PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Universidade Federal do Rio de Janeiro

Rafael Ris-Ala José Jardim

DESENVOLVIMENTO DE UM MODELO CLASSIFICADOR DE QUESTÕES PARA O CENÁRIO EDUCACIONAL BRASILEIRO FUNDAMENTADO EM CIÊNCIA DE DADOS

DISSERTAÇÃO DE MESTRADO

Instituto de Matemática

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

RAFAEL RIS-ALA JOSÉ JARDIM

DESENVOLVIMENTO DE UM MODELO CLASSIFICADOR DE QUESTÕES PARA O CENÁRIO EDUCACIONAL BRASILEIRO FUNDAMENTADO EM CIÊNCIA DE DADOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática e Instituto Tércio Pacciti, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Informática.

Orientador: Prof. Daniel Schneider, DSc.

Co-orientadora: Profa. Carla Delgado, DSc.

Rio de Janeiro
2022

CIP - Catalogação na Publicação

Jardim, Rafael Ris-Ala José
J37d Desenvolvimento de um modelo classificador de questões para o cenário educacional brasileiro fundamentado em ciência de dados / Rafael Ris-Ala José Jardim. -- Rio de Janeiro, 2022.
70 f.

Orientador: Daniel Menasché.
Coorientador: Carla Delgado.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Programa de Pós-Graduação em informática, 2022.

1. Aprendizagem de máquina. 2. Ciência de dados. 3. Classificador Automático de Questões. 4. Processamento de Linguagem Natural (PLN). 5. Orientado por dados. I. Menasché, Daniel, orient. II. Delgado, Carla, coorient. III. Título.

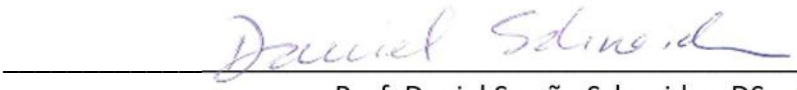
Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

RAFAEL RIS-ALA JOSÉ JARDIM

DESENVOLVIMENTO DE UM MODELO CLASSIFICADOR DE QUESTÕES PARA O CENÁRIO EDUCACIONAL BRASILEIRO FUNDAMENTADO EM CIÊNCIA DE DADOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática e Instituto Tércio Pacciti, Universidade Federal do Rio de Janeiro, como requisito parcial à obtenção do título de Mestre em Informática.

Aprovada em 31 de janeiro de 2022.



Prof. Daniel Serrão Schneider, DSc., UFRJ

Participação por videoconferência
Profa. Carla Amor Divino Moreira Delgado, DSc., UFRJ

Participação por videoconferência
Prof. Mônica Ferreira da Silva, DSc., UFRJ

Participação por videoconferência
Prof. Cláudia Lage Rebello da Motta, DSc., UFRJ

Participação por videoconferência
Prof. Marcos dos Santos, DSc., UFF

Rio de Janeiro

2022

Dedico este trabalho ao avanço científico e educacional do Brasil.

Agradecimentos

Agradeço a Deus, minha família, namorada e amigos pelo suporte.

Agradeço o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Agradeço a confiança de todos os membros do Programa de Pós-Graduação em Informática da Universidade Federal do Rio de Janeiro (PPGI/UFRJ).

“O que sabemos é uma gota; o que ignoramos é um oceano.”

Isaac Newton

RESUMO

Ciência de Dados tem obtido êxito no desenvolvimento de aplicativos para educação, inclusive na engenharia de classificadores automáticos de questões. Entretanto, estudos indicam que não há um modelo classificador que reconheça o nível de dificuldade de questões na língua portuguesa. Para superar esse desafio, esta dissertação propõe um modelo classificador automático de questões por dificuldade. Foi construído um corpus de questões de provas do Exame Nacional do Ensino Médio do Brasil, cujo grau de acerto foi validado por milhões de estudantes, via sabedoria das multidões, e um rótulo de dificuldade foi atribuído a elas. Em seguida, um *pipeline* foi projetado, no qual uma Rede Neural Convolucional recebia a matriz dos vetores das questões e previa a probabilidade das classes de dificuldade, na camada de saída. Além disso, sua performance foi aferida. Como resultado, obteve-se um classificador orientado por dados que identifica a dificuldade de questões em português. As principais contribuições desta pesquisa são: a combinação de tecnologias computacionais; o desenvolvimento de um corpus de questões validado; o desenvolvimento de um classificador automático eficaz; a comprovação empírica da performance do modelo; e o desenvolvimento de um Sistema de Apoio à Decisão (SAD).

Palavras-chave: Aprendizagem de Máquina; Ciência de Dados; Classificador Automático de Questões; Processamento de Linguagem Natural (PLN); Orientado por Dados.

ABSTRACT

Data Science has been successful in developing *e-learning* applications, including engineering of automatic question classifiers. However, studies indicate that there is no classifier model recognizes the level of difficulty of issues in the Portuguese language. To overcome this challenge, this research proposes an automatic question classifier model by difficulty. A corpus was built with exam questions from the “Exame Nacional do Ensino Médio” in Brazil, whose hit level was validated by millions of students and a difficulty label was assigned to them. Then, a pipeline was designed, in which a Convolutional Neural Network received the matrix of the question vectors and predicted the probability of the classes of difficulty, in the output layer. Besides, its performance was measured. As a result, a data-driven classifier was obtained that identifies the difficulty of questions in Portuguese. The main contributions of this research are: the combination of computational technologies; the development of a validated corpus of questions; the development of an effective automatic classifier; empirical evidence of the model's performance; and the development of a Decision Support System (DSS).

Keywords: Data-Driven; Data Science; Machine Learning; Natural Language Processing (NLP); Automatic Question Classifier.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

AI	Inteligência Artificial
BI	Business Intelligence
BNI	Banco Nacional de Itens
CH	Ciências humanas e suas tecnologias
CN	Ciência da natureza e suas tecnologias
CNN	Convolutional Neural Network
CNN	Rede Neural Convolucional
COVID-19	Doença do Coronavírus de 2019
DDDM	Tomada de Decisão Baseada em Dados
DDDM	Data-Driven Decision Making
DSS	Decision Support System
EAD	Educação a Distância
ENEM	Exame Nacional do Ensino Médio
ETL	Extract, Transform and Load
FAIR	Findable, Accessible, Interoperable and Reusable
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LC	Linguagens, códigos e suas tecnologias
LDB	Lei de Diretrizes e Bases da Educação
MEC	Ministério da Educação
MT	Matemática e suas tecnologias
NER	Named Enity Recognition
NER	Reconhecimento de Entidades Nomeadas
PISA	Programa Internacional de Avaliação de Alunos
PLN	Processamento de Linguagem Natural
POS	Part-of-speech
RQ	Research Questions
SAD	Sistema de Apoio à Decisão
TCT	Teoria Clássica dos Testes
TIC	Tecnologias de Informação e Comunicação

TOEFL	Test of English as a Foreign Language
TRI	Teoria de Resposta ao Item

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Atualmente, as Tecnologias de Informação e Comunicação (TIC) têm sido determinantes para o processo de mudança social, uma vez que representam o elemento fundamental de um novo tipo de sociedade: a sociedade da informação (PONTE, 2000). No contexto educacional, as TIC podem assumir diferentes propósitos educativos, a depender do objetivo a que se destinam.

Há alguns anos, o ensino presencial faz uso das TIC como recurso de apoio aos professores em sala de aula, que vão desde tecnologias tradicionais, como o quadro branco, por exemplo, até tecnologias digitais mais recentes, como os aplicativos de *e-learning* (GOMES, 2005).

E-learning ou aprendizado eletrônico é o processo de fornecimento de informações por meios digitais. Embora o conceito de *e-learning* tenha repercutido somente na década de 90, no século XIX já existiam cursos à distância oferecidos aos alunos, cujos materiais eram enviados por correio. No início do século XXI, com o *boom* da informática e o surgimento de novos materiais eletrônicos, como vídeos, áudios, e demais recursos da *web*, facilitou o acesso para atender às novas demandas sociais.

Websites, fóruns*, e-mails,* aplicativos de mensagens instantâneas, videoconferências, dentre outros inúmeros recursos computacionais e audiovisuais vêm atuando de forma cada vez mais presente, especialmente em se tratando de Educação à Distância (EAD). Aplicativos de *e-learning* são, portanto, sistemas que proporcionam a aprendizagem em um formato não presencial apoiada por Tecnologias de Informação e Comunicação (TIC) (KLAŠNJA-MILIĆEVIĆ et al., 2017).

Como supracitado, esses sistemas são amplamente usados na modalidade de EAD, que teve um recente aumento de demanda ocasionado pelo afastamento social provocado pela doença do coronavírus de 2019 (COVID-19) (WORLD HEALTH ORGANIZATION, 2020).

E-learning tem sido tema de desafios computacionais (G. SUGANYA, 2020), principalmente no aperfeiçoamento de técnicas avaliativas de cursos (VERDÚ et al., 2012) e sistemas de recomendação de questões conforme o perfil do aluno (NABIZADEH et al., 2012;

XIA et al., 2019). Para tal, é necessário fazer uma correta categorização das questões de provas (SILVA, BITTENCOURT e MALDONADO, 2019). Um exemplo de questão de prova do Exame Nacional do Ensino Médio (ENEM) é apresentado na Figura 1.

Figura 1: Exemplo da organização das questões de prova do ENEM[1].

A terapia celular tem sido amplamente divulgada como revolucionária, por permitir a regeneração de tecidos a partir de células novas. Entretanto, a técnica de se introduzirem novas células em um tecido, para o tratamento de enfermidades em indivíduos, já era aplicada rotineiramente em hospitais.

A que técnica refere-se o texto?

A) Vacina.
B) Biópsia.
C) Hemodiálise.
D) Quimioterapia.
E) Transfusão de sangue.

Fonte: MINISTÉRIO DA EDUCAÇÃO (2020a).

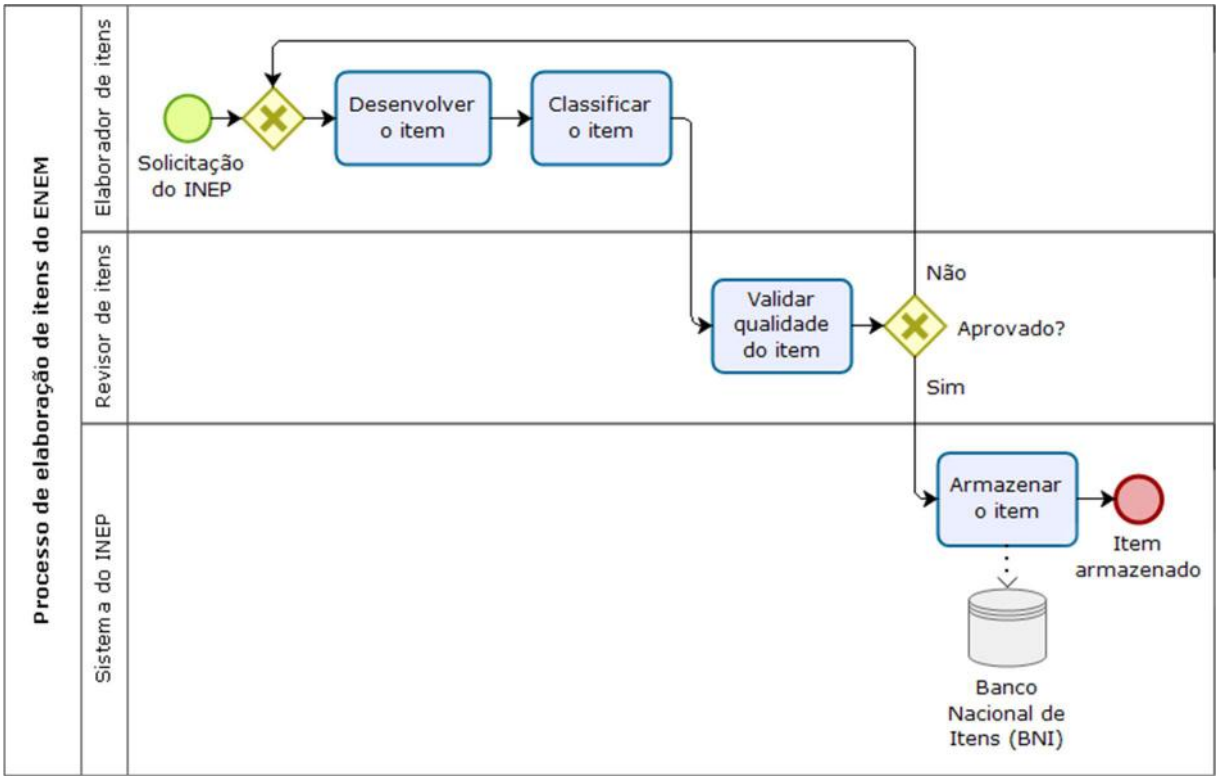
A avaliação da aprendizagem faz parte da ação pedagógica, realizando uma mediação entre ensino e aprendizagem (LUCKESI, 2014; NASCIMENTO, FAGUNDES e SOUZA, 2021). Os instrumentos de avaliação devem ser capazes de aferir o desempenho dos alunos em diferentes níveis de aquisição de conhecimento (FERRAZ e BELHOT, 2010) e, para tanto, devem ser elaborados com qualidade (STIGGINS, 2002).

Sendo assim, além de se conceber o currículo dos alunos visando atender aos objetivos educacionais (OKOYE, SUMMER e BETHARD, 2013), há também o aspecto classificatório e seletivo das provas (NETO e AQUINO, 2009), a fim de se eleger indivíduos mais aptos em processos seletivos.

A partir da exploração dos editais de chamada pública para seleção e credenciamento de elaboradores e revisores do Banco Nacional de Itens (BNI) (MINISTÉRIO DA EDUCAÇÃO, 2020b), foi possível conceber o processo de elaboração dos itens do ENEM, conforme sintetizado na Figura 2.

[1]A alternativa correta para a questão exemplificada é a letra “E”.

Figura 2: Processo simplificado da elaboração de itens da prova do ENEM.



Fonte: do autor (2022).

Um elaborador credenciado desenvolve o item da prova e o classifica de acordo com o nível de dificuldade citado anteriormente. Em seguida, um revisor valida a qualidade desse item que, sendo reprovado, é devolvido ao elaborador para reformulação e, caso seja aprovado, é armazenado no BNI.

No ENEM, o cálculo das notas dos candidatos não segue a Teoria Clássica dos Testes (TCT), segundo a qual um mesmo valor é fornecido para cada questão acertada, mas a Teoria de Resposta ao Item (TRI) (MINISTÉRIO DA EDUCAÇÃO, 2020c), um método de correção em que as questões são calibradas de forma a receberem valores de acordo com outras variáveis (MIRANDA, FERREIRA e DIAS, 2019), como é feito no Programa Internacional de Avaliação de Alunos (PISA) e no Test of English as a Foreign Language (TOEFL).

A classificação dessas questões tem se mostrado promissora com o uso do Processamento de Linguagem Natural – PLN (*Natural Language Processing - NLP*) (SILVA, BITTENCOURT e MALDONADO, 2019; LI e ROTH, 2002). O PLN é uma subárea da Inteligência Artificial que consiste no desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em alguma língua natural (PEREIRA, 2011),

cuja finalidade é capacitar os computadores para processamento e manipulação da linguagem humana, falada ou escrita.

O processamento da linguagem natural se dá mediante mineração e extração de significado dos conteúdos disponibilizados na Internet, os chamados dados não estruturados. A mineração de textos ou *text mining*, nesse contexto, oferece um conjunto de métodos que permite a navegação, organização e descoberta inteligente de informação utilizando bases de dados não estruturadas (SANTOS et al., 2014).

A análise textual utiliza algoritmos, como os de Aprendizagem Profunda (*Deep Learning*), tecnologia que utiliza as chamadas redes neurais, que em muito se assemelham ao modo de funcionamento do cérebro, ou redes neurais artificiais para a extração de conhecimento da linguagem humana. No PLN, é necessário converter informações textuais em informações vetoriais, por meio de técnicas de vetorização, agrupá-las em classes, descobrir as relações entre elas e atribuir um *label* (rótulo) a cada texto informado.

1.2 PROBLEMÁTICA

A classificação é uma das mais importantes categorias de problema de *Machine Learning*, de modo que modelos classificadores de questões podem ser gerados por algoritmos aptos a extrairem *features* e generalizarem classes a partir dos dados coletados.

Nesse aspecto, o ENEM, por se tratar de uma avaliação em larga escala, é uma confiável fonte de informação para a geração da classificação de questões. Considerando-se o contexto educacional, a criação de um modelo classificador do nível de dificuldade de questões pode apresentar aplicação prática imediata nos sistemas de *e-learning*, uma vez que fornece apoio a professores, influenciando positivamente na aprendizagem dos alunos.

Um problema endereçado pelo PLN no contexto de *e-learning* é determinar a dificuldade da questão, aspecto pelo qual se atribui um peso que contribui para o cálculo da nota do aluno. Se a prova for fácil demais, não terá cumprido sua função seletiva, pois a maioria dos candidatos passará. Se a prova for difícil demais, a maioria irá mal e poderá não haver selecionados. Por isso, para cumprir seu papel, a avaliação deve ter uma complexidade equilibrada.

Tendo em vista o processo de elaboração de avaliações para auxiliar o professor na atribuição do nível de dificuldade da questão, foi formulada a pergunta de pesquisa desta

dissertação, conforme orientações de Dias e Silva (2010): como as tecnologias da ciência de dados podem auxiliar (sistemas de *e-learning*) a classificação do nível de dificuldade de questões em português brasileiro em sistemas de *e-learning*?

1.3 OBJETIVO

O objetivo deste trabalho é desenvolver um modelo classificador automático de questões em português brasileiro por meio da combinação de tecnologias computacionais.

1.4 PÚBLICO-ALVO E RELEVÂNCIA DA PESQUISA

A educação contribui para o desenvolvimento social, econômico e cultural do indivíduo, da sociedade e do país. Em 2019, mais de 3 milhões de estudantes ingressaram nas 2608 instituições de educação superior, públicas e privadas, do Brasil (MINISTÉRIO DA EDUCAÇÃO, 2019).

O público-alvo deste projeto são os professores e os institutos de educação, que contam com tecnologias inovadoras que auxiliem na produção de melhores avaliações; e cientistas de dados interessados em aplicações práticas das pesquisas de PLN. A relevância social está em beneficiar milhões de estudantes brasileiros que poderão confiar em provas de complexidade nivelada. Além disso, a criação de um modelo automático de classificação de questões poderia se tornar uma ferramenta útil ao professor, ao possibilitar que o profissional invista o tempo que seria gasto com essa atividade realizando outras tarefas.

No que concerne à relevância acadêmica e científica, o sucesso desses modelos permitiria a implementação de sistemas de recomendação de questões, de acordo com complexidade compatível à necessidade de aprendizagem dos alunos, promovendo a aprendizagem personalizada (*personalized learning*) (G. SUGANYA, 2020; NABIZADEH et al., 2012).

1.5 ORGANIZAÇÃO

Os próximos capítulos desta dissertação estão organizados da seguinte forma: capítulo 2, em que são apresentados o referencial teórico e os trabalhos relacionados; capítulo 3, no

qual é apresentado trabalhos relacionados na academia e no mercado; capítulo 4, em que é descrita a metodologia científica de ciência de dados deste estudo; capítulo 5, na qual se consolidam os resultados obtidos; capítulo 6, em que se discutem alternativas e possibilidades consideradas na experimentação; e capítulo 7, em que se traz um resumo da pesquisa e oportunidades de melhorias futuras.

2 REFERENCIAL TEÓRICO

2.1 INTELIGÊNCIA ARTIFICIAL

Inteligência Artificial (IA) é um ramo da ciência que tem por objetivo simular a inteligência humana mediante o uso de recursos tecnológicos. O termo Inteligência Artificial ficou conhecido devido a uma fala de John McCarthy durante uma conferência sobre tecnologia em Dartmouth College (EUA), em 1956 (SILVA e MAIRINK, 2019).

No entanto, o assunto já vinha sendo debatido por cientistas como Herbet Simon e Allen Newell, considerados pioneiros ao criarem o primeiro laboratório de IA na Universidade de Carnergie Mellon (EUA), e Alan Turing, mundialmente reconhecido como o "pai da computação" e o primeiro a articular uma visão completa da IA em seu artigo *Computing Machinery and Intelligency* (1950) (GOMES, 2010).

Com o intuito de resolver problemas, criar soluções e, inclusive, tomar decisões pelo ser humano, relacionando-se com diversas áreas do conhecimento, como ética, direito, economia, medicina, ciências e tecnologia (SILVA e MAIRINK, 2019; MORAIS et al., 2020), o uso da IA visa, portanto, facilitar/revolucionar diferentes áreas do cotidiano. Isso porque, ao sistematizar e automatizar tarefas intelectuais, a IA se torna potencialmente relevante para qualquer esfera que envolva a atividade intelectual humana (GOMES, 2010).

Historicamente, é possível relacionar a Inteligência Artificial a 4 linhas de pensamento, quais sejam: i) sistemas que pensam como seres humanos (HAUGELAND, 1985); ii) sistemas que atuam como seres humanos (KURZWEIL, 1990); iii) sistemas que pensam racionalmente (CHARNIAK; MCDERMOTT, 1985); e iv) sistemas que atuam racionalmente (POOLE et al., 1998), de modo que as linhas de pensamento i e iii referem-se ao processo de pensamento e raciocínio, enquanto as linhas ii e iv dizem respeito a aspectos comportamentais (GOMES, 2010). Além disso, as linhas de pensamento i e ii estão intrinsecamente ligadas ao sucesso em relação à fidelidade ao desempenho humano, ao passo que iii e iv medem o sucesso por meio ao conceito ideal de inteligência (racionalidade) (GOMES, 2010).

Se antes a inteligência era tida como a capacidade de raciocinar, hoje entende-se como poder computacional (TEIXEIRA, 2009). Nesse sentido, uma das tecnologias utilizadas para obter a IA é o Aprendizado de Máquina (*Machine Learning*), que consiste em um método de análise de dados que automatiza a construção de modelos analíticos (MOREIRA et al., 2022)

mediante princípios do Aprendizado Indutivo (as conclusões gerais são induzidas a partir de observações específicas). Nesse contexto, a indução atua como uma das principais formas de derivação de novos conhecimentos (recurso amplamente utilizado pelo cérebro humano) e predição de eventos futuros, caracterizada, dessa forma, como o raciocínio que parte de um conceito específico e o generaliza, ou seja, da parte para o todo (MONARD e BARANAUSKAS, 2003).

Usualmente, o aprendizado indutivo é implementado por algoritmos que processam um conjunto de dados, extraindo um modelo apto a explicar ou representá-los sob determinado aspecto, a partir de 3 modalidades: supervisionada, não-supervisionada e semi-supervisionada (BRUNIALTI et al., 2015).

Na modalidade supervisionada, os algoritmos ajustam parâmetros de um modelo a partir do erro medido entre respostas obtidas e esperadas (BRUNIALTI et al., 2015), isto é, a previsão é baseada nos dados fornecidos. Nela, os dados devem ser rotulados como forma de auxiliar na tomada de decisões, a exemplo dos vetores de entrada e seus vetores-alvo correspondentes.

Na modalidade não-supervisionada, os parâmetros de determinado modelo são ajustados de acordo com a maximização de medidas de qualidade das respostas obtidas (BRUNIALTI et al., 2015), de modo a extrair significado a partir do conjunto de dados fornecido. Nesse contexto, os algoritmos buscam correlações sem nenhuma entrada externa além dos dados brutos, ou seja, apresenta-se um conjunto de vetores de entrada sem nenhum vetor-alvo correspondente, por exemplo.

Na modalidade semi-supervisionada, utiliza-se algoritmos híbridos, os quais fazem uso dos recursos de correção de erro e de maximização de medidas de qualidade de acordo com a necessidade (BRUNIALTI et al., 2015). Trata-se de um método supervisionado que utiliza algoritmos sem a necessidade de uma imensa quantidade de dados rotulados durante o treinamento.

De forma geral, na aprendizagem de máquina o aprendizado ocorre através de exemplos, ou seja, exemplos são passados e há o reconhecimento de padrões. Já o Aprendizado por Reforço (*Reinforcement Learning*) vai além, pois assume-se que não se há exemplos, ou seja, assume-se que o sistema poderá interagir com o ambiente e aprender por meio de recompensas (SILVER et al., 2021), como no treinamento de um cachorro que ganha

um biscoito quando cumpre um comando, reforçando um comportamento. Assim, a ideia do aprendizado por reforço é desenvolver um programa que poderá interagir com o ambiente, tentar coisas arbitrárias e que quando fizer a coisa certa receberá uma recompensa. Um exemplo muito utilizado é em jogos no qual o personagem é penalizado quando cai em um buraco ou recompensado quando pega uma moeda.

A modalidade supervisionada está muito relacionada com as técnicas de Aprendizagem Profunda, termo utilizado para representar o problema de treinar redes neurais artificiais que efetuem o aprendizado de características de forma hierárquica, de modo que características relativas aos níveis mais altos da hierarquia sejam formadas mediante a combinação de características de mais baixo nível (BEZERRA, 2016). As redes neurais necessitam de grandes quantidades de duas coisas: poder de computação, que permite que o programa analise os exemplos em alta velocidade, e dados, os quais “treinam” o programa para reconhecer padrões por meio de inúmeros exemplos (LEE, 2019).

Atualmente, as redes neurais ganharam destaque sob a forma de aprendizado profundo. Isso porque o cérebro humano é uma rede ampla de células interconectadas, umas curtas, outras longas, e que podem estar conectadas apenas umas às outras ou a muitas células. Por essas conexões passam os sinais elétricos, em várias taxas, os quais dão origem a disparos neurais subsequentes. Ao conhecer o modo como o cérebro opera, o homem, então, passou a aplicá-lo às máquinas, a partir da criação de redes neurais profundas que atuam como cérebros virtuais reduzidos.

Nesse sentido, a Aprendizagem Profunda é uma forma de usar grandes quantidades de dados com o objetivo de fazer com que as máquinas operem da mesma forma como o ser humano faz, e sem instruções explícitas. Trata-se, portanto, de um método que permite ao ser humano treinar uma IA para prever saídas, considerando-se, para tanto um conjunto de entradas. Apoiada pelo desenvolvimento tecnológico a nível computacional, a IA obteve resultados práticos interessantes ao desenvolver novos métodos e algoritmos, a partir da aprendizagem profunda, solucionando problemas de forma prática e eficiente (PIRES, 2017).

Em virtude do exposto, é possível depreender que a Inteligência Artificial modifica a relação do ser humano com os dados. Nesse contexto, os algoritmos de Aprendizagem Profunda simulam a capacidade de células neurais, conferindo às máquinas a capacidade de aprender na medida em que entram em contato com novas informações.

A despeito do que tem sido difundido pelo imaginário coletivo sobre a eventual substituição do homem pela máquina, a função da IA vai além da automatização de processos por si só na medida em que os resultados se mostram mais efetivos quando associadas inteligência artificial e inteligência humana. Nesse contexto, a função do professor se torna ainda mais essencial, uma vez que é ele um grande impulsionador do engajamento do aluno com a tecnologia. Isso porque o professor possui valores e habilidades humanas, como a empatia, por exemplo, que possui papel fundamental na aprendizagem e que dificilmente será reproduzido por IA.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL

Textos são dados não estruturados e as palavras apresentam uma diversidade de significado, portanto, é difícil criar regras que abranjam todas as variações linguísticas. Conforme observado nas investigações da revisão sistemática da análise de dados semânticos (GUEDEA-NORIEGA e GARCÍA-SANCHEZ, 2019), as tecnologias de semântica têm sido empregadas com sucesso, propiciando a descoberta de conhecimento útil, e uma de suas vantagens é a de "permitir processos de raciocínio e inferência lógica sobre os dados" (GUEDEA-NORIEGA e GARCÍA-SANCHEZ, 2019, p. 803).

Dentre essas tecnologias, destaca-se o PLN, que representa cada palavra como um vetor de várias dimensões. Os valores dessas dimensões são estabelecidos de acordo com a descoberta de padrões, de forma a capturar a relação entre as palavras, campo de estudo explorado por Aprendizagem Profunda.

Salas, Vidal e Martinez-Trinidad averiguaram o estado atual da Aprendizagem Profunda e ressaltaram que seu poder consiste na "capacidade de extrair atributos cada vez mais abstratos apenas dos dados de entrada" (SALAS, VIDAL e MARTINEZ-TRINIDAD, 2019, p. 1928).

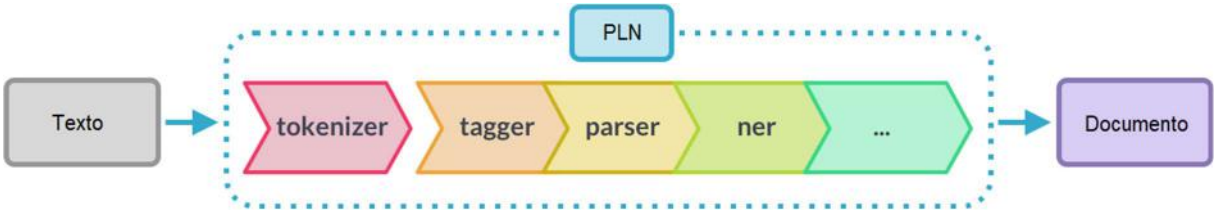
Dentre os algoritmos de Aprendizagem Profunda, destacaram a relevância das Redes Neurais Convolucionais (CNN), por utilizarem "conexões locais e pesos compartilhados, ao contrário de camadas totalmente conectadas, resultando em redes mais rápidas com menos parâmetros que são mais fáceis e rápidas de se treinar" (SALAS, VIDAL e MARTINEZ-TRINIDAD, 2019, p. 1929), muito utilizadas na detecção de objetos em imagens, como no reconhecimento de embarcações em imagens de satélite (SANTOS et al., 2020).

Foram observados alguns algoritmos de PLN, como BERT (DEVLIN et al., 2019) e GPT-3 (BROWN et al., 2020). Entretanto, neste trabalho, optou-se pela adoção da biblioteca spaCy (GITHUB, 2020a), que fornece uma variedade de anotações linguísticas da estrutura gramatical e dispõe, conforme compreendido, de uma CNN para a categorização de textos. Essa biblioteca foi adotada por disponibilizar um *pipeline* previamente treinado em português, ser de código aberto, gratuita, possuir uma boa documentação, permitir personalizar sua configuração e ter sido considerada o analisador sintático mais rápido e preciso do mundo (CHOI, TRETREAULT e STENT, 2015).

Essa CNN foi preparada para reconhecer o português a partir do "UD Portuguese Bosque v2.5", uma coletânea de textos jornalísticos contemporâneos, abrangendo sua variedade linguística, sendo o corpus composto por milhões de palavras (GITHUB, 2020b), incluindo 500 mil vetores únicos de 300 dimensões.

A spaCy fornece um *pipeline* customizável que inclui vários componentes de processamento de texto executados sequencialmente, também conhecido como *pipeline* de processamento, ilustrado na Figura 3.

Figura 3: *Pipeline* usual do spaCy.



Fonte: adaptado de GITHUB (2020a).

De forma geral, esse *pipeline* apresenta alguns componentes treinados. Primeiramente, o *tokenizer* segmenta a *string* informada em elementos (*tokens*) e os armazena em um documento. Depois disso, o *tagger* é executado para atribuir rótulos da classe gramatical (*part-of-speech* - *POS*). O *parser* realiza a análise de dependência sintática entre as palavras. Em seguida, o componente NER (*Named Enity Recognition* - Reconhecimento de Entidades Nomeadas) detecta e classifica as entidades como pessoas, lugares e organizações no texto.

Finalmente, é chamado um componente *vectorizer*, para converter o texto em vetores, havendo muitas formas de se fazer essa representação. Duas técnicas bem populares são "*one*

hot encoding" e "*word embedding*". Enquanto no caso do *one hot encoding* cada *token* (palavra) é representada por um vetor de dimensão enorme (do tamanho da quantidade de palavras do dicionário), em *word embedding* cada *token* é associada a um vetor de dimensão pequena (do tamanho da quantidade de significados). Além disso, as técnicas de *word embedding* são aplicadas de forma a incorporar os aspectos semânticos das palavras nos vetores, obtidos por meio do relacionamento entre as palavras (SRINIVASULU, 2021).

Aquelas palavras que apresentam significados semânticos semelhantes devem possuir vetores próximos, o que permite às máquinas utilizarem a proximidade dos vetores para identificar a similaridade entre textos. Como exemplo, o vetor para a palavra "raio" pode estar próximo ao do vetor da palavra "círculo" em uma dimensão e próximo a "tempestade" em outra. Alguns algoritmos bem-sucedidos de *word embedding* são Word2vec (MIKOLOV et al., 2013a; MIKOLOV et al., 2013b), GloVe (PENNINGTON, SOCHER e MANNING, 2014), FastText (BOJANOWSKI et al., 2017), ELM (PETERS et al., 2018) e BERT (DEVLIN et al., 2019).

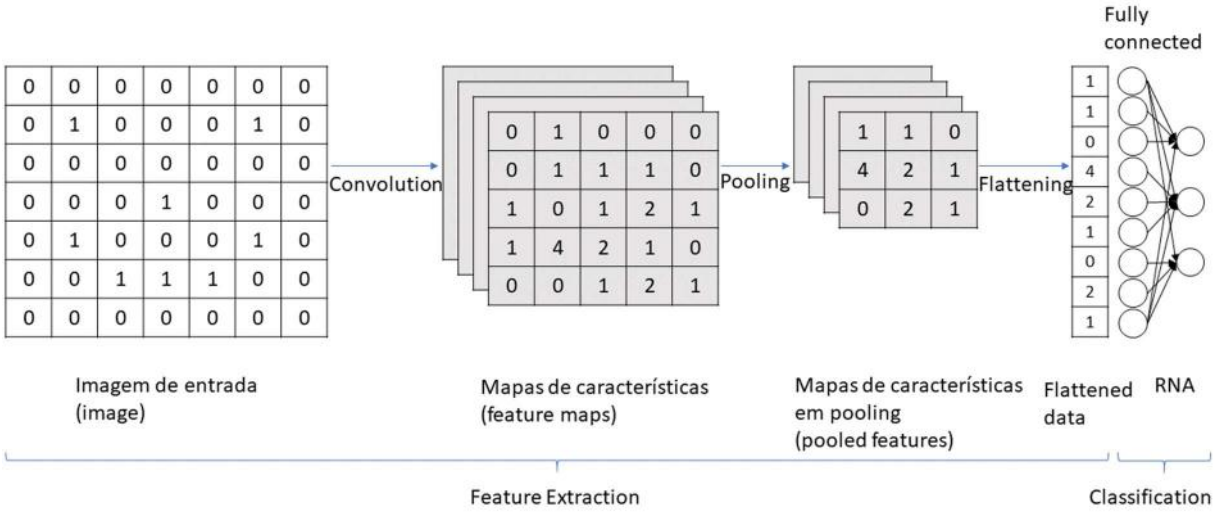
Para a realização deste trabalho, além dos componentes apresentados, foi criado o componente para a categorização de textos "textcat". O objeto "Documento" resultante é o conjunto de *tokens* com todas as informações geradas pelo *pipeline*.

2.3 REDE NEURAL CONVOLUCIONAL

As Redes Neurais Convolucionais (CNN) são modelos de Redes Neurais comumente utilizados para a classificação de imagens. Tem-se uma imagem de entrada, que é submetida à CNN, que faz uma classificação da imagem a partir da estimativa da probabilidade de a imagem pertencer a uma classe ou a outra.

Em um processo de aprendizado supervisionado, imagens previamente etiquetadas com as respectivas classes são passadas à CNN para reconhecer, aprender os padrões nessa imagem e equaliza seus parâmetros a fim de acertar a classe. Posteriormente, em uma fase de teste, novas imagens sem as respectivas classes são passadas à CNN, que tenta determinar a classe das imagens. A publicação original sobre as CNN é de Yann LeCun, de 1998 (LECUN et al., 1998). As CNN passam por 4 etapas: *Convolution*, *Pooling*, *Flattening* e *Full connected*, apresentado na Figura 4.

Figura 4: Rede Neural Convolucional



Fonte: adaptado de LECUN et al. (1998).

Como apresentado, convolução é uma operação muito utilizada no tratamento de imagens, e não é um conceito exclusivo da área de Redes Neurais. Convolução é uma operação de uma matriz por outra matriz. Nessa operação, deve-se multiplicar a matriz da imagem pela matriz do filtro (detector de características), somar os valores da matriz resultante e inserir esse resultado em um mapa de característica (*feature map*). O tamanho do movimento do detector de característica é denominado *stride*, geralmente de 1, 2 ou 3 pixels. Assim, o detector de característica vai se deslocando por toda a imagem para formar este mapa. Esse processo se repete até gerar todos os detectores de características (WU, 2017). A Figura 5 exibe a fórmula da convolução.

Figura 5: Fórmula da convolução.

$$(f * g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau$$

Fonte: SUPER DATA SCIENCE TEAM (2021).

As imagens do mundo real são não-lineares, ou seja, não seguem um padrão na mesma direção. Não existe uma linearidade, os objetos estão misturados (e possuem bordas). Quando se utiliza uma CNN, corre-se o risco de gerar muita linearidade. O propósito de se utilizar a função de ativação *Rectified Linear Unit* (ReLU) é diminuir a linearidade gerada pela convolução, removendo-se os valores negativos (KUO, 2014).

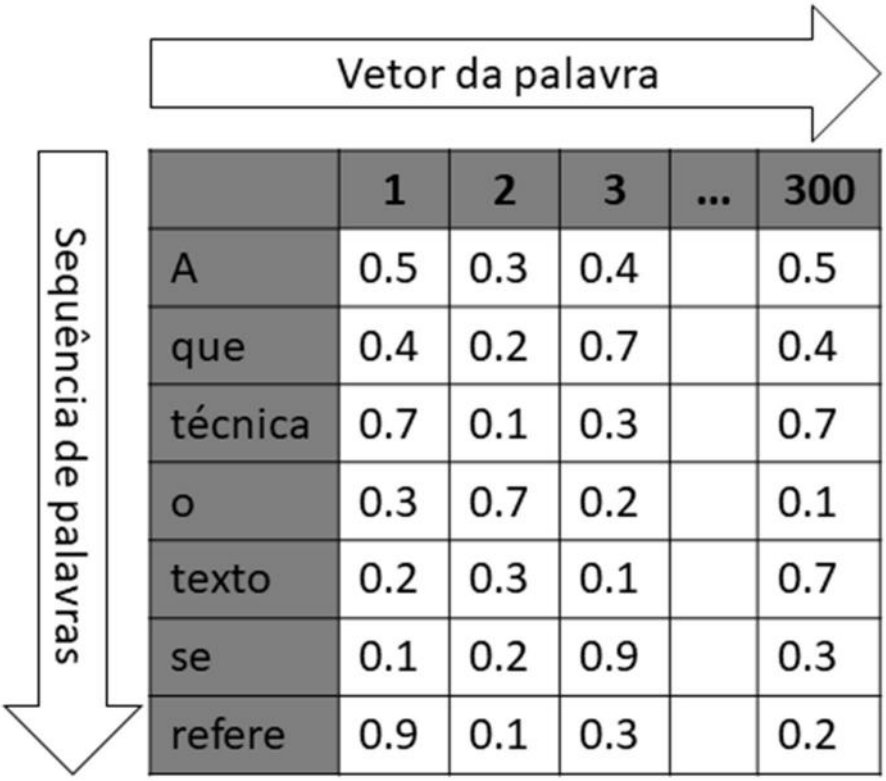
O objetivo é destacar as características mais importantes. Existem vários tipos de pooling, sendo o "*max pooling*" o mais usual. Nesse caso, é estabelecido um quadro que percorre o "mapa de característica", registrando o maior valor dentro desse quadro em um "mapa de característica em *pooling*". Isso permite preservar as características mais importantes dos objetos, reduzir o tamanho dos dados originais e, consequentemente, reduzir a quantidade de dados que serão introduzidos na Rede Neural (SCHERER, MÜLLER e BEHNKE, 2010).

Na etapa de *flattening* (achatamento), a matriz do "mapa de características em *pooling*" é transformada em um vetor. Esses valores serão as entradas da Rede Neural Artificial. Observe que todo o trabalho até agora foi para realizar um tratamento do dado a ser introduzido em uma RNA. Até poderia ser passada uma outra técnica para o processamento dos dados obtidos, mas a RNA tem uma capacidade muito boa de determinar as características relevantes na classificação das imagens.

Nessa etapa é que as redes neurais são treinadas. Em CNN, essa RNA é totalmente conectada (*fully connected*), o que significa que todos os neurônios da última camada estão conectados à camada de saída. Nem sempre uma RNA é totalmente conectada, porém, no caso das CNN, elas têm essa característica. Os dados obtidos na etapa de *flattening* são passados para a camada de entrada da RNA e processados pelas camadas ocultas, para equacionar os pesos de cada atributo. A camada de saída, por sua vez, determina a probabilidade de a imagem passada pertencer a cada umas das classes.

Conforme apresentado, as CNN foram tradicionalmente desenvolvidas para trabalharem com imagens. De forma semelhante, houve um aprimoramento para sua aplicação em textos. A diferença fica por conta de que a matriz de entrada ao invés de ser uma imagem é o texto, onde a sequência de palavras da frase são as linhas e os valores das dimensões dos vetores da palavra são as colunas, ilustrado na Figura 6.

Figura 6: Exemplo de matriz de uma frase.



	1	2	3	...	300
A	0.5	0.3	0.4		0.5
que	0.4	0.2	0.7		0.4
técnica	0.7	0.1	0.3		0.7
o	0.3	0.7	0.2		0.1
texto	0.2	0.3	0.1		0.7
se	0.1	0.2	0.9		0.3
refere	0.9	0.1	0.3		0.2

Fonte: do autor (2022).

2.4 *CROWDSOURCING*

O termo *crowdsourcing* foi utilizado pela primeira vez em 2006, por Jeff Howe e Mark Robinson, e diz respeito à junção dos termos *crowd* (multidão) e *outsourcing* (terceirização), os quais deram origem ao novo conceito de interação social, fundado na construção coletiva de soluções que beneficiem a todos (CORREIA et al., 2018). A sabedoria da multidão transpôs, nesse contexto, a lógica testemunhal, inserindo-se, a partir de então, em uma lógica imersiva em que o conhecimento ou opiniões são construídos em regime de cooperação, dentro das comunidades virtuais que disponibilizam espaço para que a multidão atribua novos usos aos seus produtos e/ou serviços (ANGELI e MALINI, 2011).

Na concepção de Howe, multidão quase sempre superará o desempenho de um grupo de funcionários, dadas as condições adequadas (SCHEE, 2009). Nesse sentido, multidão é qualquer pessoa com acesso à internet e as condições adequadas dizem respeito a uma rede virtual e motivação, que não necessariamente está ligada a algum tipo de compensação (financeira ou não), mas à satisfação estimulada pela participação (SCHEE, 2009). As

estratégias de *crowdsourcing* possuem êxito no campo da publicidade e propaganda, por exemplo, uma vez que rompem com o paradigma do *rich media*, difundindo a publicidade como um serviço tecnológico (ANGELI e MALINI, 2011).

A necessidade de inovar as estratégias de competição propicia a utilização do *crowdsourcing* pelas empresas, uma vez que se trata de um método que favorece a democratização do conhecimento (MELO et al., 2014), substituindo a realização de processos restritos a determinado grupo de especialistas, para um processo que inclui uma quantidade superior de pessoas, dentro e fora da organização (MELO et al., 2014; SCHEE, 2009).

O avanço da tecnologia e o surgimento da chamada Internet 2.0 (*web* 2.0), termo atribuído à segunda geração de comunidades e serviços que tem como princípio a ideia de *web* enquanto plataforma, com *blogs* e mídias sociais, veio acompanhado da expansão da interação entre as pessoas no espaço virtual. Isso porque se, até então, o usuário da Internet estava limitado a visualizar o conteúdo postado por terceiros ou realizar apenas buscas pela Internet, com a evolução da interação on-line, ele passou a ter a possibilidade de emitir opiniões e interagir com outras pessoas em tempo real, podendo, inclusive, ser remunerado ao enviar sugestões que contribuam para o aperfeiçoamento de serviços ou produtos (SEBRAE, 2014).

As mudanças tecnológicas oriundas das demandas provocadas pela atualidade fomentaram a competitividade do mercado, que teve que buscar soluções inovadoras, e de baixo custo, para ofertar seus produtos. A globalização do mercado propiciou o desenvolvimento de estratégias eficazes na solução de problemas, desenvolvimento de novas tecnologias, criação de conteúdos e promoção de serviços como o *crowdsourcing*, que atualmente é utilizado por inúmeras organizações, com o intuito de melhorar ou criar seus produtos (MELO et al., 2014).

Tem-se, portanto, que a força das multidões, quando utilizada por empresas como ferramenta de ações estratégicas, podem beneficiar, mediante a criação de produtos e/ou serviços, a sociedade, haja vista que a internet possibilitou a globalização da participação de pessoas em diversas áreas, somando seus conhecimentos para criar uma inteligência única.

Atualmente, existem diversas plataformas e aplicativos que fazem uso do *crowdsourcing*, como a *Wikipedia*, que hospeda conteúdos providos voluntariamente por usuários da internet, e o *Waze*, que faz uso da geolocalização dos dados, facilmente obtida

por meio dos *smartphones* (MENDES JR, 2018). Os dados coletados por meio do *crowdsourcing* podem ser extraídos pelo *Machine Learning*, uma vez que se trata de uma das técnicas de inteligência computacional que produz conhecimento a partir das informações coletadas mediante o fornecimento de uma grande quantidade de dados (MENDES JR, 2018).

Em virtude da capacidade que a multidão possui de gerar dados para o conjunto de treinamento de programas de *Machine Learning* (ABHIGNA, SONI e DIXIT, 2018), os pesquisadores adotaram o *crowdsourcing* como ferramenta para disponibilizar dados suficientes para treinar seus programas de forma eficaz (MENDES JR, 2018).

2.5 TOMADA DE DECISÃO BASEADA EM DADOS

Um dos primeiros modelos de tomada de decisão foi idealizado por von Neumann e Morgenstern (1947). A Teoria da Utilidade Esperada (*Expected Utility Theory*) compreende que o indivíduo toma sua decisão de acordo com a maior utilidade estimada, sendo a função utilidade (*utility function*) representada pela relação de um conjunto de axiomas.

A Teoria da Utilidade Esperada nem sempre corresponde à maneira como as pessoas realmente tomam decisões. Herbert Simon (1955) propôs que pessoas são racionais até certo ponto, e que o envolvimento de incertezas pode explicar decisões discrepantes sob determinadas circunstâncias, definido como Teoria da Racionalidade Limitada (*Bounded Rationality Theory*).

De acordo com Simon, o tomador de decisão geralmente seleciona a primeira solução que atende a um conjunto predefinido de restrições (SIMON, 1955). Daniel Kahneman (2011) aponta que os humanos podem se desviar da regra da racionalidade e de decisões e, em diversos casos, podem ser influenciados por características irrelevantes à atividade em evidência. Esses estudos mostram que as decisões são afetadas por circunstâncias adversas.

A fim de auxiliar a pessoa a tomar decisões assertivas, surge um agrupamento de sistema tecnológico que centraliza recursos e reduz variações na sistemática do processo de arbitragem. O termo geral para qualquer aplicativo de computador que aprimora a capacidade de uma pessoa tomar decisões é denominado Sistema de Apoio à Decisão - SAD (*Decision Support System - DSS*) (KEEN e MORTON, 1978). A exploração dessa obra e dos estudos de Power (2011) possibilitou estabelecer os tipos de estratégia que apoiam os tomadores de decisão, conforme exposto no Quadro 1.

Quadro 1: Descrição dos tipos de orientações de decisões.

Tipo	Conceito
Communication-driven (baseado em comunicação)	Utilizam o histórico de mensagens para obter um parecer.
Data-driven (baseado em dado)	Consultam uma base de dados para orientar julgamentos.
Document-driven (baseado em documento)	Pesquisam em um conjunto específico de documentos para apoiar análises.
Knowledge-driven (baseado em conhecimento)	Buscam em registros de experiências, valores e atitudes para transmitir aconselhamentos.
Model-driven (baseado em modelo)	Recorrem a funções para representar o comportamento de um sistema.

Fonte: do autor (2022).

Durante a construção de provas, o elaborador deve indicar o nível de dificuldade de cada questão. Trata-se de um processo de tomada de decisão, visto que o elaborador deve escolher uma alternativa específica. Esse problema de escolha pode apresentar variáveis diversas e ser melhor suportado com o uso de novas tecnologias que reduzam ou eliminem aspectos subjetivos.

Percebe-se que a classificação do nível de dificuldade das questões adotada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) apresenta uma abordagem baseada na experiência do professor. No entanto, a Tomada de Decisão Baseada em Dados (*Data-Driven Decision Making - DDDM*) tende a ser o processo mais factível (IKEMOTO e MARSH, 2007). A razão está no fato de que a decisão orientada por dados tende a ser mais confiável, tendo em vista sua reprodutibilidade e qualidade, quando comparadas às decisões baseadas em conhecimentos ou experiências, que são subjetivas à psiquê humana.

3 TRABALHOS RELACIONADOS

3.1 ANÁLISE NA ACADEMIA

Em uma recente revisão sistemática foi investigado o estado da arte de classificadores automáticos de questões (SILVA, BITTENCOURT e MALDONADO, 2019). Este estudo concentrou-se em avaliar 80 publicações diretamente relacionadas à temática a fim de fazer um reconhecimento do estado da arte e identificar tendências. Foram verificados os algoritmos, mais frequentes, as taxonomias adotadas, as técnicas utilizadas para identificação das características e as métricas utilizadas para avaliação dos modelos gerados. Mostrou-se evidente a relevância desses classificadores em diversas aplicações atuais. Dentre outras descobertas, a revisão apontou que a classificação de Li e Roth (2002) e a Taxonomia de Bloom (1956) são os critérios de classificação mais comuns.

Em virtude do exposto, é possível concluir que a revisão sistemática (SILVA, BITTENCOURT e MALDONADO, 2019) foi importante para discutir grande parte dos estudos disponíveis na literatura acerca de classificadores automáticos. A seguir, é descrito como o presente estudo se enquadra às *Research Questions* (RQ) tratadas nessa revisão sistemática. As questões originais encontram-se no ANEXO A, Quadro 5.

- Método computacional (RQ1): o método computacional usado para implementar o classificador se deu a partir de um único algoritmo, neste caso, o de Rede Neural Artificial, mais especificamente uma CNN. A principal contribuição deste estudo esteve focada na lógica de classificação, promovida pela adaptação do *pipeline* original;
- Taxonomia empregada (RQ2): a taxonomia adotada para a classificação foi autoral, do próprio INEP, que utiliza o critério de nível de dificuldade das questões para atribuição de peso a cada uma;
- Extração e seleção de atributos (RQ3): a técnica usada para a extração de atributos (*feature extraction*) foi o bag-of-words (BOW). Para a seleção de atributos (*feature selection*), utilizou-se a técnica de *word embedding* com o algoritmo GloVe; e
- Indicadores de resultados (RQ4): a principal métrica utilizada para medir os resultados da classificação foi a acurácia. Vale ressaltar que a aplicação primária do classificador desenvolvido é em sistemas *e-learning*.

Foi realizada uma comparação do presente estudo com trabalhos relacionados na literatura científica. Para isto, foram realizadas buscas nas bases de dados científicas a fim de selecionar artigos mais semelhantes ao contexto desta pesquisa. Os critérios para a seleção desses trabalhos foram: utilização de um modelo classificador; aplicação de PLN; uso de uma taxonomia; e contexto educacional. Os cinco principais artigos são resumidos no Quadro 2.

Quadro 2: Comparação entre trabalhos relacionados.

Estudo	**1**	**2**	**3**	**4**	**5**
Autores	H. M. Braum; S. J. Rigo; J. L. V.	K. Jayakodi; M. Bandara; D. Meedeniya	A. Osman; A. A. Yahya	B. Sun; Y. Zhu; Y. Xiao; R. Xiao; Y. Wei	R. M. Silva; R. L. S. Santos; T. A. Almeida; T. A. S. Pardo
Ano	2014	2016	2016	2018	2020
Título do trabalho	Modelo de classificação automática na Língua Portuguesa	An Automatic Classifier for Exam Questions with WordNet and Cosine Similarity	Classifications of Exam Questions using Natural Language Syntactic Features: a case study based on Bloom's Taxonomy	Automatic Question Tagging with Deep Neural Networks	Towards automatically filtering fake news in Portuguese
Algoritmos	Logistic Regression; Linear SVC; SVC (kernel linear)	WordNet; Cosine Similarity	Naïve Bayes (NB); Support Vector Machine (SVM); Logistic Regression (LogR); Decision Trees (DTs)	---------	Fact Checking
Extração de *features*	Bag-of-words (BoW); Tf-	Tag pattern generation algorithm; Grammar and	Bag-of-words (BoW), Part of Speech	PBAM (Scale); PBAM	Bag-of-words (BoW); Linguistic-

	idf; n-grams	parser generation algorithm	(PoS); n-grams	(Locally); PBAM (Fully)	based; Part of Speech (PoS); Word2Vec; FastText
Método de avaliação	Precision; Recall; F1-Score	Precision	Total Hits; Target Hits; Precision/ Recall/F-Score; Kappa; Correlation	LDA; SVM	RF; Bagging; AdaBoost; SVM; LR; NB; DT
Taxonomia	Classificador hierárquico (Li e Roth 2002)	Taxonomia de Bloom (1956)	Taxonomia de Bloom (1956)	Automatic tagging (position-based attention; keywords-based model)	Automatic, semi-automatic and manual detection of fake news
Língua	Língua Portuguesa	Língua Inglesa	Língua Inglesa	Língua Inglesa e Chinês	Língua Portuguesa
Âmbito de aplicação	Unisinos	Universidade de Wayamba	Universidade de Najran	----------	---------
Disciplina	Linguística	Informática e Sistemas de Informação	Ciência da Computação	Curso de Inglês do Ensino Fundamental e Médio.	---------
Contexto	Questões de Ensino Superior, em Língua Portuguesa, autorais e criadas com o apoio do setor de Linguística	Questões de Ensino Superior retiradas de cursos do Departamento de Informática e Sistemas de Informação	Questões de Ensino Superior, coletadas a partir de um conjunto de palestras de cursos do Programa de Ciência da Computação	Banco de questões de múltipla-escolha, rastreadas principalmente nos sites de bancos de perguntas Koolearn e Tiku.	Corpus composto por notícias falsas e verdadeiras escritas em Português brasileiro, pesquisadas na Web (FakeBr Corpus)

Fonte: do autor (2022).

O trabalho realizado por Braun, Rigo e Barbosa (2014) apresenta e avalia um modelo de classificação de questões em Língua Portuguesa. Para tanto, foram estudados modelos de classificação de questões anteriormente testados em diferentes idiomas. O modelo proposto foi testado a partir de uma carga considerável de dados, o qual foi posteriormente testado, utilizando-se a taxonomia proposta por Li e Roth (2002) com o intuito de medir seu desempenho.

A semelhança com o presente estudo está relacionada ao corpus, que também foi formado a partir de questões em Língua Portuguesa. Outra semelhança observada diz respeito a um dos métodos utilizados para extração de *features*, o bag-of-words (BOW) que, no caso do trabalho relacionado, apresentou desempenho inferior se comparado ao resultado do td-idf.

Por outro lado, o presente estudo se diferencia do trabalho realizado pelos autores, uma vez que faz uso de uma taxonomia autoral (do INEP). Além disso, a classificação das classes e subclasses utilizada no trabalho relacionado foi um classificador hierárquico (LI e ROTH, 2002), ao passo que, neste trabalho, propõe-se a utilização do classificador *data-driven*.

No que concerne ao corpus, cabe ainda ressaltar que, apesar de ambos terem sido constituídos dentro do contexto educacional, o trabalho relacionado analisou questões de nível superior, formuladas com o apoio do Setor de Linguística da Unisinos, enquanto o estudo proposto utilizou questões do ENEM, como supracitado.

No segundo trabalho, desenvolvido por Jayakodi, Bandara e Meedeniya (2016), o corpus do trabalho relacionado diz respeito a questões retiradas de exames aplicados pelo Departamento de Computação e Sistemas de Informação, Universidade de Wayamba. Apresentou o uso do algoritmo WordNet com Cosine Similarity para classificar automaticamente questões a partir dos níveis de aprendizagem propostos pela taxonomia de Bloom (1956), utilizando técnicas do PLN.

A partir disso, um conjunto de regras foram gerados utilizando o PLN e foram combinados os algoritmos WordNet e Cosine Similarity para atribuir peso a cada categoria apresentada pela taxonomia de Bloom (1956) para determinada questão do exame. A proposição de um modelo automático de classificação de questões foi objeto de estudo da presente pesquisa, motivo pelo qual se relaciona diretamente com o referido trabalho

relacionado. Além disso, ambas as pesquisas foram voltadas para atribuição de peso às questões.

Para extração de *features* foram utilizados algoritmos de geração de padrão de tag (*tag pattern generation)* e de geração de gramática e *parser.* No trabalho relacionado, os pesquisadores concluíram que a abordagem apresentada para classificação automática de questões é apropriada para ser utilizada nos exames em universidades, haja visto que a classificação manual de questões é uma atividade demorada, difícil e sujeita a erros.

Apesar de a precisão dos padrões terem sido verificadas com a ajuda de especialistas da área, que consideraram uma tarefa difícil de executar, a avaliação da estratégia de atribuição automática de peso para classificação das questões do exame identificou que, em 71% das ocasiões, os resultados foram consistentes com os fornecidos pelos especialistas, o que demonstra um resultado positivo que pode auxiliar os institutos de educação a superar desafios nesse sentido.

Nesse contexto, foi possível verificar uma relação direta com o presente estudo, visto que o modelo apresentado visa igualmente propor uma estratégia que auxilie o professor, ampliando suas capacidades. Para fins deste trabalho, a taxonomia de Bloom foi descartada devido a divergências de cunho pedagógico, motivo pelo qual se diferencia do trabalho desse relacionado. Ademais, apesar de ambos utilizarem questões de exames, o trabalho relacionado fez uso de questões de Ensino Superior em uma área específica da Informática, enquanto este estudo trabalhou com diversos domínios do Ensino Médio.

No terceiro trabalho (OSMAN e YAHYA, 2016), o objetivo foi testar e comparar diferentes algoritmos de *Machine Learning*: Naïve Bayes (NB); Support Vector Machine (SVM); Logistic Regression (LogR); Decision Trees (DTs), para classificar automaticamente questões de exames baseadas nos níveis cognitivos propostos pela Taxonomia de Bloom (1956). No que concerne ao método de extração de *features* em PLN, as semelhanças estão no fato de que tanto o trabalho relacionado quanto o presente estudo utilizaram o bag-of-words (BoW).

Ademais, durante a fase de pré-processamento, cada pergunta do conjunto de dados foi submetida a etapas semelhantes, como remoção de pontuação, remoção de termos de baixa frequência (*stopwords*), tokenização e lematização. As diferenças, por sua vez, dizem respeito à taxonomia adotada (1956) e ao corpus utilizado que, no caso do trabalho

relacionado, foi baseado em questões de Língua Inglesa formuladas no contexto de Ensino Superior, também na área de Informática.

O trabalho realizado por Sun (et al., 2018) propõe a utilização de dois modelos: modelo de atenção baseado em proposição (*position-based attention model*) e modelo baseado em palavras-chaves (*keywords-based model*) para marcar automaticamente perguntas com unidades de conhecimento. Para tanto, os pesquisadores propuseram modelos que utilizam redes neurais profundas para representar questões usando informações contextuais.

As similaridades com o estudo proposto estão, portanto, relacionadas à proposição de um modelo automático para classificação de questões, utilizando redes neurais. No caso da presente pesquisa, foi utilizada uma CNN para reconhecer a dificuldade das questões a partir do PLN, enquanto o trabalho relacionado propôs a utilização de três métodos diferentes autorais, quais sejam: PBAM (Scale), PBAM (Locally) e PBAM (Fully) para ajustar os pesos de atenção.

Os resultados apresentados pelos pesquisadores demonstraram que as respostas desempenham papel importante no reflexo das unidades de conhecimento, de modo que, no caso do trabalho relacionado, a ênfase nas respostas se mostrou útil para representar questões e prever unidades de conhecimento para diferentes disciplinas.

O último trabalho (SILVA et al., 2020) traz aspectos diferentes do estudo proposto, uma vez que não trabalha com questões, mas sim com notícias. Nele, os autores identificam que, apesar dos esforços de vários estudos sobre detecção de *fake news*, a maioria cobre apenas notícias em inglês, o que representa uma defasagem no conjunto de dados rotulados de *fake news* em outros idiomas.

Nesse sentido, não há consenso sobre as melhores estratégias de classificação de *sets* e *features* para detecção automática de notícias falsas, de modo que foi realizada uma análise abrangente dos métodos de *machine learning* para detecção de *fake news* em português. A primeira similaridade com o estudo proposto está relacionada, portanto, ao corpus do trabalho, bem como a utilização de estratégias de *machine learning*, visto que, assim como este estudo, o trabalho relacionado utilizou métodos de extração de *features* como bag-of-words (BoW) e part of speech (PoW).

No entanto, apesar de ambos os trabalhos utilizarem corpus em Língua Portuguesa, essa é também uma das principais diferenças, haja vista que o trabalho com questões de

exame se diferencia das notícias em virtude da área de aplicação. Além disso, o trabalho relacionado utilizou como métodos de avaliação para detecção automática de *fake news* RF, Bagging, AdaBoost, SVM, LR, NB e DT, que são métodos amplamente utilizados pelo *Machine Learning,* enquanto o estudo proposto fez um uso de uma CNN.

3.2 ANÁLISE NO MERCADO

Além das buscas em periódicos acadêmicos, buscou-se averiguar as soluções utilizadas no mercado educacional brasileiro para a classificação de questões de prova. Duas grandes empresas que fazem uso de novas tecnologias e comercializam soluções no setor são Super Professor[2] e Geekie[3]. Foi investigado como essas companhias realizam a classificação de questões em seus sistemas *e-learning*.

Super Professor é uma plataforma de aulas com um grande banco de questões de prova para aprendizagem dos alunos. Em contato com essa plataforma de ensino, foi informado que a curadoria do grau de dificuldade das questões é realizada de forma manual pelos professores que utilizam suas experiências de muitos anos lecionando em cursos.

Geekie é uma plataforma de aprendizagem adaptativa. Apesar da empresa utilizar tecnologia para a recomendação de questões, não fica claro como ela determina o grau de dificuldade das mesmas. Esse empreendimento não divulgou como determinam a dificuldade das questões e nem responderam ao contato.

Além dessas, a ferramenta Avaliar[4] (ELIAS et al., 2020) é um sistema que auxilia o professor na verificação da aprendizagem de cada aluno. Apesar disso, não identifica a dificuldade dessas questões. Inclusive, esta proposta pode enriquecer esse sistema colaborando com a atribuição do nível de dificuldade das questões.

Após análise de relevantes ferramentas disponíveis do setor educacional, pôde-se constatar as afirmações reunidas no Quadro 3. Percebe-se que, apesar das ferramentas não implementarem um classificador do nível de dificuldade de questões, processos seletivos como o ENEM, que é o foco de interesse de milhões de estudantes, o levam em consideração.

[2]Ferramenta Super Professor. Disponível em: https://www.sprweb.com.br/. Acesso em: 10 jul. 2021.
[3]Ferramenta Geekie. Disponível em: https://www.geekie.com.br. Acesso em: 10 jul. 2021.
[4]Ferramenta Avaliar. Disponível em: http://app.uag.ufrpe.br/avaliar/. Acesso em: 12 jul. 2021.

Portanto, isso reforça ainda mais a necessidade de se desenvolver tecnologias que integrem essa solução.

Quadro 3: Comparativo entre ferramentas do setor educacional.

Ferramenta	Tecnologia	Análise da dificuldade das questões
Super Professor	Fornece um banco de questões para treinamento dos alunos.	Manualmente.
Geekie	Sistema que recomenda questões de acordo com o perfil do aluno.	Não revelado.
Avaliar	Verifica a aprendizagem do aluno.	Não utilizam.

Fonte: do autor (2022).

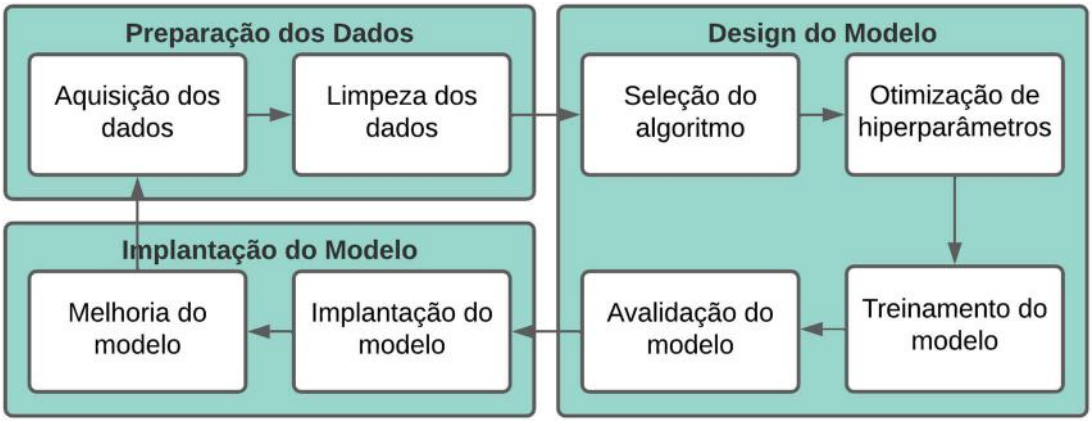
Dessa forma, este trabalho apresenta inovações tecnológicas por desenvolver soluções de apoio à docência. Essa solução pode ser aplicada no processo de elaboração de questões de prova voltada ao sistema educacional brasileiro. Conforme documentado, promove o avanço da fronteira do conhecimento sobre classificadores automáticos de questões em âmbito acadêmico e de mercado.

4 METODOLOGIA DA PESQUISA PARA O DESENVOLVIMENTO DO MODELO E DO PROTÓTIPO

A realização desta pesquisa teve como estímulo o método quantitativo de experimentação (DIAS e SILVA, 2010). A proposição é que em um ambiente controlado possa-se observar o direcionamento de milhões de alunos em cada questão do ENEM. Tornando possível, em seguida, a reprodução e a análise no cenário de pesquisa. Sendo assim, foi conduzida uma experimentação em ciência de dados.

A ciência de dados compreende a realização de estudos científicos para se aprender com os dados, conforme discutido por David Donoho (2017). Uma vez que a proposta deste projeto é a de obter conhecimento via análise dos dados de milhões de participantes do ENEM, esse trabalho tem significativa relevância na disciplina da ciência de dados. O fluxo de trabalho realizado neste estudo é apresentado na Figura 7, uma adaptação do workflow de ciência de dados de Dakuo Wang (et al., 2019).

Figura 7: Fluxo de trabalho de ciência de dados executado neste projeto.



Fonte: adaptado de WANG et al. (2019).

Este workflow possibilitou: a “Preparação dos Dados”, para a obtenção e padronização dos dados; o “Design do Modelo”, em que um algoritmo foi utilizado para a construção do modelo; e a “Implantação do Modelo”, no qual a solução desenvolvida foi implementada em um ambiente de produção. O detalhamento dos procedimentos executados é descrito a seguir.

4.1 PREPARAÇÃO DOS DADOS

A preparação dos dados consistiu na fase mais trabalhosa desta pesquisa, visto que grande importância desse trabalho estava em descobrir padrões escondidos em dados válidos. Foi necessário construir um corpus de questões com sua proporção de acerto. Como ambiente de experimentação, optou-se pela utilização das questões do ENEM.

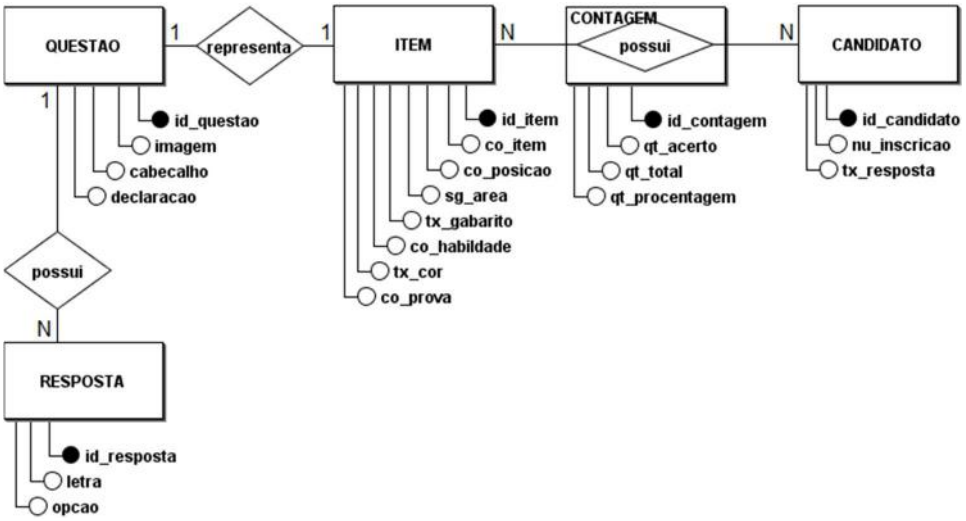
O ENEM é a avaliação mais difundida pela qual indivíduos formados no ensino médio ingressam nas universidades brasileiras. A prova do ENEM foi criada em 1998 e desde 2009 é realizada em dois dias, contando com uma redação e 180 questões objetivas, divididas nas seguintes áreas de conhecimento (BRASIL, 1996): “Ciências humanas e suas tecnologias” (CH), “Ciência da natureza e suas tecnologias” (CN), “Linguagens, códigos e suas tecnologias” (LC) e “Matemática e suas tecnologias” (MT).

Na etapa de aquisição dos dados, foram obtidos arquivos de provas e candidatos do site do INEP (MINISTÉRIO DA EDUCAÇÃO, 2020a), considerando-se a primeira aplicação, dos anos de 2016 a 2019. A partir dos arquivos comprimidos disponibilizados pelo INEP das provas do ENEM, obteve-se as tabelas que foram utilizadas para este trabalho, com base nos seguintes critérios:

- QUESTAO: tabela de questões extraídas do arquivo da prova (prova_ano.pdf). Cada questão é formada por um texto base (introdução), um comando (declaração) e alternativas de respostas (opções).
- RESPOSTA: tabela de alternativas obtidas das questões (prova_ano.pdf). Cada questão possui 5 opções de respostas.
- ITEM: tabela de itens (itens_prova_ano.csv). Cada questão da prova é representada por um item que possui informações complementares.
- CANDIDATO: tabela contendo os dados do candidato (microdados_enem_ano.csv). Cada aluno inscrito no ENEM está registrado nessa tabela, na qual está contida suas respostas.
- CONTAGEM: tabela elaborada a partir da concatenação de outras tabelas e execução de uma complexa *query* para obtenção da quantidade total de candidatos efetivos por questão e da quantidade de acertos de cada questão (contagem_ano.csv). Possui a contagem de cada item.

As tabelas obtidas foram organizadas de forma a atender o funcionamento de um banco de dados relacional, conforme processo de normalização do banco de dados (CODD, 1970). A Figura 8 apresenta o modelo conceitual dos dados obtidos, conforme notação de Peter Chen (1976).

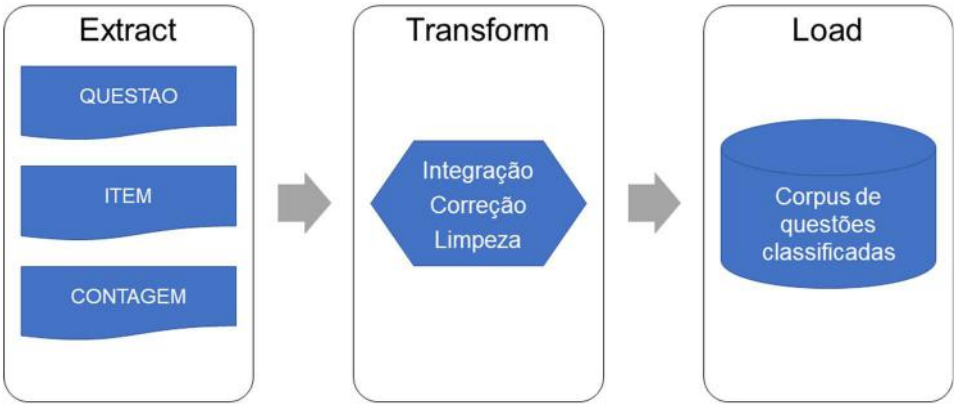
Figura 8: Modelo Entidade-Relacionamento (MER) das tabelas fornecidas pelo INEP.



Fonte: do autor (2022).

Na etapa de limpeza dos dados foi realizado um processo de ETL (*Extract, Transform and Load*) (KIMBALL e CASERTA, 2004). Assim, passou-se pelos seguintes processos: extração dos dados de diferentes formatos; transformação dos dados para garantir sua qualidade; e o carregamento em um corpus de questões, ilustrado na Figura 9.

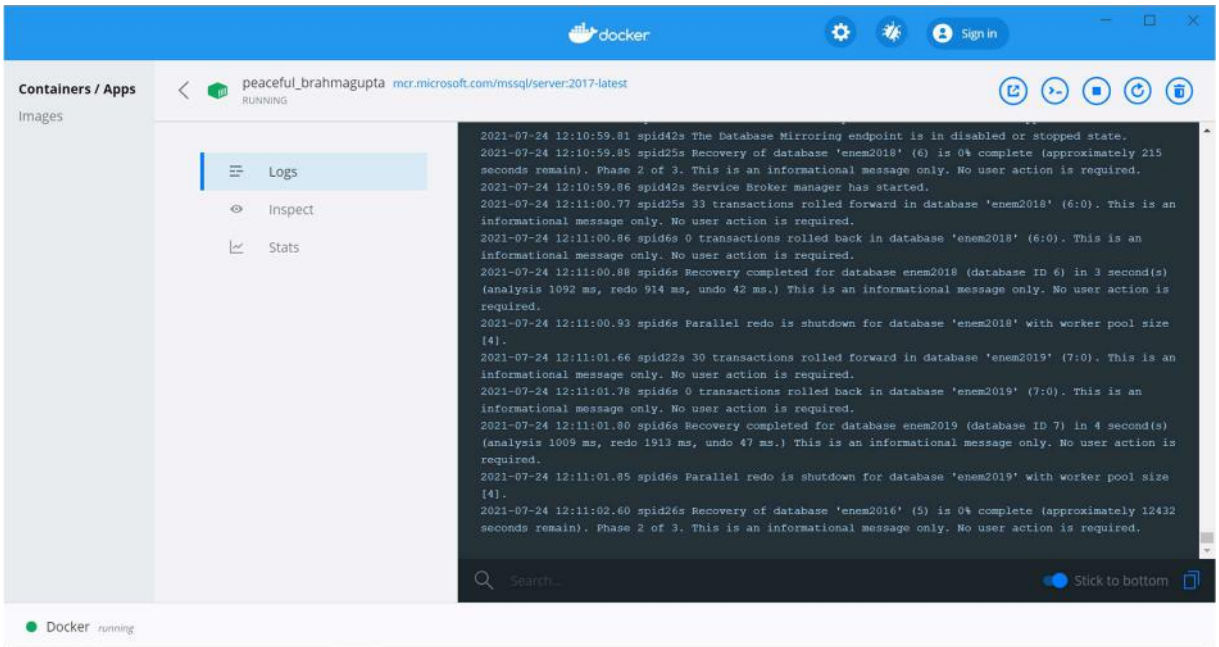
Figura 9: Processo de ETL realizado para a consolidação do corpus de questões.



Fonte: do autor (2022).

Na extração, foi utilizada a arquitetura de microsserviços em um Docker com um contêiner do Microsoft SQL Server 2017 para a obtenção dos dados. A Figura 10 exibe o Docker preparado em operação.

Figura 10: Docker em operação.



Fonte: do autor (2022).

Na transformação, foram elaboradas complexas *queries*, conforme descrito a seguir. Para a elaboração de uma *query* para se obter as questões válidas, foram consideradas as seguintes regras:

- Desconsiderar as questões das disciplinas de espanhol e inglês. A partir do ENEM de 2010, elas passaram a compor 5 das 45 questões de LC, de acordo com a escolha do candidato. Assim, devido à baixa quantidade amostral e para não haver interferência no processamento da língua portuguesa, optou-se por não as utilizar nesta experimentação.
- Desconsiderar as provas com questões substituídas. O ENEM promove algumas provas diferenciadas, nas versões ampliada, superampliada, em libras e por ledor. As provas adaptadas para ledores eram as únicas que possuíam questões substituídas dentre todas as outras versões, o que ocasionava alteração da contagem dos candidatos. Desse modo, para este experimento, foi necessário desconsiderá-las.

Para a elaboração de uma *query* para se obter a quantidade de acerto de cada questão, foram consideradas as seguintes regras:

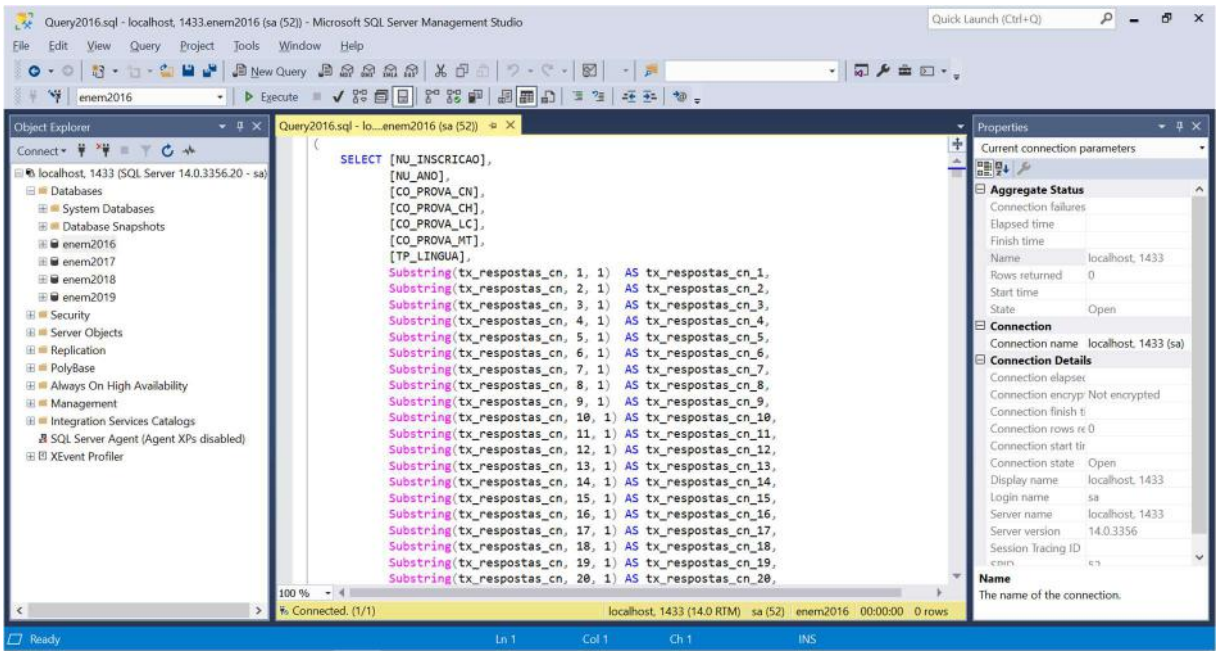
- Contabilizar que o candidato acertou a questão cuja resposta tenha sido a mesma resposta que o gabarito.
- Contabilizar que o candidato errou a questão cuja resposta tenha sido deixada em branco.
- Contabilizar que o candidato errou as questões cuja resposta tenha mais de uma marcação.

Para a elaboração de uma *query* para se obter a quantidade de candidatos efetivos, foram consideradas as seguintes regras:

- Remover da quantidade de candidatos efetivos aqueles que não compareceram no dia da prova.
- Remover da quantidade de candidatos efetivos aqueles que foram eliminados durante a prova.

Os dados foram carregados e as *queries* executadas no Microsoft SQL Server Management Studio (SSMS), demonstrado na Figura 11.

Figura 11: Microsoft SQL Server em operação.



Fonte: do autor (2022).

Para finalizar a transformação dos dados, utilizou-se a ferramenta Microsoft Power BI, de *Business Intelligence* (BI), considerando as seguintes regras:

- Eliminar as questões que foram anuladas pelo INEP.
- Eliminar as questões que continham imagem na declaração.
- Atribuir o *label* de dificuldade, considerando: as 33,333% mais acertadas como "fácil"; as 33,333% seguintes como "médio"; e as 33,333% mais erradas como "difícil". Dessa forma, foi realizada uma classificação categórica das questões e a distribuição das classes encontrava-se balanceada.

No carregamento, foram consolidadas as questões padronizadas desta proposta, constituindo o corpus de 694 questões com o nível de dificuldade validado por milhões de estudantes. Uma vez obtido esses dados, a proporção de acerto das questões foi determinada

pela quantidade de candidatos que a acertou dividida pela quantidade de candidatos que a realizou.

4.2 DESIGN DO MODELO

Todo o processo de design do modelo foi realizado em Jupyter Notebook, em um ambiente na nuvem, hospedado nos servidores da Google Colab (GOOGLE, 2021), que suporta a execução do código desenvolvido em linguagem Python em suas GPUs.

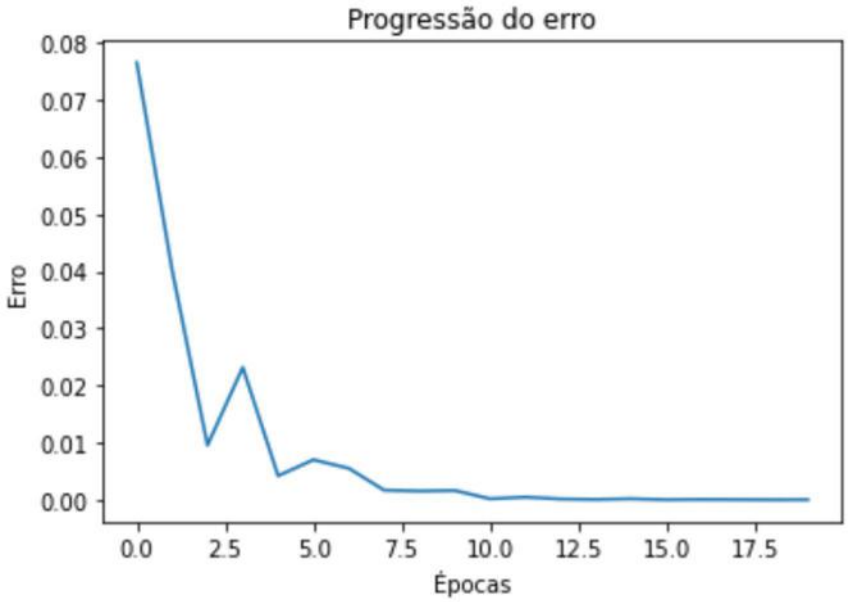
Para a etapa de seleção do modelo, foi designada a CNN pré-treinada da biblioteca spaCy, principalmente por facilitar a adaptação de seu *pipeline* para a criação de um componente para o reconhecimento da classe do texto, além das razões expostas no Referencial Teórico.

A etapa de otimização de hiperparâmetros visou maximizar a taxa de aprendizagem através da análise sistemática dos valores das variáveis dentro do conjunto de parâmetros de treinamento. Foram feitas alterações em alguns hiperparâmetros que permitiram melhorar a qualidade do modelo classificador resultante, sendo elas *batch* (lote) e *epoch* (época). “Lote” refere-se à quantidade de registros do dataset, enquanto “época” diz respeito à quantidade de vezes que todo o dataset é processado.

O experimento foi repetido diversas vezes para configurar os hiperparâmetros. Obteve-se o melhor desempenho do modelo com lotes de 30 registros aleatórios, até que todos os 555 registros fossem processados, tendo sido programado para executar por 20 épocas. A otimização se deu pela escolha desse conjunto de variáveis atingindo condições ótimas para o treinamento do modelo.

O resultado do treinamento com os hiperparâmetros otimizados é exposto na Figura 12. Nota-se que houve um grande aprimoramento da primeira para a segunda época; em seguida, houve uma diminuição gradual do erro, evidenciando a aprendizagem do modelo.

Figura 12: Progressão do erro ao longo das épocas.



Fonte: do autor (2022).

Para a etapa do treinamento do modelo, o corpus foi importado e realizado um pré-processamento em Python. Foram aplicadas as seguintes regras: remoção de pontuações; remoção de *stopwords* (conjunto de palavras consideradas irrelevantes, não existindo uma lista universal); tokenização (separação do texto em uma lista de palavras); e lematização (deflexionamento de uma palavra para determinar o seu lema, deixando os verbos no infinitivo e os substantivos e adjetivos no masculino e singular). Além disso, foi necessário fazer um tratamento das classes, no qual se realiza uma adequação das classes para o formato esperado pela biblioteca spaCy.

Iniciou-se o aprendizado supervisionado. Foi criado o componente “textcat” para o reconhecimento multiclasse, compreendendo os *labels* “fácil”, “médio” e “difícil”, e, em seguida, esses elementos foram adicionados ao final do *pipeline* original. O dataset de treinamento, contendo as questões do ENEM com o *label* de dificuldade, foi processado pela CNN, a fim de determinar os pesos do modelo.

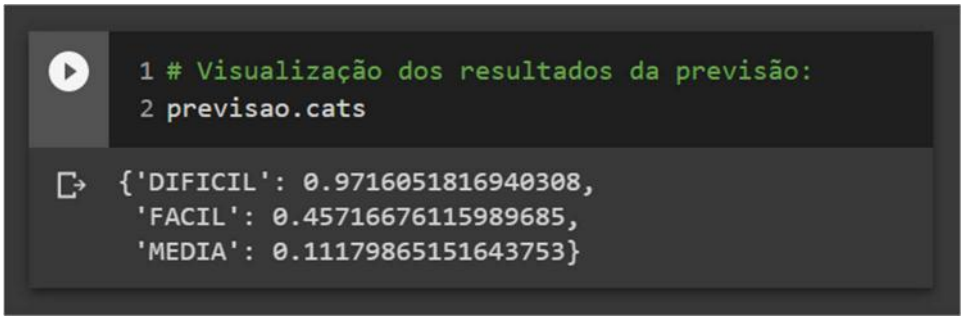
Durante o treinamento, as previsões foram comparadas aos resultados reais e um erro foi calculado. A cada processamento de um lote, os pesos do modelo eram atualizados a fim de se diminuir esse erro, o que promoveu a aprendizagem por quantas épocas o processo havia sido programado.

A etapa de validação do modelo foi realizada por técnica de *holdout*. Neste caso, o modelo foi treinado com 80% das questões e validado com os outros 20%, ou seja, 555 questões no dataset de treinamento e 139 no dataset de teste. Desta forma, foi realizada uma comparação das classes previstas com as classes reais para se verificar a acurácia do modelo.

4.3 IMPLANTAÇÃO DO MODELO

A etapa de implantação do modelo foi realizada em um ambiente de desenvolvimento externo. O modelo foi exportado com ajuda do módulo Pickle da linguagem Python e importado em um novo Jupyter Notebook sem que os datasets originais fossem carregados. Em seguida, para verificar a implantação funcional do modelo, foi realizado um teste para prever a classe de novas questões. O resultado da implantação está ilustrado na Figura 13.

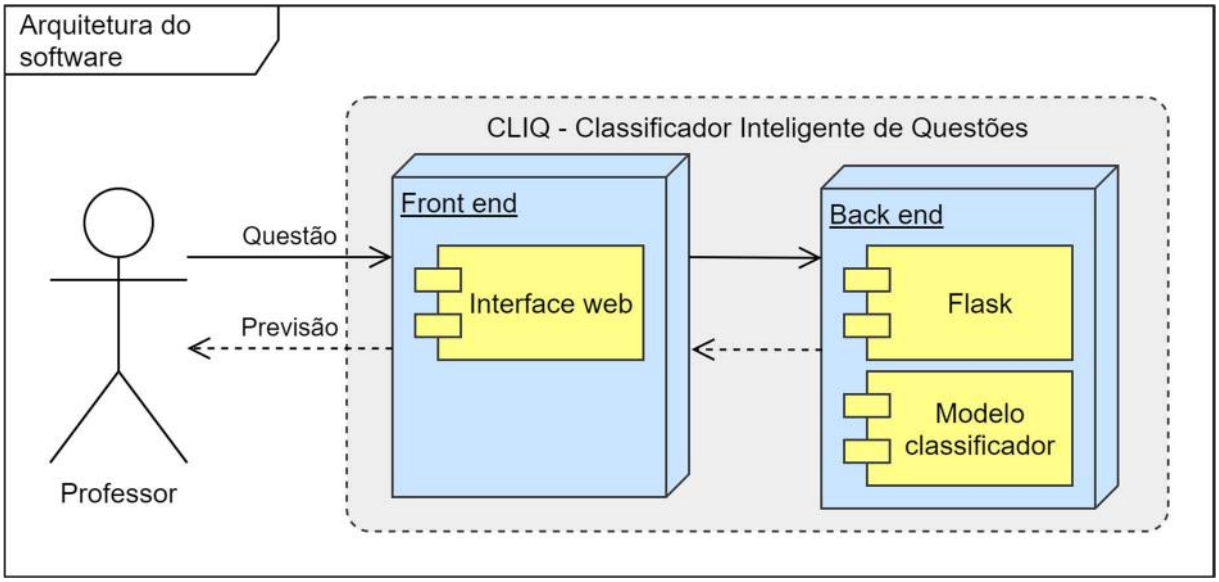
Figura 13: Demonstração do modelo classificador carregado e em execução.



Fonte: do autor (2022).

Foi idealizada a arquitetura de um protótipo “CLIQ - Classificador Inteligente de Questões” que embarcasse o modelo em ambiente de produção, a fim de demonstrar sua viabilidade. A arquitetura do software é demonstrada na Figura 14. Para seu desenvolvimento, foi empregado o editor Visual Studio Code na criação da aplicação web em Python, utilizando-se em front end as linguagens HTML, JavaScript e CSS e em back end o microframework Flask e o modelo treinado.

Figura 14: Arquitetura do sistema CLIQ em UML.



Fonte: do autor (2022).

Na etapa de melhoria do modelo, à medida que os anos avançarem e ocorrerem novas aplicações do ENEM, poderá ser alimentado com mais questões atuais classificadas. Assim, conforme se der a evolução do perfil dos estudantes, o modelo continuará a aprender e a refletir esse progresso. Outras análises também poderão ser mensuradas por um *Data Scientist* nas etapas preliminares. O modelo continuará a levar esse feedback a partir dos alunos e evoluirá junto com o perfil dos candidatos.

5 RESULTADOS

Foi constatada a quantidade de respostas válidas para cada área de conhecimento do ENEM nos respectivos anos da pesquisa. O Quadro 4 exibe a quantidade de candidatos efetivos de cada ano. Cada questão teve sua proporção de acerto calculada em relação a esse quantitativo.

Quadro 4: Total de candidatos efetivos da 1ª aplicação por ano e área de conhecimento.

Ano	CH	CN	LC	MT
2016	5.833.931	5.833.931	5.684.524	5.684.524
2017	4.693.990	4.431.288	4.693.990	4.431.288
2018	4.140.367	3.898.626	4.140.367	3.898.626
2019	3.915.806	3.703.494	3.915.806	3.703.494

Fonte: do auto (2022).

O corpus de questões construído foi publicado no repositório https://doi.org/10.5281/zenodo.5573846, Figura 15. Essa publicação está em conformidade aos princípios FAIR (*Findable, Accessible, Interoperable and Reusable*) de disponibilização de dados de pesquisa (WILKINSON et al., 2016). Sendo assim, os dados obtidos estão compartilhados para a comunidade científica e proporciona o reuso por outros *Data Scientists* para futuras experimentações.

Figura 15: Corpus de questões do ENEM.

Fonte: do autor (2022).

O código para o modelo classificador está armazenado no ambiente colaborativo do GitHub. A comunidade acadêmica e desenvolvedores podem acessá-lo no endereço https://github.com/rafaeldigital2/ModeloCLIQ, possibilitando sua utilização por diversos outros pesquisadores.

Em suma, a determinação do nível de dificuldade da questão adotou uma estratégia *data-driven*. Portanto, foi baseada nos dados provenientes do índice de acerto dos alunos, que foi utilizado para o treinamento de uma CNN, e não a tradicional alternativa, baseada na experiência do professor. Sendo assim, obteve-se um modelo classificador a partir de uma CNN *data-driven*.

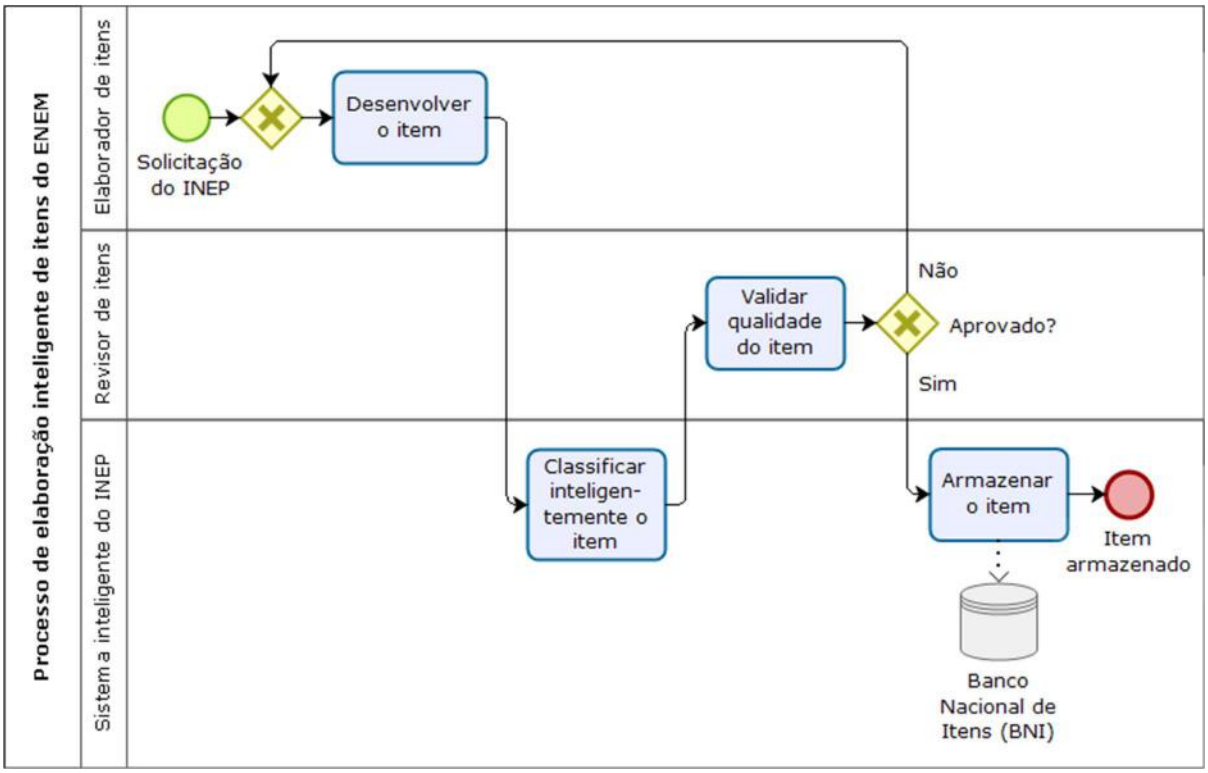
A implantação do modelo foi realizada com a construção do aplicativo "CLIQ - Classificador Inteligente de Questões". O protótipo do sistema web está em produção no endereço https://cliq-enem.herokuapp.com/. Sua interface está ilustrada na Figura 16. De acordo com o enunciado da questão informada pelo professor, o aplicativo é capaz de indicar seu nível de dificuldade.

Figura 16: Protótipo do CLIQ - Classificador Inteligente de Questões.

Fonte: do autor (2022).

Graças à eficiência da abordagem proposta, viabiliza-se o modelo classificador para determinar objetivamente o nível de dificuldade das questões. Essa tecnologia poderá compor sistemas, como o do INEP, para automatizar parte do processo de elaboração de itens do ENEM. Um esboço do processo pode ser observado na Figura 17.

Figura 17: Proposta de novo processo de elaboração de itens do ENEM.



Fonte: do autor (2022).

Obteve-se ganho de desempenho do processo devido à automatização da tarefa de classificação do nível de dificuldade das questões. Ademais, obteve-se ganho na qualidade da análise do nível de dificuldade das questões, uma vez que é possível estimar a reação dos estudantes às questões. Além disso, também é possível mensurar o desempenho destas em relação à prova em geral.

O treinamento do modelo alcançou sua eficiência com o dataset de treinamento, ou seja, o modelo convergiu para determinar corretamente a classe de todo o dataset de treinamento. Sendo assim, ele foi capaz de descobrir padrões nesses dados e identificar corretamente suas classes.

Em seguida, foi realizada a avaliação do dataset de teste por técnica de *holdout*, isso permitiu avaliar a performance do modelo com questões novas. Nesse caso, o modelo foi capaz de prever corretamente 96 questões que nunca foram apresentadas ao modelo, o que equivale a 68,57% do dataset de teste.

A estratégia *data-driven* contribui com a comprovação da relevância dos dados, eliminando-se suposições e decisões não baseadas em números. Permite-se, assim, aprimorar

o processo de elaboração de provas, como as do ENEM, bem como melhorar seus cálculos, como a TRI, a fim de torná-las mais assertivas. A solução orientada por dados torna possível classificar as questões com base em confirmações massivas via *crowdsourcing*, o que assegura previsões mais confiáveis do que a tradicional classificação, transpassada por uma subjetividade inerente.

6 DISCUSSÃO

Sabe-se que há outros critérios de classificação das questões, como a Taxonomia de Bloom (1956), as quais foram considerados nos momentos iniciais desta pesquisa. Entretanto, como suas fundamentações pedagógicas não são utilizadas pelo ENEM, essas metodologias foram descartadas.

Ao longo do desenvolvimento dessa tecnologia, a pesquisa perpassou por algumas abordagens de SAD, conforme apresentadas no Quadro 1, e propôs a melhoria delas. Para este estudo, passou-se de *knowledge-driven* para *data-driven*. Além disso, como um dos resultados deste trabalho é um modelo, estima-se que possa impulsionar também o avanço de sistemas *model-driven*. Constata-se, portanto, que no desenvolvimento de novas tecnologias ocorre uma alternância entre as abordagens mencionadas no Quadro 1. Ainda não é possível perceber se uma abordagem é superior à outra, mas nota-se que sua alternância possibilita a promoção de novas tecnologias.

Seria interessante, ainda, comparar se os níveis de dificuldades das questões atribuídas pelos elaboradores refletem o nível de acerto das questões pelos candidatos. Dessa forma, seria possível comparar a abordagem utilizada tradicionalmente com a *data-driven* proposta. Infelizmente, o INEP não disponibiliza esses níveis atribuídos. De todo modo, a proposta deste trabalho não se pautou em comparar os melhores classificadores, mas em desenvolver uma solução *data-driven*.

Apesar desta proposta ajustar a calibração da classificação do nível de dificuldade das questões, ela não interfere na fórmula da TRI. Sendo assim, a tecnologia apresentada pode ser adotada sem alterar a nota do candidato. De qualquer forma, esta proposta, aparentemente, torna o resultado de seu cálculo mais confiável, visto que o sistema é dotado de Inteligência Artificial capaz de atribuir o peso das questões.

A partir da adoção de soluções de Aprendizagem Profunda, como a desta iniciativa, é possível ao INEP aprimorar o cálculo da TRI. Sabe-se que o INEP atualmente realiza algumas etapas de “pré-testes” com os estudantes para estimar a complexidade das questões. Possivelmente, estas etapas poderão ser substituídas por sistemas inteligentes que contam com uma determinação intrínseca por meio de dados para estimar a performance dos candidatos e apoiar o cálculo do TRI.

Conforme observado nos Trabalhos Relacionados, Capítulo 3, boa parte das pesquisas de PLN são em inglês. Apesar de estar surgindo um interesse no PLN em língua portuguesa, não se identificou outros trabalhos sobre a classificação automática de questões do ENEM, o que também confere a originalidade desta pesquisa.

Os empreendimentos educativos analisados, Super Professor e Geekie, apesar de serem reconhecidas plataformas tecnológicas educacionais e utilizarem modernas soluções digitais, não aparentam possuir um sistema automatizado para a determinação de dificuldade das questões. Sendo assim, percebe-se haver uma área de negócio na qual a ferramenta CLIQ pode explorar e contribuir.

Gradualmente os enunciados das questões de cada ano do ENEM foram sendo obtidos, processados e padronizados, bem como os resultados dos candidatos, que também foram sendo apurados. Então, de forma periódica, o experimento foi repetido com os novos dados tratados. À medida que esses dados foram sendo processados no experimento, foi percebida uma melhora na performance de sua acurácia. Notou-se que a pouca quantidade de questões impactou a eficiência do aprendizado do modelo. Como se trata de uma experimentação em ciência de dados, é natural que o processo seja continuamente alimentado por novos dados. Assim, espera-se que as bases dos próximos anos do ENEM possam apoiar a evolução desse modelo.

Com os resultados obtidos na fase de validação, evidenciou-se a conquista de uma tecnologia dotada de Inteligência Artificial capaz de identificar padrões no comando das questões do ENEM. Apesar de originalmente se esperar uma maior acurácia do modelo, é válido ressaltar que esta proposta pretendia demonstrar a viabilidade conceitual de um sistema que fosse capaz de reconhecer esses padrões nas questões. Portanto, embora ainda não se tenha atingido sua máxima eficiência, foi demonstrado experimentalmente o objetivo concebido.

Sistemas recebem a denominação de “inteligentes” devido à capacidade de executarem diversas funções semelhantes à habilidade racional do ser humano, com o objetivo de solucionar problemas. No entanto, faz-se necessário destacar que, para fins desta pesquisa, não se pretende substituir, ou até mesmo diminuir a atuação do professor. Ao contrário, o uso da Inteligência Artificial na educação visa apoiar e ampliar as capacidades do professor. Além disso, o intuito é que a utilização de recursos tecnológicos facilite o

acompanhamento personalizado do estudante e influencie positivamente o processo de ensino e aprendizagem.

7 CONCLUSÃO

Respondendo à pergunta da pesquisa, este trabalho demonstrou a criação de um modelo classificador do nível de dificuldade de questões de prova por meio da combinação de tecnologias computacionais da ciência de dados. Esta pesquisa utilizou os dados de milhões de estudantes, por meio de *crowdsourcing* implícito, para validar a quantidade de acerto de cada questão de um legitimado processo seletivo, o ENEM. Esse quantitativo de acerto foi atrelado a cada questão. A partir dessa apuração, uma Rede Neural Convolucional foi treinada a reconhecer a dificuldade das questões com base no PLN do comando das questões. Foi construído um modelo preditivo que conseguiu realizar a classificação do nível de dificuldade das questões.

Nesse contexto, foi desenvolvido o sistema CLIQ para auxiliar professores e instituições de educação no processo de elaboração de avaliações por meio de uma abordagem empírica, orientada por dado (*data-driven*). A ferramenta demonstrou a viabilidade da implantação do modelo desenvolvido. Ela auxilia os professores na tomada de decisão do nível de dificuldade das questões de forma empírica, o que corresponde à relevante etapa do processo de elaboração de provas e aprimora a capacitação dos alunos. É uma ferramenta SAD que está em ambiente de produção e aberta aos professores e às instituições de ensino. Além disso, permitiu que usuários, não-programadores, façam uso da tecnologia desenvolvida por meio de uma interface amigável.

Em virtude do exposto, percebe-se que os dados, por si só, não possuem valor extrínseco, daí a necessidade de se projetar sistemas inteligentes que se aproveitem da informação e extraiam esse valor.

A solução apresentada pode ser generalizada para outros cenários com aplicações para além do INEP.

7.1 CONTRIBUIÇÕES

Este estudo colaborou com o avanço do conhecimento científico de diversas áreas. Pode-se destacar as seguintes contribuições:

1. Combinação de tecnologias e métodos computacionais modernos para a concretização deste tratado científico.
2. Desenvolvimento de um corpus de questões apuradas por milhões de estudantes via *crowdsourcing*.
3. Desenvolvimento de um modelo classificador do nível de dificuldade das questões em português brasileiro por meio de Aprendizagem de Máquina.
4. Construção de um componente classificador que pode ser exportado para outros trabalhos de Processamento de Linguagem Natural.
5. Comprovação empírica da performance do modelo proposto para a classificação de questões.
6. Desenvolvimento de um Sistema de Apoio a Decisão com interface web para classificação das questões de prova para apoio aos professores e às escolas.
7. Desenvolvimento de uma tecnologia original aplicada à real necessidade social e educacional brasileira.

Algumas publicações científicas relacionadas a este estudo foram realizadas e uma lista pode ser observada no Apêndice A. Outros trabalhos paralelos foram realizados ao longo deste mestrado e podem ser conferidos no Apêndice B.

7.2 LIMITAÇÕES DO ESTUDO

A fim de se evitar uma má codificação do modelo desenvolvido para a língua portuguesa, conforme anunciado, este trabalho não considerou as questões do ENEM das disciplinas de Inglês e Espanhol. O enunciado dessas disciplinas continha termos que poderiam enviesar o modelo.

Este estudo se limitou a utilizar as questões do ENEM dos anos de 2016 a 2019. Visto que os dados dos anos anteriores apresentam codificação diversificada para sua utilização e, até o desenvolvimento desta pesquisa, o INEP não havia publicado os dados dos alunos do ENEM de 2020.

7.3 TRABALHOS FUTUROS

Há trabalhos a serem realizados para a melhoria do modelo e percebe-se que o desenvolvimento de tecnologias para a educação é um campo de estudo promissor. O investimento em sistemas de aprendizagem adaptativa poderá promover a capacitação de milhões jovens e proporcionar o progresso de um Brasil mais próspero.

Para estudos futuros, pretende-se: atualizar o modelo desta pesquisa com os dados do ENEM dos próximos anos; aumentar sua eficiência por meio da avaliação de outros algoritmos; validar a ferramenta de acordo com a vivência dos professores e instituições de ensino; e aprimorar a ferramenta de acordo com a experiência do usuário (*User Experience - UX*).

REFERÊNCIAS

ANGELI, Rafael de; MALINI, Fábio. Crowdsourcing e colaboração na internet: breve introdução e alguns cases. In: **XVI Congresso de Ciências da Comunicação na Região Sudeste**, São Paulo – SP, 12 a 14 de maio de 2011.

ABHIGNA, B. S.; Soni, Nitasha; Dixit, Shilpa. Crowdsourcing – A Step Towards Advanced Machine. In: **Procedia Computer Science**, vol. 132, jan. 2018.

BEZERRA, Eduardo. Introdução à Aprendizagem Profunda. In: **Tópicos em Gerenciamento de Dados e Informações**. 1 ed. Porto Alegre: SBC Editora, 2016.

BLOOM, Benjamin S.; ENGELHART, Michael D.; FURST, E. J.; HILL, W. H.; KRATHWOHL, D. R. **Taxonomy of educational objectives**. **Handbook I:** Cognitive domain. New York: Ed. David McKay, 1956.

BOJANOWSKI, Piotr et al. Enriching Word Vectors with Subword Information, **Transactions of the Association for Computational Linguistics**, vol. 5, p. 135–146, 2017.

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as Diretrizes e Bases da Educação Nacional. Diário Oficial da União, Brasília. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 30 nov. 2020.

BRAUN, Henrique Maia; RIGO, Sandro José; BARBOSA, Jorge L. V. Modelo de classificação automática de questões na Língua Portuguesa. **CINTED – Novas Tecnologias na Educação**, [S. l.], v. 12, n. 2, p. 1-10, dez., 2014.

BROWN, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, Daniel M.; Wu, Jeffrey; Winter, Clemens; Hesse, Christopher; Chen, Mark; Sigler, Eric; Litwin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya; Amodei, Dario. Language Models are Few-Shot Learners. **arXiv:2005.14165**, jul., 2020.

BRUNIALTI, Lucas F. et al. M. Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual: Uma Revisão Sistemática. In: **XI Brazilian Symposium on Information System**, Goiania, GO, 26-29 mai. 2015.

CHEN, Peter Pin-Shan. The Entity-Relationship Model—toward a Unified View of Data. **ACM Transactions on Database Systems (TODS)**, v. 1, n. 1, p. 9–36, 1976.

CHOI, Jinho D.; TRETREAULT, Joel; STENT, Amanda. It depends: Dependency parser comparison using a web-based evaluation tool. In: **53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing**, vol. 1, 2015.

CLIQ – Classificador Inteligente de Questões. Disponível em: http://cliq-enem.herokuapp.com/. Acesso em: 13 jul. 2021.

CODD, E. F. A Relational Model of Data for Large Shared Data Banks. **Communications of the ACM**, vol. 13, n. 6, p. 377–387, 1970.

CORREIA, António et al. Crowdsourcing and massively collaborative science: A systematic literature review and mapping study. In: **International Conference on Collaboration and Technology**, vol. 11001, p. 133–154, 2018.

DIAS, D. S.; SILVA, M. F. Como escrever uma monografia: manual de elaboração com exemplos e exercícios. São Paulo: Atlas, 2010.

DEVLIN, Jacob et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, vol. 1, p. 4171–4186, 2019.

DONOHO, David. 50 Years of Data Science. **Journal of Computational and Graphical Statistics**, vol. 26, n. 4, p. 745–766, 2017.

ELIAS, A.; Vanderlei, I.; Andrade, M.; Gusmão, R.; & Teixeira, J. Avaliar: Sistema para Autoria e acompanhamento de recursos avaliativos, In: **XXVI Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia),** p. 97-101, 2020. https://doi.org/10.5753/webmedia_estendido.2020.13070

FERRAZ, Ana Paula do Carmo Marcheti; BELHOT, Renato Vairo. Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais, **Gestão & Produção**, vol. 17, n. 2, 2010.

G, Suganya et al. Subjective Areas of Improvement: A Personalized Recommendation. **Procedia Computer Science**, vol. 172, p. 235-239, 2020.

GITHUB. **explosion/spaCy: Industrial-strength Natural Language Processing (NLP) in Python**. [S. I.], 2020a. Disponível em: https://github.com/explosion/spaCy. Acesso em: 5 out. 2020.

GITHUB. **UniversalDependencies/UD_Portuguese-Bosque: This Universal Dependencies (UD) Portuguese treebank**. [S. I.], 2020b. Disponível em: https://github.com/UniversalDependencies/UD_Portuguese-Bosque. Acesso em: 3 dez. 2020.

GOMES, Dennis dos Santos. Inteligência Artificial: conceitos e aplicações. **Revista Olhar Científico – Faculdades Associadas de Ariquemes**, vol. 01, n. 2, ago./dez. 2010.

GOMES, Maria João. E-Learning: reflexões em torno do conceito. In: Atas do Congresso Internacional sobre Tecnologias da Informação e Comunicação na Educação, vol. 4, Braga, 2005.

GONÇALVES JR, Wanderley P.; BARROSO, Marta F. As questões de física e o desempenho dos estudantes no ENEM, **Revista Brasileira de Ensino de Física**, vol. 36, n. 1, p. 1–16, 2014.

GOOGLE. Colaboratory. Disponível em: https://colab.research.google.com. Acesso em: 10 jan. 2021.

GUEDEA-NORIEGA, Héctor H.; GARCÍA-SÁNCHEZ, Francisco. Semantic (Big) Data Analysis: An Extensive Literature Review. **IEEE Latin America Transactions**, vol. 17, n. 5, p. 796–806, 2019.

IKEMOTO, Gina Schuyler; MARSH, Julie A. Cutting Through the "Data-Driven" Mantra: Different Conceptions of Data-Driven Decision Making, In: **Yearbook of the National Society for the Study of Education**, vol. 106, n. 1, p. 105–131, 2007.

JAYAKODI, Kithsiri; BANDARA, Madhushi; MEEDENIYA, Dulani. An automatic classifier for exam questions with WordNet and Cosine similarity. In: **Moratuwa Engineering Research Conference (MERCon)**. Moratuwa, 2016, p. 12-17.

KAHNEMAN, Daniel. **Thinking, Fast and Slow**. New York: Ed. Farrar, Straus and Giroux, 2011.

KEEN, Peter G. W.; MORTON, Michael S. Scott. **Decision Support Systems: An Organizational Perspective**. Ed. Addison-Wesley, 1978.

KIMBALL, Ralph; CASERTA, Joe. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.** Indianapolis: Ed. Wiley Publishing, Inc., 2004.

KLAŠNJA-MILIĆEVIĆ, Aleksandra et al. **E-Learning Systems. Intelligent Techniques for Personalization.** 1 ed. Switzerland: Springer International Publishing, 2017.

KUO, C.-C Jay. Understanding Convolutional Neural Networks with A Mathematical Model. **arXiv:1609.04112**, vol. 2, [S. I.], 14 set. 2014.

LECUN, Yann et al. Gradient-Based Learning Applied to Document Recognition. **Proc. of the IEEE**, [S. I.], nov. 1998.

LEE, Kai-Fu. **Inteligência Artificial**: como os robôs estão mudando o mundo, a forma como amamos, nos relacionamos, trabalhamos e vivemos. [Recurso eletrônico]. Tradução: Marcelo Barbão. 1 ed. Rio de Janeiro: Globo Livros, 2019.

LI, Xin; ROTH, Dan. Learning question classifiers. In: **Proceedings of the 19th International Conference on Computational Linguistics**, vol. 1, p. 1–7, 2002.

LUCKESI, Cipriano Carlos. **Avaliação da aprendizagem escolar**. São Paulo: Ed. Cortez, 2014.

MELO, Camila Bitencourt Bezerra de et al. Crowdsourcing como uma Ferramenta à Inovação Estratégica Empresarial. **Rev. de Empreendedorismo, Inovação e Tecnologia**, vol. 1, n. 1, p. 13-24, 2014.

MENDES JR, Ricardo. *Crowdsourcing* e *machine learning*: uma revisão sistemática com discussão do uso para a participação pública dos cidadãos. **Inteligência Artificial: 3º Grupo de Pesquisa do ITS**, 2018. ITS Rio.

MIKOLOV, Tomas et al. Distributed representations ofwords and phrases and their compositionality. **Advances in Neural Information Processing Systems**, [S. I.], 16 out. 2013a.

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. In: **1st International Conference on Learning Representations - Workshop Track Proceedings**, 2013b.

MINISTÉRIO DA EDUCAÇÃO. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo da educação superior 2019**. Brasília, DF, 2019. Disponível em: http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2020/Notas_Estatisticas_Censo_da_Educacao_Superior_2019.pdf. Acesso em: 8 dez. 2020.

MINISTÉRIO DA EDUCAÇÃO. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Microdados**. Brasília, DF, 2020a. Disponível em: http://inep.gov.br/microdados. Acesso em: 3 out. 2020.

MINISTÉRIO DA EDUCAÇÃO. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Legislação referente ao Banco Nacional de Itens**, Brasília, DF, 2020b. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/bni. Acesso em: 01 dez. 2020.

MINISTÉRIO DA EDUCAÇÃO. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Teoria de resposta ao Item**. Brasília, DF, 2020c. Disponível em: https://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf. Acesso em: 6 dez. 2020.

MIRANDA, Leonídio Antonio Sousa de; FERREIRA, Andrea Claudia Freitas; DIAS, Glaecir Roseni Mundstock. Análise de conteúdo das questões de Fisiologia Humana da Prova de Ciências da Natureza e suas Tecnologias do Exame Nacional do Ensino Médio (1998-2016), **Ciência & Educação**, Bauru, vol. 25, n. 2, p. 375–393, 2019.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de Máquina. In: **Sistemas Inteligentes Fundamentos e Aplicações**. 1 ed. Barueri – SP: Manole Ltda, 2003, p. 89-114.

MORAIS, Diogo Martins Gonçalves de et al. O conceito de Inteligência Artificial usado no mercado de softwares, na educação tecnológica e na literatura científica. **Educação Profissional e Tecnológica em Revista**, v. 4, n. 2, 2020 – Rede Federal de Educação Profissional, Científica e Tecnológica.

MOREIRA, Miguel Ângelo Lellis; Gomes, Carlos Francisco Simões; dos Santos, Marcos; Júnior, Antonio Carlos da Silva; Costa, Igor Pinheiro de Araújo. Sensitivity Analysis by the PROMETHEE-GAIA method: Algorithms evaluation for COVID-19 prediction, **Procedia Computer Science**, Volume 199, 2022, Pages 431-438, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.01.052.

NABIZADEH, Amir Hossein et al. Learning path personalization and recommendation methods: A survey of the state-of-the-art. **In Expert Systems with Applications**, vol. 159, 30 nov. 2020.

NASCIMENTO, Rafaella L. S. do; FAGUNDES, Roberta A. de A.; SOUZA, Renata M. C. R. Statistical Learning for Predicting School Dropout in Elementary Education: A Comparative Study, **Ann. Data Sci.,** p. 1–28, mar. 2021.

NETO, Ana Lúcia Gomes Cavalcanti; AQUINO, Josefa de Lima Fernandes. A avaliação da aprendizagem como um ato amoroso: o que o professor pratica?, **Educação em Revista**, vol. 25, n. 2, 2009.

OKOYE, Ifeyinwa; SUMNER, Tamara; BETHARD, Steven. Automatic extraction of core learning goals and generation of pedagogical sequences through a collection of digital library resources. In: **Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries**, p. 67–76, 2013.

OSMAN, Addin; YAHYA, Anwar Ali. Classifications of Exam Questions Using Linguistically-Motivated Features: A Case Study Based on Bloom's Taxonomy. In: **The Third International Arab Conference on Quality Assurance in Higher Education (IACQA)**. Republic of Sudan, 2016.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. GloVe: Global vectors for word representation. In: **Conference on Empirical Methods in Natural Language Processing (EMNLP),** p. 1532–1543, 2014.

PEREIRA, Silvio do Lago. Processamento de Linguagem Natural. [S. I.], vol. 31. São Paulo: Universidade de São Paulo, 2011.

PETERS, Matthew E. et al. Deep contextualized word representations. In: **NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, vol. 1, p. 2227–2237, 2018.

PIRES, João Miguel Neves Gusmão**. Aprendizagem Profunda**: Estudos e Aplicações. 2017. 79f. Dissertação (Mestrado em Engenharia Informática). Universidade de Évora, Évora, 2017.

PONTE, João Pedro da. Tecnologias de informação e comunicação na formação de professores: que desafios? **Revista Iberoamericana de Educación**, [S. I.], n. 24, p. 63-90, 2000.

POWER, Daniel J. Supporting Decision-Makers: An Expanded Framework. **Challenges to Informing Clients:** A Transdiciplinary Approach, Krakow, Poland, p. 431-436. 19-22 jun. 2001.

SALAS, Joaquin; VIDAL, Flavio de Barros; MARTINEZ-TRINIDAD, Francisco. Deep Learning: Current State. **IEEE Latin America Transactions**, vol. 17, n. 12, p. 1925–1945, 2019.

SANTOS, Francisco Heider Willy dos et al. Rastreamento de embarcações em imagens satelitais utilizando metodologia multicritério para a priorização em tarefas de busca e salvamento. **Brazilian Journal of Development**, vol. 6, n. 5, p. 28245-28257, 2020.

SANTOS, Ronnie E. S. et al. Técnicas de Processamento de Linguagem Natural Aplicadas ao Processo de Mineração de Textos: Resultados Preliminares de um Mapeamento Sistemático. **Revista de Sistemas e Computação**, Salvador, vol. 4, n. 2, p. 116-125, jul./dez. 2014.

SEBRAE. Economia Criativa. **Corwdsourcing**. 2014. Disponível em: http://www.bibliotecas.sebrae.com.br/chronus/ARQUIVOS_CHRONUS/bds/bds.nsf/53db425dba9eb17208f2935a28cd1894/$File/2014_07_17_RT_Maio_EconomiaCriativa_Crowdsourcing_pdf.pdf. Acesso em: 7 jul. 2021.

SCHEE, Brian A. Vander. Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business [Book Review] 2009 Jeff Howe. **Journal of Consumer Marketing**, NY, vol. 26, n. 4, jun. 2009.

SCHERER, Dominik; MÜLLER, Andreas; BEHNKE, Sven. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In: **20th International Conference on Artificial Neural Networks (ICANN)**, Thessaloniki, Grécia, set. 2010.

SEGAL, Avi et al. A difficulty ranking approach to personalization in E-learning. **Int. J. Hum. Comput. Stud.,** vol. 130, p. 261–272, oct. 2019.

SILVA, Jennifer Amanda Sobral da; MAIRINK, Carlos Henrique Passos. Inteligência artificial: aliada ou inimiga? **LIBERTAS: Rev. Ciênci. Soc. Apl**., Belo Horizonte, vol. 9, n. 2, p. 64-85, ago./dez., 2019.

SILVA, Renato M. et al. Towards automatically filtering fake news in Portuguese. **Expert Systems With Applications**, [S. I.], vol. 146, 2020.

SILVA, Valtemir A.; BITTENCOURT, Ig Ibert; MALDONADO, José C. Automatic Question Classifiers: A Systematic Review. **IEEE Transactions on Learning Technologies**, vol. 12, n. 4, p. 485–502, 2019.

SILVER, David; SINGH, Satinder; PRECUP, Doina; SUTTON, Richard S. Reward is enough. **Artificial Intelligence**, vol. 299, n. 3535, 2021.

SIMON, Herbert A. A behavioral model of rational choice, **Quarterly Journal of Economics**, vol. 69, n. 1, p. 99–118, 1955.

SRINIVASULU, Kothuru. Health-Related Tweets Classification: A Survey. **Advances in Intelligent Systems and Computing**, vol. 1245, p. 259–268, 2021.

STIGGINS, Richard J. Assessment crisis: The absence of assessment for learning, **Phi Delta Kappan**, vol. 83, n. 10, 2002.

SUN, Bo et al. Automatic Question Tagging with Deep Neural Networks. **IEEE Transactions on Learning Technologies**, [S. I.], vol PP, n. 99, 2018.

SUPER DATA SCIENCE TEAM. https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-1-convolution-operation. Acesso em: 10 set. 2021.

TEIXEIRA, João de Fernandes. **Inteligência artificial**: uma odisseia da mente. [Recurso eletrônico]. Rio de Janeiro: Paulus, 2009.

VERDÚ, Elena et al. A genetic fuzzy expert system for automatic question classification in a competitive learning environment. **Expert Systems with Applications**, vol. 39, n. 8, p. 7471-7478, 15 jun. 2012.

VON NEUMANN, John; MORGENSTERN, Oskar. **Theory of Games and Economic Behavior**. Princeton: Princeton University Press, 1947.

WANG, Dakuo et al. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. **Proceedings of the ACM on Human-Computer Interaction,** vol. 3 (CSCW), p. 1–24, 2019.

WILKINSON, Mark D. et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship, **Scientific Data**, vol. 3, n. 160018, 2016.

WORLD HEALTH ORGANIZATION. [S. I.]. **Novel Coronavirus (2019-nCoV) Situation Report.** [S. I.]. Disponível em: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf. Acesso em: 11 fev. 2020.

WU, Jianxin. **Introduction to Convolutional Neural Networks**. LAMDA Group. National Key Lab for Novel Software Technology, Nanjing University, China, 1 mai. 2017.

XIA, Meng et al. PeerLens: Peer-inspired Interactive Learning Path Planning in Online Question Pool. In: **Conference on Human Factors in Computing Systems - Proceedings**, p. 1–12, may 2019.

GLOSSÁRIO

Deep Learning: Algoritmos de aprendizado profundo.

Ciência de Dados: Disciplina de *Data Science*.

E-learning: Ferramenta de aprendizagem eletrônica.

Label: Rótulo

spaCy: Biblioteca de anotações linguísticas.

Machine Learning: Aprendizado de Máquina

APÊNDICE A - PRODUÇÕES CIENTÍFICA CORRELATAS

Esta pesquisa propiciou a divulgação de alguns resultados correlatos em âmbito acadêmico, os quais podem ser citados:

- Apresentação do trabalho "*Data science supporting a model question classifier*" na conferência internacional "Information Technology and Quantitative Management (ITQM 2020&2021)" em Chengdu na China em 2021.
- Publicação de artigo em jornal: Rafael Jardim, Carla Delgado, Daniel Schneider. Data science supporting a question classifier model. Procedia Computer Science, Volume 199, 2022, Pages 1237-1243, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.01.157, categorizado como Q2 no Scopus e A3 no Qualis CAPES.
- Competiu ao Prêmio LF do "XXVII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)" com o trabalho "CLIQ! Classificador Inteligente de Questões Baseado em Ciência de Dados para a elaboração de provas".

APÊNDICE B - PRODUÇÕES CIENTÍFICA PARALELAS

Ao longo deste mestrado, foi desenvolvido outros trabalhos científicos paralelos, não relacionados a esta pesquisa, que podem ser citados:

- R. R. J. Jardim, M. Santos, E. Neto, E. da Silva and F. de Barros. "Integration of the waterfall model with ISO/IEC/IEEE 29148:2018 for the development of military defense system," in IEEE Latin America Transactions, vol. 18, no. 12, pp. 2096-2103, December 2020, https://doi.org/10.1109/TLA.2020.9400437.
- Rafael Jardim, Marcos dos Santos, Edgard Neto, Fernando Martins Muradas, Bruna Santiago, Miguel Moreira. Design of a framework of military defense system for governance of geoinformation. Procedia Computer Science, Volume 199, 2022, Pages 174-181, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.01.022.
- M. Â. L. Moreira, C. F. S. Gomes, M. dos Santos, M. P. Basilio, I. P. A. Costa, C. S. R. Junior, R. R. J. Jardim. Evaluation of drones for public security: a multicriteria approach by the PROMETHEE-SAPEVO-M1 systematic. Procedia Computer Science, Volume 199, 2022, Pages 125-133, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.01.016.
- DE PAULA, Natália Oliveira Barbosa de Paula; JARDIM, RAFAEL R. J.; PEREIRA, Daniel Augusto de Moura; SANTOS, M. . Estratégia de distribuição de vacinas no combate ao COVID-19: apoio à decisão a partir do método multicritério ELECTRE-MOr. In: XLI Encontro Nacional de Engenharia de Produção (ENEGEP 2021), 2021, Foz do Iguaçu. Anais Eletrônicos do Encontro Nacional de Engenharia de Produção, 2021.
- Yuri Marinho de Carvalho, & Jose Carlos Cesar Amorim, & Marcos dos Santos, & Rafael Ris-Ala José Jardim. (2021, February). APLICAÇÃO DO MÉTODO SAPEVO-M PARA INTEGRAÇÃO LOGÍSTICA INTERMODAL NA REGIÃO AMAZÔNICA: UMA CONTRIBUIÇÃO PARA O EXÉRCITO BRASILEIRO. Trabalho apresentado em Anais do LII Simpósio Brasileiro de Pesquisa Operacional.
- JARDIM, RAFAEL R. J.; OLIVEIRA, L. F.; ESTEVES, M. G. P.; DE SOUZA, J. M.; DE CASTRO, N.; ROSENTAL, R.; OLIVEIRA, C.. Designing a Collaboration Platform for Electricity Consumer Councils. In: 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), p. 452., 2019, Porto.

ANEXO A - QUESTÕES ORIGINAIS

Quadro original das “questões de pesquisa” da revisão sistemática de SILVA, BITTENCOURT e MALDONADO (2019).

Quadro 5: Questões originais da pesquisa.

Research Question	Motivation
RQ1: What computational methods are used to implement classifiers?	This question identifies the methodology adopted in the construction of the main algorithms of Machine Learning to classify questions.
RQ2: Which taxonomies were adopted for classification?	The answer to this question makes it possible to identify the classification criteria applied.
RQ3: What are the main techniques used for feature extraction and selection?	The extraction and selection of relevant features have a direct impact on the performance of classifiers.
RQ4: What are the main instruments used to measure the classification results?	To check the different ways of analyzing the efficiency of the classification process.

Fonte: SILVA, BITTENCOURT e MALDONADO (2019).