
ML	***Machine Learning***
MLP	***MultiLayer Perceptron***
MT	**Mineração de Texto**
NLTK	***Natural Language Toolkit***
OCDE	**Organização para Cooperação e Desenvolvimento Econômico**
PLN	**Processamento de Linguagem Natural**
PNA	**Política Nacional de Alfabetização**
QB	**Questão Bibliométrica**
QP	**Questão de Pesquisa**
RNC	**Redes Neurais Convolucionais**
RSL	**Revisão Sistemática da Literatura**
SGD	***Stochastic Gradiente Descent***
SVM	***Support Vetor Machine***
WOS	***Web of Science***

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A escrita é uma prática de grande importância, seja no mundo acadêmico, corporativo ou até mesmo na vivência social. Além disso, é parte imprescindível no desenvolvimento da cognição humana. A escrita não está inserida apenas nas atividades sociais, além de inserida num contexto profissional ela faz parte do processo de crescimento pessoal e de ensino-aprendizagem do indivíduo (Gomes, 2020, p.123).

O processo de ensino-aprendizagem vem se adaptando às novas tecnologias de comunicação atualmente disponíveis. Costa (2019, pg.19) afirma que mesmo quem se nega a usar a tecnologia acaba sendo envolvido por ela, então, ou se adapta ou sucumbe em alguma esfera. Assim, o papel da escola como um segmento formador e transformador pode incrementar seus processos ao incluir os meios tecnológicos disponíveis para facilitar o processo de aquisição de conhecimento e melhoria da escrita do indivíduo.

A habilidade de se comunicar por meio da escrita continua sendo primordial, a diferença é que com a tecnologia ela se expande muito rápido para diferentes lugares do mundo. Assim, escrever ainda é uma habilidade indispensável a qualquer profissional de sucesso. A mudança é que no contexto atual os conteúdos produzidos precisam ser objetivos e contundentes, e as mensagens eletrônicas possuem conteúdos valiosos. Para o estudante que almeja a entrada no ensino superior uma boa escrita da redação pode facilitar esse processo. (SQUARISI; SALVADOR, 2020).

Para que uma redação faça sentido e atinja o seu objetivo é preciso aplicar práticas que garantam o encadeamento de ideias de forma lógica e compreensível. Para tanto, os profissionais de língua portuguesa assumem uma grande responsabilidade ao manter e controlar a aprendizagem do indivíduo no quesito expressão escrita por meio de diferentes gêneros textuais (Hentz et al., 2020, pg.111). Porém, este processo demanda esforço e planejamento para o professor desenvolver no aluno as habilidades pertinentes ao desenvolvimento de um texto de qualidade elevada. Assim, a devolutiva das avaliações de inúmeros textos que ocorrem simultaneamente com diferentes turmas é uma tarefa morosa e cansativa aos docentes. Somando-se a este

fato, o Brasil tem um dos maiores índices de quantidade de alunos por turma no nível secundário em todo o mundo.

Pesquisa realizada pelo World Education Indicators (WEI), pela Unesco e pela Organização para Cooperação e Desenvolvimento Econômico (OCDE) em 2002, mostra que o Brasil tem uma média de 35,6 estudantes por classe. O país apresenta um índice bastante superior aos demais países da América Latina. Na Argentina, por exemplo, a relação é de 11,2; no Peru, de 18,5 e no Uruguai, de 14,9 alunos por turma. Dentre todos os países que participaram da pesquisa WEI, Portugal tem o menor índice: nove estudantes por sala de aula. Em seguida, vêm Luxemburgo, com 9,2; e Bélgica, com 9,7. Este estudo foi realizado em 42 nações ricas ou em desenvolvimento (INEP, 2004, s.p.).

Em complemento ao contexto indicado, dados do último censo escolar de 2018 apontaram que o Brasil teve uma melhoria no referido índice, apresentando média de 30,4 alunos por turma. Porém, tal resultado ainda é muito superior aos demais países participantes do estudo anteriormente indicado. Ademais, o país mantém uma média entre 400 e 500 alunos por professor. Isto acontece devido à número de turmas que o professor leciona, assim ele consegue complementar seu salário, desta forma, aumenta sua carga de trabalho com mais horas semanais. (LOURENCETTI, 2014, pg. 15).

Uma entrevista realizada pela instituição Universia em 2015 com o professor Adrian Chan, do Centro de Treinamento e Capacitação para o Enem, destacou que o professor pode demorar de 40 segundos a 10 minutos para corrigir uma redação, uma vez que o tempo varia de acordo com a qualidade da escrita apresentada no texto (UNIVERSIA, 2015, s.p.)

Para exemplificar os dados apresentados anteriormente, se um professor de língua portuguesa solicitar uma redação para cada um de seus 500 alunos, ele pode demorar de 16 a 41 horas consecutivas para corrigir os textos apresentados pelos alunos. Tal estimativa leva em consideração que o profissional gaste respectivamente entre 2 e 5 minutos ao avaliar cada redação. Contudo, sabe-se que não existe apenas esta forma avaliativa, o professor demanda ainda muito tempo para correção de outras atividades, tais como interpretação de texto, questionários, atividades em grupo etc. Somando-se a isto, as instituições de ensino têm exigido a intensificação da aplicação de

redações em sala de aula, devido sua importância para o processo seletivo de entrada em cursos de nível superior.

O vestibular é o processo seletivo para ingresso no ensino superior. Em geral, a avaliação pode ser elaborada pela própria instituição, seja privada ou pública, ou ainda por uma instituição independente. Em 2020, só no estado de São Paulo, 26 universidades particulares fizeram o processo seletivo (vestibular) totalmente online, estabelecendo como principal instrumento de avaliação a redação (MORALES, 2020, s.p.).

Ainda assim, o ENEM (Exame Nacional do Ensino Médio), criado em 1998, é uma das principais formas de ingresso ao ensino superior no Brasil. Só em 2019 o ENEM foi responsável por 32% das seleções em processos seletivos de entrada em cursos superiores, de acordo com o Censo da Educação Superior do Ministério da Educação (MEC). Este exame mensura o desempenho dos alunos após o término do ensino médio. No Brasil as vagas são distribuídas em 129 universidades públicas e mais de mil instituições particulares. Além disso, fora do país o ENEM é utilizado como forma de ingresso em 42 universidades em Portugal (AGÊNCIA BRASIL, 2019, s.p.).

Dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP (2020, s.p.) informam que em 2020 houve 6.121.363 inscrições. Tal quantidade mostra um crescimento em relação aos anos anteriores, visto que foram 5.095.308 inscrições em 2019 e 5.513.712 inscrições em 2018. A quantidade de inscrições no ENEM mostra a consolidação da importância desta ferramenta de seleção para ingresso no ensino superior.

A avaliação do ENEM é composta por 180 questões de múltipla escolha e conta com a redação como único item discursivo da prova. A redação é avaliada de acordo com cinco competências que devem ter sido desenvolvidas durante os anos de escolaridade do aluno, quais sejam:

1. Demonstrar domínio da modalidade escrita formal da língua portuguesa;

2. Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa;

3. Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;

4. Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação;

5. Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos (BRASIL, 2019, pg. 49).

Cada competência avaliada contabiliza até 200 pontos ao candidato autor da redação, de modo que se ele atingir o domínio das cinco competências poderá pontuar no máximo 1000 pontos como nota atribuída à redação. Contudo, o cenário de desempenho dos estudantes neste tipo de exame não vem obtendo resultados satisfatórios. Segundo o INEP (2020), de todos os inscritos no ano de 2020, 65% já concluíram o ensino médio em anos anteriores, sendo que a maioria está fazendo a prova pela segunda ou terceira vez, o que demonstra grande despreparo desses candidatos para participar deste processo avaliativo na forma de elaboração de redação.

Campos (2020) divulga que só em 2019, 143 mil estudantes tiraram nota zero na redação do ENEM e somente 53 pessoas alcançaram nota máxima em sua produção textual. Segundo o INEP (2019), os maiores motivos que levaram ao zero foram: redação em branco com 39,8%; fuga ao tema com 28,4%; cópia do texto motivador com 16,2%, dentre outros motivos. Assim, uma vez que o candidato elabore a redação, verifica-se que o maior motivo de atribuição de nota mínima está relacionado à fuga ao tema proposto. Em complemento, estudo realizado por Diana (2021, s.p.), professora e especialista em gestão de conteúdos on-line, destaca os dezesseis erros mais graves que podem ser cometidos por estudantes em uma redação. Dentre estes, fugir ao tema é o primeiro de sua lista, pois leva o aluno a uma nota muito baixa ou mesmo o cancelamento da prova do ENEM.

A fuga do tema está relacionada à competência 2 estabelecida pelo ENEM e ocorre quando o candidato escreve um texto que não traz nenhuma

referência à frase temática indicada na proposição estabelecida para a redação. Para evitar que este tipo de desvio ocorra, o candidato precisa ter uma boa interpretação da proposta estabelecida na prova, para que a redação não seja anulada pelo avaliador. Assim, se o candidato fugiu à proposta estabelecida, o avaliador não precisa seguir a correção da redação elaborada.

Neste contexto, esta dissertação visa tratar exclusivamente os desvios de escrita que tangenciam a competência 2 estabelecida pelo ENEM. O conjunto de circunstâncias relatado anteriormente ressalta a importância de estudar uma forma de auxílio aos professores para agilização do processo de avaliação de textos discursivos. Para tanto, uma alternativa é dar apoio para a criação de um sistema de correção inteligente de redações. A ideia de tal solução não é substituir totalmente o trabalho do professor na avaliação do texto, mas tonar mais ágil o processo avaliativo ao fornecer indicações e apontamentos de possíveis falhas na escrita do aluno, relativamente à fuga ao tema proposto.

Caso seja possível uma solução inteligente conseguir apontar possível fuga ao tema, o professor não precisará dar seguimento à correção da redação. Até porque, conforme afirmou o professor Adriano Chan, responsável pelo portal de ensino de redações chamado “Laboratório de Redações”, quando o texto é confuso, o avaliador pode demorar até 10 minutos para corrigir apenas uma única redação. O fato é que o professor acaba lendo o texto mais de uma vez para garantir que realmente fugiu a temática, já que este fato anula a redação (UNIVERSIA, 2015, s.p.).

Sendo assim, esta pesquisa de dissertação procura oferecer auxílio aos profissionais de língua portuguesa para avaliarem textos educacionais. Para tanto, esta pesquisa se utilizará de Inteligência Artificial (IA), um campo das Ciência da Computação que envolve uma combinação de algoritmos projetados para criar máquinas que simulem a capacidade humana de raciocinar, tomar decisões e resolver problemas (CAPGEMINI, 2017; SANTOS, 2015; KAUFMAN, 2018)

Segundo Tavares (2020), “os textos são fontes de dados relevantes e a utilização de tecnologias de IA pode auxiliar a estudar formas para capacitar o computador a entender, analisar, manipular e, potencialmente, criar a

linguagem humana por meio de textos". Sendo assim, este estudo se utilizará de algumas subáreas da IA, tais como o Processamento de Linguagem Natural (PLN), Mineração de Textos (MT), e as seguintes técnicas para classificação: Redes Neurais Convolucionais (RNC), MLP (*MultiLayer Perceptron*), árvores de decisão; Florestas Aleatórias, *Gradiente Boost, Ada Boost, Stochastic Gradiente Descent (SGD) e Support Vetor Machines* (SVM's).

O PLN é uma subárea da Inteligência Artificial que tem como objetivo a compreensão automática de linguagens humanas, de maneira que possam ser manipuladas por computadores. Em geral, essas técnicas estão voltadas aos seguintes aspectos prioritários: fonologia, morfologia, sintaxe, semântica e pragmática (GRANATYR, 2016).

Já a MT busca produzir percepções significativas a partir de dados de textos em linguagem natural para realizar um tipo especial de análise linguística que essencialmente ajuda uma máquina a 'ler' um texto (SUBRAMANIAN, 2019).

A classificação automática de textos é a tarefa de associar textos em linguagem natural a rótulos pré-definidos e engloba diferentes conceitos para de extração de informação, estes classificadores são construídos por técnicas de aprendizagem de máquina (AM) hoje em dia alcançam expressivos níveis de efetividade (KUSHMERICK e THOMAS, 2003). Nesta pesquisa os rótulos pré-definidos para classificação das redações são determinados como 'fuga' ou 'não fuga' ao tema.

Em virtude do tema tratado toda contribuição é bem-vinda, principalmente as apoiadas em IA, o que abre espaço para a aplicação de PLN, MT e técnicas inteligentes para classificação de fuga ao tema em redações. Sendo assim, neste trabalho será realizado um estudo comparativo de diferentes técnicas o que possibilita compreender aquelas que alcançam melhores resultados.

1.1 Problemática e Relevância

Com o advento da tecnologia, os sistemas de computação têm auxiliado os professores na avaliação de provas e atividades pertinentes ao processo de

ensino-aprendizagem, sobretudo voltado às questões de múltipla escolha, nas quais só existe a possiblidade de uma resposta ser verdadeira. Assim, existem atualmente diferentes plataformas que avaliam e já pontuam o desempenho do aluno em alguma atividade de forma automática. Todavia, questões dissertativas, em especial textos e redações, acabam por demandar a correção manual de um profissional humano, atividade que consome tempo e esforço consideráveis.

O processo de correção manual de redações acarreta algumas dificuldades, dentre as quais aponta-se o tempo dispendido para a correção e para a devolutiva de resposta ao aluno. Atender pontualmente as dificuldades particulares de cada discente é uma necessidade intrínseca à atividade do professor, porém, não fácil de ser cumprida a contento. O que se espera é que por meio do acompanhamento do professor de orientação de leitura, escrita, correções e feedbacks frequentes, os alunos possam progredir no quesito argumentação e escrita.

Contudo, o cenário atual mostra que os professores passam por dificuldades ao avaliar individualmente diferentes alunos. Estudo realizado por Riolfi e Igreja (2010) aponta que os professores dedicam apenas 6% do seu tempo em sala de aula para o ensino da dissertação. Nesta mesma pesquisa foi identificado ainda que em alguns casos, após a correção dos textos dos alunos, o professor comentava oralmente as redações, ignorando outros problemas textuais em suas exposições à toda a turma. (Riolfi; Igreja, 2010, pg. 321)

Os mesmos autores realizaram a análise do processo seletivo da Universidade de São Paulo – Fundação Universitária para o Vestibular (FUVEST) em 2008. Foram 11.242 candidatos e, destes, aqueles que estavam entre os aprovados estudaram exclusivamente em colégios particulares, o que correspondia a 70,9%, sendo que os alunos oriundos da escola pública não ultrapassaram 20,3% (Riolfi e Igreja, 2010, pg. 315).

Outro autor executou uma análise de um conjunto de textos que deveriam seguir os gêneros discursivo/textual, baseado nas competências do ENEM, ele identificou que apenas 13% dos alunos escrevem com estrutura

pertinente e com as características de gênero adequadas. (Striquer, 2018, pg. 73.)

Em pesquisa conduzida por Pinho et al. (2020), verificou-se que o maior problema em oferecer feedback de redações ao aluno no ano de 2020 se deveu à diversidade de plataformas diferentes pelas quais os professores receberam os textos, tais como aplicativos de mensagens, e-mail e outras plataformas digitais de ensino específicas. Outro contratempo verificado foi a maneira como os professores recebem as redações, sendo a maior incidência de textos enviados na forma de fotos do caderno (70,4%). Os docentes informam que a leitura é de difícil entendimento, por conta da letra do aluno ou da má qualidade da imagem encaminhada.

Na mesma pesquisa, os autores confirmaram que mais de 60% dos professores usam menos de 25% de seu tempo para o ensino e devolutivas de dissertações. Os motivos alegados pelos docentes são diversos, tais como excesso de turmas e alunos, além da carência em fornecer um feedback detalhado e criterioso de forma individualizada.

Além das dificuldades pontuais de professores em sala de aula, também existem problemas relativos às avaliações (dissertativas, argumentativas ou narrativas) que acontecem em larga escala, como é o caso do exame do ENEM. O processo de avaliação das redações no ENEM se dá em larga escala, sendo executado de forma manual por um grande contingente de avaliadores. Como consequência, este processo manual apresenta problemas relativos ao tempo, custo, confiabilidade e subjetividade do avaliador.

O INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) divulgou como é executada a avaliação das redações: os textos são corrigidos por mais de 5 mil avaliadores, corrigindo 150 textos a cada 3 dias, a cada 50 redações, o avaliador recebe duas delas já avaliadas por outro especialista (INEP, 2020, s.p.).

Assim, cada redação é corrigida por dois professores e eles desconhecem a nota atribuída pelo outro, além de não saberem quem é o candidato avaliado, todo esse cuidado é tomado para prestar maior confiabilidade ao processo manual executado por pares de avaliadores (INEP,

2020, s.p.). Entretanto, esse processo manual acarreta elevado tempo para divulgação dos resultados aos candidatos, além de custos altos para manter tal metodologia.

Analisando os relatos dos docentes e sabendo que o INEP ainda faz as correções de redações de forma manual, percebe-se que não há ferramentas distribuídas em larga escala nos setores privado ou público que realizem a avaliação automática de redações. Contudo, esta dissertação buscará demonstrar que o crescimento nesta área de estudo, o que poderá possibilitar num futuro próximo a disponibilização de tecnologia que auxilie os professores neste processo.

Assim, a relevância desse estudo vai ao encontro dos problemas citados anteriormente. O uso de ferramentas digitais e inteligentes pode trazer melhorias significativas ao processo de correção de textos dissertativos. Entende-se que a implantação de soluções baseadas em Inteligência Artificial, dentre as quais destacam-se como alternativas o Processamento de Linguagem Natural (PLN), a Mineração de Textos (MT) e diferentes técnicas de classificação, tais como: Redes Neurais Convolucionais (RNC); MLP (*MultiLayer Perceptron*), árvores de decisão; Florestas Aleatórias, *Gradiente Boost, Ada Boost, Stochastic Gradiente Descent (SGD) e Support Vetor Machines* (SVM's). Espera-se que tais soluções e técnicas possam auxiliar o professor na tarefa de identificar a evolução na escrita dos alunos, bem como na identificação de erros em suas redações.

Assim, a presente pesquisa procura responder à seguinte questão-problema**: Quais técnicas de IA são as melhores para identificação de fuga ao tema em redações?**

1.2 Objetivos

Uma vez estabelecido o questionamento norteador da pesquisa, o seguinte objetivo geral foi traçado:

Comparar diferentes técnicas de IA para classificação de fuga ao tema em textos e identificar aquelas que trouxeram melhores resultados

Em complemento, os seguintes objetivos específicos foram estabelecidos:

1. Identificar na literatura as técnicas de Inteligência Artificial mais utilizadas para a análise e classificação de textos;

2. Realizar a análise preliminar da base de dados aplicando PLN e MT a fim de conhecer o *corpus* utilizado, diagnosticar padrões, identificar os primeiros desvios de escrita e preparar a base de dados para a classificação das redações;

3. Selecionar e aplicar as técnicas inteligentes mais pertinentes a partir da utilização de uma base de dados com 1320 redações;

4. Avaliar os resultados dos experimentos que trazem os modelos mais adequados na identificação de fuga ao tema e, então, indicar as melhores soluções para prestação de feedback mais rápido e efetivo ao professor.

1.3 Justificativa

Como o processo de correção manual costuma trazer problemas relativos ao grande tempo necessário à sua execução, esta pesquisa pretende viabilizar melhores soluções que facilitam o trabalho dos professores e instituições na identificação prévia de desvios cometidos pelos educandos. Tal solução facilitaria o processo de ensino-aprendizagem, bem como a gestão dos conhecimentos envolvidos, sem contar o ganho no tempo despendido para fornecer feedback individual ao aluno. Trevisani (2019), argumenta que o feedback individualizado ao aluno é de extrema importância, pois o orienta e visa a ajudá-lo a aprimorar suas técnicas de redação e aumentar sua competência na escrita dissertativa de textos.

Com a utilização de inteligência artificial, aplicada a uma plataforma digital, seja em sala de aula ou à distância, se faz possível aumentar a solicitação de leituras que podem gerar resenhas ou redações, com a possibilidade de a solução inteligente a ser elaborada fornecer imediatamente um feedback individualizado a respeito das possíveis falhas na escrita do estudante. Sendo assim, a maior prática na produção textual poderá possibilitar aos alunos maiores oportunidades de performance mais elevada na avaliação

da redação do ENEM, contribuindo ainda para a diminuição da desigualdade nesse processo dentre diferentes perfis de alunos.

De acordo com Barros (2020), “ao longo da vida escolar o estudante deve exercitar sua escrita, pois a cada ano irá se tornando mais clara e desenvolta, em face da própria maturidade que apresenta”. Partindo desta premissa, é possível entender que uma solução de avaliação inteligente e automatizada poderá proporcionar que o discente exercite mais sua escrita, fazendo-o se sentir mais preparado para encarar os desafios que enfrentará após o ensino médio em sua vida acadêmica na universidade. Por consequência, também ajudará ao aluno em seu desenvolvimento para atuação futura na profissão escolhida.

Os resultados desta pesquisa poderão contribuir para acelerar o processo de correção com a automatização da detecção de alguns desvios de escrita, tal como a fuga ao tema proposto numa redação. A partir da detecção de inconsistências no texto o avaliador não deverá prosseguir com a correção, pois a prova deverá ser zerada obtendo-se, desta forma, um ganho no tempo despendido para a avaliação.

Numa perspectiva gerencial pode-se proporcionar a diminuição de custos com a contratação de professores, além de acelerar a devolutiva dos resultados aos candidatos do ENEM. A título de dimensionamento dos custos incorridos, em 2016 o INEP pagava em torno de R$ 15,00 por texto corrigido. Esses custos englobam todos os valores referentes à estrutura necessária à correção das redações, tanto o aparato físico, profissionais envolvidos na logística e em atividades administrativas, quanto de capacitação dos corretores que atuam nessa atividade. Contudo para o p0rofissional avaliador da redação em si, o valor recebido é menor, girando em torno de R$ 3,00 a R$ 3,50 por redação (PORTAL G1, 2016, s.p.).

Um fato relevante para o desenvolvimento de uma solução inteligente para correção de redações é que o MEC anunciou em 2019 que o ENEM deverá ser aplicado em formato totalmente digital até o ano de 2026. Esta mudança viabilizará e favorecerá o processo de correção automatizada com aplicação de solução inteligente, o que possibilitará a implementação comercial da solução a ser desenvolvida nesta dissertação.

Embora haja algumas iniciativas para o desenvolvimento de plataformas que fazem correções automatizadas, tais ferramentas ainda não avaliam todas as competências do ENEM ou ainda não são sendo distribuídas e aplicadas em larga escala. Um exemplo é a ferramenta CIRA, desenvolvida em 2020 numa pesquisa realizada pela USP São Carlos (USP, 2021). Tal ferramenta realiza a correção de redações, no entanto, em testes realizados por esta pesquisadora a mencionada solução excluiu redações com notas zeradas, uma vez que sua inteligência de correção automatizada tem como foco apenas aspectos ligados à correção ortográfica ou problemas de pontuação.

Outras pesquisas demonstram a intenção de alguns estados brasileiros implantarem sistemas dessa natureza, como é o caso do governo do Paraná (SEED, 2020, s.p.) e do governo de Goiás (Portal do Governo de Goiás, 2020, s.p.). Ainda assim, as plataformas citadas não estão sendo distribuídas em larga escala, sendo que os projetos estão em fase de experimentação e testes preliminares.

Ao realizar um estudo sobre as plataformas que já realizam correções automatizadas de textos dissertativos destacaram-se também projetos de outros países. Em trabalho realizado pela EdX, empresa sem fins lucrativos fundada pela Universidade Harvard e pelo MIT (Massachusetts Institute of Technology), destacou-se a criação de um programa que não só corrige as provas escritas, como também dá a nota quase automaticamente aos alunos (MARKOFF, 2013, s.p.). Outra plataforma com a mesma finalidade é “Write & Improve”, ferramenta gratuita para alunos de inglês que marca a escrita em segundos. Este serviço é fornecido em associação com o Cambridge English da Universidade de Cambridge (CAMBRIDGE ENGLISH, 2020, s.p.).

A efetividade de ferramentas de correção automatizada de textos tem experimentado evolução e ampliado seus benefícios. Anant Agarwal, presidente da EdX, destaca que o software de notas instantâneas é uma ferramenta pedagógica útil, que permite que os estudantes façam testes e escrevam redações várias vezes para melhorar a qualidade de suas respostas. Segundo Agarwal, “os alunos nos dizem que estão aprendendo muito mais com o feedback instantâneo” (MARKOFF, 2013, sp.).

Outro ponto importante a destacar é o fato de os professores de língua portuguesa estarem dispostos a aderir a novas tecnologias para correção de redações. Em pesquisa realizada por Pinho et al. (2020) junto a 30 professores, 90% se mostraram abertos ao uso de novas tecnologias voltadas à correção inteligente automatizada de redações, sendo que apenas 10% dos entrevistados manifestaram preferência por realizar a correção manual de redações. Neste mesmo estudo foi questionado aos professores sobre sua opinião a respeito da aplicação futura de uma plataforma que realize a correção automática de redações. No questionamento feito os professores podiam marcar mais de uma opção de resposta. Os resultados mais relevantes indicaram que 70% dos professores informaram que iriam ganhar mais tempo e assim poderiam aumentar a quantidade de aplicações de redações; 73% acreditam que a tecnologia ajudará muito, pois o processo de correção de redações é árduo; e outros 10% não confiariam num sistema fazendo a correção automática das redações. Não obstante, mesmo não tendo a preferência destes últimos professores, obteve-se como sugestão que uma solução inteligente automatizada poderia fornecer uma primeira correção, que deverá ser considerada na elaboração da avaliação definitiva do profissional.

A partir dos levantamentos preliminares efetuados junto à plataforma teórica disponível, bem como junto aos profissionais envolvidos, entende-se a importância da realização de pesquisas que apliquem técnicas inteligentes na educação, sobretudo no Brasil. Isto porque, segundo a Agência Brasil (2020), o país está atrasado na corrida por Inteligência Artificial e sua maior demanda é no setor negócios, deixando a educação com uma fatia mínima na aplicação de tecnologias inteligentes.

1.4 Delimitação do Tema

Este estudo enfoca a avaliação automatizada de redações com aplicação de solução inteligência artificial para tanto. No escopo em questão será tratada exclusivamente a competência 2 do ENEM, que busca compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolvimento do tema, considerados os limites estruturais do texto dissertativo-argumentativo em prosa. As demais quatro

competências do ENEM não serão consideradas de forma direta nesta pesquisa.

Para a condução dos experimentos para a elaboração de solução automatizada com aplicação de inteligência artificial serão considerados o Processamento de Linguagem Natural (PLN) e a Mineração de Textos (MT) com aplicação de Redes Neurais Convolucionais (RNC), MLP (MultiLayer Perceptron) e outras técnicas de classificação de Aprendizado de Máquina (AM).

A escolha destas técnicas ocorreu após o processo da Revisão Sistemática da Literatura (RSL) executado para a elaboração da presente dissertação. Salienta-se ainda que a presente pesquisa não levará em consideração outros ramos da IA, tais como: robótica, visão computacional e speech analytics (inteligência artificial associada a voz).

2 REFERENCIAL TEÓRICO

Neste capítulo será apresentada inicialmente a revisão sistemática da literatura com suporte a técnicas bibliométricas, em seguida é exposto o atual contexto da produção textual no Brasil e, por fim, as técnicas de Inteligência Artificial utilizadas nesta pesquisa.

2.1 Revisão sistemática da literatura com suporte de técnicas bibliométricas

Neste item são apresentados os resultados relativos à pesquisa bibliométrica sobre as principais técnicas que estão sendo empregadas para análise de textos utilizando Processamento de Linguagem Natural (PLN), Mineração de Textos (MT) e Aprendizagem de Máquina (AM). Assim, este passo da pesquisa explicita a contribuição do conhecimento científico derivado da área temática nesta dissertação, buscando ainda promover o aprofundamento ao contexto para o entendimento das técnicas que estão sendo aplicadas atualmente.

2.1.1 Revisão Bibliométrica

SOARES et al. (2016) afirmam que a bibliometria possibilita a observação de toda produção científica registrada em repositórios de dados. Esta pesquisa bibliométrica foi realizada utilizando dados do repositório do Google acadêmico, Web of Science (WoS) e Scopus com o objetivo de compreender a produção científica em nível global, além de realizar a análise dos documentos produzidos especificamente no Brasil.

Para tanto, foi realizada busca nas bases de dados mencionadas procurando evidenciar literatura científica que sustente esta pesquisa de dissertação por meio da execução de Revisão Sistemática da Literatura (RSL) para avaliar e interpretar os documentos disponíveis. A RSL foi baseada nas diretrizes propostas por Kitcheham e Charters (2007), que indicam três fases principais para a sua execução: planejamento, condução e elaboração de relatórios. A análise realizada foi de caráter quantitativo por meio das redes bibliométricas geradas a partir dos trabalhos encontrados.

2.1.2 Planejamento

Na fase do planejamento foram definidos os objetivos desta pesquisa bibliométrica, sendo: identificar os principais documentos e sua evolução no decorrer dos últimos dez anos em relação aos termos utilizados neste estudo; identificar quais são os autores mais relevantes e, por fim; relacionar os trabalhos encontrados por meio das palavras-chave a fim de entender se estes se adequam ao escopo de pesquisa em questão. Desta forma, foram levantadas três questões bibliométricas (QB), tendo como base a proposta de Salvetti (2019):

I. QB-1: Qual a evolução temporal em relação à quantidade de pesquisas realizadas entre no período de 2011 a 2020?;
II. QB-2: Quais os principais autores da área de estudo em nível mundial, visualizados por meio dos repositórios do WOS e Scopus?;
III. QB-3: Qual a relação entre os trabalhos encontrados a partir das palavras-chave estipuladas?

Após responder as três questões bibliométricas estipuladas com o objetivo de entender a pertinência e relevância da temática abordada nesta dissertação, o próximo passo voltou-se a responder a seguinte questão de pesquisa (QP):

I. QP: Quais são as técnicas inteligentes utilizadas na área da Educação nos últimos dez anos associadas à PLN, MT e AM?

No Quadro 1 é observada a indicação da estratégia de pesquisa aplicada nas bases de dados selecionadas, bem como as palavras-chave adotadas que foram diferenciadas em cada base, pois cada repositório tem uma estratégia de busca específica nas relações lógicas entre as strings a serem adotadas. Contudo, vale ressaltar que o mesmo objetivo de busca foi mantido nas strings de busca aplicadas às duas bases de dados pesquisadas. A pesquisa foi realizada em junho de 2021 e o intervalo de tempo estipulado foi os últimos dez anos (2011 a 2020).

Quadro 1: Estratégia de busca em bases de dados

Consultas	Google Acadêmico		Web of Science		Scopus	
c1	(("writing" OR "essay" OR "text analysis" OR "education") AND ("natural language processing" AND ("machine learning" OR "deep learning" OR "text mining")))	653 resultados	(("writing" OR "essay" OR "text analysis" OR "education") AND ("natural language processing" AND ("machine learning" OR "deep learning" OR "text mining")))	362 resultados	(("writing" OR "essay" OR "text analysis" OR "education") AND ("natural language processing" AND ("machine learning" OR "deep learning" OR "text mining")))	1009 resultados
c2	((("writing" OR "essay" OR "text analysis") AND "education") AND ("natural language processing" AND ("machine learning" OR "deep learning" OR "text mining"))AND ("similarity" or "cluster" or "neural network" or "classification"))	9 resultados	((("writing" OR "essay" OR "text analysis") AND "education") AND ("natural language processing" AND ("machine learning" OR "deep learning" OR "text mining"))AND ("similarity" or "cluster" or "neural network" or "classification"))	4 resultados	((("writing" OR "essay" OR "text analysis") AND "education") AND ("natural language processing" AND ("machine learning" OR "deep learning" OR "text mining"))AND ("similarity" or "cluster" or "neural network" or "classification"))	15 resultados

Fonte: Autora (2021).

O Quadro 1 exibe as respostas obtidas na busca de trabalhos efetuada nas bases de dados selecionadas. Na busca C1 o objetivo foi entender de forma mais generalista a relação da escrita com PLN, MT e AM. Com esta primeira pesquisa é esperado responder as três questões bibliométricas estipuladas. Na busca C2 os critérios adotados foram mais específicos, acrescentando-se alguns termos tais como “similaridade”, “cluster”, “classificação”, ” redes neurais” e “educação”. Estas palavras também estão

associadas à PLN, MT e AM e o objetivo foi compreender como os autores estão empregando tais técnicas em suas pesquisas.

2.1.3 Análise bibliométrica

Os sistemas utilizados para condução dos experimentos bibliométricos foram Vosviewer e Biblioshiny. Eck e Ludo (2020), criadores do Vosviewer, o definem como uma ferramenta de software para construção e visualização de redes bibliométricas, essas redes podem incluir, por exemplo, periódicos, pesquisadores ou publicações individuais. Os desenvolvedores do Biblioshiny foram Cuccurullo e Aria (2019), que explicam que este sistema possui o pacote Bibliometrix, que fornece várias rotinas para importar dados bibliográficos. Este software foi desenvolvido na linguagem de programação R, que é voltada para manipulação, análise de dados estatísticos e visualização gráfica.

Os dados analisados nas ferramentas citadas anteriormente foram retirados do repositório do Web of Science (WoS) e Scopus, estes incluem os principais métodos bibliométricos de análise.

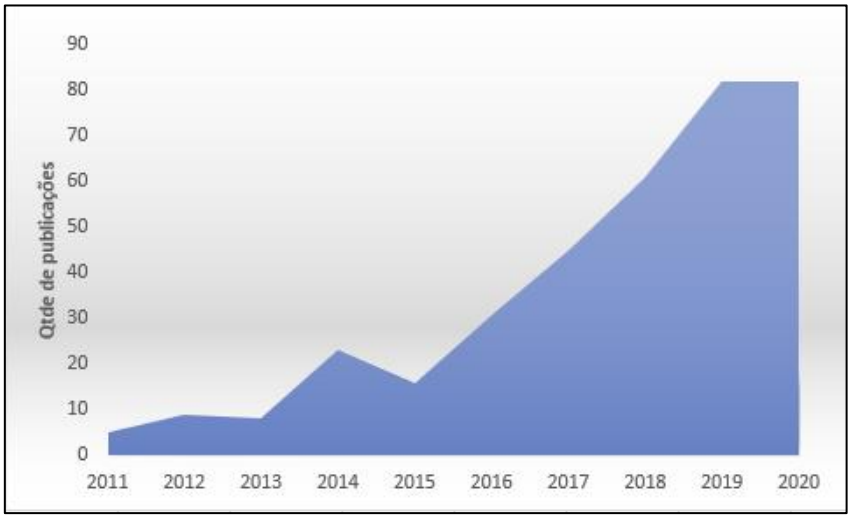
Para responder a QB-1 e compreender a evolução dos trabalhos de pesquisa na temática considerada foram avaliados os dois repositórios de dados consultados. Na Figura 1 é exposta a evolução das publicações encontradas na base do Google Acadêmico. Nesta pesquisa foi levada em consideração apenas páginas em português, pois o objetivo foi entender se nesta área de estudo o Brasil cresce no mesmo patamar que os demais países.

Figura 1: Evolução das publicações no Google Acadêmico

Fonte: Autora (2021).

Na Figura 2 é exposta a evolução das publicações encontradas na base Web of Science. Neste gráfico, assim como no anterior, foi identificado o aumento de publicações a partir do ano de 2017. Porém, como diferença verifica-se que no Web of Science a pesquisa na área de PLN, MT e AM mantém uma estabilidade de publicações a partir de 2019.

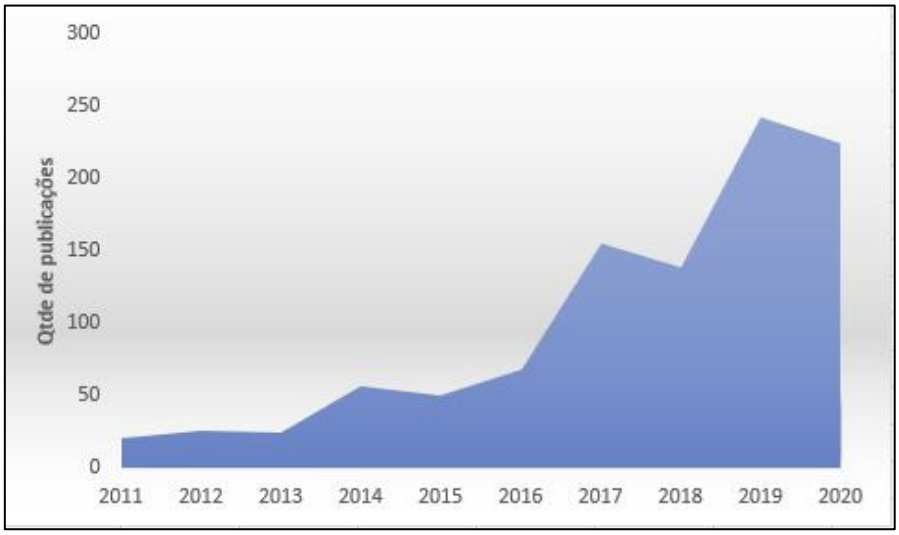
Figura 2: Evolução de publicações na Web of Science



Fonte: Autora (2021).

Na Figura 3 é exposta a evolução das publicações, que mantém o padrão ao longo dos anos. Entretanto, no repositório Scopus a quantidade de publicações a partir de 2017 é bem superior às publicações de outras bases de dados, apresentando em média o triplo de documentos.

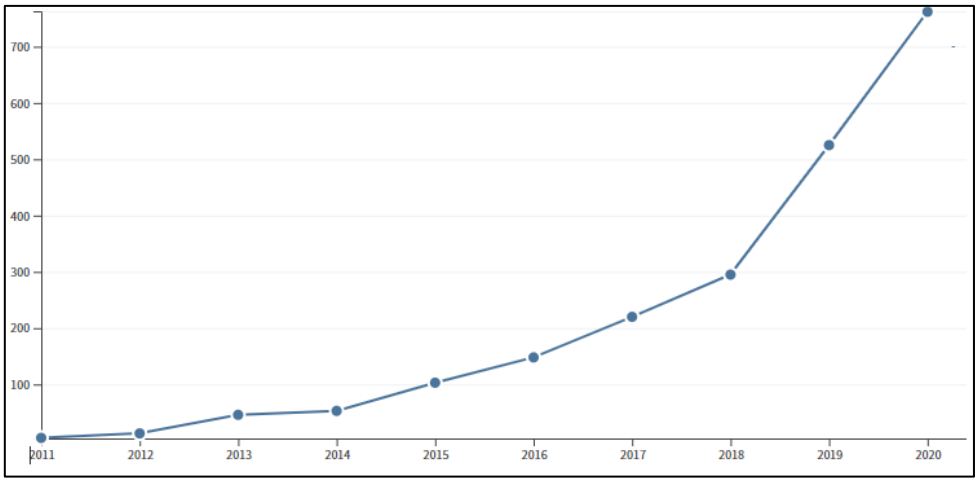
Figura 3: Evolução de publicações no Scopus



Fonte: Autora (2021).

Ainda com a finalidade de responder à QB-1, na Figura 4 é exposto o crescimento nas pesquisas em função das citações feitas nos trabalhos identificados nas bases pesquisadas. A maior incidência ocorre entre os anos de 2018 e 2020. Com base no gráfico a seguir entende-se que o investimento em pesquisa na temática abordada tem sido crescente, o que a suporte a compreensão acerca da relevância da aplicação de técnicas inteligentes para tratamento e análise de textos na atualidade.

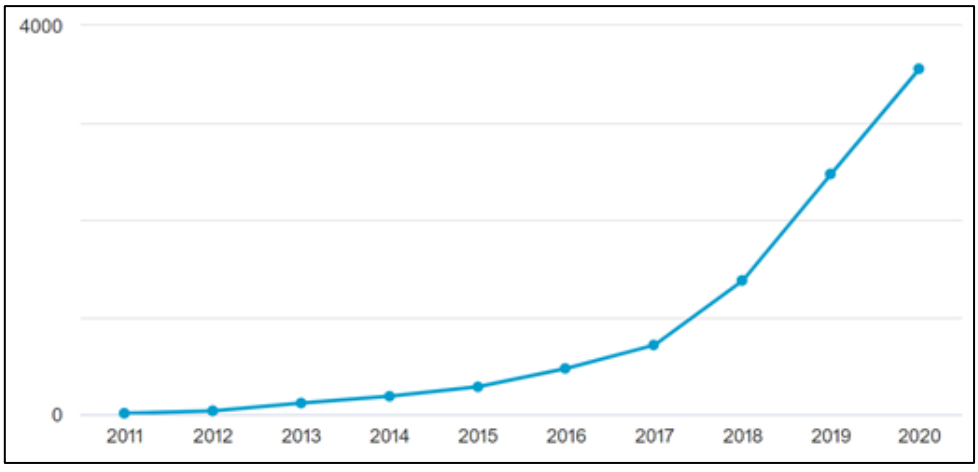
Figura 4: Evolução na citação dos artigos - Web of Science



Fonte: Autora (2021).

Na Figura 5 também se verifica um crescente aumento no número de citações na área deste estudo. O ano de 2018 também foi o de maior referência na base pesquisada.

Figura 5: Evolução na citação dos artigos - Scopus



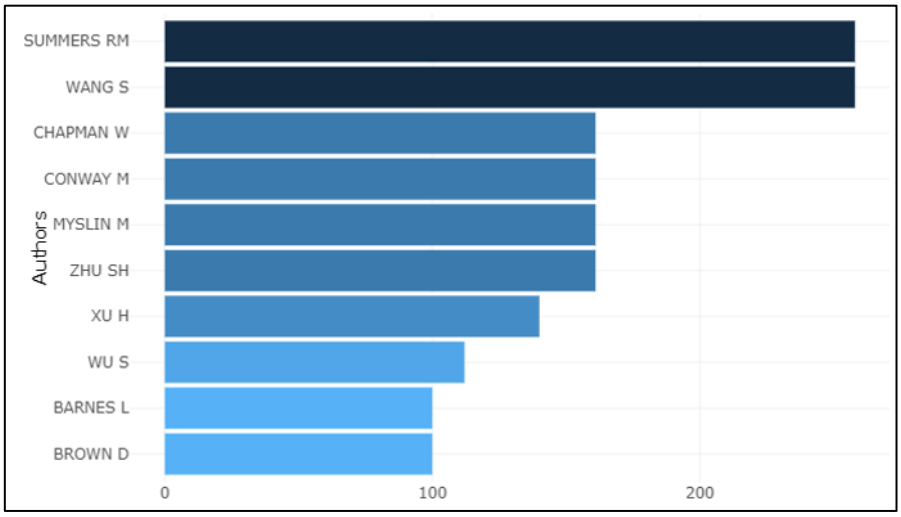
Fonte: Autora (2021).

O próximo passo foi responder à QB-2, que buscava compreender quais os principais autores da área em quantidade de citações e frequência de publicações. A análise executada demonstra que no Google Acadêmico prevalece a ocorrência de teses e dissertações, enquanto no WoS e Scopus ocorre maior frequência de publicação de trabalhos no formato de artigos. De acordo com a CAPES (2019), artigos científicos são produções mais curtas, que abordam os assuntos de forma mais atual e garantem. Assim, para os próximos gráficos apresentados será demostrada a relevância dos autores contidos no repositório no Web of Science e Scopus.

Smiraglia (2011) informa que as citações apontam os paradigmas, procedimentos metodológicos pertinentes, bem como quem são os pesquisadores de maior relevância, já a cocitação trata da frequência com que dois documentos, autores, periódicos ou países, entre outros, são citados de forma simultânea na literatura científica.

Na Figura 6 são indicados os dez autores mais citados em trabalhos publicados no repositório WoS na temática considerada nesta pesquisa. O autor mais citado foi Ronald M. Summers (EUA). Sua área de atuação é Medicina e possui grande número de artigos que tratam da área de Inteligência Artificial, sendo os artigos mais citados na área de Redes Neurais. O segundo autor mais citado foi Shijun Wang (China), também pertinente à área de Medicina, sendo que suas publicações mais citadas também utilizam Aprendizado de Máquina e Redes Neurais. A terceira autora mais citada é Wendy Chapman (EUA), com trabalhos mais citados na área de Processamento de Linguagem Natural na Medicina.

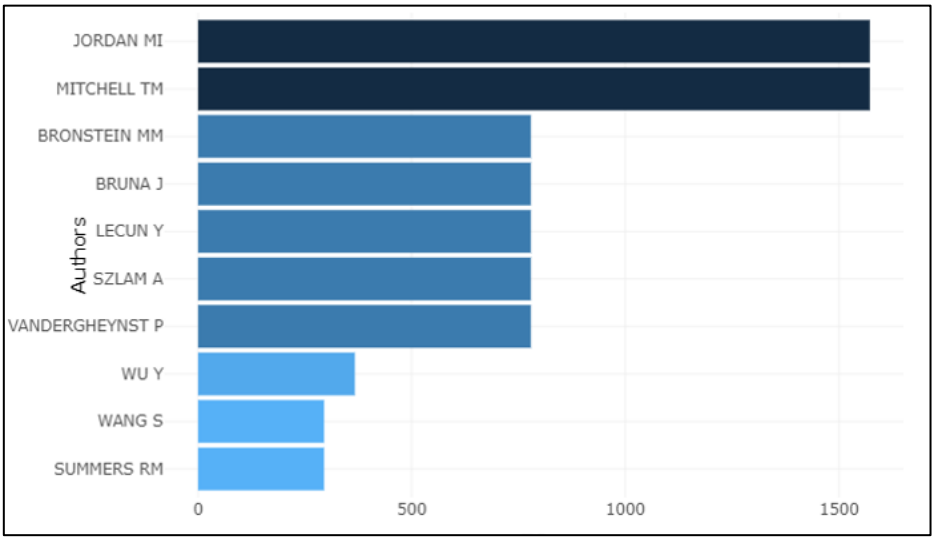
Figura 6: Autores mais citados - Web of Science



Fonte: Autora (2021).

Na Figura 7 são indicados os dez autores mais citados no repositório do Scopus. O autor mais citado foi Michael Jordan (EUA), com área de atuação em Ciência da Computação e possuidor de grande número de artigos que tratam da área de classificação de sentimentos, gradiente estocástico e *Deep Learning*. O segundo autor mais citado foi Tom Mitchell (EUA), com publicações mais citadas na área de Processamento de Linguagem Natural, com análise de sentimentos e também *Deep Learning*. O terceiro autor mais citado é Michael Bronstein (Itália), com trabalhos mais citados na área de *Deep Learning e Visão Computacional*.

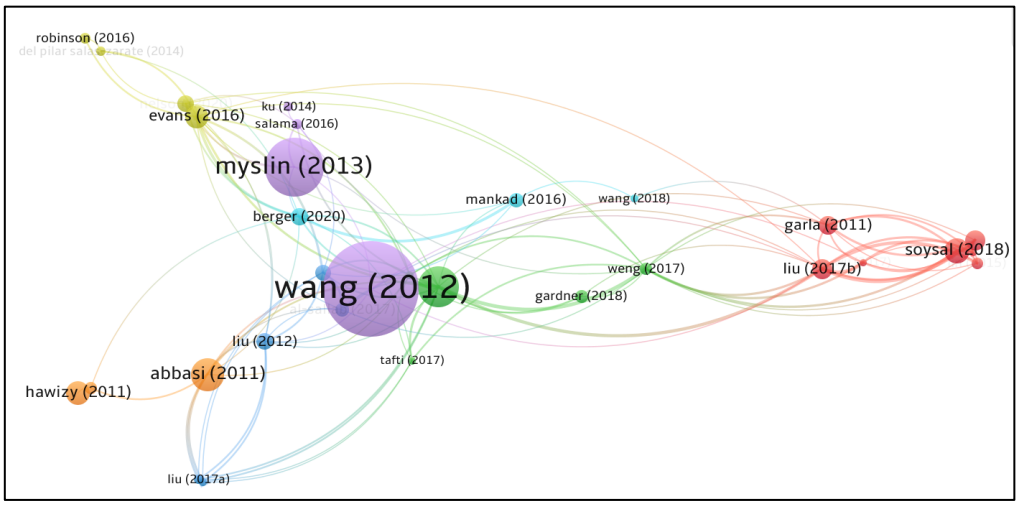
Figura 7: Autores mais citados - Scopus



Fonte: Autora (2021).

Na Figura 8 pode ser visualizado o acoplamento bibliométrico entre os autores selecionados nesta área deste estudo, de acordo com os documentos do WOS. Segundo Grácio (2016), o acoplamento bibliométrico permite conhecer as relações estruturais de conectividade, tais como a proximidade, vizinhança, associação e interlocução estabelecida entre documentos e pesquisadores. No acoplamento oriundo desta pesquisa, verifica-se a conectividade a partir das cores dos clusters estabelecidos. As figuras 8 e 9 demonstram a relação entre os autores que se relacionam por meio de citações, o tamanho das elipses demonstrar o impacto dos autores.

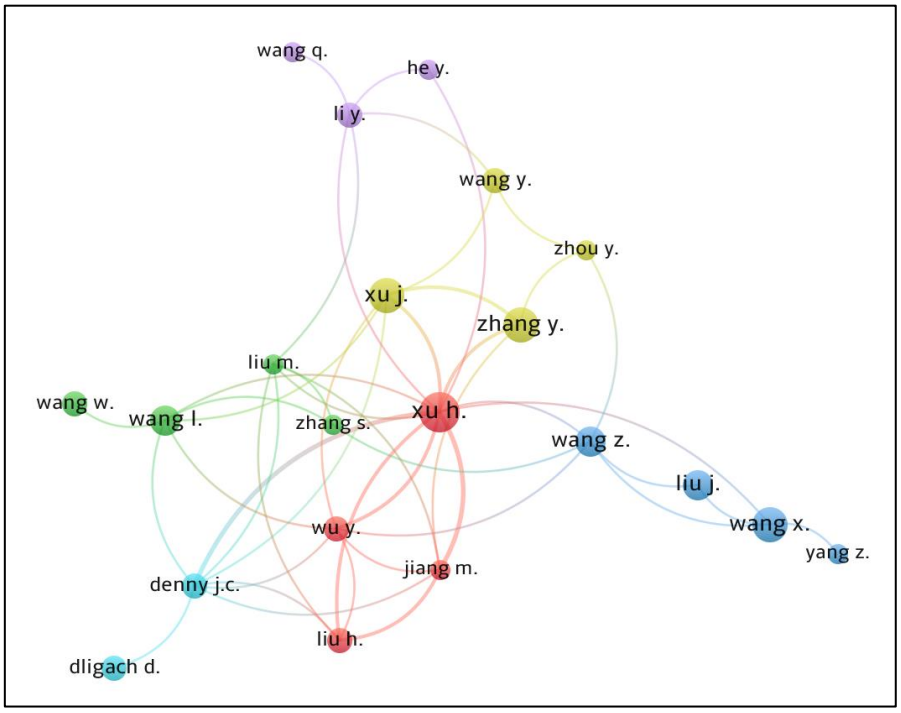
Figura 8: Acoplamento bibliométrico de autores e obras - WOS



Fonte: Autora (2021).

Na Figura 9 é demostrado o acoplamento bibliométrico entre os autores verificados no repositório do Scopus.

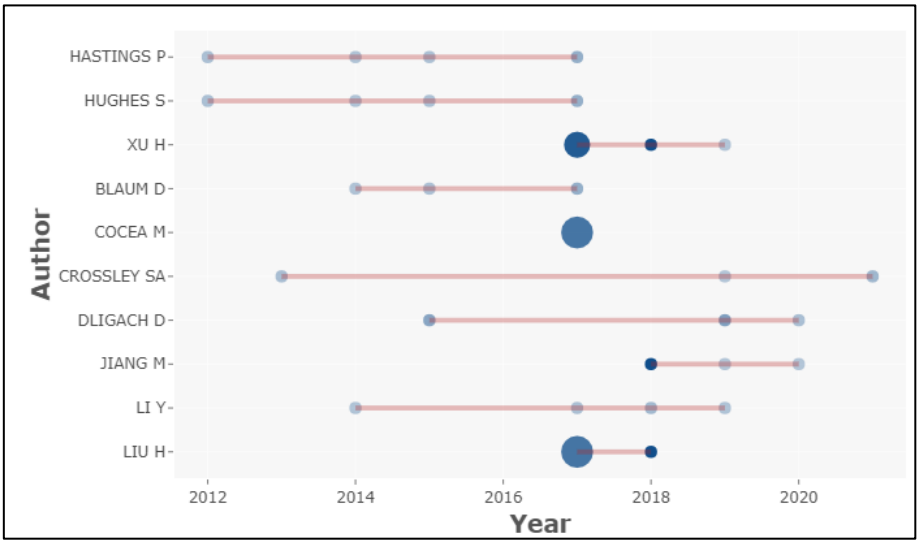
Figura 9: Acoplamento bibliométrico de autores e obras - Scopus



Fonte: Autora (2021).

Ainda com objetivo de responder a QB-2, pode ser visualizado nas Figuras 10 e 11 quais são os autores que produziram mais trabalhos no decorrer do período analisado (2011-2020).

Figura 10: Autores maior número de publicações WOS

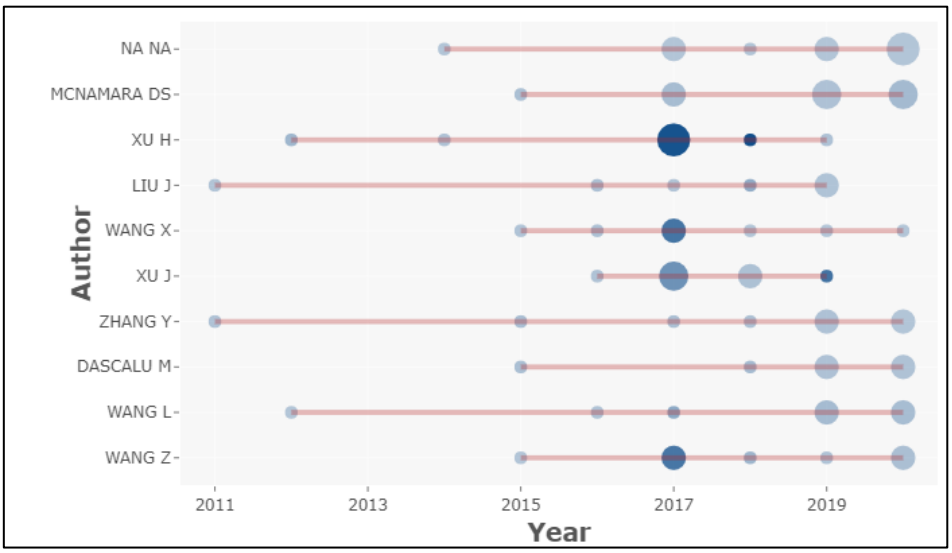


Fonte: Autora (2021).

Nas Figuras 10 e 11 estão relacionados os dez autores que mais produziram no período pesquisado. No repositório do WoS, o autor com maior

produção é Andrew Scott Crossley (EUA), que atualmente possui 159 publicações, sendo que em seus artigos prevalece o uso da linguística associada à PLN. Já no repositório do Scopus há um padrão dos autores em relação a publicação de trabalhos desde 2015. O autor que se manteve constante ao longo dos anos pesquisados foi Zhangn (EUA), que possui cerca de 36 publicações no período analisado.

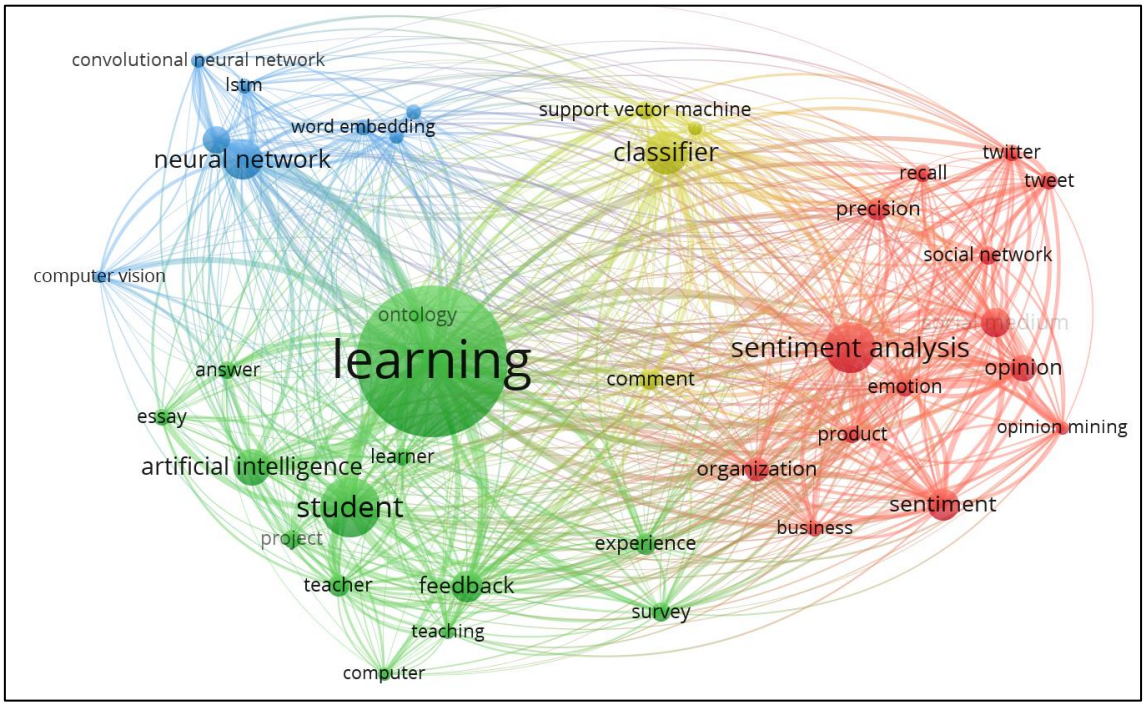
Figura 11: Autores maior número de publicações - Scopus



Fonte: Autora (2021).

A última questão bibliométrica a ser respondia (QB-3) buscava analisar a relação entre os trabalhos encontrados por meio de suas palavras-chave, visando assim identificar os temas mais tratados e a relação entre estes. Na Figura 12, elaborada pelo software VosViewer, foi analisada a ocorrência de palavras-chave entre todos os trabalhos encontrados nas bases pesquisadas.

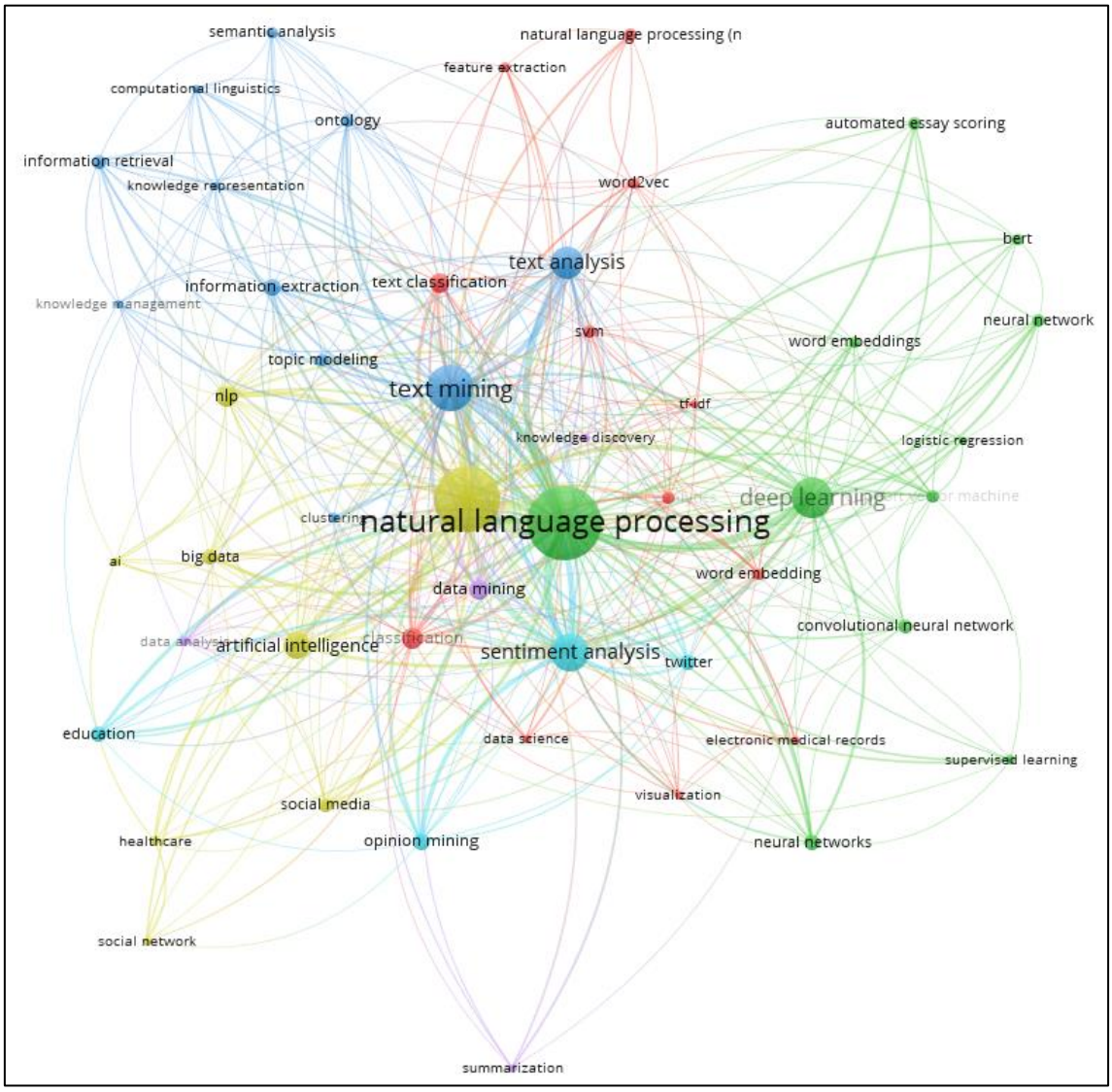
Figura 12: Relação entre os documentos por meio de Palavras Chaves - WOS



Fonte: Autora (2021).

Analisando-se as informações expostas nas Figuras 12 e 13 entende-se que os documentos encontrados nas bases de dados consultadas realmente possuem relação entre si, incluindo-se aí a temática abordada nesta dissertação. Além disso, destacam-se algumas palavras-chaves que estão relacionadas a proposta desta pesquisa, como: “*Classifier*”, “*Convolutional Neural Network*”, “support vector machine”, "Supervised Learning","Clustering”, “*essay*”, “teaching” e ”feedback”. Assim sendo, foram aplicados mais alguns filtros, conforme já apontados na condição C2 exposta no Quadro 1, com o objetivo de identificar trabalhos que especificam técnicas com a utilização de Redes Neurais, Similaridade ou Clusterização, as quais podem ser utilizadas nesta pesquisa.

Figura 13: Relação entre os documentos por meio de Palavras Chaves - Scopus



Fonte: Autora (2021) – Extraído do VosViewer

2.1.4 Elaboração do Relatório Bibliométrico

Após a condução dos experimentos é estabelecida a conclusão sobre os dados coletados para as questões bibliométricas QB-1, QB-2 e QB-3. Os principais conhecimentos extraídos da análise executada são expostos no Quadro 2.

Quadro 2: Principais resultados das questões bibliométricas aplicadas

QB-1	A análise da evolução temporal dos trabalhos publicados na temática de pesquisa desta dissertação apontou que a partir de 2017 houve aumento nas publicações e que o maior índice de citações começou a ocorrer em 2018, demonstrando que a temática ainda é recente e relevante para ser aplicada nesta pesquisa.
QB-2	Os autores mais relevantes em relação às citações foram: Ronald M. Summers (EUA), Shijun Wang (China), Wendy Chapman (EUA), Michael Jordan (EUA), Tom Mitchell, (EUA) e Michael Bronstein (Itália). Os autores mais produtivos são Andrew Scott

	Crossley (EUA) e Zhang (EUA). Assim, como sequência natural destes resultados, nesta pesquisa os trabalhos desses autores serão analisados em maior detalhamento para o aprofundamento do referencial teórico a ser estabelecido nesta pesquisa.
QB-3	Na ocorrência de palavras-chave verificou-se que, além dos termos usados nesta pesquisa, surgiram novos termos, tais como: *"Classifier", "Convolutional Neural Network", "support vector machine", "Supervised Learning","Clustering", "essay", "teaching" e "feedback"*. Todos estes termos possuem relação com a proposta desta pesquisa. Como próximo passo da RSL as técnicas utilizadas serão avaliadas com maior detalhamento.

Fonte: Autora (2020).

2.1.5 Detalhamento da RSL – Pesquisas mais relevantes e critérios de exclusão

O próximo passo da RSL voltou-se a analisar em detalhes os trabalhos filtrados a partir da *string* C2. Os resultados estão expostos nos Quadros 3, 4, 5 e 6. Neste passo foram analisados os trabalhos indicados como mais pertinentes e contributivos ao estabelecimento da plataforma teórica estipulada para esta pesquisa de dissertação.

Para a validação e aplicação nesta pesquisa foi realizada a leitura dos documentos separados como resultado da busca C2, abstraindo-se deles conceitos importantes para esta pesquisa. O Quadro 3 relaciona oito dissertações/teses encontradas no repositório Google Acadêmico. Na terceira coluna estão listadas as técnicas utilizadas para atingir os objetivos das pesquisas.

Foram excluídos desta pesquisa os trabalhos que não demonstraram relevância quanto à aplicação de experimentos computacionais utilizando-se alguma técnica associada à PLN, MT e AM. Contudo, apenas um trabalho foi excluído por não cumprir a necessidade estipulada nesta pesquisa. Desta forma, o Quadro 3 possui oito dissertações/teses.

Quadro 3: Teses e dissertações disponíveis no Google Acadêmico

Título	Conceitos principais / Autores	Procedimentos e técnicas utilizadas para alcançar os resultados
Lie-o-matic: using natural language processing to detect contradictory statements	O foco desta dissertação é fazer a relação entre documentos e detectar contradições. A pesquisa trabalha com um a base de dados que expõe opiniões do ex-presidente Donald Trump. Dissertação de Baldaia (2020)	Aprendizado Supervisionado, Classificação, Arvore de Decisão, *Support Vector Machines* (SVMs)
Text analytics in business environments: a	O objetivo desta tese é desenvolver um framework para fazer a análise de dados turísticos, com base no TripAdvisor, o autor	Construção da base de dados por Web Scraping, Classificação de textos,

managerial and methodological approach	deseja monitorar tendências de destinos ao identificar padrões em viagens e analisar os sentimentos dos usuários nos locais de estadia. Tese de Marcolin (2018)	redes neurais, análise de sentimentos, nuvens de palavras e clusterização.
Reconhecimento de entidades nomeadas na Web	O principal objetivo da dissertação é encontrar os melhores métodos de extrair Reconhecimento de Entidades Nomeadas (NER) baseados em aprendizado de máquina para a Web. Foi utilizado como foco extrair nomes de pesquisadores. Dissertação de Veneroso (2019)	Processamento de Linguagem Natural e Redes Neurais Artificiais
Fully-disentangled text-to-image synthesis	O objetivo desta dissertação é realizar a síntese de imagens em linguagem natural, trazendo informações de cores e texturas, fornecer diversidade para formas, poses, planos de fundo ou qualquer outro recurso. Dissertação de Barros (2019)	Visão computacional para extrair informações das imagens, similaridade entre textos, e Processamento de Linguagem Natural.
Hotel Revenue Management: Using Data Science to Predict Booking Cancellation	Esta tese procura desenvolver um modelo de classificação para prever os clientes que irão cancelar estadias em hotéis. Tese de António (2019)	Microsoft Azure Machine Learning Studio, *Decision Tree (BDT), Decision Forest (DF), Decision Jungle (DJ), Locally Deep Support Vector Machine (LDSVM) and Neural Network (NN)*
Um benchmark para comparação de métodos para análise de sentimentos	Essa dissertação faz uma comparação de 21 métodos e ferramentas muito utilizados na análise de sentimentos para melhor entender suas performances. Dissertação de Pollyanna Gonçalves (2015)	Avalia técnicas de Aprendizado supervisionado, Naive Bayes, aprendizado não supervisionado (identificação de padrões em rótulos)
KDC: uma abordagem baseada em conhecimento para classificação de documentos	O objetivo dessa dissertação foi classificar documentos pela associação de rótulos de classes a documentos com o objetivo de criar agrupamentos semânticos. Dissertação de Gleidson Silva (2015)	Algoritmos de classificação utilizados: algoritmos Árvore de Decisão, SVM e Naive Bayes.
Extracting keywords from tweets	O objetivo dessa dissertação é avaliar a eficácia do algoritmo YAKE para extrair informações do Twiter. O autor faz a comparação com outros algoritmos de mesma finalidade. Dissertação de Farinha (2018).	Propõe o uso do YAKE (um algoritmo de extração de palavras-chave não supervisionado) a partir de um conjunto de tweets

Fonte: Autora (2021).

Com a finalidade de continuar entendendo a aplicação das técnicas que serão utilizadas nesta pesquisa foi realizada uma busca no Google Acadêmico, mas utilizando-se as palavras em português. O objetivo foi fortalecer a base teórica deste estudo e identificar o que tem sido utilizado no Brasil em relação ao uso das técnicas associadas à PLN, AM e MT. Desta forma, foram selecionados mais treze documentos em português, conforme disposto no Quadro 4.

Quadro 4: Teses e dissertações brasileiras associadas à pesquisa

Título	Conceitos principais / Autores	Procedimentos e técnicas utilizadas para alcançar os resultados
Contribuições para acelerar o aprendizado sobre a construção de uma máquina de classificação de sentimentos utilizando processamento de linguagem natural	É proposta a criação de uma máquina de classificação de sentimentos (positivos e negativos) na língua portuguesa. Foi utilizada uma base de comentários sobre filmes na língua portuguesa do Brasil. Dissertação de Bonadia (2019).	PLN para o Pré-processamento de textos; Algoritmos de Regressão Logística; *Naive Bayes;* Árvore de Decisão; Floresta Aleatória SVM e SVM linear
Classificação de tendências políticas em notícias via mineração de texto e redes neurais sem peso	Buscou identificar a polaridade em notícias políticas em português através do processo de mineração de dados textuais com a utilização da Rede Neural sem Peso WiSARD e de uma derivação, a ClusWiSARD. Dissertação de Cavalcanti (2017).	*Crawler* para obter textos; Pré-processamento de textos; TF-IDF para verificar ocorrência de palavras; Técnicas de Rede Neural sem Peso WiSARD; ClusWiSARD; Regressão Logística; *Naive Bayes;* SVM
Método *fuzzy* para a sumarização automática de texto com base em um modelo extrativo (FSumm)	Usa técnica de sumarização extrativa ao qual se associam tarefas de Recuperação de Informação (RI) e de Processamento de Linguagem Natural (PLN). Dissertação de Goularte (2015).	Pré-processamento de textos; TF-IDF para verificar ocorrência de palavras; Lógica *Fuzzy;*
Uma aplicação de mineração de dados para recomendação social	Propõe-se a criação de uma metodologia e aplicação que realize a Mineração de Dados com informações textuais vinculados a dados e geolocalizações. Dissertação de Feitosa (2013).	CRAWLER para obter textos; Biblioteca NLTK; Biblioteca *Numpy*; e algoritmos TF_IDF; KDD; KNN; JSON
Uma plataforma distribuída de mineração de dados para big data: um estudo de caso aplicado à Secretaria de Tributação do Rio Grande do Norte.	Propõe uma plataforma distribuída de mineração de dados para a Secretaria de Tributação do Rio Grande do Norte, que possa extrair conhecimento de maneira variada, considerando as características específicas das notas fiscais eletrônicas. Dissertação de Santos (2018).	Biblioteca NLTK para PLN; algoritmo do K-means; software WEKA utilizando MLP;
Avaliação de mecanismos de suporte à tomada de decisão e sua aplicabilidade no auxílio à priorização de casos em regulações de urgências e emergências	Trabalho propôs a criação de uma proposta para suporte em prioridades para atendimento médico regulando urgência e emergência fazendo análise de risco por óbito, similaridade entre casos histórico e apoio a decisão baseado nesses históricos. Tese de Pollettini (2016).	Algoritmos do KNN; TF-IDF; PHP, software WEKA utilizando IBK, J48; *RBF Network; Multilayer Perceptron; Bayes Net; Nave Bayes;* Vote
Aprendizado Automático de Relações Semânticas entre *Tags de Folksonomias*	Propõe uma abordagem de aprendizado indutivo para a detecção de semântica entre tags de folksonomias, os atributos de aprendizado são medidas de similaridade/distância tag-tag empregadas como heurísticas por diversos trabalhos relacionados para capturar relações de parentesco entre tags. Tese de Rêgo (2016).	Abordagem do KDD*; extração de textos via crawler*; *Técnicas de* RI - Similaridade; Algorimtos *K-Means; C4.5, Random Forest, SVM, Naive Bayes*, Regressão Logística, Gripa;
Análise e caracterização de textos intencionalmente enganosos escritos em português usando	Propõe extrai e analisar características textuais em documentos intencionalmente enganoso, escritos em português, com a finalidade de classificar como verdadeiros ou enganosos.	Pré-processamento de textos; *Linguist Inquiry and Word count(LIWC); Latent Dirichlet Allocation (LDA); Vector Space Model*

métodos de processamento de textos	Dissertação de Okano (2020).	*(VSM); Random Forest, SVM, Naive Bayes, k-fold cross-validation; f1-score; Gated Recurrent; Unit (GRU)*
Representação de coleções de documentos textuais por meio de regras de associação	A abordagem proposta espera-se identificar relações entre palavras de um documento, além de reduzir a dimensionalidade, pois são consideradas apenas as palavras que ocorrem ou que coocorrem acima de uma determinada frequência para gerar as regras. Dissertação de Rossi (2011).	*Bag-of-words; 10-fold-cross-validation; Classificadores: Naive Bayes, j48, SMO (Sequencial Minimal Optimization), KNN(k-Nearest Neighbors*); Weka; Similaridade; Agrupamento.
Suporte às micro e pequenas empresas a partir da gestão baseada em evidências: construção de ferramenta computacional baseada em inteligência artificial	Propôs a construção de uma ferramenta inteligente para dar suporte as MPES na retirada de dúvidas. Dissertação de Lacerda (2018).	Pré-processamento de textos; RTextTools; TF-IDF; SVM; SLDA; BOOSTING
Agrupamento de documentos forenses utilizando redes neurais art1	Agrupa tematicamente documentos retornando dados de uma ferramenta de busca utilizada com coleções textuais forenses. Oferece ao perito uma forma organizada de obter uma visão geral do conteúdo dos documentos durante o exame pericial. Dissertação de Araújo (2012).	Pré-processamento de textos; *Normalized Mutual Information (NMI)* RNA (Rede Neural Artificial); RNA SOM (*Self Organizing Map);* RNA ART (Teoria da Ressonância Adaptativa); JAVA.
Análise da evolução da pesquisa em engenharia de transportes	O objetivo principal deste trabalho é realizar uma análise temporal da evolução das pesquisas em transporte nas últimas décadas, a partir de um aprimorado estudo cienciométrico, informétrico e bibliométrico de milhares de artigos. Tese de Stefano (2016).	Webscraping; TF-IDF; Similaridade; Node.js; WordNet; Grafos.
Interface Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em *Big Data*	Visa desenvolver uma plataforma que disponibilize serviços de PLN em Big Data, sem fins lucrativos e determinar os termos mais gerais de um documento fazendo uso da abordagem da Implicação Textual por Generalidade. Dissertação de Chitongua (2019).	*Open Web Spider; Crawlers*; Pré-processamento de textos; MySQL Server; Node.js; Medidas de Associação Assimétrica (AAM).

Fonte: Autora (2021).

O Quadro 5 apresenta os artigos selecionados a partir do critério de consulta C2 no repositório do *Web of Science*. Na terceira coluna estão relacionadas as técnicas que foram utilizadas pelos autores para alcançar os resultados expostos em seus trabalhos. Foram encontrados apenas três trabalhos que tratavam de assuntos relacionados ao tema desta dissertação.

Quadro 5: Artigos Selecionados – Web of Science

Título	Conceitos principais / Autores	Procedimentos e técnicas utilizadas para alcançar os resultados
	Este artigo utiliza de técnicas de PLN e	Além de utilizar PLN os

Computer-Based Classification of Preservice Physics Teachers' Written Reflections	AM para classificar as reflexões de professores de física em fase de formação, quando estão no início da profissão, para entender as dificuldades encontradas por eles no desafio de lecionar. Wulff et al. (2020)	autores utilizaram as seguintes técnicas de classificação: *Decision* Tree Classifier, Multinomial logistic regression, Multinomial Naïve Bayes, e SGD Classifier.
Evaluation and Comparison of Text Classifiers to Develop a Depression Detection Service	Este artigo busca aplicar técnicas de PLN e classificadores para identificar sintomas de depressão em frases. A ideia é identificar o humor do autor das frases para que seja possível a intervenção em sistemas de saúde e oferecer atendimento personalizado após a identificação. Moreno-Blanco et al. (2019)	Técnica de classificação supervisionada: *Decision* Tree Classifier e Naïve Bayes.
Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training	Este arquivo trata da aplicação automatizada para pontuação em ambientes educacionais de treinamento online. A metodologia foi testada e validada em um conjunto de dados de 173 relatórios (em língua holandesa) que os alunos criaram em um jogo aplicado à política ambiental. Westera et al. (2018)	Usa Classificação binária (Aprovado ou Reprovado). Classificadores: Linear regression – M5,Linear regression, MLP, SVR – polynomial, SVR – RBF, SVR – PUK

Fonte: Autora (2021).

O Quadro 6 contém a síntese dos trabalhos selecionados pela consulta C2 no repositório do Scopus. Na terceira coluna estão relacionadas as técnicas que foram utilizadas pelos autores para alcançar os resultados expostos em seus trabalhos. Apenas nove trabalhos tratavam de assuntos relacionados à temática desta dissertação, ou possuíam livre acesso para leitura na íntegra.

Quadro 6: Artigos Selecionados pela C2 - Scopus

Título	Conceitos principais / Autores	Procedimentos e técnicas utilizadas para alcançar os resultados
Fake News Detection Using Content-Based Features and Machine Learning	O objetivo deste estudo é determinar a aplicabilidade de várias técnicas de aprendizado de máquina para a tarefa de identificar notícias falsas. Okuhle e Betram (2020)	Os Classificadores utilizados neste estudo foram: *AdaBoost as AB, Decision* *Tree as DT, K-Nearest Neighbour as KNN, Random Forest as* *RF, Support Vector Machine as SVM and XGBoost as XGB*
Computer-based Classification of Student's Report	Este trabalho aplicou abordagens de aprendizado de máquina e processamento de linguagem natural para avaliar relatos de estudantes universitários no domínio da construção do conhecimento, o objetivo foi medir e analisar o desempenho individual dos alunos.	Neste projeto foram utilizados dois classificadores*: Support Vector Machine (SVM) e Random Forest Classifier (RFC)*

	Segarra-Faggioni e Ratte(2020)	
Quantification of students' learning through reflection on doing based on text similarity	Neste artigo, é avaliado texto redigido pelos alunos de Engenharia Aeroespacial e Mecânica sobre seu aprendizado. A ideia é calcular a simiilaridade entre o que os alunos aprenderam e o que os instrutores esperavam que os alunos aprendessem, fornecendo, assim, orientação baseada em evidências para os instrutores sobre como melhorar a entrega. Peng et al.(2020)	Neste trabalho foi utilizado *cosine distance* para cálculo de similaridade entre sentenças, além das técnicas de PLN e Mineração de textos.
Annotation-free Automatic Examination Essay Feedback Generation	Propõe uma abordagem baseada em inteligência artificial aplicada em ensino a distância para fornecer um feedback rápido ao alunos, utiliza uma combinação das técnicas de processamento de linguagem natural TextRank e similaridade semântica e cria mapas conceituais para apresentação de feedback. Altoe e Joyner (2019)	Para obter os resultadado são utilizados técnicas de PLN e as Redes Neurais Artificiais.
Automated essay scoring using ontology generator and natural language processing with question generator based on blooms taxonomy's cognitive level *Open Access*	A proposta deste artigo é criar um avaliador de resposta de questionários e gerar uma pontuação baseada na resposta do indivíduo. Contreras et al. (2019)	Este projeto trabalha com algoritmos de regressão para identificar a nota: *Linear Regression; LASSO Regression; Ridge Regression; Gradient Boosting Regression*
Towards automated evaluation of handwritten assessments	Projeto propõe avaliar respostas escritas por alunos de forma manuscrita. A proposta é avaliar as respostas dos alunos em relação à resposta de referência, usando o esquema de avaliação bidirecional que classifica a resposta como “correta” ou “incorreta”. Rowtula et al.(2019)	São utilizadas técnicas de visão computacional e OCR para extrair os textos das imagens. Em seguida são aplicadas técnicas de PLN como *Tagging POS* e NER para identificar as entidades e então fazer as classificações.
Predicting at-risk students in a circuit analysis course using supervised machine learning	O objetivo deste artigo é implementar uma abordagem baseada na web para administrar o exercício da escrita e construir um aplicativo totalmente automatizado capaz de avaliar as respostas dos alunos e fornecer feedback ao usuário na tentativa de aprimorar sua compreensão conceitual. Becker et al. (2019)	Foram utilizadas além de técnicas de PLN o aprendizado supervisionado para classificação dos textos utilizando a biblioteca do *Spacy* e classificadores do *scikit-learn,* usando uma abordagem de validação cruzada.
Identification of Semantic Patterns in Full-text Documents Using Neural Network Methods	Este artigo é dedicado ao desenvolvimento de novos abordagens para a análise de textos em linguagem natural com base no mecanismo de redes neurais, o objetivo é analisar big data, identificar padrões e construir dados algoritmos de processamento	Método Word2Vec com o uso do algoritmo o algoritmo Skip-Gram.

	baseados nos padrões encontrados. Zolotarev et al (2019)	
Application of data mining in e-Leaming systems	Neste artigo,foi utilizado métodos de mineração de dados para a classificação de sentenças em linguagem natural, teve como objetivo provar o benefício potencial do uso de métodos de mineração de dados e aprendizado de máquina para identificar essas sentenças entre um grande conjunto de dados textuais coletados. Brajković et al. (2018)	Utilizam classificadores: Naive Bayes Classifier; Decision Tree Classifier; Maximum Entropy Classifier; Support Vector Classifier; NuSCV (nSVC) and Linear SCV (lSVC) classifiers;

Fonte: Autora (2021).

Além dos trabalhos selecionados pela consulta C2, foram selecionados mais nove artigos que tratam de técnicas associadas à temática desta pesquisa porém, sem o uso da *string "education"*. O objetivo foi entender como estão sendo utilizadas as principais técnicas de IA em diferentes contextos de classificação de textos. A conclusão é que os trabalhos dispostos no Quadro 7 também relacionam a maioria das técnicas e algoritmos verificados em trabalhos desenvolvidos na área de Educação.

Quadro 7: Artigos Associados a PLN, MT e AM

Título	Conceitos principais / Autores	Procedimentos e técnicas utilizadas para alcançar os resultados
Text Classification Algorithms: A Survey	Este artigo traz visão geral dos algoritmos de classificação de texto. Cobre diferentes extrações de recursos de texto, métodos de redução de dimensionalidade, algoritmos, técnicas existentes e métodos de avaliação. O artigo disponibiliza todos os seus códigos: https://github.com/kk7nc/Text_Classification. Kowsari (2019).	Pré-Processamento de Texto; *Principal Component Analysis (PCA); Linear Discriminant Analysis (LDA); GradientBoostingClassifier; CNN; CRF; Decision_Tree; K-nearest_Neighbor; MultinomialNB; RCNN; RNN; Random_Forest; Rocchio_classification; SVM;*
Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector	Projetou um sistema de identificação de autoria em documentos, utiliza o Word2Vec para extrair as características mais relevantes de um documento e depois aplica o Classificador *Perceptron* Multicamadas (MLP) para fixar regras de classificação. Benzebouchi (2019).	*Word2Vec; MLP classifier; KNN classifier; Support Vector Machines (SVM) classifier; Logistic Regression; Random Forest; Conv. Neural Net;*
Diacritic restoration of Turkish tweets with word2vec	Faz estudo sobre a restauração de palavras no twitter para a língua formal Turca, que é uma das dificuldades importantes da normalização de texto em mídia social	*Word2vec; Bag of Words (CBOW) e Skip-Gram; Tf-idf; Support Vector Machine (SVM), Naive Bayes (NB) classifier and k-*

	para reduzir o problema de ruído. . Zeynep (2019)	*nearest neighbor (k-NN);*
Evaluating Report Text Variation and Informativeness: Natural Language Processing of CT Chest Imaging for Pulmonary Embolism	O objetivo deste estudo foi quantificar a variabilidade da linguagem em relatórios de texto livre de estudos de embolia pulmonar e avaliar a informatividade do texto livre para prever o diagnóstico, usando aprendizado de máquina como proxy para a compreensão humana. Huesch (2018).	*Utilizou o SAS Enterprise Miner (software de mineração de dados) para aplicar oText Rule Builder node;*
Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes	Comparar o desempenho de pipelines tradicionais om o de incorporação de palavras combinada com uma CNN na realização de uma tarefa de classificação identificando a Classificação Internacional de Doenças baseados no Clinical Modification (ICD-10-CM) que são os códigos de diagnóstico nas notas de alta médica. Lin (2017).	*SVMs; Random Forest; Gradient Boosting Machine; Word Embedding Combined With a Convolutional Neural Network;*
Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure	Analisa conteúdos de artigos científicos e mídias sociais relacionadas à saúde para detectar e identificar Reações adversas a drogas. Foi desenvolvida uma solução de mineração de texto inteligente e escalável em infraestruturas de big data compostas por Apache Spark, processamento de linguagem natural e aprendizado de máquina. Tafti (2017).	*bigNN; crawler;CNN;bag-of-words;word2vec; j2SE; Apache Spark; No-SQL*
Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine	Teve por objetivo construir modelos preditivos de aprendizado de máquina altamente precisos que podem ser usados para acessar ou não um artigo e classificar como uma Tria controlada randomizada como sim ou não. Cohen (2015).	*Weka; logistic regression, decision trees, K-nearest neighbors, SVMLight, SVM*
Summary Frase Classification Using Stylometry	Classifica sentenças de resumo usando um método de aprendizagem estatística que modela sentenças de acordo com uma técnica linguística que examina estilos de escrita, conhecida como Estilometria. As frases em documentos são representadas usando um novo conjunto de atributos estilométricos. Shams (2015).	*K-Nearest Neighbour; Naive Bayes classifiers; Support Vector Machine (SVM) classifie;bag-of-words;clusters;*
Selecting Attributes for Sentiment Classification Using Feature Relation Networks	Propõe um método de seleção de recurso de texto multivariado baseado em regras chamado Feature Relation Network (FRN) que considera informações semânticas e também alavanca as relações sintáticas entre recursos de n-gram. Abbasi (2011).	bag-of-words; WordNet; Decision tree models (DTMs); SVM classifier; genetic Algorithm; Recursive feature Elimination;N-Gram Feature Set;Feature Relation Network;

Fonte: Autora (2021).

Com base nos resultados da análise dos trabalhos encontrados foi respondida a QP estipulada nesta dissertação, ou seja, a questão de pesquisa que procurava compreender quais são as técnicas inteligentes utilizadas nos últimos dez anos associadas à PLN, MT e AM, em especial para análise e classificações de textos. Após a realização da leitura das teses, dissertações e artigos expostos nos quadros anteriores foram selecionadas algumas técnicas de classificação para aplicar nas redações objeto dos experimentos desta dissertação. Assim, os trabalhos pertinentes encontrados na RSL procedida serão explicados no capítulo de referencial teórico.

2.2 Produção textual no Brasil

A atual língua portuguesa como componente escolar é um constructo relativamente recente, só no final do século XIX e início do século XX que a forma da escrita reconhecida atualmente foi incluída como componente curricular nas escolas (SOARES, 2002, pg. 13).

Antes disso, a língua portuguesa era usada apenas em nível de alfabetização e, em seguida, o foco era transferido para o ensino do Latim, ressaltando que o ensino era voltado apenas para as camadas mais privilegiadas da sociedade (SOARES, 2002, pg.13).

Até a década de 1940, os alunos que tinham acesso à escola continuavam a ser de classe alta, porém, o foco de estudo era a gramática, uma vez que a produção de textos escritos ainda não era muito presente no desenvolvimento escolar do aluno (SOARES, 2002, pg.13).

Nas décadas de 1960 a 1980 a escola entra em período de democratização, no qual a linguística, comunicação, expressão e literatura são instituídas visando contribuir ao ensino do Português nas escolas, em razão dos motivos indicados, Soares (2002) entende que se a inclusão do Português é relativamente recente no currículo escolar, sendo a produção textual escrita ainda mais recente.

Nos anos atuais o desenvolvimento da escrita continua sendo um desafio, principalmente para o ensino básico. Em abril de 2019 foi instituído o decreto o nº 9.765 para a Política Nacional de Alfabetização (PNA). Tal

legislação ressalta que a educação é uma preocupação central das nações do século XXI. Contudo, os resultados obtidos pelo Brasil nas avaliações internacionais e os próprios indicadores nacionais revelam um grave problema no ensino e na aprendizagem da leitura e escrita. Assim sendo, o documento enfatiza ser necessário implementar melhores condições para o ensino e a aprendizagem das habilidades de leitura e de escrita em todo o país (MEC, 2019, s.p.).

Perante as dificuldades mencionadas, o Ministério da Educação vem discutindo estratégias baseadas em experiências que deram certo em diferentes partes do mundo. Em outubro de 2019 a instituição organizou a Conferência Nacional de Alfabetização Baseada em Evidências (Conabe). O evento desenvolveu dez eixos temáticos, sendo que a escrita estava envolvida diretamente em cinco dessas temáticas (MEC, 2019, s.p.).

Em resumo, a conferência Conabe tratou de assuntos relacionados ao momento de introdução de tecnologias no processo de leitura e aprendizagem, quais práticas de leitura e escrita a família pode aplicar em casa para complementar o que é abordado em sala de aula, dificuldades e distúrbios da leitura e da escrita e desafios na alfabetização em diferentes contextos (MEC, 2019, s.p.).

2.2.1 Obstáculos enfrentados na produção textual

Segundo Conceição (2002), a escrita deve ser uma competência adquirida nas bases da jornada escolar do aluno, ou seja, em anos anteriores ao ensino médio, porém alunos têm chegado ao ensino superior ainda sem saber redigir textos com autonomia.

Em sua pesquisa realizada por meio de observações e entrevistas com diferentes professores, a autora detectou grande dificuldade desses profissionais em ensinar a produção textual e, principalmente, em avaliar os produtos oriundos desta, em consequência dessas dificuldades, o aluno tem criado uma espécie de temor pela escrita da redação (CONCEIÇÃO, 2002, pg. 45).

Matavelli (2011) entrevistou em sua pesquisa alguns professores de Língua Portuguesa a respeito de redações. Esta modalidade representa nota decisiva para os maiores vestibulares e a tendência é de aumento nesta cobrança frente aos candidatos. Os professores pesquisados relataram que os maiores problemas voltam-se à falta de argumentos, além de erros ortográficos e gramaticais incorridos pelos alunos. Na mesma pesquisa os docentes ressaltam que programas como o MSWord, que fazem autocorreções em textos, acabam por deixar os alunos mal-acostumados. Também foi relatado o impacto causado nos alunos pelo mal uso das redes sociais, que fazem com que se distanciem da escrita formal em prol da linguagem coloquial adotada nessas redes. Os docentes informam que para obter resultados satisfatórios nas redações aplicadas aos alunos é necessário aumentar a rotina de leitura e escrita. Assim, produzir textos formais de forma semanal aumentaria a eficácia nas notas dos estudantes nas redações entregues.

Riolfi e Igreja (2010) realizaram estudo para compreender as dificuldades enfrentadas por alunos de Ensino Médio que prestaram a avaliação do processo seletivo da Universidade de São Paulo – Fundação Universitária para o Vestibular (FUVEST) em 2008.

Os autores ainda levantaram que foram 11.242 candidatos e destes, aqueles que estavam entre os aprovados estudaram exclusivamente em colégios particulares com uma margem de 70,9%, sendo 20,3% de alunos oriundos de escola pública (Riolfi; Igreja, 2010, pg. 315). Para entender o motivo da dificuldade de escrita de textos desses alunos os autores analisaram 2.434 horas de aulas de língua portuguesa em escolas públicas do estado de São Paulo, como resultado, os autores detectaram que os docentes dedicavam apenas 15% deste tempo para o ensino da escrita e, mesmo assim, não eram aplicados textos dissertativos, uma vez que para esta modalidade foram dedicados apenas 6% do tempo total. (Riolfi; Igreja, 2010, pg. 321)

Na análise descrita por Riolfi e Igreja (2010) foi identificado que em alguns casos, após a correção dos textos dos alunos, o professor comentava oralmente as redações, ignorando outros problemas textuais. Alguns itens foram apontados na pesquisa dos autores, tais como: desconhecimento das características estruturais do texto dissertativo, desconhecimento dos pré-

requisitos para a articulação lógica dos segmentos e precariedade de recursos para a construção da tese ou sua sustentação (Riolfi; Igreja, 2010, pg. 316). Assim, a percepção é que o processo avaliativo de textos produzidos pelos alunos muitas vezes não ocorre individualmente, sendo que a resolução acontece por meio de comentários orais a respeito das redações para toda a turma, como exemplificado anteriormente, esta forma de correção e devolutiva ignora os problemas de cada aluno, pois no momento da fala do professor à turma pode acontecer de o estudante não se identificar com a exposição geral. (Riolfi; Igreja, 2010, pg. 318).

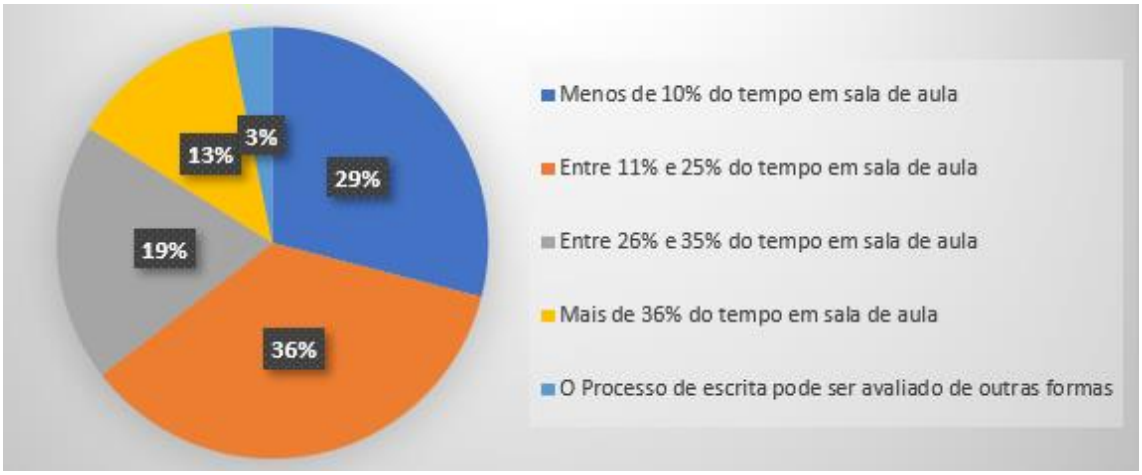
Assim sendo, as dificuldades individuais específicas na escrita de cada aluno permanecem não solucionadas apropriadamente, o que reflete nos resultados inferiores já demonstrados nas diferentes provas avaliativas para ingresso nas universidades. (Riolfi; Igreja, 2010, pg. 318).

Striquer (2018) estabeleceu análise de um conjunto de textos que deveria seguir os gêneros discursivo/textual baseados nas competências do ENEM, as redações foram produzidas por 110 alunos do último ano do Ensino Médio de quatro escolas da rede pública, de duas diferentes cidades da região norte do Paraná. As redações foram avaliadas e apenas 13% destas apresentaram estrutura adequada às características de gênero/textual exigidas. As demais redações não atenderam à proposta e apresentaram erros graves na compreensão e atendimento das cinco competências cobradas na prova de redação do ENEM. (Striquer, 2018, pg. 73.)

Outra pesquisa realizada por Pinho *et al*.(2020) junto a 30 professores de língua portuguesa de escolas públicas e privadas do estado de São Paulo procurou entender o cenário atual da aplicação de redações aos alunos. Dentre os resultados foi possível identificar como os professores recebem atualmente as redações entregues pelos alunos. As devolutivas acontecem de formas variadas, tais como o envio de fotos do caderno, arquivos de texto ou formulários eletrônicos. Assim, os docentes informaram suas dificuldades ao enviar o *feedback* individual a cada aluno. A maior incidência de reclamações está no recebimento de textos por meio de fotos (70,4%), a partir do qual os professores informam que a leitura é de difícil entendimento, notadamente por conta da letra ruim ou da má qualidade da imagem. Outro problema para um

feedback individual com maior qualidade volta-se ao número de plataformas diferentes por meio das quais os professores recebem as atividades, com 46,7% das respostas dos professores entrevistados, ocasionando assim elevada perda de tempo no processo de correção dos textos produzidos pelos alunos (PINHO *et al*., 2020).

Figura 14: Tempo em sala de aula disponibilizado para o ensino da produção textual



Fonte: (Pinho et. al, 2020)

Outro questionamento realizado na pesquisa de Pinho *et al.* (2020) foi sobre o tempo despendido pelos professores para o ensino de produção textual em sala de aula. O objetivo era confirmar os resultados de estudos já realizados anteriormente por outros autores. Os resultados podem ser visualizados na figura 14.

Ao analisar as respostas obtidas junto aos professores foi possível confirmar que mais de 60% dos docentes usam menos de 25% de seu tempo para o ensino de produção textual. Dentre os motivos alegados estão o excesso de turmas e alunos, além da carência em fornecer *feedback* detalhado e criterioso a cada aluno. Apenas 13% dos docentes informaram utilizar mais de 36% de seu tempo para o ensino da produção textual aos alunos.

Todos os obstáculos relatados anteriormente expõem as dificuldades enfrentadas pelos professores quanto ao ensino da escrita e da produção textual de redação. Tais problemas são importantes requisitos a serem considerados para a elaboração da solução inteligente automatizada vislumbrada neste estudo.

2.2.2 Avaliação de redações para ingresso no ensino superior

Em 2020, só no estado de São Paulo, 26 universidades particulares realizaram seus processos seletivos totalmente online, o instrumento de avaliação mais utilizado nesses casos foi a redação (MORALES, 2020, s.p.). Neste mesmo ano, a Pontifícia Universidade Católica de São Paulo (PUC-SP) decidiu cancelar as provas de inverno devido às contingências sanitárias implantadas por consequência da pandemia do Covid-19. Assim, a instituição considerou as notas do Enem dos anos de 2018 e 2019. Neste caso, estudantes com nota inferior a 500 pontos na redação do Enem foram desclassificados (CRUZEIRO DO SUL, 2020, s.p.).

O Exame Nacional do Ensino Médio (ENEM) foi criado em 1998 e tem o objetivo de avaliar o desempenho do estudante ao fim da escolaridade básica (ensino fundamental e ensino médio) (MEC, 2020, s.p.). Podem participar do exame alunos que estão concluindo ou que já concluíram o ensino médio em anos anteriores à edição em voga (MEC, 2020, s.p.). Para a maioria das universidades públicas o principal instrumento de avaliação aplicado para a aprovação de candidatos é o exame do ENEM, este exame é composto por cinco áreas: linguagens, códigos e suas tecnologias; ciências humanas e suas tecnologias; matemática e suas tecnologias; ciências da natureza e suas tecnologias e a redação, a redação é o único item discursivo do exame, objeto enfocado nesta pesquisa (BRASIL, 2020, s.p.).

Uma novidade divulgada em 2019 é que o ENEM já teve a aplicação digital em 2020, ainda que tenha sido realizada em modelo-piloto aplicado a um contingente de 100 mil estudantes. Nesta primeira aplicação digital, a redação ainda foi realizada de forma manual. Porém, a implantação do Enem Digital será progressiva, conforme previsão de consolidação desta modalidade a todos os candidatos na edição de 2026 (BRASIL, 2020, s.p).

Conforme as cinco competências expostas no quadro 8, a maior preocupação dos candidatos ao realizarem o ENEM é que lhes seja atribuída nota zero na redação, para que não seja atribuída a nota zero, a redação não poderá conter os seguintes erros (BRASIL, 2020, s.p):

1) Fuga total ao tema;

2) Não obediência à estrutura dissertativo-argumentativa;

3) Extensão total de até 7 linhas;

4) Cópia integral de texto(s) da Prova de Redação e/ou do Caderno de Questões;

5) Impropérios, desenhos e outras formas propositais de anulação, em qualquer parte da folha de redação; números ou sinais gráficos fora do texto e sem função clara;

6) Parte deliberadamente desconectada do tema proposto;

7) Assinatura, nome, apelido, codinome ou rubrica fora do local devidamente designado para a assinatura do participante;

8) Texto predominante ou integralmente em língua estrangeira;

9) Folha de redação em branco, mesmo que haja texto escrito na folha de rascunho. (BRASIL, 2020, s.p)

Quadro 8: Cinco competências avaliadas na redação do ENEM

Competência 1:	Demonstrar domínio da modalidade escrita formal da língua portuguesa.
Competência 2:	Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
Competência 3:	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
Competência 4:	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
Competência 5:	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Fonte: Brasil (2019).

De forma geral, os professores avaliam o desempenho dos alunos seguindo as cinco competências demostradas no Quadro 8. A fuga ao tema, demonstrada no primeiro item dos erros incorridos pelos candidatos, está inserida na competência 2 (BRASIL, 2019). A forma como a competência 2 é avaliada será descrita posteriormente.

Em 2020 os manuais de correção de redação, que até 2019 eram sigilosos; foram disponibilizados para consulta no portal do INEP (BRASIL,

2019, s.p). Com base neste manual é possível entender como é atribuída a nota à redação, cada avaliador atribuiu uma nota entre 0 e 200 pontos para cada uma das cinco competências de avaliação, a soma desses pontos comporá a nota total dada por cada avaliador, que pode chegar a 1.000 pontos, a nota final do participante será a média aritmética das notas totais atribuídas pelos dois avaliadores da redação do candidato (BRASIL, 2019, s.p.).

A fuga ao tema, avaliado na competência 2 do ENEM é um problema comum em processos avaliativos. Segundo Campos (2020), na prova do ENEM de 2019, de todas as 143 mil redações zeradas, 28,4% foram por fuga ao tema.

2.2.2.1 Avaliação da competência 2 da redação do ENEM

Todo o conteúdo demonstrado neste tópico foi extraído do material de leitura disponibilizado no processo de treinamento para avaliadores do ENEM, disponível para consulta no portal do INEP (BRASIL, 2019, s.p.). Como foco desta dissertação, a competência 2 avalia como o participante se apropria da proposta de redação, aplicando conceitos de várias áreas de conhecimento para desenvolver o tema de forma plena e consistente (BRASIL, 2019, s.p.).

A competência 2 versa ainda sobre a forma do texto dissertativo-argumentativo, devendo o candidato demonstrar conhecimento sobre os limites estruturais da tipologia textual em prosa, o INEP criou a matriz de referência da competência 2 da avaliação de redações, que pode ser visualizada no quadro 9 (BRASIL, 2019, s.p.).

A pontuação demonstrada no quadro 9 vai de 0 a 200 pontos para cada dimensão considerada, sendo que o aluno recebe zero ponto quando a redação foge ao tema, ou seja, quando nem o assunto mais amplo, nem o tema proposto foram desenvolvidos.

Quadro 9: Matriz de referência da competência 2

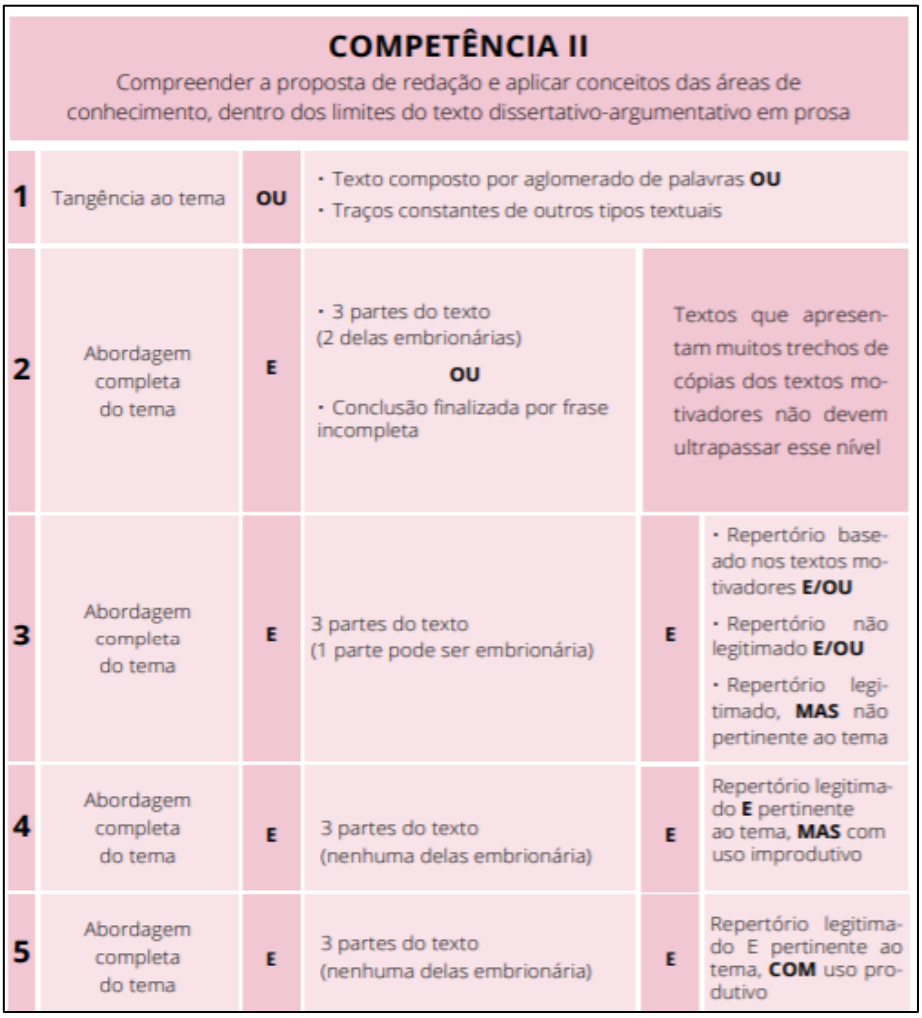
200 pontos	Desenvolve o tema por meio de argumentação consistente, a partir de um repertório sociocultural produtivo e apresenta excelente domínio do texto dissertativo-argumentativo.
160 pontos	Desenvolve o tema por meio de argumentação consistente e apresenta bom domínio do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
120 pontos	Desenvolve o tema por meio de argumentação previsível e apresenta domínio mediano do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
80 pontos	Desenvolve o tema recorrendo à cópia de trechos dos textos motivadores ou apresenta domínio insuficiente do texto dissertativo-argumentativo, não atendendo à estrutura com proposição, argumentação e conclusão.
40 pontos	Apresenta o assunto, tangenciando o tema, ou demonstra domínio precário do texto dissertativo-argumentativo, com traços constantes de outros tipos textuais.
0 ponto	Fuga ao tema/não atendimento à estrutura dissertativo-argumentativa. Nestes casos a redação recebe nota zero e é anulada.

Fonte: Brasil (2019).

Além de disponibilizar a grade específica da competência 2, o INEP ainda exemplifica formas de avaliar para o corretor, como no caso do tema de 2018 (“Manipulação do comportamento do usuário pelo controle de dados na internet”), espera-se que o candidato compreenda e aborde o tema de forma completa (BRASIL, 2019, s.p.). Ou seja, o candidato deve apresentar o controle de dados na internet, além da manipulação do comportamento e/ou suas consequências, bem como os efeitos e os exemplos do ato de manipular o usuário da internet. (BRASIL, 2019, s.p.).

O avaliador do Enem recebe uma grade específica para avaliação da competência 2, conforme exemplificada na Figura 15. A função desta grade específica da competência 2 é tornar objetivo o processo de correção do avaliador da redação (BRASIL, 2019, s.p.). Assim, a referida grade apresenta, de maneira detalhada, todos os elementos que precisam ser identificados para a classificação das redações em cada um dos níveis da competência 2 (BRASIL, 2019, s.p.).

Figura 15: Grade específica para avaliação da competência 2



COMPETÊNCIA II

Compreender a proposta de redação e aplicar conceitos das áreas de conhecimento, dentro dos limites do texto dissertativo-argumentativo em prosa

1	Tangência ao tema	**OU**	• Texto composto por aglomerado de palavras **OU** • Traços constantes de outros tipos textuais		
2	Abordagem completa do tema	**E**	• 3 partes do texto (2 delas embrionárias) **OU** • Conclusão finalizada por frase incompleta	Textos que apresentam muitos trechos de cópias dos textos motivadores não devem ultrapassar esse nível	
3	Abordagem completa do tema	**E**	3 partes do texto (1 parte pode ser embrionária)	**E**	• Repertório baseado nos textos motivadores **E/OU** • Repertório não legitimado **E/OU** • Repertório legitimado, **MAS** não pertinente ao tema
4	Abordagem completa do tema	**E**	3 partes do texto (nenhuma delas embrionária)	**E**	Repertório legitimado **E** pertinente ao tema, **MAS** com uso improdutivo
5	Abordagem completa do tema	**E**	3 partes do texto (nenhuma delas embrionária)	**E**	Repertório legitimado E pertinente ao tema, **COM** uso produtivo

Fonte: Brasil (2019).

Para facilitar este trabalho, o INEP informa em que casos os estudantes não pontuam, usando como exemplo a temática “Manipulação do comportamento do usuário pelo controle de dados na internet”, a instituição define: “textos que abordem exclusivamente tecnologia, mídia ou outros assuntos, sem sequer mencionar internet ou qualquer elemento do universo da internet, não terão abordado sequer o assunto mais geral proposto para a redação e deverão, por consequência, ser avaliados como ‘fuga ao tema’ com a atribuição de nota zero nesta competência”.

Assim como o ENEM fornece dicas aos avaliadores para a identificação que o aluno fugiu ao tema indicado, este processo pode também ser disponibilizado para uma solução inteligente automatizada para correção de redações quanto à fuga ao tema proposto.

Assim, o papel do professor em sala de aula pode ser intermediado por meio de tecnologias que facilitem esse acompanhamento, permitindo ao professor dedicar maior tempo e esforço ao ensino da redação, bem como proporcionar feedback individualizado a cada aluno. A cada dia a sociedade se torna mais digital e as cobranças em relação à escrita tem aumentado, além de se configurar num diferencial do candidato para a entrada em cursos superiores oferecidos pelas universidades.

2.2.3 Pesquisa de Soluções Inteligentes na área da Educação

A tecnologia inserida no contexto educação experimentou grande evolução a partir da década de 1990, quando os microcomputadores permitiram a geração de textos eletrônicos. Estudo realizado por Araújo (2011) discutiu os movimentos na evolução na educação no último século e a incorporação de Tecnologias de Informação e Comunicação para entender se as tecnologias podem ajudar a promover maior qualidade e êxito na educação.

Algumas soluções inteligentes e outros mecanismos computacionais têm sido pesquisados em prol de auxiliar professores no processo de correção e identificação de problemas no aprendizado da produção textual.

Outros experimentos sobre a análise automática de coesão textual em redações foram realizados por Nobre e Pellegrino (2010). Em seus estudos, os autores identificaram de forma automática problemas de coesão em 90% dos textos argumentativos e dissertativos analisados no experimento conduzido. Os resultados da solução automatizada aplicada no experimento foram compatíveis às notas atribuídas em correções feitas por avaliadores humanos.

Esses autores afirmam ainda que a correção realizada por um programa de computador não sofre interferências externas, tais como fadiga e alteração de humor, permitindo assim avaliar e analisar sempre de forma equânime. Entretanto, percebe-se a necessidade de revisão das expressões regulares visando detectar problemas não identificados pela solução computadorizada. Assim, o processo automatizado diminui a carga de trabalho do avaliador humano e se mostra uma ferramenta para apoio ao processo de correção executado por avaliadores humanos.

Santos (2017) desenvolveu sua pesquisa para melhorar a qualidade em avaliação automática de textos dissertativos utilizando Processamento de Linguagem Natural (PLN) e redes neurais. Em seu experimento o autor procurou tratar dos desvios das redações de forma genérica, sem avaliar especificamente cada competência, sendo que a rede neural aplicada deveria acertar a pontuação de 0 a 1000. Para tanto, foram avaliados 18 temas de redações, com a indicação dos resultados de cada temática tendo sigo gerada de forma separada. O melhor resultado alcançado neste experimento atribuiu notas para as redações com uma taxa de erro de 100 pontos.

Cândido e Webber (2018) entendem que é um desafio tratar a coesão de um texto de forma automática. Não obstante, em sua pesquisa os autores descrevem as possibilidades de se tratar com assertividade a coerência e coesão de redações com uso de ferramentas de PLN. O estudo realizado por eles utiliza os elementos linguísticos e técnicas computacionais para realizar a avaliação da redação. Os experimentos por eles realizados compararam a análise executada por um software e as avaliações feitas por dois especialistas humanos. Foram encontrados resultados convergentes em 70% dos casos analisados no experimento. Considera-se que tais resultados iniciais são promissores para o desenvolvimento de solução para a avaliação automática de redações, abrindo-se então novas possibilidades de pesquisa.

Passero (2018) propôs um projeto especificamente para detecção de fuga ao tema nas redações utilizando técnicas de PLN e AM. O autor implementou modelos de detecção de fuga ao tema considerando-se as técnicas de análise textual, empregando para tanto a similaridade semântica textual e algumas técnicas como regressão linear e máquinas de vetores de suporte. Em seus experimentos foram utilizados 2.151 casos de redações sem fuga ao tema, além de doze exemplos de redações com fuga ao tema. Os melhores resultados desse experimento foram obtidos utilizando-se o algoritmo KFF-A, com acurácia média entre 81,13% e 96,76%. O autor relata que seu algoritmo ainda apresenta uma taxa de falsos positivos elevada (4,24%), aquela que detecta que a redação teve uma fuga ao tema, quando na verdade não teve. Neste caso, a presença do avaliador humano ainda seria indispensável.

Ramisch (2020) investigou especificamente a recorrência de desvios de natureza sintática nas redações e as eventuais correlações com determinados atributos linguísticos das sentenças elaboradas. Contudo, em sua pesquisa foram eliminadas as redações anuladas ou com fuga ao tema. Na pesquisa foram utilizados os softwares Parser e UDPipe para a extração dos atributos linguísticos e, posteriormente, o software WEKA para a aplicação do aprendizado de máquina. Os melhores resultados obtidos pelo corpus de teste foi com o algoritmo Regressão Logística, que alcançou 75,62% de acerto.

Bittencourt Júnior (2020) propõe em seu estudo a avaliação automática de redações e utiliza em seus experimentos as redes neurais profundas. Para o processo de aprendizado foi utilizado um conjunto de redações com 18 temas diferentes. O estudo procurou avaliar as cinco competências estipuladas pelo ENEM. Como resultados, aponta-se a proposição de uma nova arquitetura Multitema, com base na hipótese de que as características aprendidas pela rede aplicada para a correção de determinado tema poderiam ajudar a aprimorar o desempenho de outros temas, mensurando assim os resultados obtidos a cada avaliação automatizada.

2.3 Inteligência Artificial (IA)

A inteligência Artificial refere-se a um campo de conhecimento associado à linguagem, inteligência, raciocínio, aprendizagem e resolução de problemas. Os estágios de desenvolvimento da IA, bem como as expectativas de resultados com a sua aplicação variam entre os diferentes campos e suas aplicações (KAUFMAN, 2018).

Um sistema com IA cobre vários conceitos e processos, ele deve ser capaz de reconhecer padrões de comportamento, tendências e gerar suposições futuras baseadas nos dados analisados, entende-se que este processo é denominado aprendizado de máquina Capgemini (2017).

Androutsopoulou (2018) relata em sua pesquisa que a IA já está em diversos domínios, como: saúde, tributação e educação. Segundo Müller (2018), a aplicação de IA na educação tem sido amplamente discutida, embora ela atenda a um número limitado de cenários de aprendizagem, já que as máquinas inteligentes operam nos limites de seu sistema.

Assim, sistemas inteligentes aplicados à educação devem fornecer suporte para os professores e melhorar o seu trabalho. Em complemento, Müller (2018) argumenta que as pesquisas de IA na educação são promissoras, ao passo que as máquinas vão se ajustando às necessidades individuais de cada profissional.

A Inteligência Artificial tem sido amplamente utilizadas para apoiar e melhorar a qualidade da tomada de decisões e solução de problemas, ela se utiliza de diferentes técnicas para fornecer informações baseadas em grandes volumes de dados, dentre elas a Mineração de Textos (MT), o Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina, este engloba as técnicas inteligentes de classificação, entre elas as redes neurais artificiais. Estes conceitos serão tratados nos próximos tópicos. (EGGERS WILLIAM et al., 2017; HARIRI, 2019).

2.3.1 Mineração de Textos

A Mineração de Texto, também conhecida por Text Mining, Text Processing ou ainda Text Analytics, é um processo semiautomatizado para extração de conhecimento de fontes de dados não-estruturados (CAFFARO FILHO, 2020, s.p.). Aproximadamente 80% a 90% de todos os dados corporativos apresentam-se em algum tipo de formato não-estruturado, a exemplo de textos, além disso, estima-se que o volume de dados corporativos não-estruturados dobra de tamanho a cada 18 meses (CAFFARO FILHO, 2020, s.p.).

Morais e Ambrósio (2007) definem mineração de textos como um processo de descoberta de conhecimento que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras.

A Mineração de Textos envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma direta (SOUZA 2019; GONÇALVES, 2012).

A mineração de textos busca extrair padrões interessantes e não-triviais de conhecimento a partir de textos. Os sistemas de mineração de textos baseiam-se em rotinas de pré-processamento, algoritmos para descoberta de padrões e elementos para apresentação dos resultados. A mineração de textos é um processo que utiliza algoritmos capazes de analisar coleções de documentos em forma de texto - tais como arquivos PDF, páginas Web e documentos XML (SOUZA 2019; GONÇALVES, 2012).

Considerando a natureza dos dados e as aplicações as quais está frequentemente relacionado, o processo de descoberta de conhecimento em dados textuais, denominado Mineração de Textos (MT), muitas vezes combina técnicas de Recuperação de Informação (RI), Aprendizado de Máquina (AM) e Processamento de Linguagem Natural (PLN), ao longo de suas etapas (MARTINS et al., 2003).

Para esses autores, o processo de Mineração de Textos é semelhante ao processo de Mineração de Dados (MD). Porém, enquanto MD trabalha com dados estruturados, o processo de MT trabalha com dados não estruturados, geralmente na forma de textos ou documentos, havendo, portanto, um tratamento diferenciado em algumas etapas do processo (MARTINS et al., 2003).

Segundo Caffaro Filho (2020), a diferença está na natureza dos dados analisados: enquanto os dados estruturados encontram-se em tabelas de bancos de dados, os dados não-estruturados apresentam-se em forma de documentos de Word, arquivos PDF, fragmentos de texto, arquivos XML etc.

Portanto, a mineração de textos é uma extensão da mineração de dados, e pode ser definida como um processo de extração de informações desconhecidas e úteis de documentos textuais escritos em linguagem natural, como a maioria das informações são armazenadas em forma de texto, a mineração de textos possui alto valor comercial, podendo ser aplicada em diferentes áreas (PEZZINI, 2016). O autor entende que algumas técnicas são essenciais no processo de Mineração de Texto, quais sejam:

1) Processamento de Linguagem Natural: para melhorar o entendimento da linguagem natural através de técnicas para processar textos rapidamente;

2) Recuperação de Informação: utiliza métodos e medidas estatísticos ou semânticos para automaticamente processar o texto de documentos para encontrar quais documentos possuem a resposta para a questão (mas não a resposta em si); e

3) Extração de Informação: possui como principal objetivo buscar partes relevantes de um texto em um documento e extrair informações específicas destas partes. Possui um conceito mais limitado da compreensão da linguagem natural.

Aranha e Passos (2007) definem algumas etapas essenciais do processo de Mineração de Textos:

1. Coleta de documentos: estes podem vir de diferentes fontes, tais como livros e documentos, que podem ser obtidos na internet. Para facilitar o acesso a esses documentos, várias ferramentas de apoio têm sido desenvolvidas utilizando técnicas de PLN e AM. (Aranha e Passos, 2007, pag.4)

2. Pré-processamento: os documentos são preparados para serem representados em um formato adequado para serem submetidos aos algoritmos de extração automática de conhecimento. Ela é responsável por obter uma estrutura, geralmente, no formato de uma tabela atributo-valor, que represente o conjunto de documentos (Aranha e Passos, 2007, pag.4);

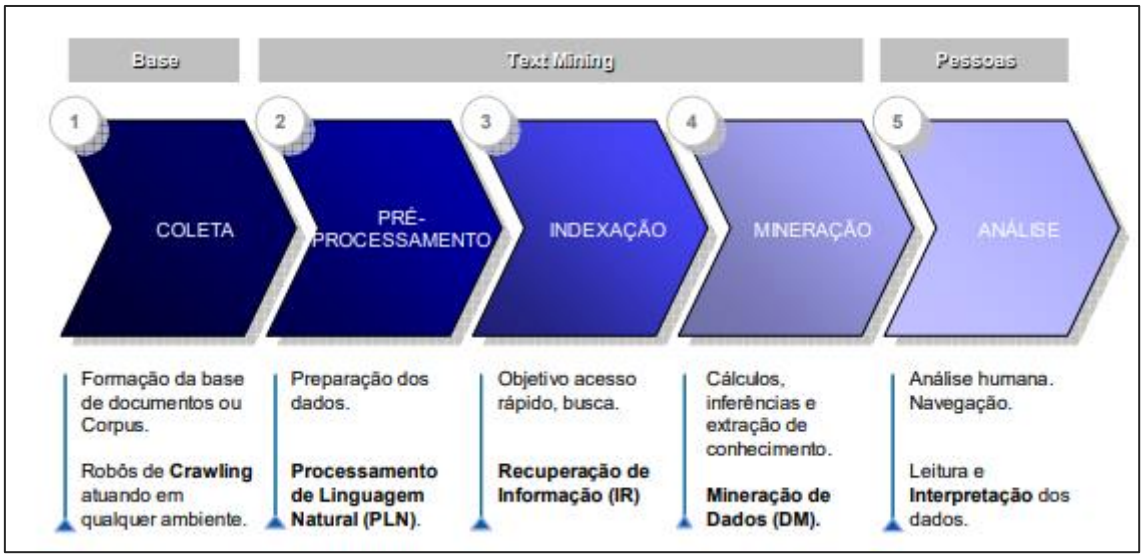
3. Indexação: é o processo que organiza todos os termos adquiridos a partir de fontes de dados, facilitando o seu acesso e recuperação. Uma boa estrutura de índices garante rapidez e agilidade ao processo, tal como funciona o índice de um livro (Aranha e Passos, 2007, pag.4);

4. Mineração: responsável pelo desenvolvimento de cálculos, inferências e algoritmos e que tem como objetivo a extração de conhecimento, descoberta de padrões e comportamentos que possam surpreender (Aranha e Passos, 2007, pag.4); e

5. Análise: é a última etapa, deve ser executada por pessoas que normalmente estão interessadas no conhecimento extraído e que devem tomar algum tipo de decisão apoiada no processo de Mineração de Texto (Aranha e Passos, 2007, pag.4).

A Figura 16 expõe as cinco etapas da metodologia do processo de Mineração de Textos proposta por Aranha e Passos (2007).

Figura 16: Metodologia de Mineração de Textos proposta por Aranha e Passos

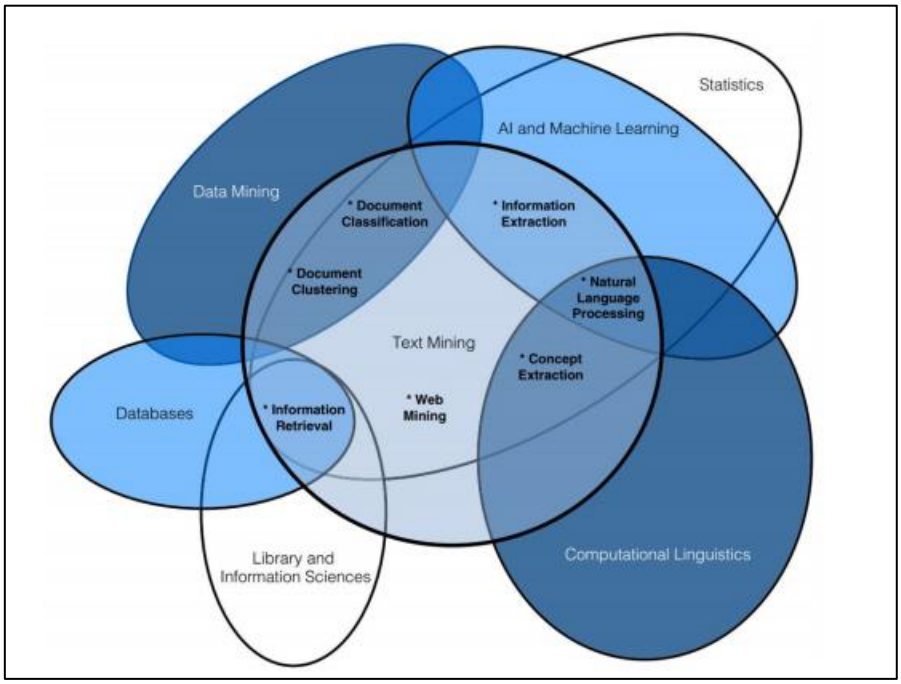


Fonte: Aranha e Passos (2007)

Segundo Aranha e Passos (2007), a metodologia é composta por cinco etapas, que são explicadas em três fases: coleta dos dados; mineração de texto, que incluem o pré-processamento dos dados, e por fim; indexação e Mineração de Dados. Por último vem a fase da avaliação humana para ler e interpretar os dados retirados do corpus e extrair o conhecimento adquirido.

A Figura 17 define a mineração de texto com mais precisão, exemplificando os seis campos que estão conectados, dentre os quais PLN e Extração de Informações, que estão relacionadas à Inteligência Artificial (IA) e Aprendizado de Máquina (AM). Os demais campos também estão relacionados à *Data Mining, Databases e* Linguística (MORAIS, 2021 *apud* Miner, 2012).

Figura 17: Relação Mineração de Textos com PLN e AM



Fonte: Morais (2021 apud Miner, 2012).

Dentre os tópicos citados na Figura 17 alguns serão tratados no referencial teórico desta dissertação, tais como PLN e Técnicas de IA para classificar documentos e extrair informações de uma base de dados.

2.3.2 Processamento de Linguagem Natural

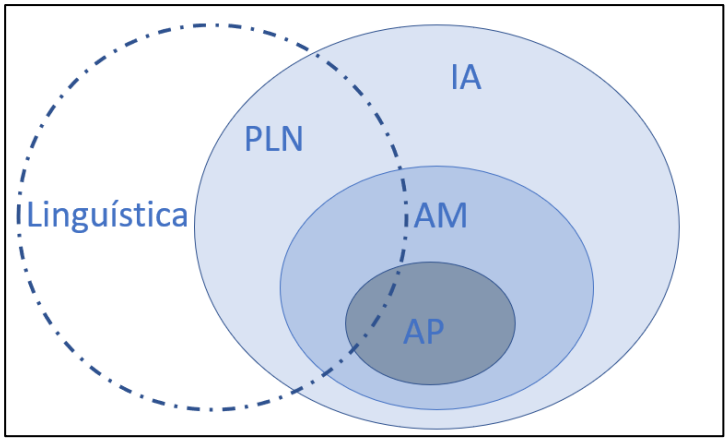
O Processamento de Linguagem Natural (PLN) é uma subárea da Inteligência Artificial que estuda a comunicação humana por métodos computacionais. Assim, busca-se converter a linguagem natural humana em uma representação formal, de forma que se torne mais facilmente manipulável por máquinas. Muitas aplicações de PLN são baseadas em modelos de linguagem que definam uma distribuição de probabilidade sobre sequências de palavras, caracteres ou bytes em uma linguagem natural (GOODFELLOW, 2016; CONEGLIAN, 2020).

O PLN é uma ferramenta para que os computadores analisem, compreendam e extraiam significado da linguagem natural de uma forma inteligente e útil. O PLN combina Inteligência Artificial (IA) e linguística computacional para que computadores e humanos possam conversar apropriadamente. A PLN capacita programas de computador a compreenderem

conteúdo não estruturado, utilizando IA e aprendizado de máquina para fazer derivações e dar contexto à linguagem, de forma similar ao cérebro humano (BANERJEE ,2020).

Barnerjee (2020) mostra de forma hierárquica que a PLN e o Aprendizado de Máquina (AM) se enquadram na categoria mais ampla de Inteligência Artificial. Ou seja, o estudo da Linguística está inserido na PLN e pode contemplar ainda outras três áreas: IA, AM e AP. As arquiteturas e algoritmos de Aprendizado Profundo (AP) fizeram importantes avanços na PLN como o reconhecimento de entidade nomeada, marcação de classes gramaticais, processamento de voz, tradução e classificação de textos. A hierarquia indicada por Barnerjee (2020) pode ser visualizada na Figura 18.

Figura 18: Hierarquia da PLN dentro da IA



Fonte: Adaptado de Barnerjee (2020).

A PLN busca padrões e indicativos que auxiliem na compreensão do texto em análise. Assim, os estudos de PLN e AM convergem cada vez mais devido à grande quantidade de dados que é gerada diariamente, sendo por meio desses dados que o computador aprende (GOODFELLOW, 2016; CONEGLIAN, 2020).

A PLN facilita a interação entre humanos e máquinas com o uso de linguagem natural. Por linguagem natural entende-se as palavras ou textos que são utilizados no dia a dia para a comunicação entre indivíduos. O grande desafio do PLN é transformar textos e falas de pessoas em conjuntos de dados capazes de serem lidos para o desenvolvimento de análises e aplicação de algoritmos de aprendizado de máquina (PRATES, 2019).

Mesmo com o avanço no relacionamento homem-máquina, a comunicação via linguagem natural continua sendo um desafio, como a questão de criar programas capazes de interpretar mensagens codificadas em linguagem natural e decifrá-las para a linguagem de máquina, com o passar dos anos houve muitas pesquisas e desenvolvimentos nos mais diversos ramos do processamento de linguagem natural (RODRIGUES, 2017).

A PLN tem experimentado grandes transformações nos últimos anos, dentre as quais destaca-se a geração de massa de dados não estruturados, especialmente no formato de texto, que tem proporcionado o surgimento de diferentes áreas de atuação. O Quadro 10 exemplifica as principais áreas de atuação de PLN existentes, conforme indicado por Prates (2019) e Stefanini (2019). O Quadro 9 exibe a gama de possibilidades para aplicação de técnicas de PLN. Contudo, os textos, imagens ou áudios precisam ser tratados para a aplicação em cada área indicada.

Quadro 10: Áreas de Atuação da PLN

Áreas	Descrição
Sistemas de respostas a perguntas de usuários	Chatbots que permitem que o usuário digite o texto de forma aberta (Prates; Stefanini, 2019).
Traduções feitas por máquinas	Softwares e aplicativos que fazem tradução instantânea utilizando diferentes línguas (Prates; Stefanini, 2019).
Reconhecimento de voz e diálogos	As aplicações de diálogo buscam produzir diálogos entre um ser humano e a máquina de forma fluida e coerente. O exemplo mais comum são aplicativos de GPS, assistentes de busca ou tutores inteligentes (Prates; Stefanini, 2019);
Classificação de documentos	É possível categorizar documentos quando não se conhece as classes possíveis via algoritmos de aprendizado de máquina, sejam eles não supervisionado ou supervisionado (Prates; Stefanini, 2019).
Reconhecimento de textos em imagens	Um exemplo é como identificar números de placas de veículos e enviar multas de excesso de velocidade de forma automática, por exemplo (Prates; Stefanini, 2019);
Análises de sentimento em textos	Método muito comum em pesquisas que contém campos abertos, como pesquisas de clima organizacional ou análise de sentimentos em redes sociais (Prates; Stefanini, 2019);.
PLN nas buscas do dia a dia	A área de recuperação das informações corresponde principalmente aos mecanismos de busca na WEB (Prates; Stefanini, 2019);

Fonte: Adaptado de Prates (2019) e Stefanini (2019).

2.3.2.1 Tratamentos de textos para análises

Dado que textos são a matéria prima para o PLN e sabendo que é preciso entender alguns conceitos para recuperar informações nos textos, serão demonstrados neste tópico as tarefas que são executadas antes que o conhecimento seja extraído de uma base de dados em forma de texto (EVANGELISTA et. al, 2020; RODRIGUES, 2017; PRATES, 2019).

Estas tarefas são necessárias, pois abstraem e estruturam a língua portuguesa escrita, deixando apenas o que seja informação relevante, esse pré-processamento reduz o vocabulário e torna os dados menos esparsos, característica imprescindível ao processamento computacional, para esses autores, os itens de maior importância no tratamento de textos para análise são comentados na sequência. (EVANGELISTA et. al, 2020; RODRIGUES, 2017; PRATES, 2019)

Corpus

Corpus é um conjunto de documentos, cada documento é um pedaço de texto, independentemente do tamanho, uma única frase ou um texto completo pode representar um documento (PRATES, 2019, s.p.). Na prática um documento pode ser um comentário em uma rede social, uma ata de reunião ou artigo de blog, por exemplo (PRATES, 2019, s.p.).

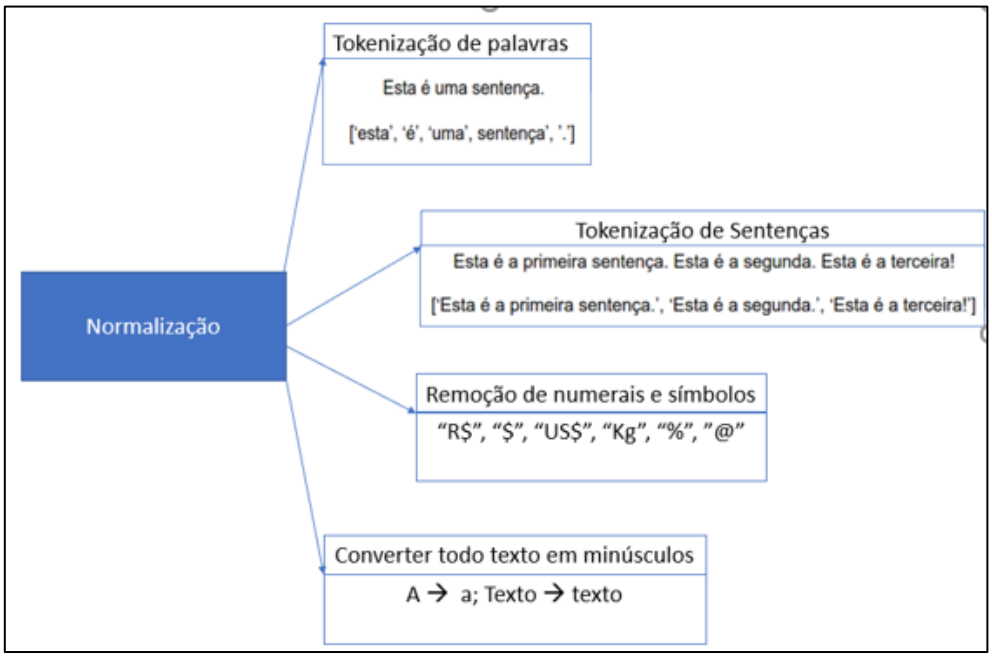
Normalização

A normalização abrange tratativas como a tokenização, ou seja, a transformação de letras maiúsculas para minúsculas, remoção de caracteres especiais, remoção de tags HTML/Javascript/CSS, dentre outras ações, caso este processo não seja aplicado, os algoritmos podem tratar palavras iguais como sendo diferentes apenas por apresentarem a letra inicial como maiúscula, ou um erro de acentuação, por exemplo (RODRIGUES, 2017; PRATES, 2019).

Segundo Evangelista et al. (2020), a tokenização, também chamada de ‘segmentação de palavras’, é responsável por quebrar uma determinada sequência de caracteres de um texto, ou seja, ela determina onde as palavras de um texto iniciam e terminam, transformando cada palavra em um token. Os tokens são listas geradas a partir de um corpus tokenizado.

Na Figura 19 estão demonstrados alguns dos processos que levam a entender por que a ação de normalização é importante, pois executá-la permite a estruturação do texto, já que os processamentos seguintes atuam em cima de unidades sentenciais e lexicais.

Figura 19: Normalização dos dados



Fonte: Adaptado de Rodrigues (2017) e Prates (2019)

Remoção de stop words

Uma das tarefas muito utilizadas no pré-processamento de textos é a remoção de stop words. Esse método consiste em remover palavras muito frequentes no texto, tais como "a", "de", "o", "da", "que", "e" e "do", dentre outras, pois na maioria das vezes não são informações relevantes para a extração de conhecimento (EVANGELISTA *et al*., 2020; RODRIGUES, 2017; PRATES, 2019).

Stop words são palavras irrelevantes consideradas ruído no texto. Essas palavras ocupam espaço desnecessário nos textos, além de tomar um tempo valioso no processamento dos dados. Buscando-se evitar o impacto negativo da ocorrência dessas palavras para a análise de textos, tais elementos devem ser removidos antes do processamento do texto. O NLTK (Natural Language Toolkit) desenvolvido em Python possui uma lista de palavras irrelevantes em 16 idiomas diferentes (BANERJEE ,2020).

Stemização ou Lematização

O processo de stemização consiste em reduzir uma palavra ao seu radical. Por exemplo, a palavra “meninas” se reduziria a “menin”. No caso de verbos, eles são reduzidos a sua forma no infinitivo. Assim, as palavras “tiver”, “tenho”, “tinha” e “tem” são formas do mesmo lema “ter” (RÊGO, 2016).

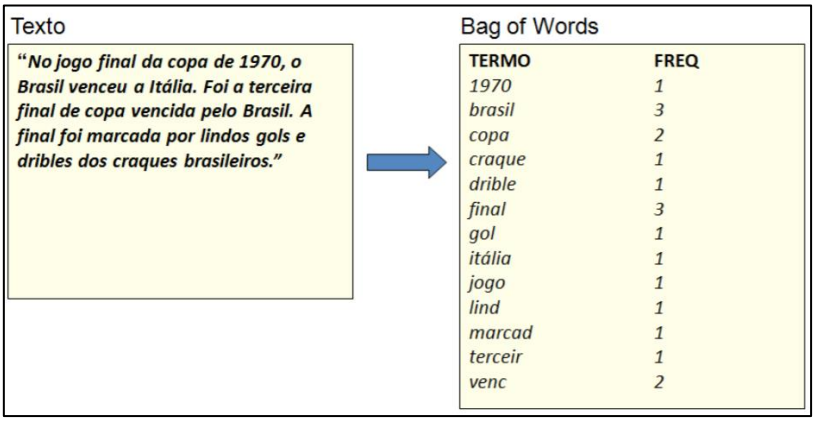
A lematização reduz a palavra ao seu lema original, que é a forma no masculino e singular, no caso de verbos, o lema é o infinitivo conforme já indicado, a vantagem de aplicar a stemização ou lematização é clara: busca-se a redução de vocabulário e abstração de significado (RODRIGUES, 2017, s.p.).

Bag of words

Depois de realizado o processo de tokenização o próximo passo trata da unificação de todas as palavras pertencentes a um texto em um vetor. Na etapa inicial do Processamento de Linguagem Natural a gramática, a ordem das palavras, a estrutura do texto e a pontuação são ignorados, sendo tudo unificado da forma como está originalmente escrito, criando-se a assim uma única bag of words (bolsa de palavras) (PRATES, 2019). Este é um vetor que registra o número de ocorrências (frequência) de cada palavra distinta do documento. As decisões tomadas pelo algoritmo são sempre baseadas nestes valores de frequência da ocorrência dos termos (GONÇALVES, 2012).

Na Figura 20 demostra-se como é gerado o vetor de frequência do bag of words montado após a remoção de stop words e a operação de stemming, conforme os três processos explicados nos anteriormente.

Figura 20: Bag of Words gerado após remoção de stop words e stemming



Fonte: Gonçalves (2012).

Padding de palavras

Um pré-processamento importante de ser realizado é o padding (preenchimento) do array de palavras, uma vez que esse esteja tokenizado, isso acontece como uma forma de deixar todas as palavras em uma mesma dimensão, ainda que na prática elas tenham tamanhos diferentes (CARNEIRO, 2020).

Supondo que em uma base de dados exista um conjunto de cores [‘roxo’, ’lilás’, ‘violeta’, ‘amarelo’]. Após a tokenização, esse conjunto fica da seguinte forma: [8, 6, 12, 2]. O tamanho desse vetor é de 4 posições, uma posição a mais do que este outro vetor: [‘vermelho’, ‘amarelo’, ‘preto’] = [1,2,3]. Portanto, deve-se fazer o padding com o valor 0 (zero) no segundo vetor, isto para manter sempre o mesmo tamanho, ficando da seguinte forma: [1,2,3,0]. (CARNEIRO, 2020).

Frequência dos termos (contagem de palavras)

Após mapear a presença de todas as palavras de cada documento o próximo passo é contá-las. Nesta etapa é calculada a frequência que cada palavra aparece no texto. Porém, um dos principais pontos de atenção nesta etapa é que ao contar todas as palavras da forma que elas aparecem no texto (sem realizar nenhum tratamento), predispõe com que muito ruído seja levantado. (PRATES, 2019, s.p.).

Para evitar tal situação, antes deverão ser executados procedimentos de limpeza nas palavras, tais como a normalização, a stemização e a eliminação de stop words, que já foram explicados anteriormente (PRATES, 2019, s.p.).

Marcação Gramatical (POS) e Análise Sintática

Trata-se de uma tarefa básica na linguística de corpus. O objetivo é atribuir características morfossintáticas a cada palavra em uma frase de acordo com o seu contexto, essa tarefa também pode ser aplicada em sentenças e parágrafos, conforme exemplos dispostos na Figura 21 (EVANGELISTA *et al.,* 2020).

Sucessora natural da marcação gramatical, a análise sintática fornece uma árvore de dependência como saída de cada palavra componente de um corpus, seu objetivo é prever, para cada sentença ou cláusula, uma representação abstrata das entidades gramaticais e suas relações (EVANGELISTA *et al.,* 2020).

Figura 21: Marcação Gramatical

```
frase --> sujeito, predicado.
sujeito --> artigo(G), substantivo(G).
predicado --> verbo, artigo(G), substantivo(G).
artigo(m) --> [o] | [os].
artigo(f) --> [a] | [as].
substantivo(m) --> [gato] | [gatos] | [rato] | [ratos].
substantivo(f) --> [gata] | [gatas] | [rata] | [ratas].
verbo --> [caçou] | [caçaram].
```

Fonte: Adaptado de Evangelista et al. (2020).

A linguagem humana não envolve somente o entendimento das palavras. Assim, é preciso que a máquina consiga interpretar a fala quando a palavra tem duplo sentido, quando a organização de palavras em uma frase não está de acordo com a gramática, o tom de voz e outras circunstâncias específicas. A partir deste ponto são necessárias as aplicações de técnicas inteligentes para coletar os resultados esperados. Tais técnicas inteligentes são aplicadas após todo o processo de tratamento de dados, conforme anteriormente indicado nesta seção (STEFANINI, 2019).

Após o tratamento dos textos, os algoritmos que aplicam PLN precisam ter um raciocínio voltado à geração de respostas às perguntas invisíveis, manipulando assim o conhecimento existente com técnicas de inferência. Isto deve ser realizado com base em análise detalhada das tecnologias atuais e dos desafios de cada aspecto envolvido. Assim, diferentes técnicas podem ser aplicadas para que sejam obtidos resultados adequados a cada tipo de necessidade (ZHOU et al., 2019). O autor argumenta ainda que nos últimos cinco anos, houve um rápido desenvolvimento da PLN, sendo o progresso mais recente na estrutura de PLN baseado em redes neurais, notadamente a partir de três perspectivas: modelagem, aprendizagem e raciocínio.

Muitos mecanismos modernos podem ser utilizados neste processo. Na seção de aprendizagem podem ser incluídas aprendizagem supervisionada,

semi-supervisionada e não supervisionada; aprendizagem multitarefa; transferência de aprendizagem; e ainda aprendizagem ativa (ZHOU et al., (2019).

As redes neurais estão sendo utilizadas com muita frequência para resolver tarefas relacionadas à PLN e funciona muito bem para tarefas supervisionadas, nas quais há dados rotulados abundantes para o aprendizado de redes neurais. Essas técnicas sofreram uma grande evolução ao longo do tempo e estão sendo utilizadas com muita frequência para resolver tarefas relacionadas a classificação de textos. (MUÑOZ-VALERO et al., 2020; Zhou et al. 2019))

2.3.3 Técnicas de Inteligência Artificial (IA)

As técnicas de IA podem ser aplicadas em diferentes áreas de estudo, uma vez que sua aplicação traz inúmeros benefícios. Assim, é cada vez mais frequente sua aplicação para obtenção de respostas nas mais diversas áreas do conhecimento. Em geral, tais técnicas podem resolver problemas cada vez mais complexos trazendo assim eficiência, significado e agilidade (PREUSS et al., 2020; RUSSO, 2020; LUDERMIR, 2021).

As principais técnicas inteligentes atualmente estão inseridas no contexto de Aprendizado de Máquina (AM), esta área é dedicada ao estudo de algoritmos de previsão e inferência, estes buscam simular em computadores o cérebro enquanto máquina de aprendizado. O AM inclui técnicas estatísticas para permitir que máquinas aperfeiçoem ao máximo suas tarefas com base nos dados extraídos por experiência, assim os algoritmos podem aprender com estes dados, identificar padrões e tomar decisões com pouca intervenção humana. (MUYLAERT, 2020; RUSSO, 2020; BIANCHI, 2020)

No contexto de aprendizagem de máquina, existem os seguintes tipos de aprendizado, conforme indicados por Carvalho (1994) e Waltrick (2020):

- **Aprendizado Supervisionado**: o modelo deverá ser, literalmente, ensinado sobre o que deve ser feito. Neste sentido, deverá ser fornecido um conjunto de dados rotulados para o modelo aprender o que é cada

classe/categoria e esses dados serão particionados entre porções para treinamento e teste.

Esse tipo de aprendizado é, geralmente, aplicado quando o objetivo é prever ocorrências futuras. Além disso, é possível utilizar as técnicas de classificação e regressão.

- **Aprendizado Não Supervisionado** (auto-organização): quando não existe um agente externo indicando a resposta desejada para os padrões de entrada. Diferentemente do aprendizado anterior, aqui será fornecido um conjunto de dados não rotulados e não se ensina ao modelo qual é o objetivo final. Alguns exemplos de técnicas que podem ser aplicadas neste aprendizado são: associação, redução de dimensão e clusterização.

- **Aprendizado por Reforço**: quando um crítico externo avalia a resposta fornecida. É usado nos casos em que o problema não é, basicamente, relacionado a conjunto de dados, mas você tem um ambiente para lidar, como um o cenário de um game ou uma cidade onde circulam carros autônomos. Utiliza o método ‘tentativa e erro’, no qual o acerto equivale a uma recompensa, enquanto o erro equivale a uma punição.

Os classificadores têm a tarefa de organizar objetos entre diversas categorias e, para tanto, o modelo analisa o conjunto de dados fornecidos, sendo que cada dado já contém um rótulo indicando a qual categoria ele pertence, a fim de ‘aprender’ como classificar novos dados. Na classificação, os algoritmos que implementam esse processo são chamados de classificadores (RAMOS et al., 2018; HAN et al., 2011).

Affonso et al. (2010) definem que a classificação de textos é uma “técnica utilizada para atribuir automaticamente uma ou mais categorias predefinidas”. A aplicação mais comum é a indexação de textos, mineração de textos, categorização de mensagens, notícias, resumos e arquivos de publicações periódicas. Nos sistemas computacionais, o processo de classificação envolve técnicas para extração das informações mais relevantes de cada categoria, além da utilização destas informações para ensinar o sistema a classificar corretamente os documentos.

A Aprendizagem Profunda (AP), do inglês *Deep Learning* é um ramo do AM que baseado em um conjunto de algoritmos que tentam modelar abstrações de alto nível de dados, algumas de suas representações são inspiradas na interpretação do processamento de informações e padrões de comunicação em um sistema nervoso (Premlatha, 2019; Barnerjee 2020; BIANCHI, 2020).

A rede neural é uma das arquiteturas da (AP) que tem sido aplicada em diferentes áreas, dentre elas a PLN para reconhecimento de entidade nomeada, marcação de classes gramaticais, processamento de voz, tradução e classificação de textos. (Premlatha, 2019; Barnerjee 2020; BIANCHI, 2020)

Os próximos tópicos irão tratar das técnicas aplicadas para a solução do problema identificado nesta pesquisa.

2.3.3.1 Redes Neurais

Segundo Carvalho (1994), Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento por meio da experiência, uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento; já o cérebro de um mamífero pode ter muitos bilhões de neurônios.

A propriedade mais importante das redes neurais é a habilidade de aprender sobre seu ambiente e, com isso, melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado aos seus pesos, denominado treinamento (CARVALHO, 1994). O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma determinada classe de problemas (CARVALHO, 1994).

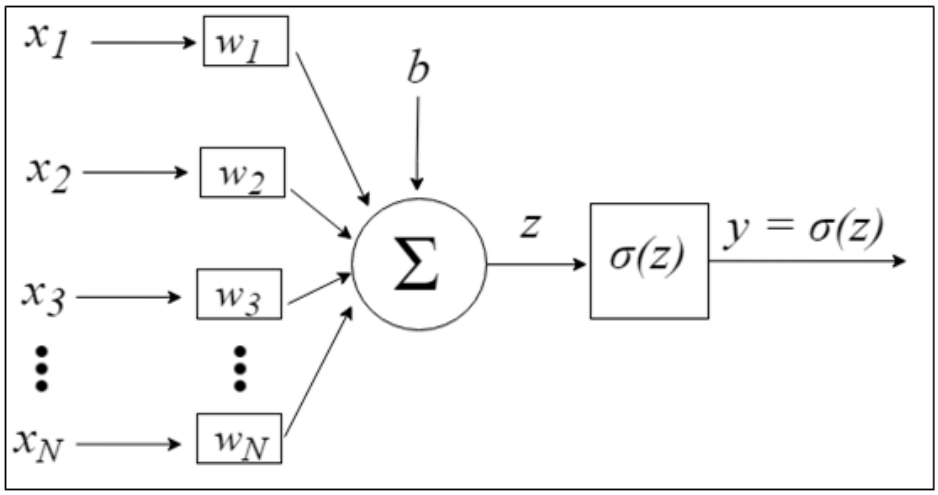
Inspirando-se no funcionamento dos neurônios biológicos do sistema nervoso dos animais, estabeleceu-se na área da Inteligência Artificial o modelo computacional de um neurônio (LEITE, 2018).

Na figura 22 é observado os sinais de entrada no neurônio, representados pelo vetor x = [x1, x2, x3, …, xN]. Ao chegarem ao neurônio, são multiplicados pelos respectivos pesos sinápticos, que são os elementos do

vetor w = [w1, w2, w3, …, wN], gerando assim o valor z, comumente denominado potencial de ativação (LEITE, 2018).

O termo b correspondendo tipicamente ao ‘bias’ (viés) é um parâmetro adicional na Rede Neural que é aplicado para ajustar a saída junto da soma ponderada das entradas para o neurônio. O valor z passa então por uma função matemática de ativação σ, com a característica de ser não linear, e responsável por limitar tal valor a um determinado intervalo, produzindo o valor final de saída y do neurônio (LEITE, 2018).

Figura 22: Modelo computacional de um neurônio



Fonte: Leite, 2018

Em 1989, Yann LeCun combinou redes neurais convolutivas com Backpropagation para ler os dígitos ‘manuscritos’. A ideia era tomar uma grande quantidade de dígitos manuscritos, conhecidos como exemplos de treinamento e, em seguida, desenvolver um sistema que pudesse aprender com esses exemplos de treinamento (Deep Learning Book, 2021)

a) MLP (Multilayer Perceptron)

A MLP é uma rede neural que possui mais de uma camada de neurônios. Em casos em que não há a possibilidade de uma única reta separar os elementos, a MLP gera um plano de classificações. As redes de múltiplas camadas distinguem-se das redes de camada simples pelo número de camadas intermediárias, ou seja, aquelas entre a camada de entrada e a camada de saída. Assim, esta arquitetura possui uma ou mais camadas

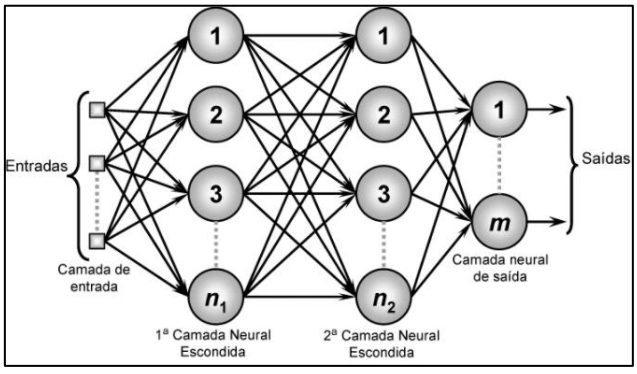
ocultas, que são compostas por neurônios computacionais, também chamados de neurônios ocultos (AFFONSO et al.,2010; LEITE, 2018).

O treinamento de uma rede MLP insere-se no contexto de aprendizado de máquina supervisionado, no qual cada amostra de dados utilizada apresenta um rótulo informando a qual classificação ela se encaixa, a ideia geral é fazer com que a rede aprenda os padrões referentes a cada tipo de coisa (cada classe) e, assim, quando uma amostra desconhecida for fornecida à rede, ela seja capaz de estabelecer a qual classe tal amostra pertence (LEITE, 2018).

O algoritmo utilizado para treinamento da MLP é chamado backpropagation ou retropropagação, sendo composto por quatro passos: inicialização, ativação, treinamento dos pesos e iteração. A ideia do algoritmo backpropagation é, com base no cálculo do erro ocorrido na camada de saída da rede neural, recalcular o valor dos pesos do vetor da camada última camada de neurônios (MOREIRA, 2018; LEITE, 2018).

Dessa forma, é possível proceder para as camadas anteriores, de trás para a frente, ou seja, atualizar todos os pesos das camadas a partir da última até atingir a camada de entrada da rede realizando, para tanto, a retropropagação do erro obtido pela rede (MOREIRA, 2018; LEITE, 2018). A arquitetura da MLP pode ser visualizada na Figura 23.

Figura 23: Arquitetura da MLP



Fonte: Moreira (2018).

Moreira (2018) afirma que quando não se atinge o resultado esperado, aumentar o número de camadas e neurônios nem sempre é a melhor opção. Isto porque esta rede possui algumas limitações, uma vez que ao se aumentar

muito o número de camadas e neurônios, a rede tende a ficar com um número de parâmetros muito elevado e, com isso, tão pesada ao ponto do hardware do equipamento não conseguir processar os dados e a rede não convergir.

b) Redes Neurais Convolucionais

A Rede Neural Convolucional ou Neural Convolucional (RNC) é um algoritmo de Aprendizado Profundo que pode captar uma entrada, atribuir importância (por meio de pesos e vieses que podem ser aprendidos) a vários aspectos, sendo capaz de diferenciar um do outro.

As Redes Neurais Convolucionais (RNCs) são responsáveis por avanços na classificação de imagens, configurando-se no núcleo da maioria dos sistemas de visão por computador atuais, desde a marcação automática de fotos do Facebook até carros autônomos. Mais recentemente, tem-se aplicado RNCs em problemas de Processamento de Linguagem Natural, para os quais vem se obtendo resultados promissores (BRITZ, 2015),

Rodrigues (2018) destaca que as RNCs necessitam de grande quantidade de dados rotulados para a extração de padrões. Elas se mostram muito eficazes para resolver problemas de classificação, apresentando-se como uma alternativa viável aos métodos tradicionais para esse tipo de problema.

A arquitetura e funcionamento das RNCs aplicada à linguagem natural é voltada a tarefas de classificação textual, como análise de sentimentos, categorização de notícias e detecção de spam, entre outros (CARNEIRO, 2020).

A seguir é explicado o processo de RNC. Uma vez que todo o pré-processamento nos textos brutos foi realizado, como os tópicos destacados anteriormente no tópico de PLN (normalização, remoção de stop words, padding de palavras e outros), é então gerada a representação matricial das palavras do corpus.

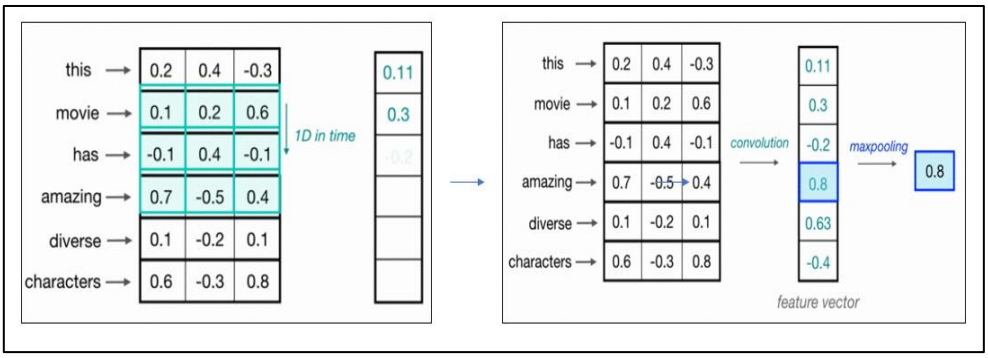
O primeiro quadro a esquerda da Figura 24 mostra que cada frase do documento ou da base de dados se transforma em uma representação matricial. A quantidade de linhas dessa matriz é basicamente a quantidade de

palavras existentes na frase e a quantidade de colunas ou dimensões (assim como o valor de cada dimensão) é criado automaticamente pelo algoritmo de word embedding, quando se chega a este ponto, o algoritmo de redes neurais convolucionais pode então ser aplicado (CARNEIRO, 2020).

A ideia de word embedding é criar uma representação contextualizada das palavras, ao mesmo tempo que possibilita a redução da dimensionalidade em relação ao método anterior, a aplicação dessa técnica, que também é feita utilizando-se aprendizado de máquina, identifica a relação entre as palavras a partir de sua vizinhança, colocando-as em um vetor no qual cada dimensão representa um contexto específico (CARNEIRO, 2020).

Assim, as palavras parecidas ocupam dimensões próximas umas das outras devido à sua dependência mútua (CARNEIRO, 2020).

Figura 24: Representação matricial das palavras e processo de convolução



Fonte: Sharma (2020) e Carneiro (2020).

O processo gera um output (saída) com a mesma quantidade de linhas que a matriz de input (entrada), no caso da Figura 26, de 6 linhas. A camada de convolução aqui (onde a aplicação do kernel é realizada) é chamada de 1D-conv, pois o output é um array (de uma dimensão), diferentemente da convolução para imagens, onde o output é uma matriz, portanto, um 2D-conv (SHARMA, 2020; CARNEIRO, 2020).

O output pode ser chamado de vetor de características, pois a aplicação dos kernels possui justamente o objetivo de extrair e sumarizar características principais encontradas em uma matriz (SHARMA, 2020; CARNEIRO, 2020).

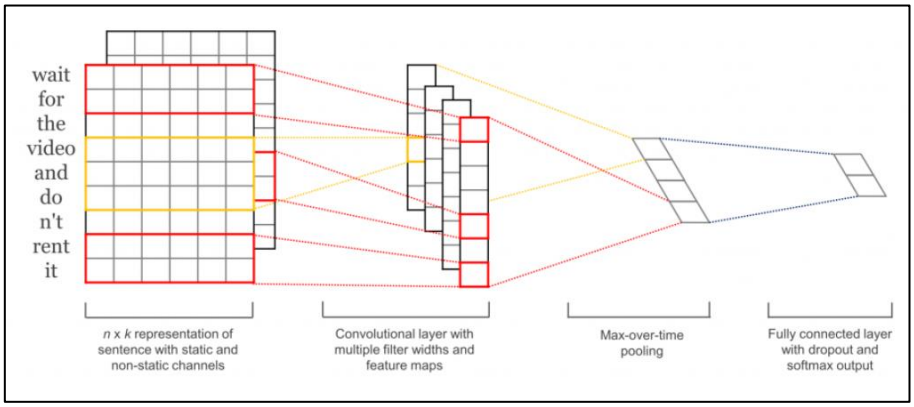
No estudo de Carneiro (2020), o max pooling ao final do processo de convolução tem por objetivo extrair o maior valor do vetor de características,

como foi o caso do 0,8; sendo basicamente a representação da palavra que teve maior destaque na frase (SHARMA, 2020; CARNEIRO, 2020).

Em seu estudo, Kim (2014) avaliou uma arquitetura RNC em vários conjuntos de dados de classificação em textos, no trabalho foi realizado a categorização de tópicos. A arquitetura RNC atingiu um desempenho muito bom em todos os conjuntos de dados e um novo estado da arte em alguns.

Britz (2015) explica que no processo de RNC exemplificado na Figura 25, as operações de convoluções e agrupamento perdem informações sobre a ordem local das palavras, de modo que a marcação de sequência como em Marcação de PoS ou Extração de entidade é um pouco mais difícil de se ajustar em uma arquitetura RNC. O autor ainda ressalta que esta arquitetura funciona bem para textos longos (como resenhas de filmes), mas seu desempenho em textos curtos (como tweets) não é claro.

Figura 25: Redes neurais convolucionais para classificação de sentenças



Fonte: Kim (2014).

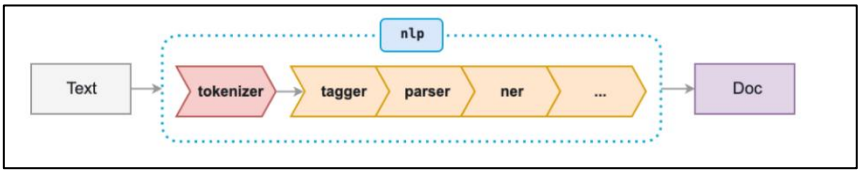
O processo de aplicação das RNCs pode ser realizado por meio da utilização da Biblioteca Spacy. SpaCy é uma biblioteca de código aberto gratuita para Processamento de Linguagem Natural (PLN) avançado em Python. Ela foi projetada especificamente para uso em produção e ajuda a criar aplicativos que processam e ‘entendem’ grandes volumes de texto. Ela pode ser aplicada para a construção de sistemas de extração de informações ou para a compreensão de linguagem natural (SPACY IO, 2021).

O ambiente Spacy NER usa uma estratégia de incorporação de palavras aplicando recursos de subpalavra e incorporação de Bloom e Rede Neural Convolucional 1D (RNCs), conforme demostrado a seguir: (SHARMA, 2020).

Incorporação de Bloom: é semelhante à incorporação de palavras e representação otimizada de espaço maior. Ele dá a cada palavra uma representação única para cada contexto distinto em que ela se encontra.

1D RNC: é aplicado sobre o texto de entrada para classificar uma frase / palavra em um conjunto de categorias predeterminadas.

Figura 26: Funcionamento do Spacy



Fonte: (Sharma, 2020)

O processo de funcionamento do Spacy está exposto na Figura 26, para gerar o documento que passará pela rede convolucional do Spacy é necessário aplicar o pré-processamento dos dados utilizando PLN anteriormente a partir das seguintes características (SHARMA, 2020):

1. Ele simboliza o texto, ou seja, frase de entrada dividida em palavras ou incorporação de palavras

2. As palavras são então divididas em características e, em seguida, agregadas a um número representativo

3. Esse número é então alimentado para uma estrutura neural totalmente conectada, que faz uma classificação com base no peso atribuído a cada recurso dentro do texto.

Já o processo demonstrado na Figura 27 exibe como funciona o processo de treinamento e geração do modelo de classificação da RNC do Spacy (SHARMA, 2020), com as seguintes características:

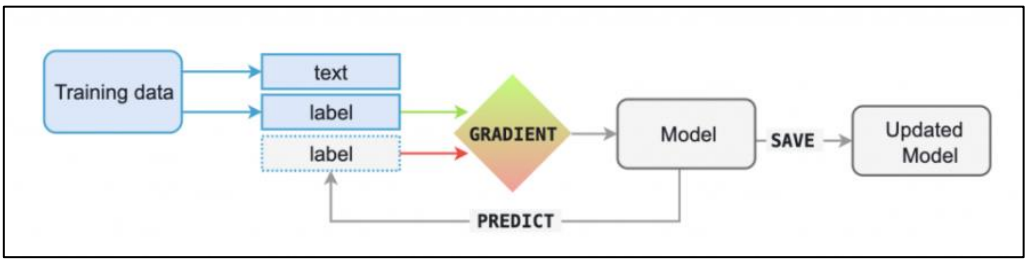
1. Dados de treinamento: os dados de treinamento do modelo utilizam um percentual de toda a base, este pode variar de acordo com a necessidade, normalmente entre 70% e 80%;

2. Texto: texto de entrada para o qual o modelo deve prever um rótulo.

3. Rótulo: o rótulo que o modelo deve prever, também chamada de target, normalmente a coluna de classificação;

4. Gradiente: Calcula como alterar os pesos para melhorar as previsões. (Compara o rótulo de predição com o rótulo real e ajusta seus pesos para que a ação correta tenha uma pontuação mais alta na próxima vez.) Geração do Modelo para aplicação no software.

Figura 27: Como treinar o Spacy



Fonte: Sharma (2020).

c) Árvores de Decisão

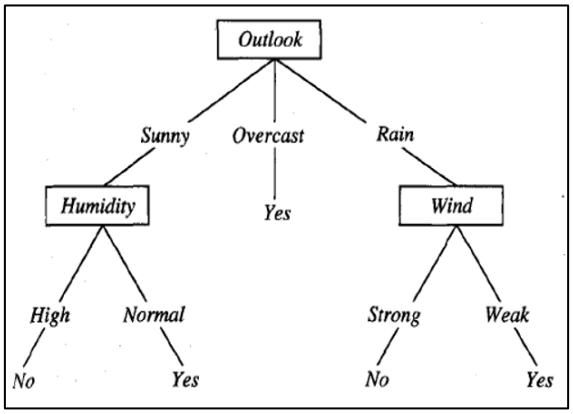
Árvores de decisão é um algoritmo de aprendizado supervisionado, utilizados em tarefas de classificação e regressão. A árvore de decisão é um método para aproximar funções-alvo com valores discretos, em que a função aprendida é representada por uma árvore de decisão. As decisões são tomadas com base num conjunto de regras do tipo 'se-então' *('if-then'*) (MITCHELL, 1997).

Árvore de decisão é um dos algoritmos de inferência indutiva mais populares, sendo considerado robusto a ruídos nos dados. O algoritmo utiliza o viés indutivo e preferência por árvores menores por meio da navalha de occam, uma vez que este prefere as hipóteses mais curtas (mais simples), que se ajustam aos dados (MITCHELL, 1997).

As árvores de decisão representam uma das formas mais simplificadas de um sistema de suporte a decisão. A partir de um conjunto de dados o algoritmo cria uma representação do conhecimento ali embutido, em formato de árvore. A biblioteca do Scikit-learn do python já tem implementado um algoritmo para a geração das árvores de decisão. A biblioteca envolve todos os cálculos matemáticos envolvidos para gerar as previsões. (PESSANHA, 2019)

O funcionamento da Árvore de Decisão visa formar ‘caminhos’ que vão dividindo os dados em pequenos grupos. Essa divisão acontece com base nas características dos dados para que, ao final, após o processo de treinamento, novos registros são inseridos e o modelo poderá predizer em qual classe se encaixa melhor (BERLATTO, 2021).

Figura 28: Arquitetura da Árvore



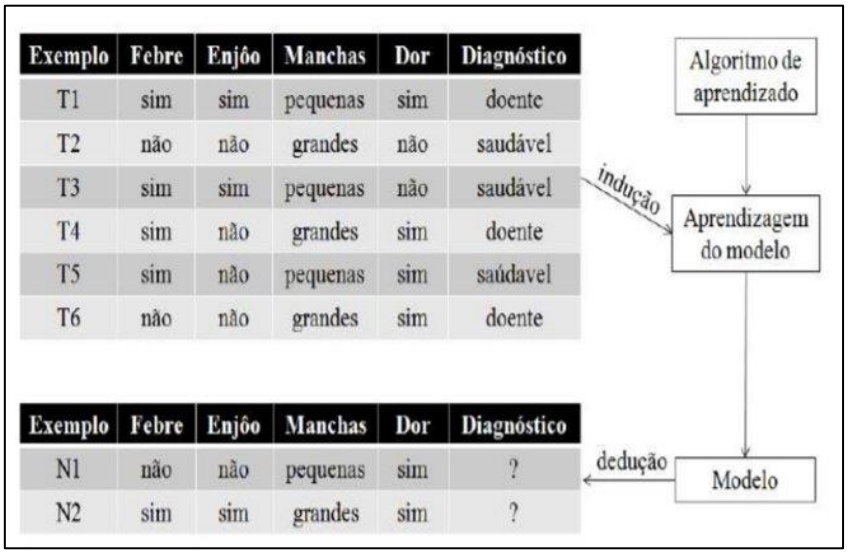
Fonte: Mitchell (1997).

Na figura 28 é possível visualizar o exemplo fornecido por Mitchell (1997), no qual o algoritmo faz a previsão da ação do usuário em jogar ou não jogar tênis baseado das variáveis do ambiente, tais como sol, chuva, humidade etc. No exemplo acima, o algoritmo da árvore determina, por exemplo, como ação positiva que o usuário irá jogar quando o tempo estiver nublado ou quando estiver ensolarado e umidade normal ou quando está chovendo, porém com vento fraco.

Assim, para o processo de classificação a base de dados deverá possuir uma coluna-alvo com valores de saída discretos ou categóricos (pode-se usar como exemplo classificação booleana 1 ou 0 - true ou false, sim ou não, como no exemplo da figura 28, informando assim a possibilidade de uma pessoa jogar tênis). Os valores discretos podem estar relacionados às variáveis que implicam um diagnóstico médico (doente ou saudável) ou a uma análise de um texto, que pode ser classificado, além de diferentes categorias (política, esporte, tecnologia, etc.).

A figura 29 exemplifica como funciona o processo de indução de um classificador e, posteriormente, sua dedução.

Figura 29: Indução de um classificador e dedução das classes para novas amostras



Fonte: ZUBEN e ATTUX - DCA/FEEC/Unicamp, 2016

Na primeira tabela os rótulos são conhecidos e é utilizado um algoritmo de classificação para construir um modelo de predição. Depois que o modelo é gerado novos dados podem ser inseridos para o processo de dedução, sendo que na segunda tabela os rótulos não são conhecidos. O objetivo desse algoritmo é encontrar o atributo que gera a melhor divisão dos dados, ou seja, o subconjunto com maior pureza (ZUBEN; ATTUX, 2016).

d) Florestas Aleatórias

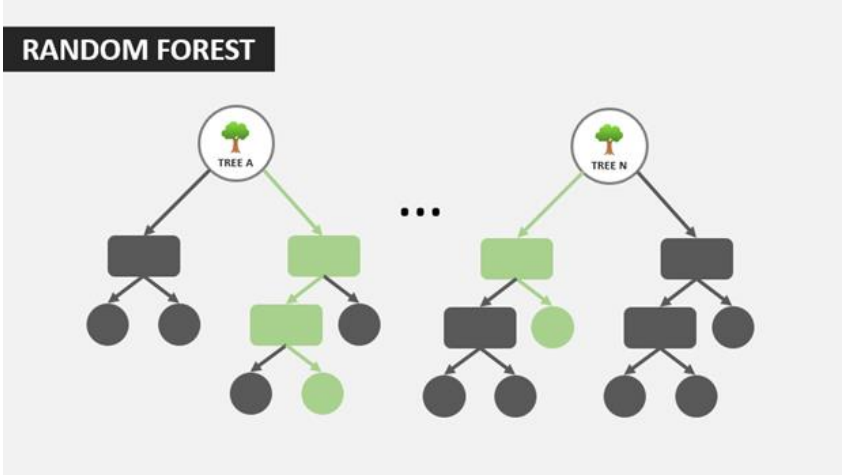
Floresta Aleatória (*Random Forest*) é um algoritmo de aprendizagem supervisionada que cria uma floresta de modo aleatório, a 'floresta' criada é uma combinação (ensemble) de árvores de decisão, na maioria dos casos treinados com o método de bagging, a ideia principal do método de bagging é que a combinação dos modelos de aprendizado aumenta o resultado geral (COSTA DA SILVA, 2018).

A floresta aleatória é um algoritmo de aprendizado de máquina que constrói um conjunto baseados em árvore, uma vez que ele obtém um voto das previsões de cada classificador. Esse algoritmo funciona muito bem quando existem muitos atributos a serem avaliados, pois seleciona aleatoriamente os atributos para dividir cada nó. Tal característica o torna mais robusto para o tratamento de dados com ruídos (SANTOS, 2019).

O algoritmo de Floresta Aleatória propõe a criação de várias árvores de decisão baseadas em subconjuntos aleatórios de uma base de dados. Dessa forma, é possível classificar um documento baseado não só em uma árvore, mas, sim em um conjunto de árvores (‘floresta’) (ALBUQUERQUE, 2019).

O *overfitting* (termo usado em estatística para descrever quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados) pode acontecer para essa técnica (ALBUQUERQUE, 2019). Nesse caso, o que ocorre é que uma árvore de decisão pode ficar tão profunda, ou seja, com tantos atributos, que nenhuma instância de teste possa ser corretamente classificada. Para esse problema, a solução é ‘podar’ a árvore em determinados pontos, aumentando assim a generalidade de sua decisão (ALBUQUERQUE, 2019).

Figura 30: Representação Floresta Randômica



Fonte: Pessanha, 2019

A figura 30 exemplifica a Floresta Randômica, com os respectivos passos para criação do algoritmo, quais sejam: realizar a seleção aleatória de alguns recursos; identificar o recurso mais adequado para a posição do nó raiz; gerar nós filhos e, por fim; repetir os passos anteriores até que se atinja a quantidade de árvores desejadas. Assim que o modelo é gerado, as previsões são feitas a partir de ‘votações’.

Posteriormente, cada mini árvore toma uma decisão a partir dos dados apresentados, a decisão mais votada é a resposta do algoritmo (PESSANHA,

2019). A biblioteca do Scikit-learn do Python já tem implementado um algoritmo para a geração das Florestas Randômicas

e) Gradiente Boosting

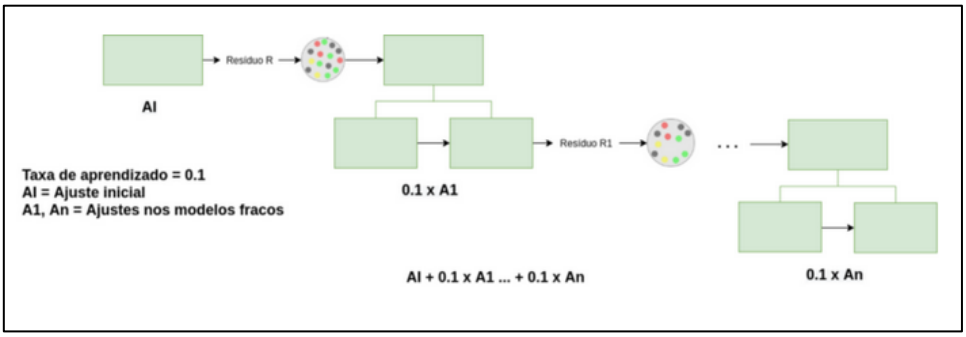
O Gradient Boosting é uma generalização do boost para funções de perda diferenciáveis arbitrárias. Esse algoritmo é um procedimento preciso e eficaz que pode ser usado para problemas de regressão e classificação em uma variedade de áreas, incluindo a classificação de pesquisa na Web ou ecologia, por exemplo. (Biblioteca Scikit Learn - Gradient Boosting, 2021)

O módulo do sklearn fornece métodos para classificação e regressão por meio de árvores de decisão incrementadas por gradiente. A biblioteca GradientBoostingClassifier suporta classificação binária e multiclasse. A classificação com mais de 2 classes requer a indução de n_classes de árvores de regressão a cada iteração e, portanto, o número total de árvores induzidas é igual. Para conjuntos de dados com um grande número de classes, é altamente recomendável usar como alternativa a n_classes * n_estimators (Biblioteca Scikit Learn - Gradient Boosting, 2021).

O algoritmo Gradient Boosting está incluído no grupo de classificadores Ensemble. Esse classificador utiliza uma combinação de resultados de preditores fracos com o objetivo de produzir um melhor modelo preditivo. Na técnica de Boosting, cada classificador fraco é treinado com um conjunto de dados, de forma sequencial e de uma maneira adaptativa, pela qual um modelo base depende dos anteriores e, ao final, são combinados de uma maneira determinística (SILVA, 2020).

Isso se dá de forma diferente de algoritmos de Bagging (outro algoritmo Ensemble tradicional), nos quais os preditores fracos são treinados de forma individual e paralela, e ao final são combinados seguindo um processo determinístico de média, como uma votação (SILVA, 2020).

Figura 31: Funcionamento do Gradiente Boosting



Fonte: Silva (2020).

A figura 31 demonstra o funcionamento do algoritmo. Cada modelo gerado sofre um ajuste que gera um resíduo, sendo este calculado pela distância entre o que foi previsto e o valor real. Um próximo modelo é criado e ajustado com base no resíduo gerado pelo modelo anterior.

Essas interações são repetidas por um determinado número de vezes, buscando-se assim minimizar o resíduo gerado pelos modelos fracos, ou seja, até que a distância entre o previsto e o valor real seja a menor possível, o modelo final é a soma dos ajustes de todos os modelos fracos (SILVA, 2020). A biblioteca do Scikit-learn do python já tem implementado um algoritmo para a geração das previsões do Gradiente Boosting.

f) Ada Boost

O módulo *Sklearn Ensemble* inclui o algoritmo de *boosting* popular conhecido como *AdaBoost*, introduzido em 1995 por Freund e Schapire. O princípio básico do *AdaBoost* é ajustar uma sequência de aprendizes fracos (ou seja, modelos que são apenas ligeiramente melhores do que suposições aleatórias, como pequenas árvores de decisão) em versões repetidamente modificadas dos dados (Biblioteca Scikit Learn - AdaBoost, 2021)

As previsões neste algoritmo são então combinadas por meio de uma maioria ponderada de votos (ou soma) para produzir a previsão final. As modificações de dados em cada uma das chamadas iterações de reforço consistem na aplicação de pesos w1, w2, ..., wn para cada uma das amostras de treinamento (Biblioteca Scikit Learn - AdaBoost, 2021).

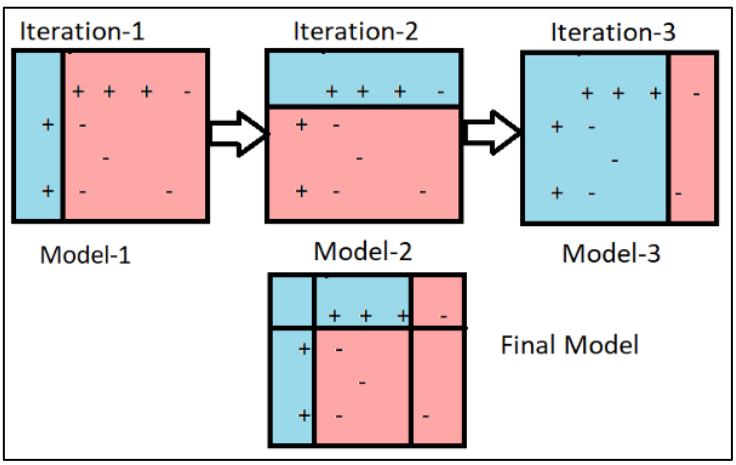
Para cada iteração sucessiva, os pesos da amostra são modificados individualmente e o algoritmo de aprendizagem é reaplicado aos dados reponderados. À medida que as iterações prosseguem, os exemplos difíceis de prever recebem uma influência cada vez maior (Biblioteca Scikit Learn - AdaBoost, 2021).

AdaBoost foi o primeiro algoritmo de *boosting* realmente bem-sucedido desenvolvido para classificação binária. É o melhor ponto de partida para entender o impulso. Os métodos modernos de *boosting* são baseados no AdaBoost, principalmente as máquinas de *boosting* de gradiente estocástico (BROWNLEE, 2016).

O AdaBoost pode ser usado para impulsionar o desempenho de qualquer algoritmo de aprendizado de máquina. É melhor usado com classificadores fracos. Estes são modelos que alcançam precisão logo acima do acaso em um problema de classificação (BROWNLEE, 2016).

As previsões são feitas calculando a média ponderada dos classificadores fracos. Para uma nova instância de entrada, cada aluno fraco calcula um valor previsto como +1,0 ou -1,0. Os valores previstos são ponderados por cada valor de estágio do aluno fraco. No Ada boosting, os modelos básicos são dependentes uns dos outros de tal forma que um modelo é criado tendo em mente o erro cometido pelo anterior para que o erro de treinamento seja reduzido gradualmente, tal esquema pode ser visualizado na figura 32, que expõe o modelo final após 3 iterações com o erro reduzido (BROWNLEE, 2016).

Figura 32: Funcionamento do Ada Boosting



Fonte: Brownlee, 2016

Na previsão para o modelo de conjunto é considerada a soma das previsões ponderadas. Se a soma for positiva, a primeira classe é predita; se negativa, a segunda classe é predita (BROWNLEE, 2016).

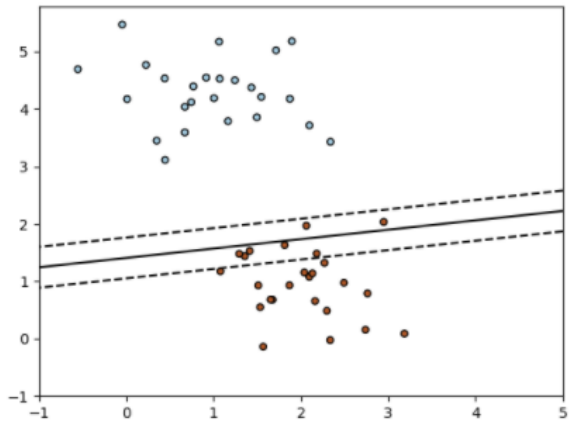
g) Stochastic Gradient Descent (SGD)

A SGD é um algoritmo simples mas eficiente, empregado para ajustar classificadores lineares e regressores sob funções de perda convexa, como máquinas de vetor de suporte (linear) e regressão logística. Embora o SGD já exista na comunidade de aprendizado de máquina há muito tempo, ele recebeu uma quantidade considerável de atenção recentemente no contexto do aprendizado em grande escala (Biblioteca Scikit Learn - *Stochastic Gradient Descent*, 2021).

O SGD foi aplicado com sucesso a problemas de aprendizado de máquina esparsos e em grande escala, frequentemente encontrados na classificação de texto e no processamento de linguagem natural. As vantagens da aplicação deste algoritmo são a eficiência e facilidade de implementação, pois oferece muitas oportunidades para ajuste de código. A classe SGDClassifier implementa uma rotina de aprendizagem de gradiente descendente estocástico simples que suporta diferentes funções de perda e penalidades para a classificação (Biblioteca Scikit Learn - SGD, 2021).

A classe *SGD Classifier* implementa uma rotina de aprendizagem de descida gradiente estocástica simples que suporta diferentes funções de perda e penalidades para a classificação. A figura 33 mostra o limite de decisão de um SGDClassifier treinado com a perda de dobradiça, equivalente a um SVM linear (PATLOLLA, 2017).

Figura 33: limite de decisão de um SGDClassifier



Fonte: Biblioteca Scikit Learn – SGD Classifier (2021).

O Classificador SGD pode funcionar tão bem quanto a regressão logística. Para tanto, é necessário realizar a otimização de hiper parâmetros, uma vez que essa adequação permite que o algoritmo execute uma pesquisa exaustiva de grade em um modelo, o que dá ao usuário a flexibilidade de especificar o conjunto de validação, pontuação métrica e opcionalmente plotar as pontuações sobre a grade de hiper parâmetros inseridos (PATLOLLA, 2017).

Como outros classificadores, o SGD deve ser equipado com duas matrizes: uma matriz X de forma (n_samples, n_features) contendo as amostras de treinamento e uma matriz y de forma (n_samples) contendo os valores alvo (rótulos de classe) para as amostras de treinamento (PATLOLLA , 2017).

h) Support Vector Machines (SVM)

As SVMs são um conjunto de algoritmos de aprendizado supervisionado usados para classificação, regressão e detecção de outliers (uma observação que se diferencia tanto das demais observações que levanta suspeitas de que aquela observação foi gerada por um mecanismo distinto). As SVMs são eficazes em espaços dimensionais elevados e quando o número de dimensões é maior do que o número de amostras (Biblioteca Scikit Learn - SVM, 2021)

As SVMs analisam os dados e reconhecem padrões, considerando como entrada um conjunto de dados para predizer, para cada entrada dada,

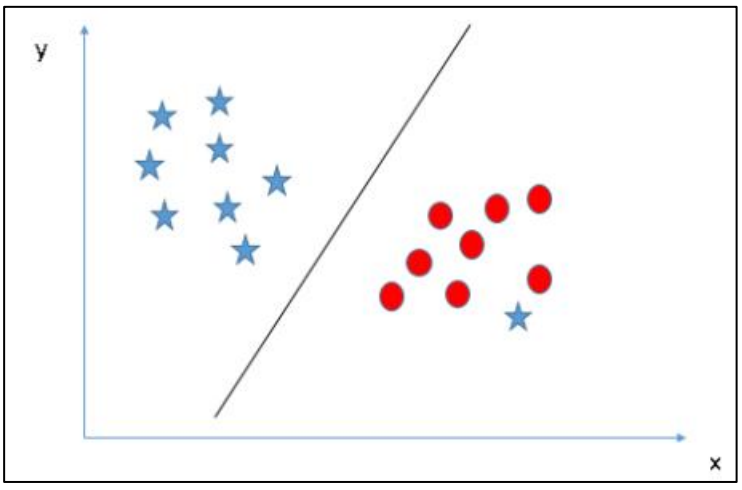
qual de duas possíveis classes a entrada faz parte, o que faz da SVM um classificador linear binário não probabilístico. Usa um subconjunto de pontos de treinamento na função de decisão (chamados de vetores de suporte), portanto, também é eficiente em termos de memória (Biblioteca Scikit Learn - SVM, 2021).

Assim, os algoritmos são considerados versáteis no qual diferentes funções do Kernel podem ser especificadas para a função de decisão. Kernels comuns são fornecidos, mas também é possível especificar kernels personalizados (Biblioteca Scikit Learn - SVM, 2021).

As SVMs são classes capazes de executar binários e classificação multi-classe em um conjunto de dados. Nesse algoritmo, plota-se cada item de dados como um ponto no espaço n-dimensional (onde n é o número de recursos que se tem), com o valor de cada recurso sendo o valor de uma determinada coordenada, então, executa-se a classificação encontrando o hiperplano que diferencia muito bem as duas classes (SOUZA, 2019).

O SVM tem um recurso para ignorar valores discrepantes e encontrar o hiperplano que tem margem máxima, portanto, pode-se dizer que SVM é robusto para *outliers*, na figura 34 o *outlier* está localizado ao lado das elipses vermelhas (SOUZA, 2019).

Figura 34: Identificação de *outlier* pela SVM



Fonte: Souza (2019.)

Os algoritmos de aprendizagem de máquina (SVM) têm como objetivo a determinação de limites de decisão que produzam uma separação ótima entre

classes por meio da minimização dos erros. Ele consiste numa técnica computacional de aprendizado para problemas de reconhecimento de padrão (VAPNIK, 1995).

Essa classificação é baseada no princípio de separação ótima entre classes, tal que se as classes são separáveis, a solução é escolhida de forma a separar o máximo as classes, o SVM já foi utilizado na área de sensoriamento remoto com relativo sucesso (BROWN *et al.*, 2000; MELGANI; BRUZZONE, 2004).

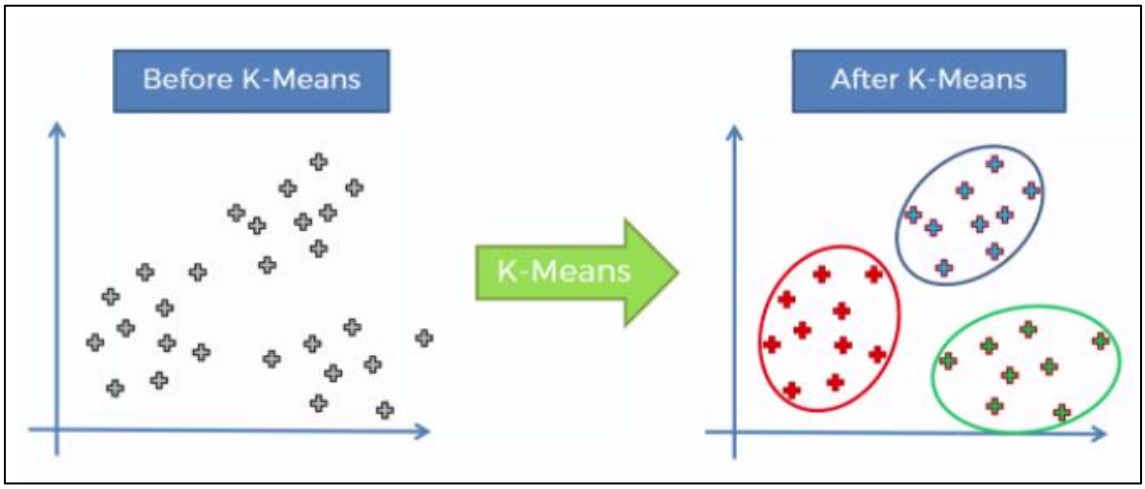
i) K-Means

O K-Means é um algoritmo que utiliza aprendizado não supervisionado, ou seja, que não trabalha com dados rotulados. O objetivo desse algoritmo é encontrar similaridades entre os dados e agrupá-los conforme o número de grupos passados pelo argumento k. Ele encontra k clusters diferentes no conjunto de dados. O centro de cada cluster é nominado centroide e apresenta a média dos valores neste cluster (HONDA, 2017).

A similaridade do algoritmo K-Means é calculada com base em uma função de distância, que pode estar relacionada à própria distância euclidiana ou a alguma medida de similaridade como um coeficiente de correlação. As coordenadas dos objetos para a plotagem dos gráficos podem ser obtidas via escalonamento multidimensional (RÊGO, 2016; FERNANDEZ E MARQUES, 2019).

O objetivo é encontrar ou descobrir alguma relação ou equivalência em sua base de dados e, posteriormente, agrupá-los conforme um número de *clusters* (grupos) informado (ou não) previamente ao seu algoritmo. Além disso, o algoritmo se baseia em cálculos de distância entre pontos e, através dessa métrica, é definido a qual grupo o dado vai pertencer. A figura 35 demonstra a divisão dos *clusters* após o processo de classificação (SILVA, 2020).

Figura 35: K-mena



Fonte: Silva (2020).

O escalonamento multidimensional (MDS) é uma técnica de interdependência que permite mapear distâncias entre objetos. A técnica é apropriada para representar graficamente n elementos em um espaço de dimensão menor do que o original, levando-se em conta a distância ou à similaridade que os elementos têm entre si. (FÁVERO, 2009)

Nesse tipo de representação, quanto mais próximos estiverem os objetos, mais semelhantes entre si eles serão. Ressalte-se que o escalonamento multidimensional é uma técnica exploratória usada para obtenção de avaliações comparativas entre objetos, de forma a identificar comportamentos não observados em outras análises HAIR et al. (2009).

2.3.4 Métricas de avaliação de Desempenho

Na classificação, duas etapas principais são realizadas. Na primeira etapa é gerado o modelo que aprende por meio do treinamento dos dados, normalmente utilizando 70% a 80% da base de dados. Na segunda etapa, os dados separados são testados, entre 30% e 20% da base para estimar o desempenho, mensurando-se assim os acertos do modelo (RAMOS et al., 2018; HAN et al., 2011).

Após o processo de classificação é necessário avaliar o desempenho do classificador, para tal são utilizadas algumas métricas. Para distinguir entre a

classe real e a classe prevista, são usados rótulos (P - Positivo, N - Negativo) para as previsões de classe produzidas por um modelo.

Segundo Ramos et al. (2018) e Guamá (2019), dado um classificador e uma instância a classificar, há quatro resultados possíveis:

a) VP (Verdadeiro Positivo), quando o rotulo avaliado é verdadeiro e modelo trouxe como resultado um valor positivo – indicando acerto do modelo;

b) FN (Falso Negativo), quando houve um erro do modelo, previu a classe negativo quando o valor real era positivo;

c) VN (Verdadeiro Negativo), quando o rótulo avaliado é negativo e o modelo trouxe como resultado um valor negativo – indicando acerto do modelo;

d) FP (Falso Positivo), quando o rótulo avaliado é negativo e o modelo trouxe como resultado um valor positivo, indicando assim erro do modelo.

Figura 36: Matriz de Confusão

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Rodrigues, (2019)

A matriz de confusão visualizada na Figura 36, demonstra como as nomenclaturas citadas anteriormente ficam dispostas. Destas, são retiradas informações outras métricas de avaliação como: Acurácia, Precisão, Recall e F1-score demonstradas no quadro 11.

Quadro 11: Métricas de Avaliação

Métrica	Descrição	Fórmula
Acurácia	Indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente;	$\frac{VP + VN}{VP + FV + FP + FN}$

Precisão	Dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas;	$\frac{VP}{VP + FP}$
Recall	Dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas;	$\frac{VP}{VP + FN}$
F1-Score	Média harmônica entre precisão e recall.	$\frac{2 * Precisão * Recall}{Precisão + Recall}$

Fonte: Adaptado - Rodrigues, (2019)

As métricas podem ser utilizadas em diferentes classificadores, como os que foram demonstrados nos tópicos anteriores: redes neurais convolucionais; Multilayer Perceptron (MLP) classifier; decision tree classifier; random forest classifier; gradiente boosting classifier; adaboosting classifier; Support Vector Machine (SVM) classifier.

3 MÉTODO E MATERIAIS

Neste capítulo são abordados o método e materiais utilizados para realização desta pesquisa, assim como as fases dos experimentos computacionais executados.

3.1 Caracterização da pesquisa

A metodologia de pesquisa nesta dissertação foi definida como pesquisa aplicada e experimental, executada por meio da aplicação dos algoritmos e mensuração dos resultados obtidos em experimentos desenvolvidos pela autora. Barros e Lehfeld (2014) definem que a pesquisa aplicada, também chamada de prática, é aquela na qual o pesquisador busca orientação prática à solução de problemas encontrados na realidade.

Já a pesquisa experimental estabelece caminhos para a realização do experimento, sendo que as tarefas devem ser estruturadas para melhor controle dos instrumentos necessários. Após a efetivação dos experimentos chega-se à análise e discussão dos resultados. Barros e Lehfeld (2014)

3.2 Base de dados e Plataforma de Ensaios

Para a condução dos experimentos foi utilizada uma amostra com 1.320 redações distribuídas em 119 temas diferentes. Este *corpus* foi extraído da base de dados disponível no repositório do Portal UOL (2019) de fonte aberta e também da plataforma desenvolvida por Pinho *et al.* (2020) com o objetivo de montar um repositório de redações corrigidas por diferentes professores e níveis de alunos. Além disso, todas as redações componentes do corpus escolhido já terem sido corrigidas por professores, possuindo, portanto, as notas e comentários atribuídos pelos docentes ao final do processo de correção.

Quanto à sua estrutura, a base de dados possui 12 colunas, conforme exemplificado no Quadro 12. As colunas consideradas para a condução dos primeiros experimentos foram: redação, tema e fuga. Sendo que a última (fuga) foi responsável pelo processo de classificação, consistindo no atributo alvo deste estudo.

Quadro 12: Estrutura da base de dados usada na fase inicial dos experimentos

REDAÇÃO	TEMA	TITULO	TEXTO MOTIVADOR	Nota comp 1	Nota comp 2	Nota comp 3	Nota comp 4	Nota comp 5	Total	Comentário avaliador	Fuga
Texto com a redação	Tema da Redação	Título fornecido pelo aluno	Texto fornecido na proposta falando sobre o assunto da temática	0 a 200	0 a 200	0 a 200	0 A 200	0 a 200	0 a 1000	Comentário do avaliador	Sim ou não

Fonte: Autora (2021)

A distribuição das notas na base de dados pode ser visualizada na Tabela 1, considerando-se que o objetivo inicial desta pesquisa é analisar se a redação fugiu ao tema proposto. Houve 230 redações classificadas como ‘fuga ao tema’, representando 17% da base de dados total considerada para o experimento.

Tabela 1: Distribuições das notas das redações

Redações Avaliadas		
Qtde	**Nota**	**Percentual**
266	0 pontos.	20,1%
	(Obs.: 230 redações foram classificadas como fuga ao tema).	
33	20 a 100 pontos	2,5%
172	120 a 300 pontos	13%%
293	320 a 500 pontos	22,2%
252	520 a 700 pontos	19,2%
192	720 a 900 pontos	14,5%
112	920 a 1000 pontos	8,5%
Total:1320		**Total: 100%**

Fonte: Autora (2021).

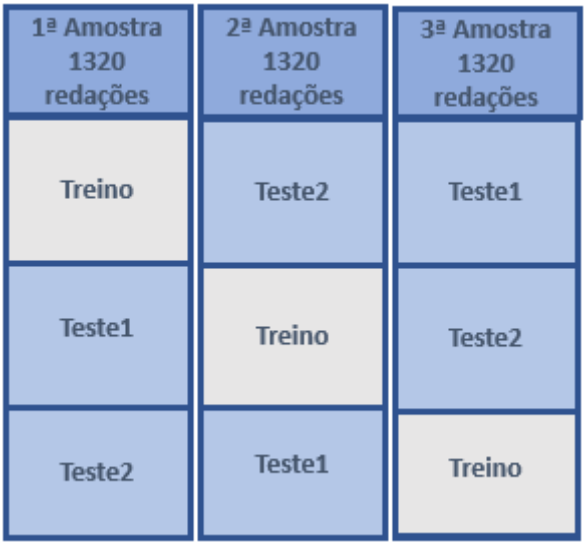
3.2.1 Distribuição dos dados

No processo de separação da base de dados houve uma preocupação em realizar o processo de validação cruzada, com o objetivo de evitar que apenas uma porção de dados de treino e teste pudessem ser muito parecidas. Isto porque, neste caso, quando houvesse testes com novos dados muitos diferentes do modelo treinado os resultados seriam insatisfatórios.

Trabalhar com diferentes distribuições possibilita diminuir os riscos de vícios de trabalhar com apenas uma amostra. Desta forma os dados foram separados em 3 conjuntos diferentes seguindo a ideia demonstrada na figura

37, na qual foram gerados 3 grupos diferentes para realização dos experimentos.

Figura 37: Demonstração da Validação Cruzada para o processo de treinamento e testes



Fonte: Autora (2021).

Estes grupos foram embaralhados para gerar as 3 amostras, assim foram realizados 3 experimentos com cada classificador. A proposta era utilizar os mesmos conjuntos de dados, tantos nos classificadores do Sklearn como na RCN. Desta forma, foi possível realizar a comparação dos resultados entre as técnicas, já que utilizaram as mesmas distribuições. Este processo permite evitar a variância, além de possibilitar entender se os experimentos realizados trouxeram a mesma média de resultados com diferentes combinações.

3.2.2 Arquitetura computacional

A arquitetura computacional estabelecida para os experimentos executados foi composta de uma plataforma experimental cuja principal linguagem utilizada foi o Python versão 3.7. O computador utilizado para processar os experimentos possui uma Placa de Vídeo GPU Geforce GTX 1660 Dual com 6GB de memória e interface de 192 bits, processador Intel Core i5 9400F 2.90GHz com 6 núcleos, 16GB da RAM e SSD de 480GB.

Os experimentos preliminares realizados empregaram as seguintes bibliotecas do Python:

- Spacy: para o processo de normalização e classificação de texto;
- NLTK: para auxílio na normalização do texto;
- Pandas: para a manipulação da base de dados das redações e análise dos dados;
- Numpy: para álgebra linear e operações com matrizes, quando gerado vetores para entrada na Rede Neural;
- Scikit-learn: para extração de atributos e algoritmos de aprendizado de máquina;
- Matplotlib: para a visualização dos dados por meio de gráficos e histogramas.
- TensorFlow: bibliotecas para trabalhar com a entrada na forma de texto, como *strings* de texto bruto ou documentos e conversão de valores discretos em numéricos.

3.3 Etapas da pesquisa e fluxo de atividades

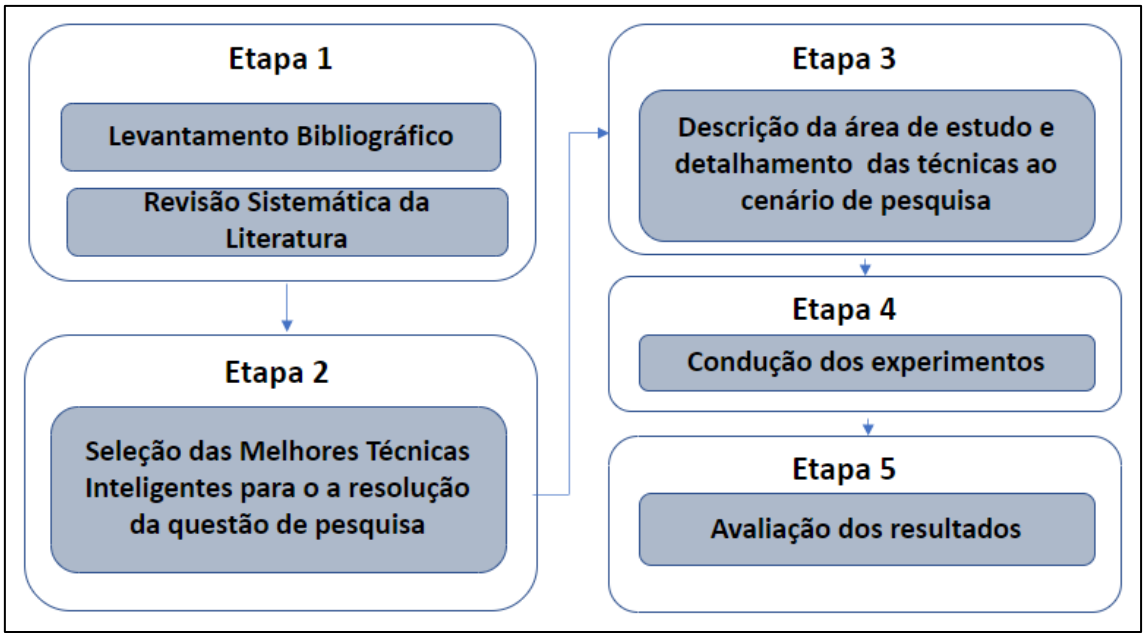
Com a finalidade de alcançar o objetivo deste estudo (desenvolver e validar uma solução para identificação e análise de fuga ao tema em textos dissertativos empregando o processamento de linguagem natural e mineração de textos baseados nas competências do ENEM) foram definidas as seguintes etapas do modelo conceitual de experimentos:

1. Revisão Sistemática da literatura empregando técnicas bibliométricas sobre o uso de técnicas inteligentes para avaliação de textos;
2. Seleção das melhores técnicas a serem empregadas para identificação de desvios de escrita em textos educacionais;
3. Descrição da área de estudo demonstrando o contexto ao qual o presente projeto de pesquisa está inserido, bem como o detalhamento das técnicas utilizadas nos experimentos;
4. Condução dos experimentos empregando-se as técnicas selecionadas;

5. Avaliação dos resultados dos experimentos com a identificação do desempenho de cada técnica e dos respectivos impactos para a identificação de fuga ao tema e, posteriormente, viabilização da criação de solução para correção inteligente automatizada de redações.

As etapas descritas anteriormente foram inseridas na Figura 38, que expõe cada etapa conduzida nesta pesquisa, desde o levantamento bibliográfico até a avaliação dos resultados.

Figura 38: Etapas da pesquisa



Fonte: Autora (2021).

Além das etapas definidas para a pesquisa também foi gerado fluxo de atividades para a elaboração dos experimentos, conforme apresentado na Figura 39. A sequência desenvolvida é constituída inicialmente pelas redações em língua portuguesa para avaliação de desvios estabelecidos pela competência 2 do ENEM (fuga ao tema).

No processo dos experimentos a serem conduzidos nesta pesquisa foram realizados os processos pertinentes à mineração de textos, processamento de linguagem natural e, por fim, treinamento na base utilizando-se as redes neurais convolucionais Multilayer Perceptron (MLP). Além das redes neurais, o treinamento também foi realizado com outras técnicas e algoritmos de classificação, tais como: Árvores de Decisão, Florestas

Aleatórias, Gradiente *Boosting, Ada Boost, Stochastic Gradiente Descent* (SGD) e *Support Vector Machines (SVMs)*. Os resultados esperados após a aplicação do modelo e das técnicas e algoritmos indicados volta-se à obtenção da classificação da fuga ao tema como positiva ou negativa. Este processo leva em consideração as redações sob análise para gerar a futura solução buscada.

Figura 39: Fluxo de Atividades

Fonte: Autora, 2021

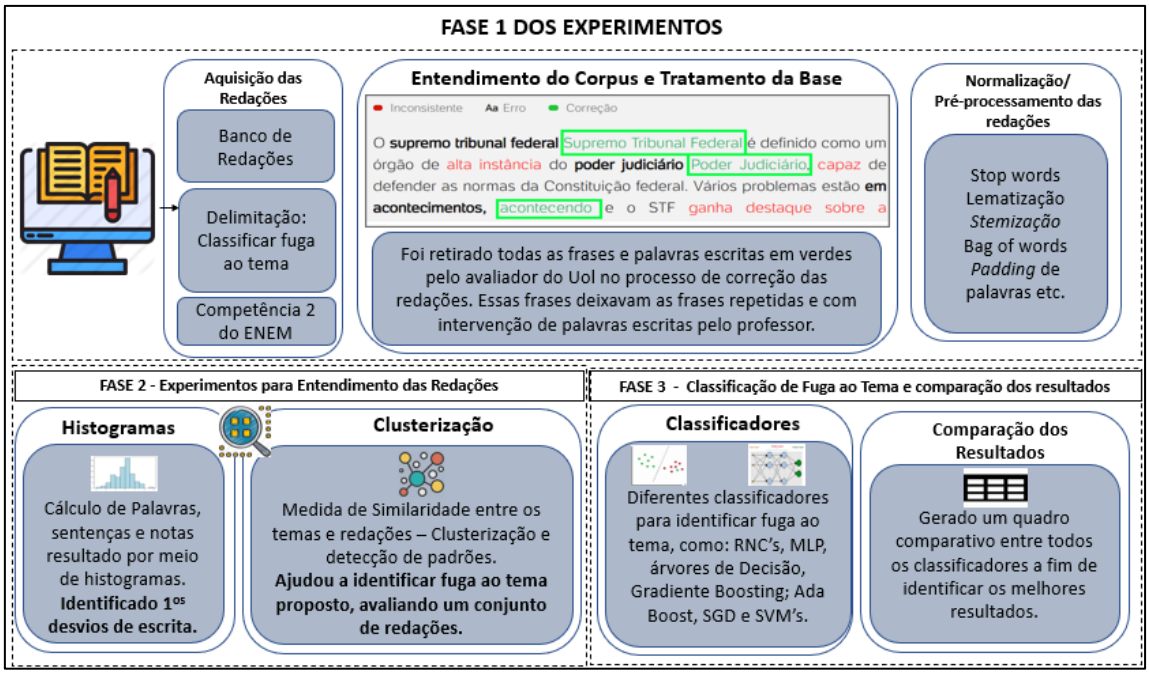
3.4 Detalhamento dos experimentos

Depois de entender o fluxo dos experimentos, buscou-se detalhar a sequência destes, conforme exposto na Figura 32. Analisando-se a fase 1 do esquema exposto na figura, entende-se que inicialmente será utilizada uma base de dados com 1.320 redações para identificar desvios de escrita na competência 2. Em seguida estas redações passaram por uma análise da pesquisadora, nesta etapa foi identificado que elas possuíam palavras e frases repetidas, estas, foram escritas no processo de correção pelo avaliador

(pessoa física) do Uol, todas as frases e palavras estavam destacadas pela cor verde, assim elas foram retiradas e os textos voltaram a sua forma original.

Dando sequência a análise da figura 40, após a retirada dos textos redundantes o próximo passo foi aplicar as primeiras técnicas de PLN e Mineração de Textos para normalizar os documentos e prepará-los para as primeiras análises gráficas, clusterização e elaboração dos modelos de treinamento.

Figura 40: Sequência de experimentos



Fonte: Autora (2021).

Na fase de normalização dos dados foi criada uma função em Python que faz toda a tratativa necessária. Nesta fase foi utilizada a biblioteca do Spacy e criada uma função para deixar todos os caracteres em minúsculo, realizar o processo de tokenização, lematização, remoção de stop words e retirada de caracteres especiais das redações do corpus em análise.

A figura 41 exemplifica como ficaram as frases após o processo de normalização anteriormente explicado. Neste exemplo, o tema de redação normalizado fica em minúsculo, sem caractere especial e sem stop words, que seriam as palavras 'ao' e 'ou'. Também pode ser visualizado o processo de lematização. Neste processo, efetivamente, há ação de deflexionar uma

palavra para determinar o seu lema, como pode ser visto na palavra ‘fumo’, que passou para ‘fumar’ e também na palavra ‘combate’, que se tornou ‘combater’.

Figura 41: Visualização dos textos após normalização

Tema redação sem normalizar: Combate ao fumo: autoritarismo ou dever do governo

Tema redação normalizado: combater fumar autoritarismo dever governar

Fonte: Autora (2021).

Na segunda fase foram avaliados os primeiros resultados para identificação de desvios de escrita e compreensão mais aprofundada da base trabalhada. Nesta fase já foi possível extrair informações importantes para o desenvolvimento de um sistema inteligente.

Na última fase dos experimentos foram então aplicadas as técnicas de classificação, que estabelecerão a classificação das redações sob análise. Em seguida, foi gerado um comparativo entre todas as técnicas aplicadas, de modo a evidenciar as particularidades da aplicação de cada uma delas no experimento.

Quadro 13: Detalhamento dos experimentos

Etapas / bibliotecas / experimentos realizados	
1)	Importação das bibliotecas necessárias para iniciar os experimentos, como *Spacy*, Pandas, *Numpy*, *MatplotLib, Scikit-learn* etc.
2)	Extração da base de dados com 1320 redações
3)	Pré-processamento dos dados, fazendo uma limpeza na base com o processo de normalização, remoção de stop words, remoção de caracteres especiais, s*temização* e *lematização*, *padding* de palavras;
4)	Antes da do processo de treinamento do modelo procurou-se colher informações a respeito dos textos gerando alguns histogramas para entender a base de dados.
5)	Ainda na fase de entendimento da base de dados e procurando formas de auxiliar os professores para saber as temáticas que os alunos possuem maior dificuldade de escrita, foi então realizado o processo de similaridade entre as redações e suas temáticas, em seguida foi comparado os clusters gerados com as notas inseridas pelos docentes.
6)	Para o processo de treinamento foram utilizadas as redações com notas superiores a 499 pontos e as redações com fuga ao tema. O total de redação para o processo de treino foram 628 redações e para o processo de teste foram 209 redações. (As redações de teste correspondem a 33% das redações treinadas)
7)	Após divisão da base em treino e teste o primeiro classificador testado foi a rede neural convolucional. Para a entrada da Rede neural no Processo de Treino, o primeiro passo foi transformar a variável target ‘fuga’ em valores numéricos 0 e 1, na qual a fuga ao tema passou a ser representado por 1 e a não fuga por 0.
8)	Na sequência todos os textos precisaram ser vetorizados passando pelo processo de *padding* (criado um vetor de números que representam cada palavra). Nesta fase, cada redação e proposta do tema foi vetorizado e trabalhados para ficarem na mesma dimensão. Para a rede Neural Convolucional, este processo ocorreu utilizando a

biblioteca ‘*tfds’* do ‘*tensorflow_datasets’.*
9) Uma vez que toda preparação dos dados foi realizada o próximo passo foi a configuração do modelo determinando as camadas de entrada, filtros, neurônios, a quantidade de documentos que seriam avaliadas para a atualização dos pesos e a quantidade de épocas para execução do treinamento.
10) Após o processo de Treinamento o próximo passo foi utilizar o modelo gerado para as 209 redações separadas para a fase de teste.
11) Após a aplicação do Primeiro Teste, foram geradas a matriz de confusão com os resultados e acurácia do experimento.
12) Outro teste foi realizado agora com as redações que ainda não tinham sido treinadas, ou seja, aquelas com notas inferiores a 500.
13) Para os próximos classificadores foi também foi realizado o processo de normalização das redações, assim como na rede neural convolucional foi necessário fazer nova representação dos valores textuais (discretos), os classificadores utilizados "só entendem números". Assim foi necessário converter os dados brutos, que estão em formato de texto, para um formato numérico. Isto deve acontecer antes passar as redações para o classificador, o processo de vetorização para os classificadores do *Scikit-learn* foi realizado utilizando o *TfidfVectorizer*.
14) As redações que já haviam sido divididas em treino e teste foram então configuradas em cada classificar do *Scikit-learn utilizando os* hiper parâmetros definidos na própria documentação da Biblioteca
15) Os resultados dos experimentos serão demonstrados no próximo capítulo.

Fonte: Autora (2020).

No Quadro 13 foram expostas as etapas, bibliotecas e passos realizados para cada fase dos experimentos executados.

4 APRESENTAÇÃO DOS RESULTADOS

Neste capítulo são apresentados os primeiros resultados da pesquisa. Para tanto, inicialmente será demonstrada a fase de normalização dos dados e, em seguida, a fase de análise inicial dos histogramas, detecção de padrões por meio da clusterização e os resultados das classificações de fuga ao tema resultantes dos experimentos.

4.1 Análise a partir de Histogramas

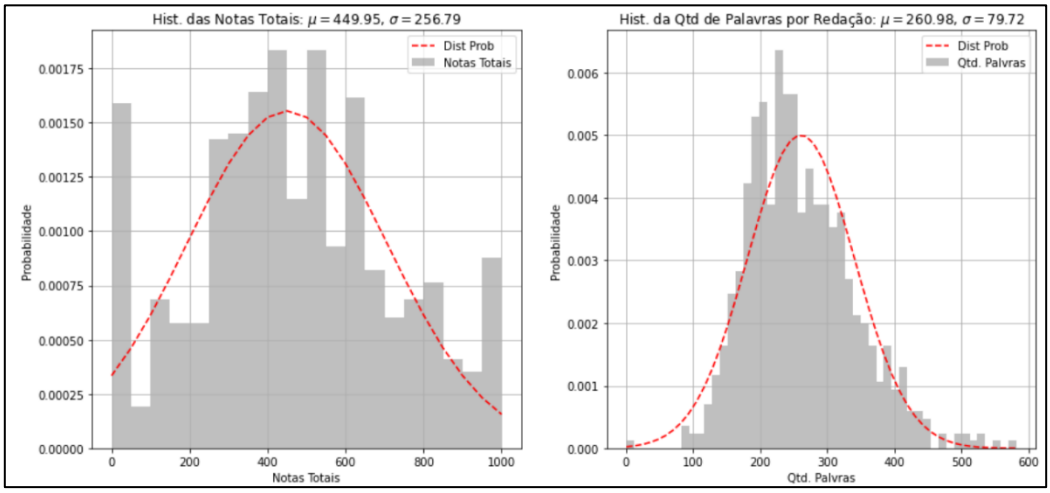
Os primeiros resultados expostos neste capítulo são demonstrados por meio de histogramas gerados após a normalização dos dados. A biblioteca utilizada para plotar os gráficos foi o Matplotlib, sendo que eles foram demonstrados como resultados, pois já indicaram possíveis desvios de escrita nas primeiras análises efetuadas.

É importante ressaltar que estes são resultados de avaliações corrigidas pelos professores do Uol, bem como o conteúdo das redações (contagem de palavras, sentenças e notas) avaliadas. Estes gráficos poderão ser inseridos de forma mais interativa em um futuro sistema para uso na área de educação voltado para o acompanhamento pelo professor. Os indicadores expostos podem facilitar o diagnóstico e evolução dos estudantes em relação à atividades de produção textual.

Inicialmente esses resultados não eram foco do estudo, já que o objetivo desta pesquisa é analisar fuga ao tema. Contudo, tais informações são úteis para comprovar que a aplicação de PLN e MT pode ser útil em outras competências a serem aprofundadas em futuros experimentos, além de trazer importantes observações acerca da base de dados avaliada.

No primeiro histograma da Figura 42 é possível visualizar o histórico das notas aplicadas, tendo sido identificadas as notas médias aferidas na faixa entre 256 e 449 pontos, em sua maioria. No segundo histograma da Figura 42 é possível visualizar a quantidade de palavras por redação, sendo que a maioria das redações se situou entre 200 e 400 palavras na produção textual elaborada pelos autores.

Figura 42: Análise das notas e da frequência de palavras por redação

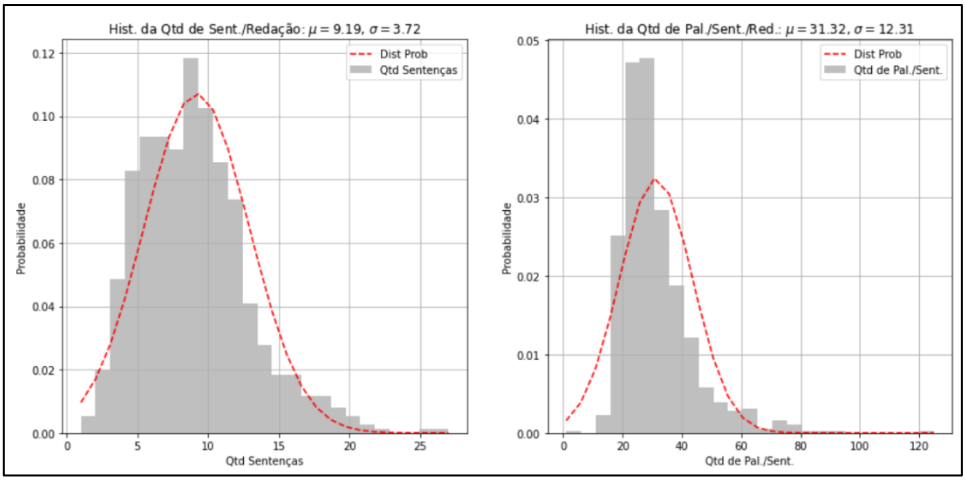


Fonte: Autora (2021).

No primeiro histograma da Figura 43 é demonstrada a quantidade de sentenças nas redações consideradas, com média de dez sentenças por redação. Entretanto, também foi identificado que há redações com menos de duas sentenças, o que aponta para textos com quantidade de linhas insuficiente para a avaliação da redação ou com problemas de falta de pontuação adequada.

O segundo histograma da Figura 43 apresenta um cruzamento entre as sentenças e palavras por sentenças, cujo resultado ficou entre 12 e 31 palavras por sentença. Porém, foram encontradas sentenças com até 120 palavras, o que também pode indicar o uso inadequado de pontuação, como vírgulas ou ponto final.

Figura 43: Quantidade de sentenças por redação e de palavras por sentença



Fonte: Autora, 2021

Algumas análises podem ser levantadas a partir dos histogramas expostos, preponderantemente voltadas aos critérios diferenciados que apontam para a possibilidade de estabelecimento de correlação entre a nota atribuída numa redação e a quantidade de sentenças. Tal achado poderia auxiliar o avaliador a responder às seguintes perguntas durante o processo de avaliação: Há um padrão entre a quantidade de sentenças e a nota aplicada pelo avaliador? Uma sentença muito longa indica a falta de pontuação?

Neste caso, podem ser agregadas outras técnicas com o objetivo de indicar ao avaliador em quais redações há a possibilidade de erros de pontuação e, consequentemente, problemas de coesão textual. Na aplicação deste experimento é possível avaliar hábitos de escritas, como por exemplo: regras de pontuação, repetição de palavras e sentenças e coesão textual. Tais achados estão em consonância com os estudos de Bazelato e Amorim (2010), Epstein e Reategui (2015) e Nobre e Pellegrino (2010).

4.2 Detecção de padrões

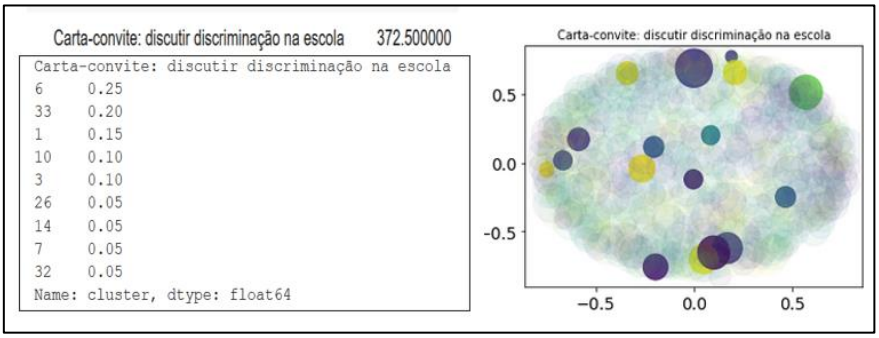
Num segundo momento, após o pré-processamento, tratamento dos dados e plotagem dos histogramas, foi realizado o processo para detecção de padrões entre as redações analisadas. Para tanto, este procedimento aplicou técnicas de clusterização, que podem auxiliar os professores a entenderem as temáticas que os alunos têm maior dificuldade na escrita, ou seja, aquelas nas quais os alunos obtiveram notas mais baixas.

Inicialmente todos os temas foram ordenados em ordem decrescente de acordo com o total da nota atribuída. Uma vez ordenados os temas, foram aplicados as técnicas de vetorização com a matriz TF-IDF da biblioteca do Scikit-learn para verificação da quantidade de termos por documento. Desta forma, foi possível compreender quais foram as palavras mais relevantes para cada tema de pesquisa.

Assim, todas as redações foram divididas em clusters, que são exatamente os temas trabalhados nas redações em análise. Em seguida foi realizado o processo de verificação de similaridade entre os temas, para saber

se as palavras destacadas como maior relevância num tema ficaram classificadas dentro de seu próprio cluster. Este processo mostrou os resultados apresentados na Figura 44, que expõe um exemplo da análise do cluster 33; na Figura 45, que expõe um exemplo da análise do cluster 31 e, por fim; na Figura 46 que expõe um exemplo da análise do cluster 20.

Figura 44: Análise do cluster 33

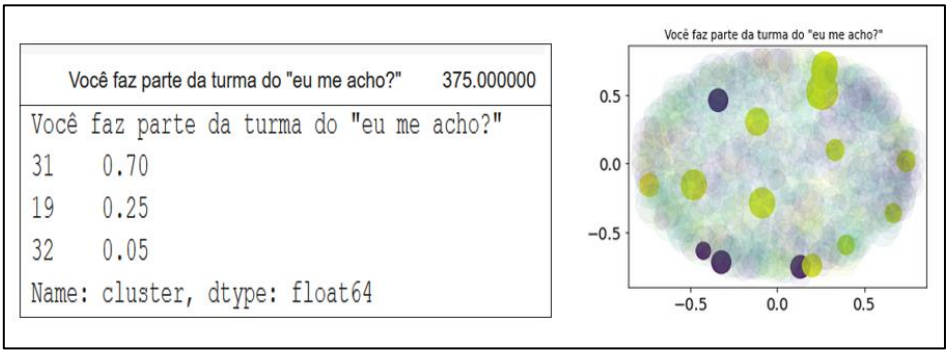


Fonte: Autora (2021)

Ao realizar a análise dos clusters que obtiveram as notas mais baixas entende-se que o professor pode utilizar essa solução para diagnosticar o quanto os alunos desviaram do tema solicitado. A título de exemplo, no cluster 33 (Figura 44) foi possível verificar que apenas 20% dos alunos versaram de forma totalmente adequada sobre o tema indicado. Tal resultado demonstra que houve elevado índice de fuga parcial ao tema, já que seus textos foram classificados em outros temas, que não o tema 33. A média das notas do tema 33 foi de apenas 372 pontos, sendo que a maior média identificada no corpus analisado foi de 640 pontos.

Este fator pode indicar um ponto de atenção para que o professor retome essa temática junto aos seus alunos, como prática de feedback da avaliação feita na correção dos textos.

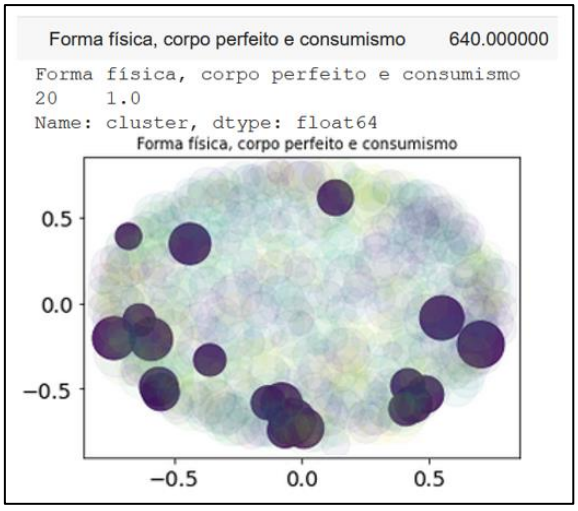
Figura 45: Análise do Cluster 31



Fonte: Autora (2021).

Já o cluster 31 (Figura 45) aponta que 70% dos alunos ficaram consoantes ao tema proposto, sendo que as redações dos demais 30% dos alunos também foram classificadas em outros clusters (temas). A imagem expõe à direita como os alunos também tiveram média baixa de notas, pois quanto menor o tamanho da elipse, menor a nota aplicada. Assim, o professor pode entender que também deve retomar esse tipo de tema com seus alunos, devido às notas baixas atribuídas.

Figura 46: Análise do cluster 20



Fonte: Autora, 2021

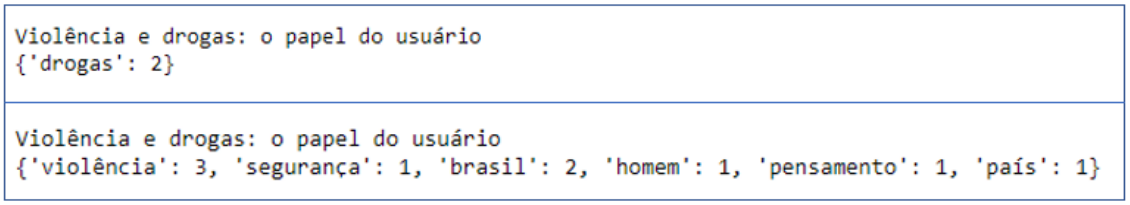
Ao procurar entender os resultados apresentados do cluster 20 (Figura 46), percebe-se que todos os alunos foram classificados dentro do mesmo cluster, indicando assim que não houve fuga parcial ou total ao tema proposto. Consequentemente, este cluster foi aquele no qual os alunos obtiveram a maior

média das notas. Neste caso, o docente não precisar (re)trabalhar esse tema novamente com seus alunos.

Outro experimento realizado com o objetivo de detectar padrões nas redações que indicavam fuga ao tema foi a aplicação da técnica de marcação gramatical, também conhecida como *pos tagging* da biblioteca do Spacy. Neste experimento foram extraídas tanto da proposta da redação (texto motivador e tema) como do da redação, os verbos, substantivos e adjetivos, considerados os termos que fornecem mais significado à escrita.

Após obtenção dos termos com maior valor semântico de cada texto, estes foram então comparados com o objetivo de encontrar as palavras que o aluno escreveu que tinham aderência à proposta de tema para a elaboração da redação.

Figura 47: Exemplo de Correlação entre Redação e Proposta da Redação – Redações com fuga ao tema



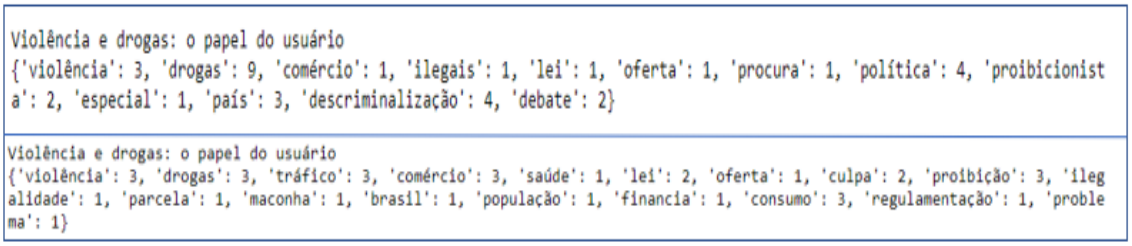
Fonte: Autora (2021).

A Figura 47 demonstra duas redações classificadas como fuga ao tema, assim como sua relação entre a redação elaborada pelo aluno e a proposta de tema intitulada “Violência e Drogas: o papel do usuário”. Ao analisar a imagem é possível identificar na primeira redação da Figura 47 que apenas a palavra ‘drogas’ apareceu duas vezes em toda a redação, demonstrando que o aluno não usou mais nenhuma palavra relativa à proposta que foi indicada pela instituição aplicadora da redação.

A segunda redação da Figura 47 demonstra que o aluno usou seis palavras que podem ter relação com a temática previamente indicada, palavras estas que se encontravam na proposta da redação. Porém, foi possível verificar que o aluno não usou as palavras ‘drogas’ ou ‘usuário’, bem como a maioria das outras palavras que tiveram relação com a proposta e que poderiam ter sido empregadas em diferentes contextos, não apenas nessa temática.

Na Figura 48 são exemplificadas duas redações classificadas sem fuga ao tema, tendo a primeira recebido nota 700 e a segunda recebido nota 900.

Figura 48: Exemplo de Correlação entre Redação e Proposta da Redação – Redações sem fuga ao tema



Fonte: Autora (2021).

Ao analisar os resultados das duas redações expostas na Figura 48, que possuem a mesma temática abordada na Figura 47 anterior, foi possível identificar que, além de ter muitas palavras relacionadas ao texto motivador, tais palavras têm maior relação com o tema, a exemplo de: ‘descriminalização’, ‘ilegais’, ‘drogas’, ‘política e ‘tráfico’, dentre outras ocorrências.

Tais resultados podem ser disponibilizados para o professor em conjunto com a indicação da probabilidade de fuga ao tema, que será exposta nos classificadores dos próximos tópicos. Desta forma, além da indicação do percentual, o docente terá também a indicação das palavras que mais aderiram à proposta da redação.

Essa primeira fase dos experimentos proporcionou informações relevantes aos professores sobre o desempenho de suas turmas e alunos, identificando aqueles que estão com maiores dificuldades na produção textual, bem como temas com maior dificuldade de assimilação por parte dos estudantes. Esta solução pode ser facilmente ser implementada num sistema de gerenciamento de turmas, proporcionando assim benefícios tanto para o docente, quanto para a escola.

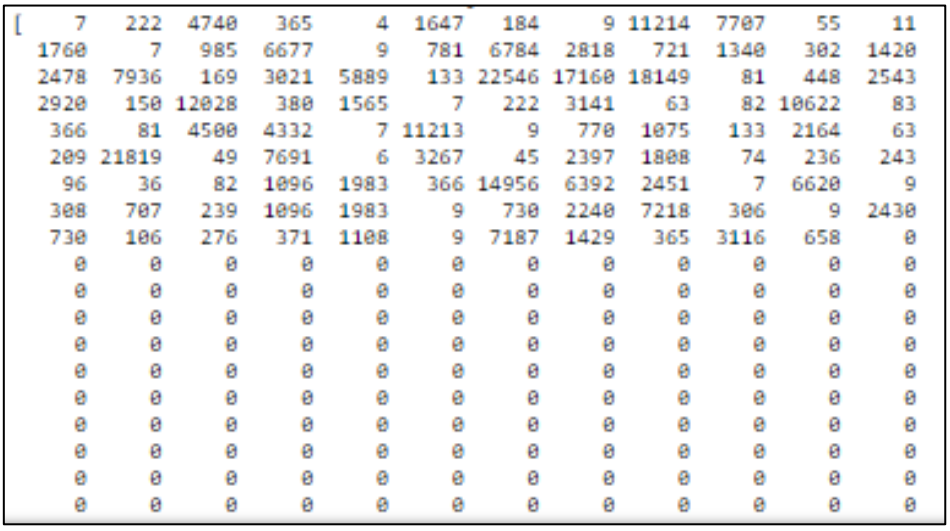
4.2.1 Preparação da base para os classificadores

Para as classificações uma nova fase foi iniciada. Após o processo de normalização das redações ainda foi necessário adaptar os textos aos classificadores, uma vez que estes não reconhecem valores discretos (textos).

Assim sendo, foi necessário aplicar o processo de padding e vetorização dos textos antes de iniciar o treinamento do modelo. Uma melhor compreensão deste conceito pode ser encontrada no exemplo indicado no tópico Bag of words e padding de palavras no referencial teórico desta dissertação, tópico 2.3.2.

Na Figura 49 é demonstrada uma redação convertida em números. Neste processo cada palavra assume uma codificação. Todos os classificadores considerados nos experimentos utilizaram dessa técnica para treinar seus modelos e, posteriormente, realizar a classificação de fuga ao tema nas redações.

Figura 49: Padding das sentenças



```
[    7   222  4740   365     4  1647   184     9 11214  7707    55    11
  1760     7   985  6677     9   781  6784  2818   721  1340   302  1420
  2478  7936   169  3021  5889   133 22546 17160 18149    81   448  2543
  2920   150 12028   380  1565     7   222  3141    63    82 10622    83
   366    81  4500  4332     7 11213     9   770  1075   133  2164    63
   209 21819    49  7691     6  3267    45  2397  1808    74   236   243
    96    36    82  1096  1983   366 14956  6392  2451     7  6620     9
   308   707   239  1096  1983     9   730  2240  7218   306     9  2430
   730   106   276   371  1108     9  7187  1429   365  3116   658     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
```

Fonte: Autora (2021).

A dimensão dos vetores criados foi de 510 palavras, o que significa que após o processo de normalização efetuado, a maior redação ou texto motivador encontrado possuía 510 palavras. Assim, o algoritmo completa os vetores com zero para deixá-los todos na mesma dimensão antes de começar o treinamento dos próximos modelos.

Uma vez que os dados foram tratados, o próximo passo é aplicar os modelos de classificação. Para os próximos classificadores foram gerados os modelos aplicando-se as duas bases de teste separadas, assim como ocorreu com na RNC. Na sequência as métricas de avaliação utilizadas foram a acurácia e matriz de confusão de cada modelo gerado.

4.3 Classificação de Fuga ao Tema usando as RNCs

Os resultados deste classificador foram dispostos de forma separada dos posteriores, pois ele se utiliza de outras técnicas para sua aplicação, utilizando a biblioteca do Spacy, conforme exemplificada no referencial teórico, tópico 2.3.3.1 item (b) desta dissertação, além de utilizar um método de vetorização diferente dos classificadores do *Scikit Learn*.

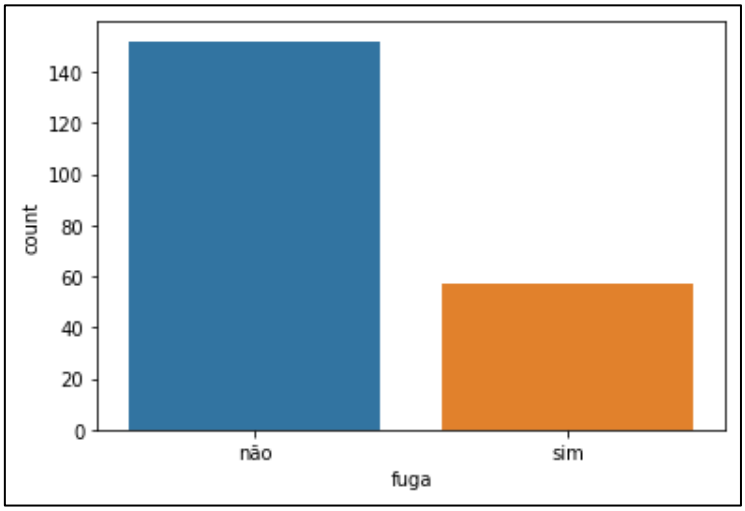
Na terceira fase dos experimentos, após a realização de todo o processo de pré-processamento das redações por meio das normalizações, remoção de stop words, remoção de caracteres especiais, stemização e lematização e padding de palavras, conforme exposto no tópico de métodos e materiais, a base foi finalmente separada entre conjunto de treinamento e conjunto de teste.

Para o treinamento da RNC foram utilizadas 628 redações, o que equivale a 47% das redações da base em análise. A escolha das redações para treino e teste do modelo levou em consideração a nota real atribuída pelos professores a cada redação. Os melhores resultados foram obtidos quando as redações tinham notas superiores a 499 pontos. Este critério também levou em consideração que essa normalmente é a margem que as universidades utilizam como critério de eliminação dos candidatos (CRUZEIRO DO SUL, 2020).

A configuração do modelo concebido utilizou a biblioteca do keras do tensorflow, com os seguintes parâmetros: emb_dim = 200; nb_filters = 700; ffn_units = 1000; batch_size = 32; dropout_rate = 0.2; nb_epochs = 40. Os valores iniciais dos parâmetros foram fornecidos em um curso oferecido por Granatyr (2020) na Plataforma IA Expert. Nos experimentos seguintes os parâmetros foram ajustados de acordo com os resultados apresentados.

Após o processo de treinamento e a geração do modelo, o próximo passo foi o teste com as 209 redações restantes, que equivalem a 33% da base de treinamento. O balanceamento entre redações que fugiram e não fugiram ao tema pode ser visualizado na Figura 50. As redações que fugiram ao tema equivaleram a 17% do total analisado. Na fase de treinamento também foi trabalhado o mesmo percentual de redações com fuga ao tema.

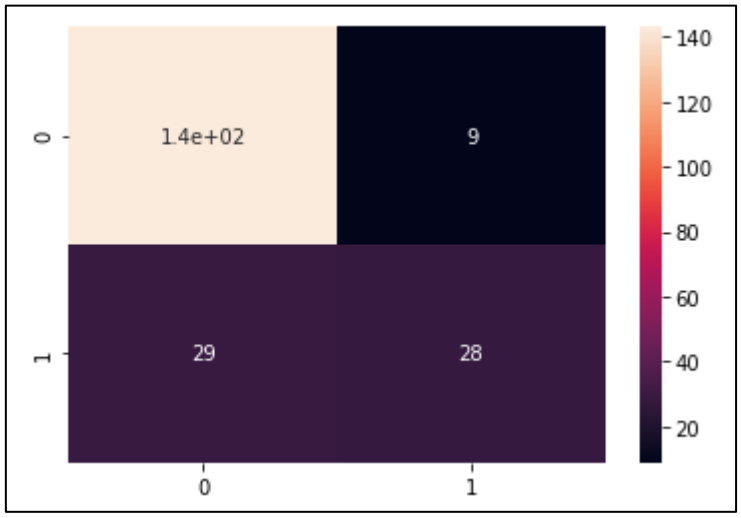
Figura 50: Distribuição da base de teste



Fonte: Autora, 2021

Após a aplicação da base de teste no modelo, os primeiros resultados podem ser visualizados na Figura 51, na qual está representada a Matriz de Confusão, com os acertos e erros do modelo.

Figura 51: Matriz de Confusão – Primeira base de Teste



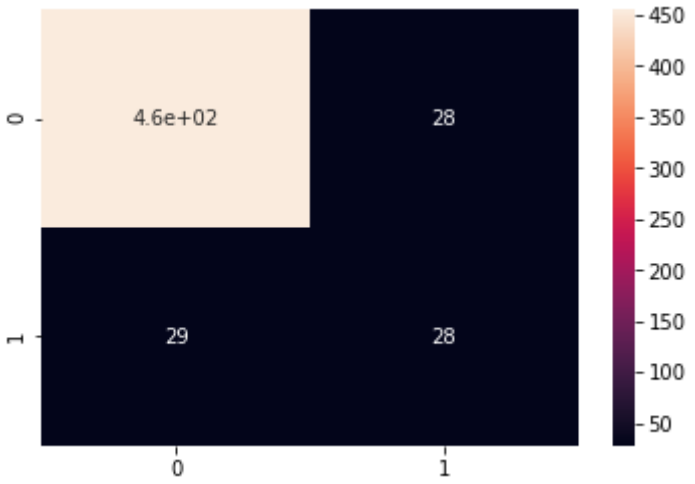
Fonte: Autora (2021).

Ao realizar a análise da Figura 51 foi identificado que das 152 redações que não tiveram fuga ao tema, o modelo classificou como fuga 9 redações indicando 5,9% de erro, valor este que representa os falsos positivos. Na segunda linha da matriz de confusão havia 57 redações com fuga ao tema. Destas, o algoritmo classificou como fuga 28 redações, o que equivale a 49% de acerto, este valor representa os verdadeiros positivos.

Neste primeiro teste a acurácia (taxa de acertos) ficou em aproximadamente 81,8%, comparando-se os resultados da solução implementada nos experimentos com os resultados reais oriundos das correções efetuadas pelos professores.

Um segundo teste foi realizado com as redações com notas inferiores a 500, ou seja, normalmente aquelas que são desconsideradas por grande parte das universidades para aprovação dos candidatos. Os resultados podem ser visualizados na Figura 52. Este novo teste foi realizado com as outras 540 redações que não foram utilizadas na fase de treinamento.

Figura 52: Matriz de Confusão - Segunda base de Teste



Fonte: Autora, 2021.

No segundo teste, realizado com as redações com notas inferiores, obteve-se acurácia de 89,4%. Outra análise foi realizada a partir das previsões para saber se era possível tirar alguma informação dos falsos positivos incorridos, já que o sistema informou que 28 redações foram classificadas como fuga ao tema, mas na verdade não apresentavam fuga ao tema de fato, a taxa de erros em relação aos falsos positivos da segunda base de teste foi de 5,7%.

A Tabela 2 consolida os resultados obtidos no processo de classificação da RNC. Apesar de ter obtido resultados de acurácia superiores a 80% nas duas bases de dados, a maior atenção está em relação aos falsos positivos, ou seja, quando o sistema afirma que a redação teve fuga ao tema de forma

equivocada. A taxa de erro ficou entre 5,9% e 5,7% na classificação de fuga, quando na verdade não houve. Ainda assim é importante avaliar se estas redações podem ter fugido parcialmente à temática proposta, o que também irá subtrair nota do estudante.

Tabela 2: Resultados consolidados da Matriz de Confusão - RNC

PRIMEIRA BASE DE TESTE COM REDAÇÕES DE NOTAS SUPERIORES A 500 e 57 REDAÇÕES COM FUGA AO TEMA					
CLASSIFICADOR *RNC*	ACURÁCIA	VP (Verdadeiro Positivo)	FN (Falso Negativo)	FP (Falso Positivo)	VN (Verdadeiro Negativo)
	81,8%	49,1%	50,9%	5,9%	94,1%
SEGUNDA BASE DE TESTE COM REDAÇÕES DE NOTAS INFERIORES A 500 e 7 REDAÇÕES COM FUGA AO TEMA					
CLASSIFICADOR *RNC*	ACURÁCIA	VP (Verdadeiro Positivo)	FN (Falso Negativo)	FP (Falso Positivo)	VN (Verdadeiro Negativo)
	89,4%	49,1%	50,9%	5,7%	94,3%

Fonte: Autora, 2021

A Tabela 3 indica os resultados de outras métricas de avaliação, tais como *precision, recall* e *F1-score* para cada uma das classes analisadas: ‘1’ para fuga ao tema e ‘0’ para não fuga.

Tabela 3: Resultados consolidados Precisão, Recall e F1-Score - RNC

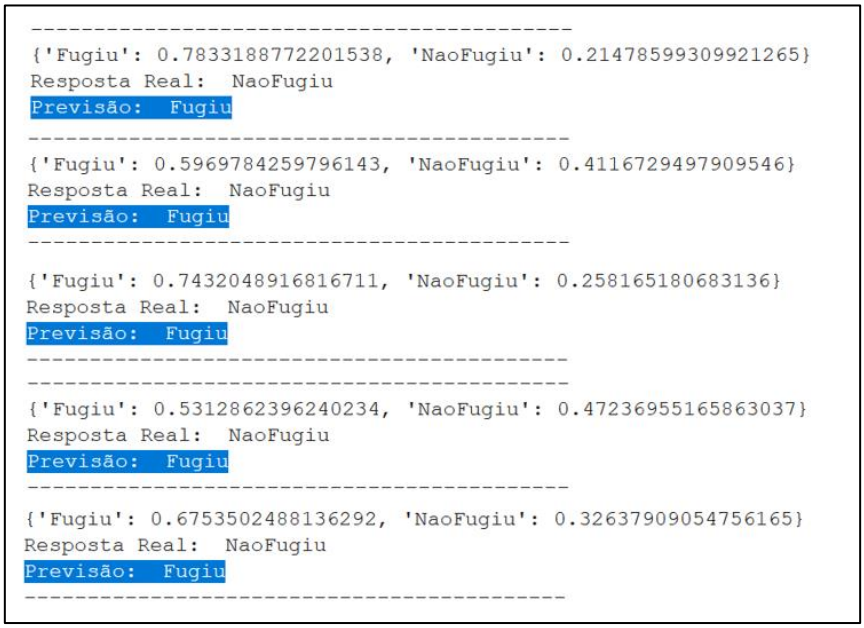
PRIMEIRA BASE DE TESTE COM REDAÇÕES DE NOTAS SUPERIORES A 500 e 57 REDAÇÕES COM FUGA AO TEMA				
		Precisão	Recall	F1-Score
CLASSIFICADOR *RNC*	0	94%	94%	94%
	1	50%	49%	50%
SEGUNDA BASE DE TESTE COM REDAÇÕES DE NOTAS INFERIORES A 500 e 7 REDAÇÕES COM FUGA AO TEMA				
		Precisão	Recall	F1-Score
CLASSIFICADOR *RNC*	0	94%	94%	94%
	1	50%	49%	50%

Fonte: Autora, 2021

Os resultados apresentados da precisão indicam todas as classificações de classe positivo que o modelo fez, e quantas estão corretas. Já o Recall dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas e, por fim, o F1-Score faz uma média harmônica entre as outras duas. Essas métricas também apontam um maior grau de acerto para a classe negativa das redações, aquelas que não tinham fuga ao tema.

Na Figura 53 é indicado que o algoritmo aplicado também ficou em dúvida. Um padrão verificado foi que os valores inferiores a 80% indicando que havia fuga ao tema, na verdade não deveriam ser classificados dessa forma. Assim, o professor pode utilizar dessas previsões para facilitar seu processo de avaliação das redações.

Figura 53: Exemplos de Previsões de falsos positivos



Fonte: Autora, 2021

A figura 53 ainda explicita a importância de disponibilizar estes dados ao professor, pois desta forma, mesmo que o sistema classifique a fuga, o avaliador humano conseguirá identificar se algoritmo ficou na dúvida e ainda associar os resultados demonstrados no tópico anterior, quando é possível identificar as palavras que mais aderiram a proposta.

Após o processo de classificação utilizando as redes neurais convolucionais, o próximo passo foi testar outros classificadores, conforme demonstrado nos próximos tópicos.

4.4 Classificação de Fuga ao Tema aplicando outros classificadores selecionados

Nesta subseção são apresentados os resultados referentes aos classificadores de aprendizado de máquina da biblioteca do Scikit Learn.

Os classificadores selecionados foram explicitados no tópico 2.3.3.1 desta dissertação, sendo estes: MLP (*MultiLayer Perceptron*), árvores de decisão; Florestas Aleatórias, *Gradiente Boost, Ada Boost, Stochastic Gradiente Descent (SGD) e Support Vetor Machines* (SVM's).

4.4.1 Resultados dos Classificadores do Scikit Learn

Neste tópico são demonstrados os resultados apresentados nos classificadores de AM do Scikit Learn, após nova normalização das redações. Para todos os classificadores foram utilizados a mesma base da RNC. As redações foram divididas em treino e teste, utilizando os mesmos critérios de divisão da RNC. Deste modo foi possível comparar os resultados obtidos.

A primeira métrica utilizada para avaliar os resultados foi a matriz de confusão, avaliando como cômputo a classe real e a classe prevista, com uso dos rótulos 'P – Positivo' e 'N – Negativo'. Outra métrica avaliada foi a acurácia, que mede o percentual de acertos em relação a quantidade de ocorrências, as outras métricas avaliadas foram precisão, recall e f1-score, assim como no classificador anterior.

A Tabela 4 expõe os resultados de cada classificador utilizando o mesmo critério na RNC, ou seja, notas atribuídas pelo avaliador humano superiores a 500 pontos.

Tabela 3: Classificadores Scikit Learn – Consolidado da Matriz de Confusão

PRIMEIRA BASE DE TESTE COM REDAÇÕES DE NOTAS SUPERIORES A 500 e 57 REDAÇÕES COM FUGA AO TEMA					
CLASSIFICADOR	ACURÁCIA	VP (Verdadeiro Positivo)	FN (Falso Negativo)	FP (Falso Positivo)	VN (Verdadeiro Negativo)
MLPClassifier	78%	33%	67%	4,6%	95,4%
DecisionTreeClassifier	74,6%	14%	86%	2,6%	97,4%
RandomForestClassifier	72,7%	0%	100%	0%	100%
SGDClassifier	78%	47%	53%	11%	89%
SVM (SVC)	72,2%	0%	100%	0%	100%
GradientBoostingClassifier	74,6%	51%	49%	16%	84%
AdaBoostClassifier	77%	44%	56%	11%	89%

Fonte: Autora (2021).

Para identificar os melhores resultados da primeira base de teste deve-se levar em consideração a acurácia aliada aos VP e VN, além da classificação que apresentou o menor erro do FP, ou seja, aquele resultado que apontou que

a redação teve fuga ao tema quando, na verdade não tinha. Isto porque este item é o que anularia a prova do aluno, devendo assumir, portanto, a menor taxa de erro.

Assim, levando-se em consideração tais informações, o classificador que apresentou os melhores resultados foi o *GradientBoostingClassifier,* com maior acerto de fuga ao tema (VP) de 51%, taxa de erro dos FP de 16% e acurácia de 74,6,7%. Outros classificadores que tiveram bons desempenhos foram o *SGDClassifier* e o *MLPClassifier*, com as seguintes taxas de erro 11% e 4,6%, respectivamente, apresentando os dois classificadores a acurácia de 86%.

A Tabela 5 expõe os resultados da segunda base de teste, ou seja, aquelas que tiveram notas atribuídas pelo avaliador humano abaixo de 500 pontos.

Tabela 4: Resultados dos Classificadores Scikit Learn – Segunda base de teste

SEGUNDA BASE DE TESTE COM REDAÇÕES DE NOTAS INFERIORES A 500 e 57 REDAÇÕES COM FUGA AO TEMA					
CLASSIFICADOR	ACURÁCIA	VP (Verdadeiro Positivo)	FN (Falso Negativo)	FP (Falso Positivo)	VN (Verdadeiro Negativo)
MLPClassifier	91%	33%	67%	3,3%	96,7%
DecisionTreeClassifier	82%	14%	86%	9%	91%
RandomForestClassifier	89,4%	0%	100%	0%	100%
SGDClassifier	86,6%	47%	53%	8,7%	91,3%
SVM (SVC)	89,4%	0%	100%	0%	100%
GradientBoostingClassifier	72,9%	51%	49%	24%	76%
AdaBoostClassifier	71,2%	44%	56%	25%	75%

Fonte: Autora (2021).

Os resultados relativos àquelas redações com notas abaixo de 500 pontos expostos na Tabela 5 expuseram a mesma ordem de hierarquia dos classificadores com melhor desempenho. Os resultados desta tabela, apesar de tratar as redações de notas inferiores, obtiveram resultados similares à base de dados anterior. A maior diferença foi verificada nos FP que, neste caso, obteve como classificadores com menor erro nos Falsos Positivos o MLP e *SGDClassifier, com respectivamente* 3,3% e 8,7%.

A Tabela 6 demonstra os resultados de outras métricas de avaliação para auxiliar no entendimento dos resultados do modelo.

Tabela 5: Métricas Precisão, Recall e F1-Score

Consolidado de Classificadores Scikit Learn

MLPClassifier	Precisão	Recall	F1-Score
Classe 0 (não fuga)	79%	95%	87%
Classe 1 (Fuga ao tema)	73%	33%	46%
DecisionTreeClassifier	*Precisão*	*Recall*	*F1-Score*
Classe 0 (não fuga)	75%	97%	85%
Classe 1 (Fuga ao tema)	67%	14%	23%
RandomForestClassifier	*Precisão*	*Recall*	*F1-Score*
Classe 0 (não fuga)	*73%*	*100%*	*84%*
Classe 1 (Fuga ao tema)	*0%*	*0%*	*0%*
SGDClassifier	*Precisão*	*Recall*	*F1-Score*
Classe 0 (não fuga)	*82%*	*89%*	*85%*
Classe 1 (Fuga ao tema)	*61%*	*47%*	*53%*
SVM (SVC)	*Precisão*	*Recall*	*F1-Score*
Classe 0 (não fuga)	*73%*	*100%*	*84%*
Classe 1 (Fuga ao tema)	*0%*	*0%*	*0%*
GradientBoostingClassifier	*Precisão*	*Recall*	*F1-Score*
Classe 0 (não fuga)	*82%*	*84%*	*83%*
Classe 1 (Fuga ao tema)	*54%*	*51%*	*52%*
AdaBoostClassifier	*Precisão*	*Recall*	*F1-Score*
Classe 0 (não fuga)	*81%*	*89%*	*85%*
Classe 1 (Fuga ao tema)	*61%*	*44%*	*51%*

Fonte: Autora (2021).

Ao avaliar os resultados da Tabela 6, os classificadores com melhores resultados e que mantiveram valores similares nas métricas de precisão, recall e F1-Score foram o *SGDClassifier e GradientBoostingClassifier*. Isto porque

para a Classe Negativo (não fuga) ficaram acima de 80% e os acertos da classe Positiva (fuga ao tema) mantiveram a média acima de 50%.

O Classificador MLP manteve a ‘precisão’ nas duas classes (positivo e negativo) acima de 70%. Já no item ‘recall’, para a classe positivo ele obteve um acerto de apenas 33%. O F1-Score fez uma média harmônica entre os resultados dos dois anteriores e para a classe positivo também se manteve baixo de 50%.

4.5 Avaliação dos Resultados

Levando-se em consideração os experimentos realizados nesta pesquisa, iniciando por aqueles que ainda não utilizaram os classificadores, já foi possível identificar algumas possibilidades para a avaliação da escrita e/ou argumentação dos alunos em redações. Partindo-se deste princípio, entende-se que é possível mensurar o número de redações aderentes à proposta de tema informada, o que pode trazer importante conhecimento ao avaliador ou docente em relação à evolução dos alunos na produção textual.

Outro resultado importante ao identificar padrões entre a redação e a proposta do tema foi a possibilidade de identificar as palavras mais aderentes a temática fornecida.

Estes experimentos ainda podem ser aprimorados e aplicados num sistema para avaliar a evolução dos alunos do decorrer de seus estudos acadêmicos, proporcionando assim ao docente conhecer as dificuldades dos alunos numa turma.

Na avaliação dos resultados obtidos nos classificadores da RNC foi identificado maior ganho em relação aos classificadores do Scikit Learn, tanto em relação à acurácia, quanto em relação aos resultados de falsos positivos, métricas de ‘precisão’, 'recall' e 'F1-Score'. Apesar dos melhores resultados terem sido obtidos com a RNC, deve-se destacar os bons resultados obtidos com os classificadores Scikit Learn.

Na avaliação do melhor modelo é importante entender como eles seriam aplicados numa solução real. Dessa forma, o percentual de erro do FP deve

ser mínimo, assim como o percentual de VP deve ser alto, o que significa que o sistema identificou a fuga ao tema, proposta desta pesquisa de dissertação.

Para analisar os resultados foram desconsiderados os algoritmos que não conseguiram identificar a fuga ao tema, ou seja, aqueles que apresentaram a taxa de VN de 100%, significando que estes não conseguiram atingir o objetivo proposto da pesquisa, quais sejam: *RandomForestClassifier* e SVM (SVC).

Na sequência, o algoritmo que teve a melhor acurácia foi a RNC, com resultados de até 89% de acurácia e taxa de FP de apenas 5,7%. Contudo, caso se avalie a taxa de VP, aquela em que o algoritmo acertou a fuga ao tema, os melhores resultados ocorreram com o *GradientBoostingClassifier,* com 51% e acertos na classe Positivo, contudo, a sua taxa de Falsos Positivos ficou em média de 20%. Outro Classificador que obteve resultados melhores em relação aos Falsos Positivos foi a MLP, com no máximo 4,6% de erro, taxa de VP de 33% e acurácia entre 78% e 90%.

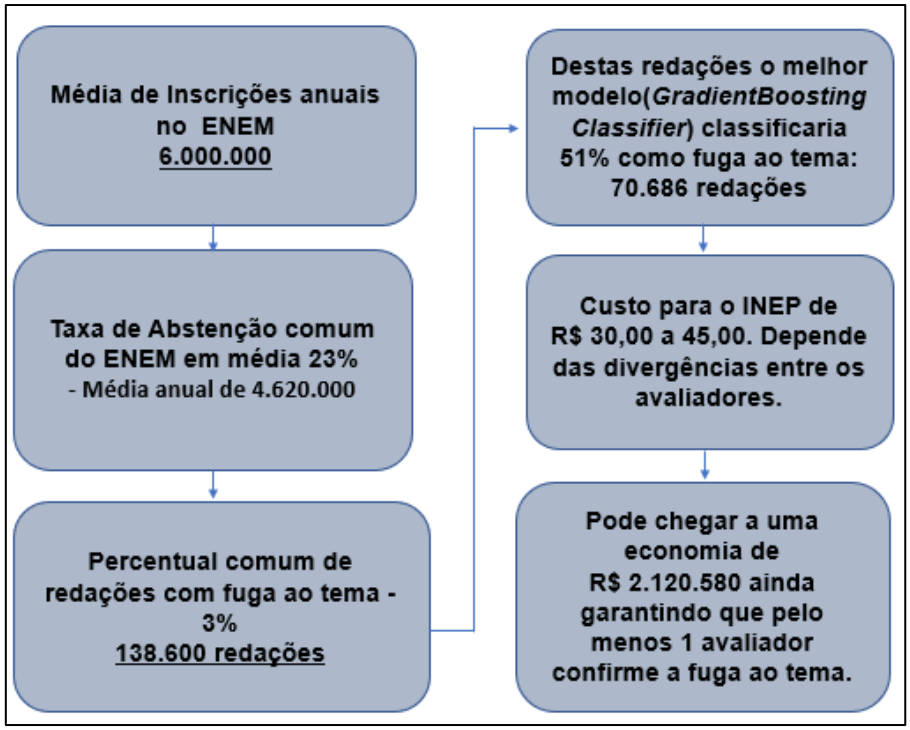
Contudo, tais métricas ainda podem ser melhoradas quando forem aplicadas a partir de uma base de dados maior e com mais exemplos de fuga ao tema, em especial a RNC. Segundo Rodrigues (2018), ela obtém melhores resultados com uma grande base de exemplo.

Levando-se em consideração essas taxas e a quantidade de redações que são corrigidas anualmente pelo ENEM, em média 4 milhões de acordo com o portal do INEP (2019), a solução apresentada já poderia indicar de antemão as redações que provavelmente fugiram ao tema, economizando assim o tempo e o custo da correção da redação. Já em sala de aula, a solução delineada pode ajudar a agilizar a devolutiva do professor e tornar seu trabalho menos desgastante ao ter um sistema que indica possíveis falhas de escrita.

A Figura 54 exemplifica uma possível economia de aproximadamente R$ 2 milhões de reais com a aplicação deste primeiro modelo testado. O modelo ainda pode ser melhorado com mais exemplos de fuga ao tema, contudo, já com esta solução implementada é percebido que mesmo a redação sendo avaliada por um corretor humano para garantir que realmente haja fuga ao

tema nas redações separadas pelo modelo, a instituição INEP ainda poderia economizar com o gasto de até dois avaliadores humanos.

Figura 54: Simulação de economia com modelo aplicado ao ENEM



Fonte: Autora, 2021

Para a sala de aula o ganho é em relação ao tempo e menos desgaste do professor nas avaliações dos textos. A ideia é aplicar o modelo apresentado numa ferramenta que indique ao professor a probabilidade de fuga ao tema e ainda trazer as palavras mais aderentes a proposta.

Nas pesquisas realizadas não foram encontrados outros autores que utilizaram da mesma técnica utilizada nesta pesquisa, por meio da classe ‘fuga ao tema’. A pesquisa mais próxima deste trabalho foi a de Passero (2018), que analisou especificamente fuga ao tema. O autor obteve ótimos resultados com uma acurácia de 96,76% e FP de 4,24%. Contudo o autor não disponibilizou em sua pesquisa a taxa de VP, aquelas que identificaram a fuga ao tema, fator crucial desta pesquisa.

A plataforma CIRA, desenvolvida em 2020 numa pesquisa realizada pela USP São Carlos (USP, 2021) também tem o objetivo de corrigir redações

automaticamente, contudo o autor excluiu em sua análise a verificação de fuga ao tema. Assim, na aplicação que está disponível para uso há uma grande taxa de erros quando avaliadas as redações com notas consideradas baixas. Nos testes realizados, redações que deveriam ter nota 0 (zero) por fuga ao tema, foram avaliadas com notas superiores a 400.

Outra pesquisa como a de Ramisch (2020) também avaliou redações, contudo sua proposta foi encontrar problemas de desvios sintáticos, atingindo uma acurácia de 75,6% de acerto.

Desta forma, entende-se que os resultados apresentados nesta pesquisa trazem importantes contribuições à evolução do estudo desta área de pesquisa acadêmica. Assim, a partir dos resultados preliminares aqui expostos é possível vislumbrar as primeiras prerrogativas de benefícios da solução ora desenvolvida e a ser validada para auxílio ao trabalho de docentes e avaliadores durante o processo de correção de textos produzidos por alunos ou candidatos.

5 CONCLUSÕES

A aplicação das técnicas mencionadas nos experimentos executados nesta pesquisa buscou responder a indicação de quais das técnicas de IA comparadas apresentam melhores resultados para a identificação de fuga ao tema em redações. Assim, entende-se que esta pesquisa respondeu à questão-problema formulada, uma vez que após a aplicação de diferentes classificadores foi possível identificar aqueles que trazem melhores resultados ao identificar a fuga ao tema nas redações, indicando ainda o percentual de acerto do modelo concebido nesta pesquisa.

Os experimentos trouxeram resultados promissores tanto nas RCNs como nos classificadores do *Scikit Learn*. O algoritmo que teve a melhor acurácia foi a RNC, com resultados de até 89,4% de acurácia e taxa de FP de apenas 5,7%. Contudo, caso se avalie a taxa de VP, aquela em que o algoritmo acertou a fuga ao tema, os melhores resultados ocorreram com o *GradientBoostingClassifier*, com 51% de acertos na classe Positivo, não obstante sua taxa de Falsos Positivos tenha sido, em média, de 20%. Outro classificador que obteve resultados melhores em relação aos Falsos Positivos foi a MLP, com no máximo 4,6% de erro, taxa de VP de 33% e acurácia entre 78% e 90%

Desta forma, o objetivo proposto de comparar diferentes técnicas de IA para classificação de fuga ao tema em textos e identificar aquelas que trouxeram melhores resultados foi alcançado. A solução desenvolvida nesta dissertação possibilita a geração de informações e conhecimento úteis aos avaliadores de textos educacionais para a identificação de desvios de escrita e possível fuga ao tema proposto para a elaboração de redação, problemas que, uma vez incorridos, acarretam notas insuficientes aos alunos.

Os resultados dos testes realizados demonstram assertividade de 89,4% de acurácia. Os primeiros resultados já possibilitam a criação de uma aplicação para fornecer um *feedback* automático como suporte ao professor ou avaliador de textos, o que contribuirá para diminuir o tempo demandado para correção, além de prestar melhor auxílio às instituições, professores e alunos.

A solução ora desenvolvida visa diminuir a desigualdade nos processos seletivos, oferecendo maior oportunidade de aprendizagem independente da instituição em que o aluno estude. Em complemento, a solução desenvolvida nesta dissertação, além de proporcionar a possibilidade de treinar e aperfeiçoar a qualidade na escrita, proporciona ainda retornos mais rápidos aos envolvidos no processo de ensino-aprendizagem, ou seja, professores e alunos.

Para as instituições de ensino com elevada carga de textos produzidos, a solução provida nesta dissertação viabiliza a diminuição dos custos e melhora da qualidade das avaliações de textos, já que o papel do professor passa da auditoria nas correções, para a comprovação da efetividade da correção. Além disso, o fator fadiga do docente seria diminuído, já que atualmente o profissional responsável corrige em torno de 50 redações ao dia, no caso da avaliação de textos produzidos no ENEM. Isto porque, além de contar com um sistema que indique prováveis erros, a solução validada nesta dissertação facilitaria em muito o trabalho dos avaliadores e professores.

A aplicação das primeiras técnicas empregadas nos experimentos realizados já permitiu identificar padrões no conteúdo de cada redação analisada, bem como encontrar relações de similaridade entre as redações sobre um determinado tema em específico, possibilitando assim a classificação da fuga ao tema.

Os resultados encontrados nos experimentos realizados demonstraram que existem temas de redação com elevada coesão nos textos produzidos, ou seja, grande parte das redações analisadas encontra-se num mesmo *cluster* (agrupamento) nas análises realizadas. Porém, também foi possível verificar que há outros temas que apresentaram maior dispersão, indicando que seus textos continham conteúdos que fugiram ao tema indicado na prova.

5.1 Contribuições do estudo

As principais contribuições deste estudo buscam permitir ao avaliador, professor ou empresas que aplicam processos seletivos avaliar as redações com menor esforço, otimizando assim o trabalho e reduzindo o tempo e o custo

do processo de avaliação de textos dissertativos. Esta colaboração pode ser primordial na aplicação do ENEM digital, proporcionando assim ao avaliador auxílio na identificação das falhas de escrita, minimizando interferências como fadiga e alteração de humor do avaliador, sintomas estes que podem afetar a correção de um texto dissertativo.

O cumprimento do objetivo proposto nesta pesquisa denota ainda contribuição sob a perspectiva acadêmica, ao servir de base para estudos que, uma vez alinhados aos conhecimentos dos profissionais de ensino, possam gerar novas abordagem que possibilitem capacitar os alunos a redigirem textos coesos ao tema proposto.

Com isso, a desigualdade entre alunos de escolas públicas e privadas pode ser diminuída, uma vez que ter uma plataforma que possibilite o treino mais frequente, como afirma Barros (2020), pode facilitar não apenas o trabalho do professor, como também proporcionar futuramente uma escrita mais clara e desenvolta, possibilitando assim maior maturidade neste processo.

Os experimentos realizados neste estudo trazem uma contribuição na esfera da pesquisa acadêmica na temática abordada, principalmente por se tratar de análise de textos em português, uma vez que a grande parte das pesquisas disponíveis têm experimentos aplicados a textos em língua inglesa. Dessa forma, a solução validada nesta dissertação contribui para a evolução das técnicas aplicadas, além de proporcionar a adaptação de classificação para diferentes campos de estudo nas áreas da educação ou corporativa.

5.2 Limitações da pesquisa

A pesquisa experimental executada apresenta algumas limitações, dentre as quais destaca-se que ainda não foi possível identificar todas as possibilidades que acarretam a atribuição de nota 0 (zero), de acordo com as regras do Exame Nacional do Ensino Médio (ENEM). Isto porque a presente pesquisa restringiu-se especificamente ao fator ‘fuga ao tema’ para avaliação de redações.

Outra questão que pode limitar os resultados apresentados neste estudo refere-se à base de dados utilizada. Pretende-se adicionar à esta mais redações, notadamente com maior ocorrência de desvios de escrita. Vislumbra-se para tanto, principalmente, aqueles desvios de escrita que acarretaram nota zerada por fuga ao tema, uma vez que no *corpus* deste estudo havia apenas 230 redações com fuga ao tema diagnosticadas pelo avaliador. No processo de aprendizado de máquina sabe-se que são necessários muitos dados para treinar o modelo em desenvolvimento. Para tanto, é primordial o enriquecimento da base de dados, incorporando-se a ela redações do próprio ENEM ou de escolas públicas e particulares que as aplicam como rotina no processo de ensino-aprendizagem.

Ressalte-se ainda que a necessidade de se ter uma base de dados rotulada também é um fator limitador no processo de classificação supervisionada. Neste caso, para a elaboração de um *corpus* maior será necessário ter o *feedback* de correções para assim compreender se houve o desvio de escrita com relação à fuga ao tema, fato que pode deixar o processo da criação da base de dados mais demorado.

5.3 Contribuições para área

Baseadas nas limitações citadas anteriormente relativas à base de dados, a autora desta dissertação projeto criou uma plataforma para colher redações e disponibilizar auxílio aos professores no processo avaliativo de textos dissertativos. O sistema atualmente está em teste, com acesso por meio do seguinte endereço: www.agirweb.com.br/redacao. Cinco professores especialistas da área de língua portuguesa estão contribuindo para esta pesquisa, todos atuantes em escolas públicas do estado de São Paulo, parte da base de dados foi utilizada a partir deste sistema.

Vale ressaltar ainda que em 2021 esta proposta foi apresentada a uma empresa especialista em desenvolvimento de software, que se propôs a investir e desenvolver a plataforma para disponibilizar gratuitamente às escolas de SP para a geração de base de dados mais ampla para o treinamento do modelo. A parceria com a empresa já foi consolidada, tendo o desenvolvimento

da ferramenta já se iniciado. O objetivo é criar uma plataforma inteligente que avalie as cinco competências indicadas pelo ENEM.

Além da solução apresentada nesta pesquisa, este estudo proporcionou uma atenção maior às outras competências do ENEM que já estão em fase de pesquisa para agregar um melhor resultado à solução desenvolvida nesta dissertação, o que poderá proporcionar ao professor maior ganho de tempo em suas correções e, como consequência, o aluno terá uma resposta mais rápida quanto à avaliação do texto por ele produzido.

5.4 Sugestão de pesquisas futuras

Visando ampliar a qualidade dos resultados dos classificados analisados nesta pesquisa, sugere-se a testagem de outras técnicas, tais como *Ensemble Methods in Machine Learning*, que podem trazer as melhores características de cada classificador em um único modelo. Isto porque essa técnica combina vários modelos básicos para produzir um modelo preditivo ideal.

Outra proposta seria dar continuidade ao desenvolvimento de experimentos voltados à identificação de outros desvios em textos que obtiveram nota 0 (zero) como resultado da avaliação. Vislumbra-se ainda a possibilidade de aplicação de técnicas de PLN para trabalhar as demais competências consideradas pelo ENEM na avaliação de redações.

REFERÊNCIAS

ABBASI, A., FRANCE, S., ZHANG, Z.; CHEN, H. (2011). **Selecting Attributes for Sentiment Classification Using Feature Relation Networks**. IEEE Transactions on Knowledge and Data Engineering, 23(3).

AFFONSO, Emmanuel T. F. ; SILVA, Alisson M. ;SILVA, Michel P. ; RODRIGUES, Thiago M. D. ; MOITA, Gray F. **Uso Redes Neurais Multilayer Perceptron (MLP) em Sistema de Bloqueio de Websites Baseado em Conteúdo**. Mecánica Computacional Vol XXIX. Publicadado em 15/11/2010.

AGÊNCIA BRASIL. **Enem é um dos principais instrumentos de acesso ao ensino superior**. Publicado em 31/10/2019. Disponível em: https://agenciabrasil.ebc.com.br/educacao/noticia/2019-10/enem-e-um-dos-principais-intrumentos-de-acesso-ao-ensino-superior. Acesso em: 19/03/2021.

ALBUQUERQUE, Rotsen Diego Rodrigues de. (2019). **Estudo Comparativo de Algoritmos de Classificação Supervisionada para Classificação de Polaridade em Análise de Sentimentos**. Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco. Recife, 2019

ALTOE, F.; JOYNER D.(2019), "**Annotation-free Automatic Examination Essay Feedback Generation**," *2019 IEEE Learning With MOOCS (LWMOOCS)* , 2019, pp. 110-115, doi: 10.1109 / LWMOOCS47620.2019.8939630.

ANDROUTSOPOULOU, Aggeliki; Karacapilidis Nikos; Loukis Euripidis; Charalabidis Yannis (2019). **Transforming the communication between citizens and government through AI-guided chatbots**. Government Information Quarterly, Volume 36, Issue 2, 2019, Pages 358-367, ISSN 0740-624X, https://doi.org/10.1016/j.giq.2018.10.001.

ANTÓNIO, Nuno Miguel da Conceição (2019). **Hotel Revenue Management: Using Data Science to Predict Booking Cancellations**. Thesis(Department of Information Science and Technology). Instituto Universitário de Lisboa.

ARANHA, Christian; PASSOS, Emmanuel. **A Tecnologia de Mineração de Textos. Revista Eletrônica de Sistemas de Informação**, [S.l.], v. 5, n. 2, ago. 2006. ISSN 1677-3071. Disponível em: <http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171>. Acesso em: 13 ago. 2021. doi:https://doi.org/10.21529/RESI.2006.0502001.

ARAÚJO, Georger Rommel Ferreira de (2012). **Agrupamento de documentos forenses utilizando redes neurais art1**. 2011. xxiii, 131 f. Dissertação (Mestrado em Engenharia Elétrica)-Universidade de Brasília, Brasília, 2012.

ARAÚJO, U. **A quarta revolução educacional: a mudança de tempos, espaços e relações na escola a partir do uso de tecnologias e da inclusão social**. ETD - Educação Temática Digital, v. 12, n. , p. 31-48, 2011.

BALDAIA, Beatriz Souto de Sá (2020). **Lie-o-matic: using natural language processing to detect contradictory statements**. Dissertação (Mestrado - Computing Engineering)- Universidade do Porto, Portugal, 2020.

BANERJEE, Dibyendu (2020). **Natural Language Processing (NLP) Simplified : A Step-by-step Guide**. Publicado em: 14/04/2020. Data Science Foudation. Disponível em: https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide. Acesso em: 15/04/2021

BARROS, Aidil Jesus da Silveira; LEHFELD, Neide Aparecida de Souza. **Fundamentos de Metodologia Científica**. 3. ed. São Paulo: Pearson Prentice Hall, 2014.

BARROS, Jussara (2020). **Desenvolvendo uma Boa Escrita**. Portal Mundo Educação. Disponível em: https://mundoeducacao.uol.com.br/educacao/desenvolvendo-uma-boa-escrita.htm. Acesso em: 10/06/2020.

BARROS, Rodrigo Coelho (2019). **Fully-disentangled text-to-image synthesis**. Dissertação (Programa de Pós-Graduação em Ciência da Computação). Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS).

BAZELATO, B. S.; AMORIM, E. C. F. A bayesian classifier to automatic correction of portuguese essays. In: Congresso Internacional de Informática Educativa (TISE), XVIII. **Anais...** Porto Alegre: Centro de Computação e Comunicação para a construção do Conhecimento, 2013, p. 1-13.

BECKER J.P.,SIOR E.,Hoy J.,KAHANDA I.(2019).**Predicting at-risk students in a circuit analysis course using supervised machine learning**. ASEE Annual Conference and Exposition, Conference Proceedings15 June 2019 126th ASEE Annual Conference and Exposition: Charged Up for the Next 125 Years, ASEE 2019, Tampa, 15 June 2019 - 19 June 2019, 157034

BENZEBOUCHI, N. E., N. Azizi, N. E. Hammami, D. Schwab, M. C. E. Khelaifia and M. Aldwairi,(2019) **"Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector**," 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD), Istanbul, Turkey, 2019, pp. 371-376, doi: 10.1109/SSD.2019.8894872.

BERLATTO, Leonardo. **Como Funciona uma Árvore de Decisão**. Publicado em: 25/04/2021. Disponível em: https://medium.com/data-hackers/como-funciona-uma-%C3%A1rvore-de-decis%C3%A3o-be3eba5918a1#9112.

Acesso em: 11/05/2021

BIANCHI, Alexandre. **As classificações dos algoritmos de Machine Learning**. Publicado em 27/05/2020. Disponível em: https://www.viceri.com.br/insights/as-classificacoes-dos-algoritmos-de-machine-learning. Acesso em: 09/05/2021.

BITTENCOURT JÚNIOR, José Adenaldo Santos. **Avaliação automática de redação em língua portuguesa empregando redes neurais profundas**. 2020. 100 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Goiás, Goiânia, 2020.

BONADIA, Graziella Cardoso (2019). **Contribuições para acelerar o aprendizado sobre a construção de uma máquina de classificação de sentimentos utilizando processamento de linguagem natural** . 2019. 1 recurso online (126 p.). Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP.

BRAJKOVIĆ, E.; RAKIĆ, K.; KRALJEVIĆ G., "**Application of data mining in e-Leaming systems**," *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 2018, pp. 1-5, doi: 10.1109/INFOTEH.2018.8345536.

BRASIL(2019). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **A redação no Enem 2019: cartilha do participante**. Brasília, 2019. Disponível em: http://inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/6736715. Acesso em 20/10/2019

BRASIL(2019). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Entenda como é calculada a nota do Enem**. Disponível em: http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/84461-entenda-como-e-calculada-a-nota-do-enem. Acesso em 10/06/2020

BRITZ, Denny (2015). **Understanding Convolutional Neural Networks for NLP**. Publicado em 07/11/2015. Disponível em:

http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/. Acesso em: 31/03/2021

BROWN, M.; Lewis, H.G.; Gunn, S.R**. Linear spectral mixture models and support vector machines for remote sensing**. IEEE Transactions on geoscience and remote sensing, v. 38, n. 5, p. 2346-2360. 2000

BROWNLEE, Jason (2016). **Boosting and AdaBoost for Machine Learning**. Machine Learning Matery. Disponível em: https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/. Acesso em 10/06/2021

CAFFARO FILHO, Roberto. **Mineração de Texto – O que é? Como aplicar?.** Hup Data Solutions. Publicado em 04/05/2020. Disponível em: https://hupdata.com/mineracao-de-texto-o-que-e-como-aplicar/. Acesso em 30/03/2021

Cambridge English. **Plataforma Write&Improve**. Desenvolvido por ELiT. Disponpivel em: https://writeandimprove.com/. Acesso em: 11/06/2020.

CAMPOS, Lorraine Vilela (2020). **Mais de 143 mil participantes tiraram zero na redação do Enem 2019**. Portal Brasil Escola – UOL. Disponível em: https://vestibular.brasilescola.uol.com.br/enem/mais-143-mil-participantes-tiraram-zero-na-redacao-enem-2019/347183.html. Acesso em 10/06/2020

CÂNDIDO, T.; WEBBER, C. (2018). **Avaliação da Coesão Textual: Desafios para Automatizar a Correção de Redações.** RENOTE - Revista Novas Tecnologias na Educação ISSN 1679-1916. 1-10. Doi: 10.22456/1679-1916.86013.

CAPGEMINI (2017). **Unleashing the potential of Artificial Intelligence in the Public Sector**. Disponível em: https://www.capgemini.com/consulting/wp-content/uploads/sites/30/2017/10/ai-in-public-sector.pdf. Acesso em 30/03/2021

CARNEIRO, Alvaro Leandro Cavalcante. **Redes Neurais Convolucionais para processamento de linguagem natural**. Publicado em 07/07/2020. Disponível em: https://medium.com/data-hackers/redes-neurais-convolucionais-para-processamento-de-linguagem-natural-935488d6901b. Acesso em: 31/03/2021

CARVALHO, A. C. P. F. de; FAIRHURST, M. C. and D. L. Bisset. (1994**) An Integrated Boolean Neural Network for Pattern Classification**. Pattern Recognition Letters, Vol. 15, pages 807-813, August 1994. ISSN: 0167-8655

CAVALCANTI, Rafael Dutra (2017). **Classificação de tendências políticas em notícias via mineração de texto e redes neurais sem peso**. Dissertação(Mestrado) Universidade Federal do Rio de Janeiro - Instituto De Matemática, Rio de Janeiro, RJ.

CHITONGUA, Fátima Joana Dantas Gonçalves(2019). **Interface Ubíqua, Interoperativa e Escalável para uma Plataforma de Serviços PLN em Big Data**. Dissertação (Engenharia Informática) - UNIVERSIDADE DA BEIRA INTERIOR, 2019.

COHEN, AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, Yu PS(2015). **Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine**. J Am Med Inform Assoc. 2015 May;22(3):707-17. doi: 10.1093/jamia/ocu025.

CONCEIÇÃO, R. I. S. (2012). **O ensino de produção textual e a (re)construção da competência discursiva do aluno**. Trabalhos Em Linguística Aplicada, 40(1). Recuperado de https://periodicos.sbu.unicamp.br/ojs/index.php/tla/article/view/8639351

CONEGLIAN, Caio Saraiva (2018). **Recuperação da informação com abordagem semântica utilizando linguagem natural: a inteligência artificial na ciência da informação**. Tese (Ciência da Informação - FFC) – Universidade Estadual Paulista UNESP – SP.

CONTRERAS, Jennifer O.; HILLES, Shadi M. S.; ABUBAKER, Zainab Binti. Automated Essay Scoring using Ontology. **Generator and Natural Language Processing with Question Generator based on Blooms Taxonomy's Cognitive Level**. International Journal of Engineering and Advanced Technology Open AccessVolume 9, Issue 1, Pages 2448 – 2457, October 2019

COSTA, Tatiane Olívia Riffel da (2019). **Estudo sobre a contribuição dos aplicativos de celular na produção textual escolar de alunos do ensino médio**. Universidade Federal de Santa Catarina. Linguagem e Educação a Distância. Publicado em: 02/07/2019. Disponível em: https://repositorio.ufsc.br/handle/123456789/199564. Acesso em 11/06/2020.

CUCCURULLO, Corrado; ARIA, Massimo (2020). **Bibliometrix: uma breve história**. Disponível em: https://www.bibliometrix.org/About.html. Acesso em 23/10/2020.

Deep Learning Book (2021). **Deep Learning Book, 2021**. Data Science Academy. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 14/04/ 2021.

DIANA, Daniela Biason gomes (2021). **Os 16 maiores erros de redação cometidos pelos estudantes**. Publicado em 08/01/2021. Disponível em: https://www.todamateria.com.br/erros-de-redacao/. Acesso em 29/03/2021

ECK, Nees Jan van; WALTMAN, Ludo (2020). **Sobre o Software VosViewer**. Disponível em: https://www.vosviewer.com/. Acesso em: 23/10/2020

EGGERS WILLIAM, D. Schatsky, P. Viechnicki(2017). **AI-augmented government using cognitive technologies to redesign public sector work**. Deloitte University Press (2017). Disponível em: https://www2.deloitte.com/content/dam/insights/us/articles/3832_AI-augmented-government/DUP_AI-augmented-government.pdf. Acesso em: 30/03/2021

EPSTEIN, D.; REATEGUI, E. B. Uso de mineração de textos no apoio à compreensão textual. **RENOTE**, v. 13, n. 1, p. 1-10, 2015.

EVANGELISTA, João; SILVA, Ellen; SASSI, Renato. (2020). **Enriquecimento de Base de Dorks Com Processamento de Linguagem Natural**. 6. 10763-10780. 10.34117/bjdv6n3-085.

FARINHA, André Filipe Neves (2018). **Extracting keywords from tweets**. Dissertação (Mestrado em Engenharia Informática - Internet das Coisas). IPT - ESTT - Escola Superior de Tecnologia de Tomar

FÁVERO, Luiz Paulo; BELFIORE, Patrícia; SILVA, Fabiana Lopes e CHAN, Betty Lilian(2009). **Análise de Dados: Modelagem Multivariada para Tomada de Decisões**. Rio de Janeiro: Elsevier - Editora Campos - 2009.

FEITOSA, Rodrigo Miranda(2013)**. Uma aplicação de mineração de dados para recomendação social**. 2013. 154 f. Dissertação (Mestrado em Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luís, 2013.

FERNANDEZ, Pedro J.; MARQUES, Paulo C. F (2019). **Data Science, Marketing & Business**. Editora: Insper. Publicado em Maio/2019. Disponível em: https://datascience.insper.edu.br/. Acesso em: 03/05/2021

GLEIDSON SILVA, Antônio Cardoso da (2015). **KDC: uma abordagem baseada em conhecimento para classificação de documentos**. Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2015.

GOMES, M. de F. C. (2020). **A PNA e a unidade dialética afeto-cognição nos atos de ler e escrever**. Revista Brasileira De Alfabetização, 1(10). https://doi.org/10.47249/rba.2019.v1.368

GONÇALVES, Eduardo Corrêa (2012). **Mineração de Texto - Conceitos e Aplicações Práticas**. Publicado em: Novembro/2012 - SQL Magazine, v. 105, 2012, p. 31-44. Disponível em: https://www.researchgate.net/publication/317912973_Mineracao_de_texto_-_Conceitos_e_aplicacoes_praticas. Acesso em: 15/04/2021

GOODFELLOW, Ian; YOSHUA Bengio et. al (2016). **Deep learning**, volume 1. MIT press Cambridge, 2016. ix, 2, 18, 22

Goularte, Fábio Bif (2015). **Método fuzzy para a sumarização automática de texto com base em um modelo extrativo (FSumm).** Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2015.

Grácio, M. C. C. (2016). **Acoplamento bibliográfico e análise de cocitação: revisão teórico-conceitual**. Encontros Bibli: Revista eletrônica De Biblioteconomia E Ciência Da informação, 21(47), 82-99. https://doi.org/10.5007/1518-2924.2016v21n47p82

GRANATYR, Jones (2016). **Processamento de Linguagem Natural com NLTK e Python**. Publicado em 23/08/2016. Disponível em: https://iaexpert.academy/2016/08/23/ferramentas-para-ia-processamento-de-linguagem-natural-com-nltk-e-python/. Acesso em 29/03/2021.

GRANATYR, Jones (2020). **Processamento de Linguagem Natural com Deep Learning**. Curso realizado em novembro de 2020. Disponível em: https://iaexpert.academy/courses/processamento-linguagem-natural-deep-learning-transformer/. Acesso em 10/12/2020.

GUAMÁ, Juliana. **Métricas de avaliação de classificadores**. Publicado em: 16/03/2019. Disponível em: https://medium.com/pyladiesbh/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-de-classificadores-6aadc3dacd51#:~:text=A%20matriz%20de%20confus%C3%A3o%20%C3%A9,dos%20resultados%20fique%20mais%20claro. Acesso em: 09/05/2021

HAIR Jr., J.F.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E. & TATHAM, R.L. **Análise multivariada de dados**. 6.ed. Porto Alegre, Bookman, 2009. 688p.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques: concepts and techniques**. Elsevier, 2011. ISBN 0123814804.

HARIRI, R. H., Fredericks, E. M., & Bowers, K. M. (2019). **Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data***, 6(1).* doi:10.1186/s40537-019-0206-3

HENTZ, Maria Izabel de Bortoll; GUIMARÃES, Ana Maria de Mattos; CARNIN, Anderson (2020). **Evidências do ensino da escrita em textos de alunos do Ensino Médio: um olhar para o(s) impacto(s) do agir docente no trabalho com redação para a prova do ENEM.** VI Encontro Internacional de Reflexão sobre a Escrita . Indagatio Didactica, vol. 12. https://doi.org/10.34624/id.v12i2.17457.

HONDA, Hugo (2017). **Introdução Básica à Clusterização**. Laboratório de Aprendizado de Máquina em Finanças e Organizações (LAMFO) - UNB. Disponível em: https://lamfo-unb.github.io/about/. Acesso em: 27/10/2020

HUESCH, M.D.; Cherian, R.; Labib, S.; Mahraj, R. (2018). **Evaluating Report Text Variation and Informativeness: Natural Language Processing of CT Chest Imaging for Pulmonary Embolism**. Journal of the American College of Radiology 15(3 Pt B): 554-562. DOI: 10.1016/j.jacr.2017.12.017

Jornal Cruzeiro do Sul (2020). **Universidades privadas de SP adotam vestibular online e nota do Enem**. Publicado em: 09/06/2020. Disponível em: https://www.jornalcruzeiro.com.br/brasil/universidades-privadas-de-sp-adotam-vestibular-online-e-nota-do-enem/. Acesso em: 29/03/2021

KAUFMAN, Dora (2018). **A inteligência artificial irá suplantar a inteligência humana?** E-book - Estação das Letras e Cores, 2019. ISBN: 978-85-68552-90-2

Kim, Y. (2014). **Convolutional Neural Networks for Sentence Classification**. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1746–1751.

KITCHEHAM, B.; CHARTERS, S (2007). **Guidelines for performing Systematic Literature Reviews in Software Engineering**. Staffordshire: Elsevier.

KOWSARI, K.; JAFARI Meimandi, K.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D (2019). **Text Classification Algorithms: A Survey**. *Information* 2019, *10*, 150.

KUSHMERICK, N. & THOMAS, B. 2003**. Adaptive information extraction: Core technologies for information agents, Lecture Notes in Computer Science**, Vol. 2586, Springer. DOI:10.1007/3-540-36561-3_4

LACERDA, Daniel Pacheco (2018). **Suporte às micro e pequenas empresas a partir da gestão baseada em evidências: construção de ferramenta computacional baseada em inteligência artificial**. Dissertação(Mestrado em Engenharia de Produção e Sistemas) – Unisinos.

LEITE, Tiago M. **Redes Neurais, Perceptron Multicamadas e o Algoritmo Backpropagation**. Publicado em 10/05/2018. Disponível em: https://medium.com/ensina-ai/redes-neurais-perceptron-multicamadas-e-o-algoritmo-backpropagation-eaf89778f5b8. Acesso em: 10/05/2021.

LIN, C; HSU, CJ; LOU, YS; YEH, SJ; LEE, CC; SU, SL; CHEN, HC (2017). **Artificial Intelligence Learning Semantics via External Resources for**

Classifying Diagnosis Codes in Discharge Notes. J Med Internet Res 2017;19(11):e380. DOI: 10.2196/jmir.8344

LOURENCETTI, G. do C (2014). **A baixa remuneração dos professores: algumas repercussões no cotidiano da sala de aula**. Revista de Educação Pública, [S. l.], v. 23, n. 52, p. 13-32, 2014. DOI: 10.29286/rep.v23i52.1422. Disponível em: https://periodicoscientificos.ufmt.br/ojs/index.php/educacaopublica/article/view/1422. Acesso em: 19 mar. 2021.

LUDERMIR, Teresa Bernarda. **Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências**. Estud. av., São Paulo , v. 35, n. 101, p. 85-94, Apr. 2021 . Available from <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142021000100085&lng=en&nrm=iso>. access on 09 May 2021. Epub Apr 19, 2021. https://doi.org/10.1590/s0103-4014.2021.35101.007.

MARCOLIN, Carla Bonato (2018). **Text analytics in business environments: a managerial and methodological approach**. Tese (Ciências Sociais Aplicadas) - Universidade Federal do Rio Grande do Sul(2018).

MARKOFF, John (2013). **Software corrige redações**. Publicado em 16/04/2013. Disponível em: http://www.observatoriodaimprensa.com.br/e-noticias/ed742-software-corrige-redacoes/. Acesso em: 11/06/2020.

MARTINS, Claudia Aparecida et al. (2003). **Uma Experiência em Mineração de Textos Utilizando Clustering Probabilístico Clustering Hierárquico**. 2003. Instituto de Ciências Matemáticas e de Computação. São Carlos: Universidade de São Paulo.

MATAVELLI, Renan(2011). **Professores apontam dificuldades dos alunos na hora da redação**. Publicado em 28/03/2011. Disponível em: http://www.metodista.br/rronline/noticias/cidades/2011/10-1/temporario/redacao-e-a-maior-dificuldade-dos-vestibulandos. Acesso em: 25/10/2019

MEC (2019). **MEC realiza conferência para discutir estratégias de alfabetização no Brasil**. Publicado em: 22/10/2019. Disponível em: http://portal.mec.gov.br/component/tags/tag/5?start=60. Acesso em: 15/04/2021

MELGANI, F.; Bruzzone, L. **Classification of Hyperspectral Remote Sensing Images with Support Vector Machines**. IEEE Transactions on Geoscience and Remote Sensing, vol. 42, No. 8, August 2004.

MITCHELL, T. M.. **Machine Learning**. McGraw-Hill, New York, 1997

MORAIS, Alaydes Mikaelle de (2021). **Technological Trajectory Analysis In Wind And Solar Energy From Text Mining Techniques In Patents And Papers**. Dissertação (Informática e Gestão do Conhecimento). Universidade Nove de Julho 2021.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. **Mineração de textos.** Goiás: UFG, 2007. (Relatório Técnico. Instituto de Informática). Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf. Acesso em 25 de janeiro de 2019.

MORALES, Juliana(2020). **Com a pandemia, como será o vestibular das principais particulares?** Guia do Estudante. Publicado em 23/06/2020. Disponível em: https://guiadoestudante.abril.com.br/universidades/com-a-pandemia-como-sera-o-vestibular-das-principais-particulares/. Acesso em: 31/10/2020.

MOREIRA, Sandro. **Rede Neural Perceptron Multicamadas**. Publicado em 24/12/2018. Disponível em: https://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9. Acesso em: 10/05/2021

MORENO-BLANCO, Diego; OCHOA-FERRERAS, Borja; GÁRATE, Francisco; SOLANA SÁNCHEZ, Javier; SÁNCHEZ-GONZÁLEZ, P.; GÓMEZ AGUILERA, Enrique. (2019). **Evaluation and Comparison of Text Classifiers to Develop a Depression Detection Service**. 10.1007/978-3-030-31635-8_146. Conferência: 15th Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON) Local: UNESCO World Heritage Univ, Coimbra, PORTUGAL Data: SEP 26-28, 2019

MÜLLER, Sarah; Bergande, Bianca, and Brune, Philipp. (2018). **Robot Tutoring: On the Feasibility of Using Cognitive Systems as Tutors in Introductory Programming Education: A Teaching Experiment**. In Proceedings of the 3rd European Conference of Software Engineering Education (ECSEE'18). Association for Computing Machinery, New York, NY, USA, 45–49. DOI:https://doi.org/10.1145/3209087.3209093

MUÑOZ-VALERO, David ; Luis Rodriguez-BenitezORCID; Luis Jimenez-Linares; Juan Moreno-Garcia(2020). **Using Recurrent Neural Networks for Part-of-Speech Tagging and Subject and Predicate Classification in a Sentence**. International Journal of Computational Intelligence Systems. Publication Date 2020/06. https://doi.org/10.2991/ijcis.d.200527.005.

MUYLAERT, Renata (2020). **Pandemia do novo coronavírus, Parte 6: inteligência artificial (NLP)**. Disponível em:

https://marcoarmello.wordpress.com/2020/08/19/coronavirus6/.Acesso em 27/07/2021.

NOBRE, J. C. S.; PELLEGRINO, S. R. M (2010). ANAC: **um analisador automático de coesão textual em redação**. *In*: Brazilian Symposium on Computers in Education - SBIE), 2010. **Anais.** SBC, 2010, p. 1-12.

OKANO, Émerson Yoshiaki(2020). **Análise e caracterização de textos intencionalmente enganosos escritos em português usando métodos de processamento de textos.** Dissertação (Mestrado em Computação Aplicada) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2020. doi:10.11606/D.59.2020.tde-29062020-171001. Acesso em: 2020-10-31.

OKUHLE, Ngada; BETRAM, Haskins (2020). "**Fake News Detection Using Content-Based Features and Machine Learning**," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411638.

OLIVEIRA, Elida (2020). **Nota do Enem serve como vestibular em universidades públicas e privadas, além de dar acesso a bolsas e financiamentos**. Publicado em: 23/11/2020. Disponível em: https://g1.globo.com/educacao/enem/2020/noticia/2020/11/23/nota-do-enem-serve-como-vestibular-em-universidades-publicas-e-privadas-alem-de-dar-acesso-a-bolsas-e-financiamentos-confira-as-opcoes.ghtml. Acesso em: 19/03/2021.

PASSERO, Guilherme (2018). **Detecção de Fuga ao Tema em Redações de Língua Portuguesa**. Dissertação (Mestrado) - Computação Aplicada - UNIVALI - Universidade do Vale do Itajaí – SC

PATLOLLA, Vinay (2017). **How to make SGD Classifier perform as well as Logistic Regression using parfit**. Publicado em 29/11/2021. Disponível em: https://towardsdatascience.com/how-to-make-sgd-classifier-perform-as-well-as-logistic-regression-using-parfit-cc10bca2d3c4. Acesso em: 21/10/2021

PENG, S.;MING, Z.; ALLEN, J.K.; SIDDIQUE, Z.; MISTREE, F. (2020).**Quantification of students' learning through reflection on doing based on text similarity**. Proceedings of the ASME Design Engineering Technical ConferenceVolume 32020 Article number V003T03A009ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, IDETC-CIE 2020, Virtual, Online, 17 August 2020 - 19 August 2020, 164730

PESSANHA, Cinthia (2019). **Random Forest: como funciona um dos algoritmos mais populares de ML**. Publicado em 20/11/2019. Disponível em:

https://medium.com/cinthiabpessanha/random-forest-como-funciona-um-dos-algoritmos-mais-populares-de-ml-cc1b8a58b3b4. Acesso em 10/06/2021

PEZZINI, A. (2017). **Mineração De Textos: Conceito, Processo e Aplicações**. REAVI, v. 5, n. 8, p. 01-13, dez., 2016ISSN: 2316-4190, DOI: 10.5965/2316419005082016058. Disponível em: https://www.revistas.udesc.br/index.php/reavi/article/view/6750/6415. Acesso em: 30/03/2021

PINHO, CM, VANIN, A.S.; BELAN; P.; NAPOLITANO, D. **Uma ferramenta on-line para ensino de Redação, baseada nos critérios avaliativos do ENEM** . Anais do KM Brasil 2020 – 15º Congresso Brasileiro de Gestão do Conhecimento. Publicado em: 2020 p. 599-615. ISSN: 1678-1546.

PNA – MEC (2019). **Política Nacional de Alfabetização (PNA).** Publicado em 11/04/2019. Disponível em: http://portal.mec.gov.br/images/banners/caderno_pna_final.pdf. Acesso em 15/04/2021

POLLETTINI, Juliana Tarossi. **Avaliação de mecanismos de suporte à tomada de decisão e sua aplicabilidade no auxílio à priorização de casos em regulações de urgências e emergências.** 2016. Tese (Doutorado em Clínica Médica) - Faculdade de Medicina de Ribeirão Preto, University of São Paulo, Ribeirão Preto, 2016. doi:10.11606/T.17.2017.tde-30032017-101723. Acesso em: 2020-10-23.

POLLYANNA GONÇALVES, de Oliveira (2015). **Um benchmark para comparação de métodos para análise de sentimentos**. Dissertação(Pós-Graduação em Ciência da Computação). Universidade Federal de Minas Gerais

Portal Biblioteca Spacy IO. **Natural Language Processing**. Disponível em: https://spacy.io/. Acesso em: 10/12/2020

Portal do Governo de Goiás (2020). **Plataforma faz correção automática de redação**. Publicado em 13/10/2020. Disponível em: https://www.goias.gov.br/servico/103-tecnologia/123257-centro-de-intelig%C3%AAncia-artificial-desenvolve-plataforma-que-corrige-reda%C3%A7%C3%A3o.html. Acesso em: 29/03/2021

Portal G1 (2016). **Corretores de redação do Enem avaliam em média 74 textos por dia**. Publicado em 16/09/2016. Disponível em: https://g1.globo.com/educacao/enem/2016/noticia/corretores-de-redacao-do-enem-avaliam-em-media-74-redacoes-por-dia.ghtml. Acesso em: 29/03/2021

Portal INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2020). **Entenda como é calculada a nota do Enem**. Disponível em:

http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/84461-entenda-como-e-calculada-a-nota-do-enem. Acesso em 10/06/2020

Portal INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2018). **Histórico ENEM**. Disponível em: http://portal.inep.gov.br/enem/historico. Acesso em: 19/03/2021.

Portal INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2020). **65% dos inscritos no Enem já concluíram o ensino médio em anos anteriores**. Disponível em: http://portal.mec.gov.br/index.php?option=com_content&view=article&id=90701:65-dos-inscritos-no-enem-ja-concluiram-o-ensino-medio-em-anos-anteriores&catid=418&Itemid=86. Acesso em 10/06/2020.

Portal INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2018). **DADOS DO CENSO ESCOLAR: Ensino Médio brasileiro tem média de 30 alunos por sala**. Disponível em: http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/dados-do-censo-escolar-ensino-medio-brasileiro-tem-media-de-30-alunos-por-sala/21206. Acesso em: 19/03/2021.

Portal INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2004). **Brasil tem maior número de alunos por professor no nível secundário**. Disponível em: http://inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/brasil-tem-maior-numero-de-alunos-por-professor-no-nivel-secundario/21206. Acesso em: 19/03/2021.

Portal Universia (2015). **Entrevista com ex-corretor de redação.** Disponível em: https://www.universia.net/br/actualidad/orientacion-academica/corretor-redaco-do-enem-leva-cerca-2-minutos-prova-diz-professor-1132810.html**.** Acesso em: Acesso em: 19/03/2021.

PRASATH, Gokul (2019). **Diferença entre aprendizado de máquina, inteligência artificial e PNL**. Publicado em: 21/08/2019. Disponível em: https://medium.com/@cs.gokulprasath98/difference-between-machine-learning-artificial-intelligence-and-nlp-d82ba64a7f32. Acesso em: 15/04/2021

PRATES, Wlademir Ribeiro(2019). **Introdução ao Processamento de Linguagem Natural(PLN)**. Publicado em: 01/08/2019. Disponível em: https://cienciaenegocios.com/processamento-de-linguagem-natural-nlp/. Acesso em: 30/30/2021

PREMLATHA, KR (2019). **What is AI? - In a simple way.** Publicado em: 05/02/2019. Disponível em: https://www.aitimejournal.com/@premlatha.kr/what-is-ai-in-a-simple-way. Acesso em: 15/04/2021

PREUSS, Evandro; BARONE, Dante Augusto Couto; HENRIQUES, Renato Ventura Bayan. **Uso de Técnicas de Inteligência Artificial num Sistema de Mesa Tangíve**l. *In*: WORKSHOP DE INFORMÁTICA NA ESCOLA, 26. , 2020, Evento Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 439-448. DOI: https://doi.org/10.5753/cbie.wie.2020.439.

RAMISCH, Renata (2020). **Caracterização de desvios sintáticos em redações de estudantes do ensino médio: subsídios para o processamento automático das línguas naturais**. Dissertação (Mestrado) - Programa de Pós-Graduação em Linguística – UFSCar - Câmpus São Carlos - SP.

RAMOS, Jorge Luis Cavalcanti et al. **Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD**. Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE), [S.l.], p. 1463, out. 2018. ISSN 2316-6533. Disponível em: https://br-ie.org/pub/index.php/sbie/article/view/8107/5798. Acesso em: 09 maio 2021. doi:http://dx.doi.org/10.5753/cbie.sbie.2018.1463.

RÊGO, A. S. da C (2016). **Aprendizado automático de relações semânticas entre tags de folksonomias.** Tese (Doutorado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Paraíba, Brasil, 2016.

RIOLFI, Claudia Rosa; IGREJA, Suelen Gregatti da (2010). **Ensinar a escrever no ensino médio: cadê a dissertação?.** Educ. Pesqui., São Paulo , v. 36, n. 1, p. 311-324, Apr. 2010 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1517-97022010000100008&lng=en&nrm=iso>. Acesso em: 11/06/2020. https://doi.org/10.1590/S1517-97022010000100008.

RODRIGUES, Diego Alves Rodrigues (2018). **Deep Learning e Redes Neurais Convolucionais: Reconhecimento Automático de Caracteres em Placas de Licenciamento**. Centro De Informática Universidade Federal da Paraíba – Dissertação (Mestrado).

RODRIGUES, Jéssica (2017). **O que é o Processamento de Linguagem Natural?** - **Como interpretar mensagens codificadas em linguagem natural e decifrá-las para a linguagem de máquina.** Publicado em 14/07/2017. Disponível em: https://medium.com/botsbrasil/o-que-%C3%A9-o-processamento-de-linguagem-natural-49ece9371cff. Acesso em: 31/03/2021

RODRIGUES, Vitor. **Métricas de Avaliação - quais as diferenças?.** Publicado em 12/04/2019. Disponível em: https://medium.com/@vitorborbarodrigues/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c#:~:text=O%20F1%2DScore%20%C3%A9%20simplesmente,e%20recall)%20em%20alguma%20situa%C3%A7%C3%A3o.&text=Ou%20seja%2C%20quando%20tem%2Dse,ou%20o%20recall%20est%C3%A1%20baixo. Acesso em 01/06/2021.

ROSSI, Rafael Geraldeli (2011). **Representação de coleções de documentos textuais por meio de regras de associação**. 2011. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2011. doi:10.11606/D.55.2011.tde-31082011-125648. Acesso em: 2020-10-23.

ROWTULA , V.; OOTA, SR e JCV, "**Towards Automated Evaluation of Handwritten Assessments**," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 426-433, doi: 10.1109/ICDAR.2019.00075.

RUSSEL, Stuart; NORVIG Peter, **Inteligência Artificial: Uma Abordagem Moderna**. Trad. da 3a. Ed.Campus- Elsevier Editora. 2013

RUSSO, Inês Filipa Duarte (2020). "**O impacte da inteligência artificial na sustentabilidade ambiental : uma agricultura sustentável**". Dissertação de Mestrado. Universidade de Lisboa. Instituto Superior de Economia e Gestão.

SALVETTI, Nilson (2019). **PROÁGIL-29110: Processo ágil aderente à norma ISO/IEC 29110 baseado em Scrum e princípios Lean**. Tese (Programa de Pós-Graduação em Informática e Gestão do Conhecimento) - Universidade Nove de Julho - UNINOVE.

SANTOS JÚNIOR, Jário José dos(2017). **Modelos e técnicas para melhorar a qualidade da avaliação automática para atividades escritas em língua portuguesa brasileira**. 2017. 76 f. Dissertação (Mestrado em Informática) – Instituto de Computação, Programa de Pós-Graduação em Informática, Universidade Federal de Alagoas, Maceió, 2018.

SANTOS, Diego Soares dos (2018)**. Uma plataforma distribuída de mineração de dados para big data: um estudo de caso aplicado à Secretaria de Tributação do Rio Grande do Norte**. 2018. 70f. Dissertação (Mestrado Profissional em Engenharia de Software) - Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Natal, 2018.

SANTOS, Felipe Martins dos (2015). **Subáreas da Inteligência Artificial do ponto de vista computacional**. Publicado em 23/10/2015. Disponível em: https://iascblog.wordpress.com/2015/10/23/subareas-da-inteligencia-artificial-do-ponto-de-vista-computacional/. Acesso em: 15/02/2021

SANTOS, Keila Barbosa Costa dos. **Categorização de textos por aprendizagem de máquina**. 2019. 85 f. Dissertação (Mestrado em Modelagem Computacional de Conhecimento) – Instituto de Computação, Programa de Pós Graduação em Modelagem Computacional de Conhecimento, Universidade Federal de Alagoas, Maceió, 2019.

SCIKIT LEARN – *AdaBoost* (2021). **AdaBoost**. Disponível em: https://scikit-learn.org/stable/modules/ensemble.html#adaboost. Acesso em: 11/05/2021

SCIKIT LEARN - *Gradient Boosting* (2021). **Gradient Tree Boosting**. Disponível em: https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting. Acesso em: 11/05/2021

SCIKIT LEARN – SGD (2021). **Stochastic Gradient Descent**. Disponível em: https://scikit-learn.org/stable/modules/sgd.html#sgd. Acesso em: 11/05/2021

SCIKIT LEARN – SVM (2021). Support Vector Machines. Disponível em: https://scikit-learn.org/stable/modules/svm.html#svm-classification. Acesso em: 11/05/2021

SCIKIT LEARN (2021). **Supervised learning**. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning. Acesso em: 11/05/2021

SEED - Secretaria de Estado da Educação e do Esporte (2020). **Seed vai implantar ferramenta on-line de correção de redação para alunos**. Publicado em: 23/07/2020. Disponível em: http://www.educacao.pr.gov.br/Noticia/Seed-vai-implantar-ferramenta-line-de-correcao-de-redacao-para-alunos. Acesso em: 29/03/2021

SEGARRA-FAGGIONI, Veronica; RATTE, Sylvie (2020). **Computer-based Classification of Student's Report.** In 2020 12th International Conference on Education Technology and Computers (ICETC'20). Association for Computing Machinery, New York, NY, USA, 33–36. DOI:https://doi-org.ez345.periodicos.capes.gov.br/10.1145/3436756.3437017

SHAMS, R.; Mercer RE. (2015). **"Summary Frase Classification Using Stylometry"**, 2015 IEEE 14th International Conference on Machine Learning and applications (ICMLA) , Miami, FL, 2015, pp. 1220-1227, doi: 10.1109 / ICMLA.2015.181.

SHARMA, Nikita. How to create custom NER in Spacy. Publicado em 30/11/2019. Disponível em: https://nikkisharma536.medium.com/how-to-create-custom-ner-in-spacy-cfcd531f8773. Acesso em: 31/03/2021

SILVA, Jonhy (2020). **Uma breve introdução ao algoritmo de Machine Learning Gradient Boosting utilizando a biblioteca Scikit-Learn**. Publicado em: 22/06/2020. Disponível em: https://medium.com/equals-lab/uma-breve-introdu%C3%A7%C3%A3o-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-311285783099. Acesso em 21/10/2021.

Silva, Lucas Santos (2020). **Algoritmo K-Means na Prática com Dados Determinados**. Publicado em 12/06/2020. Disponível em: https://medium.com/@englucsantosilva/algoritmo-k-means-na-pr%C3%A1tica-75f4ca656bbc. Acesso em: 10/06/2021

SMIRAGLIA, R.P (2011). **ISKO 11’s Diverse Bookshelf: an editorial**. Knowledge Organization, v. 38, n.3, p. 179-186, 2011.

SOARES, Magda (2002). **Português na Escola. História de uma disciplina curricular**. In: M.Bagno(org.) Liguística da Norma. São Paulo, Loyola.

SOARES, Patrícia Bourguignon et al (2016)**. Análise bibliométrica da produção científica brasileira sobre Tecnologia de Construção e Edificações na base de dados Web of Science**. Ambient. constr., Porto Alegre , v. 16, n. 1, p. 175-185, Jan. 2016 . Available from <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212016000100175&lng=en&nrm=iso>. access on 19 Oct. 2020. https://doi.org/10.1590/s1678-86212016000100067.

SOUZA, Alex. (2019). **Algoritmo SVM (Máquina De Vetores De Suporte).** Publicado em 10/04/2019. Disponível em: https://blogdozouza.wordpress.com/2019/04/10/algoritmo-svm-maquina-de-vetores-de-suporte-a-partir-de-exemplos-e-codigo-python-e-r/. Acesso em: 10/06/2021

SOUZA, Vanessa Faria de; PERRY, Gabriela Trindade (2019). **Mineração de Texto em Moocs: Análise da Relevância Temática de Postagens em Fóruns de Discussão**. RENOTE - Revista Novas Tecnologias na Educação. DOI: https://doi.org/10.22456/1679-1916.99471

SQUARISI, Dad; SALVADOR, Arlete (2020). **A arte de escrever bem: um guia para jornalistas e profissionais do texto**. 9ª Edição - São Paulo: Contexto, 2020. 144p.

STEFANINI - Group (2019). **PNL: entenda o que é o processamento de linguagem natural**. Disponível em: https://stefanini.com/pt-br/trends/artigos/o-que-e-processamento-de-linguagem-natural. Acesso em 20/05/2021

STEFANO, Ercilia de.(2016). **Análise da evolução da pesquisa em engenharia de transportes**. Tese (Programa de Engenharia de Transportes)-Universidade Federal do Rio de Janeiro(UFRJ).

STRIQUER, Marilúcia dos Santos Domingos(2018). **Produção textual dos alunos concluintes da educação Básica: uma análise do estabelecimento da coerência em redações do Enem**. Pensares em Revista. Publicado em 02/08/2018. Disponível em: https://www.e-publicacoes.uerj.br/index.php/pensaresemrevista/article/view/34666. Acesso em: 11/06/2020. ISSN 2317-2215. DOI: https://doi.org/10.12957/pr.2018.34666.

SUBRAMANIAN, Dhilip. **Text Mining in Python: Steps and Examples**. Publicado em 22/08/2019. Disponível em: https://pub.towardsai.net/text-mining-in-python-steps-and-examples-78b3f8fd913b. Acesso em 29/03/2021

TAFTI, P. A.; BADGER J.; LaRose E.; SHIRZADI E.; MAHNKE A.; MAYER J.; YE Z.; PAGE D.; PEISSIG P. (2017). **Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure**. JMIR Med Inform. 2017 Dec 8;5(4):e51. doi: 10.2196/medinform.9170. PMID: 29222076; PMCID: PMC5741828.

TAVARES, José Fernando. **O que a Inteligência Artificial pode fazer pelos editores**. Publicado em 03/03/2020. Disponível em: https://www.publishnews.com.br/materias/2020/03/03/o-que-a-inteligencia-artificial-pode-fazer-pelos-editores. Acesso em 13/04/2021

TREVISANI, Fernando de Melo (2019). **A importância do feedback na visibilidade da aprendizagem. Portal Desafios da Educação**. Publicado em 27/02/2019. Disponível em: https://desafiosdaeducacao.grupoa.com.br/feedback-na-aprendizagem/. Acesso em 10/06/2020.

USP- São Carlos. **USP desenvolve ferramenta de correção automática de redações**. Publicado em 10/03/2021. Disponível em: http://www.saocarlos.usp.br/usp-desenvolve-ferramenta-de-correcao-automatica-de-redacoes/. Acesso em 29/03/2021

VAPNIK, V. **The Nature of Statistical Learning Theory**. New York: Springer-Verlag, 1995.

VENEROSO, Joao Mateus de Freitas (2019). **Reconhecimento de entidades nomeadas na Web**. Dissertação (DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO), Universidade Federal de Minas Gerais.

WALTRICK, Camila (2020). **Machine Learning — O que é, tipos de aprendizagem de máquina, algoritmos e aplicações**. Publicado em: 07/05/2021. Disponível em: https://medium.com/camilawaltrick/introducao-machine-learning-o-que-e-tipos-de-aprendizado-de-maquina-445dcfb708f0. Acesso em: 31/03/2021

WESTER, Wim; DASCALU, Mihai; KURVERS, Hub; RUSETI, Stefan ; TRAUSAN-MATU, Stefan (2018). **Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training**. Computers & Education Volume 123, August 2018, Pages 212-224 https://doi.org/10.1016/j.compedu.2018.05.010

WESTERA, W., DASCALU, M., KURVERS, H., RUSETI, S., & Trausan-Matu, S. (2018). **Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training**. Computers & Education, 123, 212–224. doi:10.1016/j.compedu.2018.05.010

WULFF, P., BUSCHHÜTER, D., WESTPHAL, A. et al. **Computer-Based Classification of Preservice Physics Teachers' Written Reflections**. J Sci Educ Technol 30, 1–15 (2020). https://doi-org.ez345.periodicos.capes.gov.br/10.1007/s10956-020-09865-1

ZEYNEP, Ozer; Ilyas, Ozer, Findik, Oguz. (2018). **Diacritic restoration of Turkish tweets with word2vec**. Engineering Science and Technology, an International Journal. 21. 10.1016/j.jestch.2018.09.002.

ZHOU, Ming, Nan Duan; Shujie Liu; Heung-Yeung Shum (2020). **Progress in Neural NLP: Modeling, Learning, and Reasoning**. Engineering: Volume 6, Issue 3, March 2020, Pages 275-290. ISSN 2095-8099, https://doi.org/10.1016/j.eng.2019.12.014.

ZOLOTAREV, O.a; SOLOMENTSEV Y.b; KHAKIMOVA A.c; CHARNINE M.d.(2019). **Identification of semantic patterns in full-text documents using neural network methods**. CEUR Workshop Proceedings, Volume 2485, Pages 276 - 2792019 29th International Conference on Computer Graphics and Vision, Graphi Con 2019, Bryansk, 23 September 2019 - 26 September 2019, 153360

ZUBEN, Fernando J. Von; ATTUX, Romis R. F. (2016). **Árvores de Decisão**. Disponível em: https://pdfslide.tips/documents/arvoresdecisao.html. Acesso em: 21/10/2021. DCA/FEEC/Unicamp