Automatizando a extração das competências do futuro com uso da Inteligência Artificial

José Roberto Madureira Junior
https://orcid.org/0000-0002-8059-8983

Adani Cusin Sacilotti
https://orcid.org/0009-0001-3845-7036

Reginaldo Sacilotti
https://orcid.org/0009-0004-7215-8306

Resumo – O presente artigo aborda a automatização da extração de habilidades essenciais no mercado de trabalho, a partir de relatórios sobre o assunto. Para isso, serão apresentadas as tecnologias de Processamento de Linguagem Natural (PLN) e Inteligência Artificial (IA) oferecidas como um serviço no modelo de *Cloud Computing*. O objetivo deste artigo é apresentar um método capaz de extrair as habilidades de relatórios que são essenciais no mercado de trabalho. Muito tem sido discutido e várias previsões têm sido apresentadas sobre o mercado de trabalho do futuro, no entanto, pouco se tem falado sobre métodos para automatizar a extração das habilidades dos muitos relatórios sobre o assunto. Esta pesquisa oferece uma contribuição para a extração desse tipo de informação de forma rápida e automatizada. No que diz respeito aos leitores desses relatórios, um método automatizado pode contribuir para a identificação de insights. O estudo baseou-se em dados não estruturados para a realização da análise e extração dos insights, apresentando também o estado da arte do Processamento de Linguagem Natural. São destacados exemplos do estado atual e possíveis caminhos para a evolução da tecnologia de Processamento de Linguagem Natural, bem como dos serviços de *Cloud Computing*, com foco em serviços de Inteligência Artificial relacionados às tarefas de Processamento de Linguagem Natural. Concluiu-se que a adoção de serviços de Inteligência Artificial no modelo de *Cloud Computing* pode trazer vantagens importantes para os projetos de Processamento de Linguagem Natural. Por fim, apresenta-se uma visão geral de como o método de extração de informações foi desenvolvido, com a intenção de oferecer uma visão arquitetônica da construção de soluções desse tipo, buscando auxiliar os usuários a superarem os desafios na criação de soluções semelhantes para diferentes áreas de negócio, identificando as habilidades do futuro.

Palavras-chave: Arquitetura de Sistemas. Habilidades. Inteligência Artificial.

Abstract - *This article discusses the automation of extracting essential skills in the job market from reports on the subject. To this end, technologies of Natural Language Processing (NLP) and Artificial Intelligence (AI) offered as a service in the Cloud Computing model will be presented. The aim of this article is to introduce a method capable of extracting the skills from reports that are crucial in the job market. Much has been discussed, and several predictions have been made about the future job market. However, little has been said about methods to automate the extraction of skills from the many reports on the subject. This research provides a contribution to the extraction of this type of information in a fast and automated manner. Regarding the readers of these reports, an automated method can contribute to identifying insights. The study was based on unstructured data for analysis and insight extraction, also presenting the state of the art in Natural Language Processing. Examples of the current state and potential paths for the evolution of Natural Language Processing technology and Cloud Computing services, with a focus on Artificial Intelligence services related to Natural Language Processing tasks, are highlighted. It was concluded that the adoption of Artificial Intelligence services in the Cloud Computing model can bring significant advantages to Natural Language Processing projects. Finally, an overview of how the information extraction method was developed is presented, with the intention of offering an architectural view of constructing such solutions, aiming to assist users in overcoming challenges in creating similar solutions for different business areas while identifying the skills of the future.*

Keywords: *System Architecture*. *Skills*. *Artificial Intelligence.*

1 Introdução

Desde a década de 1960, a automatização de tarefas com base em dados tem evoluído constantemente com a introdução de novos recursos, tais como sistemas de planejamento e sistemas robóticos cada vez mais avançados. Com essa evolução, a automação tem sido capaz de executar ações que vão além das tarefas usuais, permitindo que robôs se se ajustem e forneçam respostas de acordo com as alterações do ambiente em que estão inseridos. Esse avanço tecnológico tem revolucionado a forma como as tarefas são executadas em várias áreas, sendo um exemplo de como a automação tem impactado positivamente a eficiência e a produtividade em diversos campos (Laprade *et al.,* 2019).

A automação inteligente, impulsionada pela Inteligência Artificial (IA), está trazendo consigo uma nova onda de aperfeiçoamento nos processos automatizados. A incorporação da inteligência na automação está revolucionando a maneira como as pessoas utilizam a tecnologia e está auxiliando as organizações a desenvolverem serviços e produtos mais personalizados, além de reduzir custos e ampliando o êxito.

A automação impulsionada pela IA tem o potencial de reduzir a dependência da intervenção humana em tarefas operacionais rotineiras, permitindo a escalabilidade dos processos. À medida que mais tarefas são automatizadas, as pessoas são liberadas para se dedicarem a tarefas de maior valor, aumentando a eficiência e a produtividade (Laprade *et al*., 2019).

Atualmente, as técnicas de IA estão impulsionando o crescimento dos negócios em diversos setores, apesar do conteúdo muitas vezes pouco prático divulgado pela mídia *mainstream*. Essas técnicas têm sido amplamente empregadas para aprimorar a experiência em dispositivos móveis, realizar análises e extrair *insights* a partir de dados capturados por sensores, aprimorar robôs industriais e comerciais, impulsionar avanços em veículos autônomos e até mesmo para a criação de assistentes digitais. É importante destacar que essas mudanças estão apenas no começo e que é possível que venhamos a presenciar experiências ainda mais envolventes, contínuas e conversacionais. Nesse cenário, as empresas que não adotarem processos mais inteligentes estarão em grande desvantagem competitiva no mercado contemporâneo, caracterizado por uma rápida evolução.

O mercado de IA está passando por uma expansão significativa, com diversas pesquisas indicando uma demanda progressiva por ferramentas que empreguem IA para previsão, análise e automação. Um exemplo da capacidade da IA pode ser observado em sua contribuição econômica, que, segundo Pacete (2022), “[...] deve gerar US$ 13 trilhões de dólares no mundo até 2030”.

A verdade é que a IA tem se consolidado como uma das fundamentais tecnologias adotadas por empresas de vários setores. No entanto, é válido reconhecer que existem preocupações legítimas sobre o futuro da IA e suas inferências sociais (Markiewicz e Zheng, 2020, p. 4).

Um estudo que indica no mesmo sentido é o de Laprade *et al.* (2019), que afirma que essa transformação já está em andamento, com cerca de 45% das organizações pesquisadas enfrentando dificuldades para localizar as habilidades necessárias. Essa mesma pesquisa também revela que a maioria dos funcionários planeja adquirir qualificações adicionais para adquirir as habilidades necessárias para o futuro.

A carência de talentos é uma das principais preocupações para as organizações atualmente, e os executivos estão cientes da importância da lacuna de habilidades. Novas necessidades se destacam quanto as habilidades estão surgindo constantemente, enquanto outras habilidades estão se transformando em obsoletas, e essas modificações estão ocorrendo em um ritmo acelerado (Laprade *et al*., 2019).

Conforme as organizações procuram encontrar os talentos essenciais, observa-se um aumento no alinhamento entre as demandas de novas necessidades e a formação, a fim de suprir as necessidades de novas habilidades no mercado de trabalho.

Segundo Laprade *et al.* (2019) são aferidos que nos próximos três anos mais de 120 milhões de trabalhadores nas 12 maiores potencias econômicas mundiais precisarão passar por reciclagem ou requalificação. Número esse que é maior do que a força de trabalho combinada do Brasil e Canadá.

Nesse sentido, as habilidades do futuro, geralmente mencionadas em diversos relatórios, representam as habilidades que serão necessárias no mercado de trabalho em um futuro. Exemplos dessas habilidades incluem resolução de problemas, tomada de decisão e pensamento analítico.

Considerando a lacuna de habilidades essenciais no novo mercado de trabalho, especialmente impulsionado pela IA, este projeto visa desenvolver uma abordagem automatizada a extração das habilidades do futuro a partir de relatórios técnicos sobre o assunto, utilizando IA e Processamento de Linguagem Natural (PLN, em inglês *Natural Language Processing*). Além disso, serão abordados os conceitos principais, princípios arquiteturais e o estado atual das tecnologias fundamentais para a criação desse método.

2 Objetivo

O propósito deste artigo é elaborar um método automatizado para realizar a extração das habilidades do futuro a partir de relatórios técnicos sobre o tema, utilizando IA e PLN. Igualmente, o objetivo é proporcionar uma visão dos conceitos essenciais, princípios arquiteturais e o estado atual das tecnologias fundamentais para a criação desse método. Para isso, os objetivos secundários são fundamentais, a saber:

a) Analisar o estado atual da Inteligência Artificial no contexto da extração de *insights* de dados complexos, abordando suas características, cenários, contextos, capacidades e propósitos de uso;

b) Identificar os desafios relacionados à extração de conhecimento e examinar as tecnologias relacionas com essa atividade;

c) Realizar a construção de um protótipo de solução empregando a IA para a extração das competências do futuro.

3 Método

Ao longo de muitos séculos, o trabalho e a humanidade têm sido continuamente transformados pelas evoluções tecnológicas. A revolução industrial conduziu consigo a automação dos processos industriais e mudanças significativas na sociedade. Durante esse processo de industrialização, aconteceram rupturas sociais e econômicas, resultando na desvalorização de habilidades que antes eram valiosas, enquanto novas e inesperadas habilidades assumiram seu lugar.

Atualmente, a quarta Revolução Industrial, impulsionada por tecnologias como Internet das Coisas (IoT), IA e robótica, apresenta um potencial transformador comparável ao da energia elétrica, do processo de produção industrial e das telecomunicações em suas respectivas eras. Essas tecnologias abrem a

possibilidade de proporcionar novas possibilidade de vida, oferecendo condições superiores de trabalho e aumentando a longevidade da população. No entanto, isso depende da preparação das pessoas e das instituições em relação às habilidades necessárias para esse novo mercado de trabalho, onde são exigidas habilidades distintas (Autor *et al.*, 2020).

Diversos relatórios são publicados na mídia com o objetivo de apresentar de forma consolidada as habilidades que serão tendência em um futuro próximo. Nesse sentido, foram escolhidos cinco relatórios de empresas e instituições de pesquisa renomados na área. Esses cinco relatórios representam uma amostra de um conjunto maior de relatórios e pesquisas que abordam esse tema. Além disso, vale ressaltar que os relatórios selecionados possuem estruturas bastante diversificadas, apresentando informações em forma de figuras, gráfico e texto.

Com base nisso, o presente artigo tem como objetivo construir um método para identificar as habilidades necessárias para o futuro. Esse método utiliza IA e PLN para automatizar a extração dessas habilidades com base em relatórios técnicos que abordam o tema das competências. Os relatórios definidos para essa extração foram (Bakhshi *et al.*, 2017; MGI, 2017; O*NET, 2022; OECD, 2021; WORLD, 2020):

- *The Future of Jobs Report 2020,* realizado pela *World Economic Forum*;
- *OECD Skills Outlook* 2021: *Learning for Life,* realizado pela *OECD*;
- *The Future Of Skills: Employment In 2030,* realizado pela *Pearson*;
- *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation,* realizado pela *McKinsey Global Institute;*
- *The O*NET Content Model*, realizado pela *O*NET*.

Nos relatórios selecionados, é realizada uma a extração de todo o texto, e em seguida, fazendo uso de um serviço de IA em um ambiente de *Cloud Computing*, são identificadas as habilidades por intermédio de uma técnica de *Named-Entity Recognition* (NER). Após a identificação das habilidades, elas são registradas em um arquivo junto com informações como um identificador único, o relatório ao qual a habilidade pertence e a precisão da sua identificação.

Em seguida, o arquivo contendo todas as habilidades consolidadas a partir dos relatórios é utilizado para criar uma visualização interativa dos dados. Por meio dessa visualização, é possível selecionar um relatório em particular e visualizar as

informações relacionadas a ele. Isso permite uma interação mais detalhada com os dados.

Vale destacar, que qualquer profissional interessado no método de extração, como gestores de recursos humanos, executivos, pesquisadores e outros, podem adicionar novos relatórios ou versões atualizadas para extração das habilidades do futuro. No entanto, é necessário que os documentos estejam no formato PDF e escritos em inglês. Essa flexibilidade permite a inclusão de novas fontes de informações relevantes para a reconhecimento das habilidades requeridas.

4 Referencial Teórico

No mundo atual, caracterizado por uma ampla interconexão em rede, os dados e o informação fluem livremente, de forma rápida e abrangente, transcendendo diversas fronteiras. Observa-se que empresas, comunidades científicas e entidades governamentais têm experimentado diversas vantagens decorrentes dessa revolução, no entanto, também se deparam com uma série de desafios.

Os avanços significativos na Tecnologia da Informação e Comunicação (TIC) impulsionaram um considerável aumento na quantia de dados gerados e compartilhados. Consequentemente, levando a um crescimento substancial nos esforços para desenvolver sistemas, tecnologias e técnicas capazes de extrair valor desses dados. Nesse cenário, a capacidade de extrair informações valiosas dos dados transformando-se em um fator crucial tanto para a academia, a indústria, quanto para os órgãos governamentais (Zaghloul e Trimi, 2017).

Diversas empresas estão adotando abordagens de análise e processamento inteligente para realizar a extração de *insights* dos dados textuais coletados de várias fontes e em diferentes idiomas. É observado que, na maioria das vezes, esses dados são incompletos, imperfeitos e desestruturados, o que torna a tarefa cada vez mais complexa. Práticas, técnicas e metodologias voltadas para a extração de informações valiosas dos dados têm provocado mudanças nos tipos de dados que são extraídos e analisados, começando inicialmente com dados estruturados e, posteriormente, abrangendo também os dados desestruturado.

A análise de texto possui algumas áreas em desenvolvimento; são elas (Zaghloul e Trimi, 2017):

- Resposta a perguntas utilizando técnicas de PLN e aprimoramento da Interação Humano-Computador (IHC) na recuperação de informações;
- Extração e classificação de opiniões expressas em diversas fontes de dados, incluindo redes sociais, usando análise de sentimentos e técnicas de mineração de opinião;
- Identificação dos principais temas em grandes conjuntos de dados de texto desestruturado;
- Extração automatizada de informações estruturadas a partir de documentos.

O principal objetivo deste artigo é extrair as habilidades mencionadas em cinco relatórios técnicos, utilizando uma abordagem que inclui o NER. Mais especificamente, propõe-se um extrator de entidades para a categoria de habilidades, que consiste em um conjunto definido de identificadores. O método de extração proposto demonstra um aumento na precisão das entidades extraídas em comparação com o método inicialmente utilizado, contribuindo assim para a literatura existente.

Na sequência serão apresentadas as técnicas de extração de entidades nomeadas (*Named-Entity Recognition*), onde destacamos suas principais características e diferencias de cada delas. São evidenciadas as técnicas relacionadas a bibliotecas especificas das linguagens de programação como, por exemplo, *SpaCy, NLTK, OpenNLP e CoreNLP*. Igualmente ressaltamos as técnicas ligadas a serviços de *Cloud Computing* aplicados à PLN como, por exemplo, *Amazon Comprehend*, *Natural Language Understanding*, *Natural Language AI* e *Cognitive Services.*

4.1 Named-Entity Recognition

Com a evolução da capacidade de interpretação da linguagem natural por parte dos computadores, surgem novas oportunidades, como a melhoria dos mecanismos de busca, o aperfeiçoamento das interfaces de aplicativos e a interatividade dos assistentes pessoais, bem como o desenvolvimento de técnicas e ferramentas para extrair *insights* em documentos (Zaghloul e Trimi, 2017).

A área de PLN engloba a compreensão da linguagem humana em diversas tarefas, como tradução automática, resposta a perguntas, recuperação e classificação de informações. A PLN tem sido cada vez mais inserida em uma

ampla gama de aplicações, como análise de inteligência, compreensão de máquina, leitura de documentos, e análise de dados de redes sociais, entre outras. Além disso, à medida que a utilização do PLN cresce, há uma diversidade cada vez maior de domínios explorados, abrangendo áreas como farmacologia, notícias, direito, biomedicina e química. A variedade de idiomas também está em ascensão, alinhada com o objetivo de longo prazo da PLN de desenvolver algoritmos capazes de ler e adquirir conhecimento de texto automaticamente.

As aplicações que empregam um nível avançado de PLN contam com a extração de entidades e os relacionamentos entre essas entidades, além do uso de *machine learning*, a fim de aprimorar a extração de informações. A maneira mais eficaz e comum de obter o significado do texto é por meio das entidades presentes no texto, uma vez que são essas entidades que conferem sentido às informações contidas. A extração de entidades geralmente é realizada por meio de técnicas de combinação. Sistemas que utilizam essa extração têm como objetivo identificar os elementos em um texto que pertencem a uma categoria específica de entidades definidas, juntamente com seus respectivos relacionamentos (Zaghloul e Trimi, 2017).

A principal abordagem para fazer a extração de entidades, conhecida como NER, consiste em identificar automaticamente os elementos ou nomes presentes no texto e classificá-los em um conjunto pré-definido de categorias. As entidades mais comumente utilizadas incluem valores numéricos, pessoas, organizações, lugares e datas. À medida que os domínios de aplicação se expandem, surge a necessidade de ampliar igualmente as categorias de entidades, como tempo, equipamentos, armas, plantas, animais, proteínas e genes. Uma alternativa reside na extração de relacionamentos, que envolve identificar duas entidades que possuem uma associação no texto analisado. Essa análise de relacionamentos desempenha um papel fundamental em uma ampla variedade de aplicações, incluindo leitura e compreensão de texto por máquina, análise de inteligência e mídias sociais.

As abordagens de NER podem ser categorizadas em regras ou estatísticas, e atualmente há uma abordagem híbrida que combina ambos os métodos, conhecida como NER Híbrido. A abordagem fundada em regras é a mais antiga forma de realizar o NER, na qual as regras são definidas manualmente. Sistemas de PLN baseados em regras determinam a presença de uma entidade e sua

classificação com base em regras que podem se apoiar em gramática, dicionários geográficos com nomes de pessoas e empresas, entre outros exemplos.

Em sistemas de PLN baseados em regras, as ontologias também desempenham um papel importante ao definir e identificar um conjunto de categorias relacionadas. Elas são especialmente úteis em categorias que possuem entidades altamente especializadas, como por exemplo, entidades relacionadas à biomedicina, que têm um número limitado de membros. Nesses casos, o desempenho do algoritmo é determinado pela qualidade das regras e pela precisão das ontologias utilizadas (Zaghloul e Trimi, 2017).

Para realizar a extração das habilidades pode ser utilizado várias bibliotecas recorrentes na literatura, neste contexto serão apresentadas as 4 principais, são elas *SpaCy, NLTK, OpenNLP e CoreNLP*, (Pinheiro *et al*., 2021).

O *spaCy* se destaca especialmente em tarefas de extração de informações em larga escala, sendo altamente eficiente. Sua facilidade de instalação e a simplicidade e produtividade de sua API tornam o *spaCy* uma escolha popular para projetos em *Python*. Com o spaCy, é possível realizar análises sofisticadas e obter resultados precisos de forma eficiente, proporcionando uma experiência de desenvolvimento agradável para os usuários (Explosion, 2022).

O NLTK é uma das principais plataformas de PLN em *Python*, que oferece um conjunto abrangente de bibliotecas para tarefas como processamento de texto para classificação, marcação, lematização e *tokenização*. O NLTK é um projeto gratuito e de código aberto, amplamente utilizado e apreciado pela comunidade de desenvolvedores. É compatível com os sistemas operacionais *Windows*, *Mac OS X* e *Linux*, o que permite que os usuários aproveitem suas funcionalidades em diferentes ambientes (NLTK, 2021).

O *CoreNLP* é uma biblioteca de PLN desenvolvida em Java, que possibilita aos usuários obter anotações linguísticas para textos. Essas anotações incluem informações como entidades nomeadas, partes do discurso e valores numéricos, além de informações relacionadas a tempo. Recentemente, o *CoreNLP* expandiu seu suporte para oito idiomas, permitindo um processamento mais abrangente e eficiente em diferentes contextos linguísticos (Stanford, 2021).

A biblioteca *Apache OpenNLP* tem como objetivo ser um conjunto de ferramentas em linguagem Java para o processamento de textos em linguagem natural, com base em técnicas de aprendizado de máquina. Essa biblioteca oferece

suporte a diversas funções clássicas de PLN, como agrupamento, *tokenização* e segmentação de frases. Essas funcionalidades são essenciais para o desenvolvimento de aplicações avançadas de processamento de texto em uma ampla variedade de idiomas (Apache, 2021).

Uma alternativa simplificada ao uso das bibliotecas consiste na utilização de serviços de *Cloud Computing* aplicados à PLN, onde são apresentados a seguir os serviços dos principais fornecedores de *Cloud Computing*.

O *Amazon Comprehend* é um serviço de *Cloud Computing* oferecido pela *Amazon Web Services* (AWS) que permite a extração de *insights* de textos sem a necessidade de pré-processamento e treinamento utilizando técnicas de PLN. Esse serviço é capaz de processar qualquer texto no formato UTF-8 e realizar a extração de frases-chave, o reconhecimento de entidades, detecção de idioma, análise de sentimentos e outros elementos comuns em um texto (Amazon, 2022).

De forma similar, a IBM oferece o *Natural Language Understanding* é um serviço de *Cloud Computing* que faz uso de técnicas avançadas de PLN e *machine learning* para extrair informações valiosas de textos. Ele oferece recursos como entidades, análise de palavras-chave, sentimentos e relações, além de possibilitar a resposta a perguntas com base em modelos personalizados (IBM, 2022).

Já o *Natural Language AI* é um serviço de *Cloud Computing* fornecido pelo Google, projetado para gerar *insights* a partir de documentos desestruturados utilizando técnicas avançadas de *machine learning*. Esse serviço oferece recursos poderosos para extração, análise e armazenamento de textos, permitindo também o treinamento de modelos personalizados e a integração de PLN em aplicações por meio de uma API simplificada (Google, 2022).

Por fim, a plataforma *Cognitive Services* da Microsoft Azure foi criada para permitir que desenvolvedores e cientistas de dados possam facilmente incorporar recursos avançados de compreensão de linguagem, reconhecimento visual e pesquisa em suas aplicações. Isso permite que as aplicações entendam e acelerem a tomada de decisões complexas. Dentro da categoria de serviços *Cognitive Services,* destaca-se o serviço *Text Analytics*, que é especialmente útil para a extração de conteúdo de texto. O *Text Analytics* permite analisar e extrair *insights* significativos a partir de textos desestruturado (Microsoft, 2021a):

- Os recursos pré-configurados são funcionalidades que podem ser facilmente utilizadas para criar uma API com recursos padronizados.

Basta enviar os dados adequados e utilizar as respostas geradas pelo sistema;

- Os recursos personalizáveis permitem treinar um modelo de IA utilizando as ferramentas da *Microsoft* para criar uma API com recursos diferenciados.
- O *Text Analytics* possui os seguintes recursos principais:
    - A detecção de idioma é um recurso que permite analisar um texto e identificar o idioma em que foi escrito;
    - As respostas automáticas a perguntas são recursos pré-configurados que permitem apresentar respostas a perguntas extraídas do texto de entrada;
    - A análise de sentimentos e opiniões é um recurso que fornece rótulos de sentimento, como positivo, negativo e neutro, para partes específicas ou para todo o texto. Esse recurso permite obter informações sobre as opiniões expressas no texto, ajudando a compreender a polaridade emocional associada a determinadas partes do conteúdo;
    - Extração de frases-chave é um recurso que permite a uma aplicação extrair palavras-chave e frases-chave importantes de um texto;
    - NER: esta capacidade é pré-configurada para identificar várias categorias;
    - NER personalizado: habilita a capacidade de treinar o serviço para reconhecer e extrair novas entidades utilizando como fontes de treinamento dados desestruturado.

Outra característica distintiva do *Text Analytics* é a facilidade de integrar serviços relacionados à fala. Esse serviço foi selecionado para a construção do extrator devido à sua compatibilidade com *Python* e à capacidade de identificar a entidade habilidade (Microsoft, 2021a).

5 Resultados e Discussão

A principal motivação por trás da construção do extrator foi desenvolver um método capaz de identificar habilidades de forma eficiente, sem depender de um

pré-treinamento com uma lista específica de habilidades ou utilizar documentos marcados como base para o treinamento.

Esse método se concentra exclusivamente na entidade habilidade para realizar a extração, e essa escolha foi feita devido à necessidade de extrair especificamente as habilidades relevantes para o futuro mercado de trabalho. O objetivo é extrair informações de relatórios técnicos que abordam as habilidades necessárias para o futuro. Além disso, esse extrator pode ser utilizado em sistemas de recursos humanos, auxiliando nas atividades de recrutamento e seleção de candidatos.

O método proposto destina-se a pesquisadores, executivos, gestores de recursos humanos e outros profissionais interessados em extrair e analisar as habilidades mencionadas em relatórios e documentos no formato PDF[1]. No estudo, foram utilizados cinco relatórios como base para a extração das habilidades. Esses relatórios serviram como amostra representativa para validar a eficácia do método e garantir sua aplicabilidade em diferentes contextos, sendo eles (Bakhshi *et al.*, 2017; MGI, 2017; O*NET, 2022; OECD, 2021; WORLD, 2020):

- *The Future of Jobs Report 2020,* realizado pela *World Economic Forum*;
- *OECD Skills Outlook* 2021: *Learning for Life,* realizado pela *OECD*;
- *The Future Of Skills: Employment In 2030,* realizado pela *Pearson*;
- *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation,* realizado pela *McKinsey Global Institute;*
- *The O*NET Content Model*, realizado pela *O*NET*.

Esses relatórios terão suas habilidades extraídas pelos sistemas NER. A seguir será apresentado o fluxo da aplicação, conforme pode ser visto na Figura 1.

[1] Portable Document Format (PDF) é um tipo de arquivo criado com o objetivo de ser aberto em qualquer hardware e qualquer sistema operacional.

Figura 1 - Fluxo do Extrator de Habilidades.

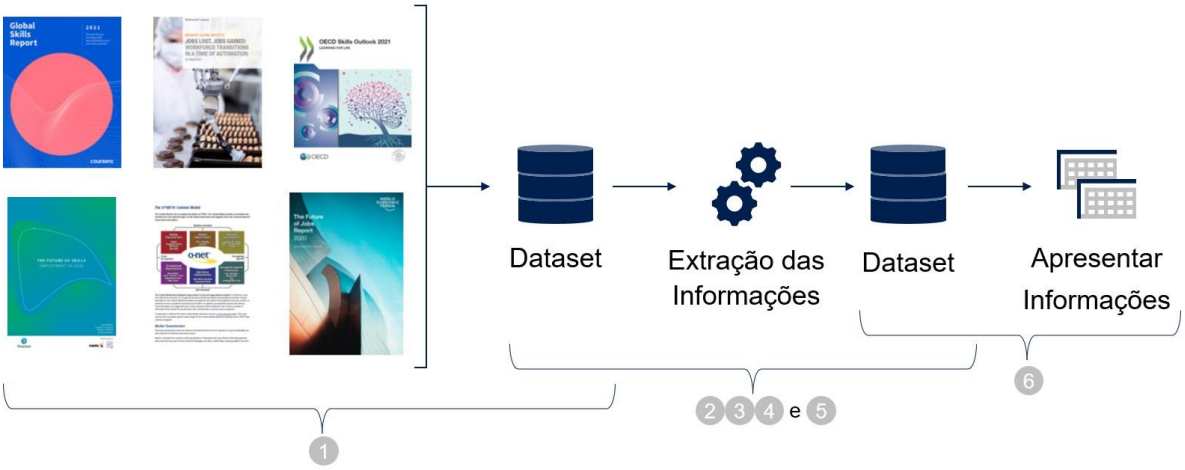

Imagem inspirada em: Bordin *et al.* (2013).

A Figura 1 apresenta uma descrição de todas as fases do fluxo, devidamente identificadas com números que correspondem às etapas a seguir descritas:

1 Inserção dos relatórios desejados para a extração das habilidades em um *dataset*[2] localizado em um repositório de documentos desestruturado;

2 Após os relatórios serem adicionados ao *dataset*, o extrator estabelece uma conexão e inicia o processo de extração do texto dos diferentes documentos contidos nele;

3 Com a conclusão da extração nos documentos, o texto é submetido para realizar a extração das entidades presentes nele;

4 Após a extração das habilidades, é criado um arquivo CSV[3] que contém três colunas. A primeira coluna identifica o relatório de onde a entidade foi extraída, a segunda coluna contém a habilidade extraída e a terceira coluna indica a precisão da identificação da habilidade extraída;

5 O extrator salva o arquivo CSV em outro *dataset*;

6 A leitura do arquivo CSV ocorre e, a partir dos dados contidos nele, é criada uma visualização dinâmica.

[2] Os datasets ou conjuntos de dados são o principal insumo dos processos de análise de dados.
[3] O arquivo CSV é um arquivo de texto com formato específico para possibilitar o salvamento dos dados em um formato estruturado de tabela.

5.1 Arquitetura do Método de Extração

O método inicial tinha uma lógica idêntica à versão final, porém possuía uma arquitetura substancialmente diferente da arquitetura final. Isso ocorria porque não era baseado no serviço de *Cloud Computing*, como ilustrado na Figura 2.

Figura 2 - Arquitetura proposta para o método.

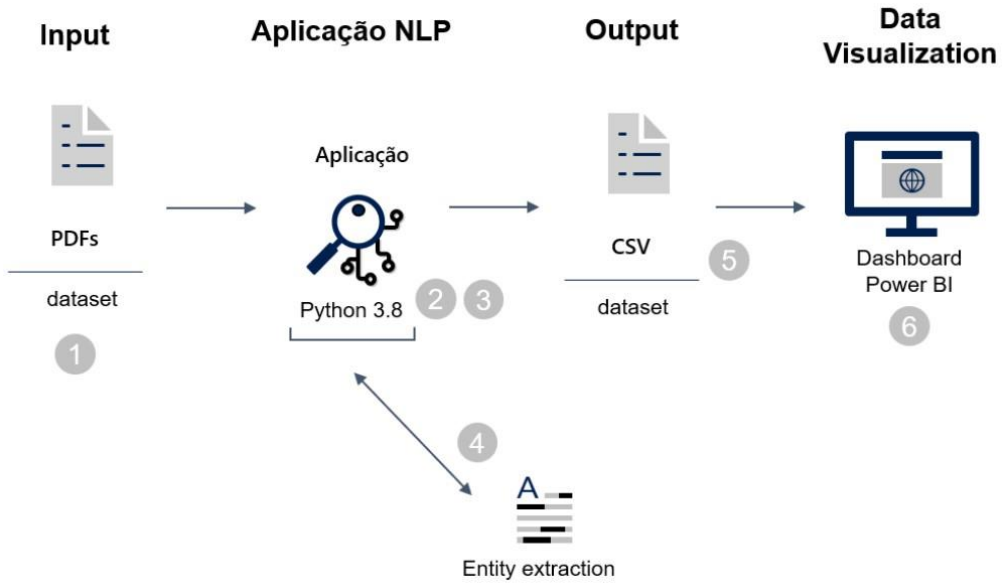

Imagem inspirada em: Microsoft (2021b).

O método principal para extrair as habilidades é ilustrado na Figura 2 e será detalhado a seguir (Explosion, 2022):

1 Os arquivos PDF dos relatórios são inseridos e armazenados em um diretório chamado *input*, localizado no mesmo diretório da aplicação;

2 A execução do aplicativo *extraction_skills_textanalytics.ipynb* ocorre para estabelecer a conexão com o diretório *input*, permitindo o carregamento dos relatórios armazenados nesse diretório*;*

3 Após carregar os relatórios na aplicação, é realizado o processo de extração de todo o texto contido nos PDFs, incluindo aqueles presentes em tabelas, gráficos e figuras;

4 A aplicação utiliza o texto extraído dos PDFs como parâmetro para um método criado em *Python*, que tem a finalidade de extrair as habilidades. O resultado desse método é retornado como as habilidades extraídas;

5 A aplicação gera um arquivo chamado *OutputFile.csv* com duas colunas: uma contendo o nome do relatório do qual as habilidades foram extraídas, e

outra contendo as habilidades extraídas. Esse arquivo é armazenado no diretório chamado *output*, localizado no mesmo diretório da aplicação.

6 Para visualizar os dados de forma dinâmica, um *dashboard* é criado utilizando o *Microsoft Power BI*. Esse painel permite apresentar os dados contidos no arquivo *output.csv* de maneira interativa e visualmente atraente.

Para realizar a extração das habilidades, foi utilizada a biblioteca *spaCy* em *Python*. No entanto, essa biblioteca não possui um treinamento prévio específico para identificar habilidades. Portanto, foi necessário realizar um treinamento personalizado usando um arquivo CSV que continha diversos exemplos de habilidades. Uma questão adicional que surgiu foi relacionada ao tamanho do arquivo de treinamento. Quanto maior o tamanho do arquivo, maior era a redução no desempenho. Além disso, a aplicação ainda não alcançava um número satisfatório de habilidades identificadas. (Explosion, 2022; Li, 2021).

Dentro deste contexto, uma nova arquitetura foi desenvolvida utilizando serviços de *Cloud Computing*. O método de extração adota a arquitetura proposta, conforme ilustrado na Figura 3.

Figura 3 - Arquitetura proposta para o método.

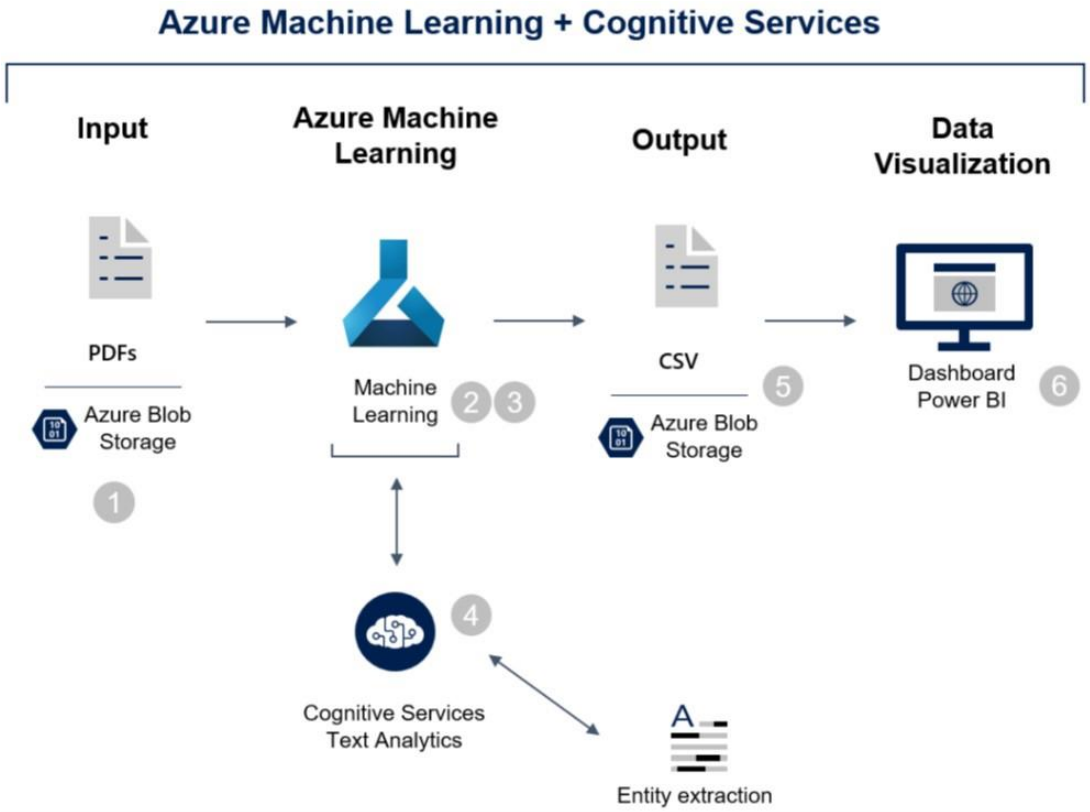

Imagem inspirada em: Microsoft (2021b).

A Figura 3 ilustra a arquitetura do método proposto, e os diferentes componentes são descritos a seguir:

1 Os arquivos PDFs dos relatórios são adicionados a um contêiner chamado *input*, localizado em uma instância do serviço *Azure Blob Storage*;
2 Uma instância do serviço *Azure Machine Learning* executa a aplicação chamada *extraction_skills_textanalytics.ipynb*. Essa aplicação estabelece a conexão com o *Azure Blob Storage* para carregar os relatórios armazenados no serviço*;*
3 Após carregar os relatórios na aplicação, é realizado o processo de extração de todo o texto contido nos PDFs, incluindo aqueles presentes dentro de tabelas, gráficos e figuras;
4 A aplicação utiliza o texto extraído dos PDFs como parâmetro e o envia para uma API criada com os *Cognitive Services Text Analytics*. Essa API retorna uma lista de habilidades extraídas daquele trecho de texto, juntamente com sua respectiva acurácia;
5 A aplicação gera um arquivo chamado *OutputFile.csv* com três colunas: o nome do relatório do qual a habilidade foi extraída, a habilidade extraída e a acurácia da habilidade. Esse arquivo é armazenado em um contêiner chamado *output* dentro de uma instância do *Azure Blob Storage*;
6 Para visualizar os dados de forma dinâmica, foi criado um *dashboard* utilizando o *Microsoft Power BI*. Esse painel é responsável por apresentar os dados de maneira interativa e visualmente atraente.

Nesse novo contexto, a arquitetura baseada em serviços de *Cloud Computing* elimina a necessidade de realizar um pré-treinamento com documentos reais semelhantes. Isso resulta em uma redução de 64,68% no tempo necessário para realizar a extração, ao mesmo tempo em que aumenta em 268,10% o número de habilidades identificadas com o método atualizado e as modificações feitas em sua arquitetura.

5.2 Data Visualization

Na área de Visualização de Dados (*Data Visualization*, em inglês), foi empregado o Power BI, que oferece a vantagem de criar visualizações interativas.

A seguir, apresentamos um exemplo da visualização geral do Dashboard, mostrada na Figura 4.

Figura 4 - Dashboard do Projeto.

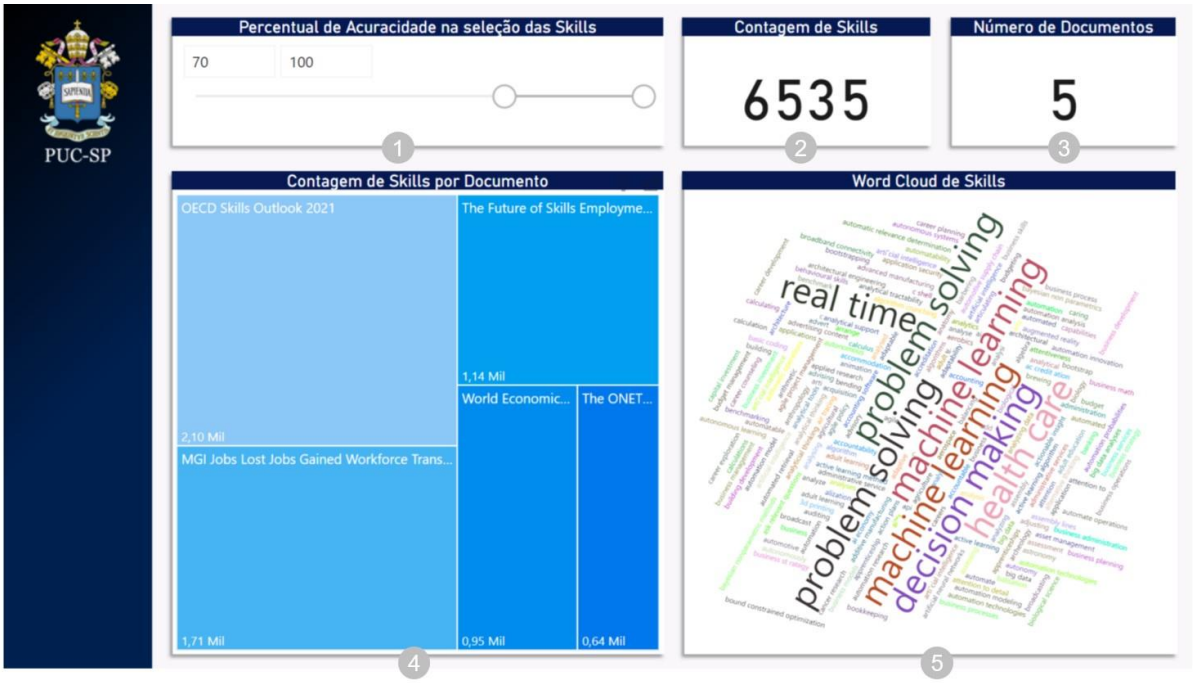

Imagem inspirada em: Microsoft (2021b).

Na Figura 4, é possível visualizar o dashboard criado para apresentar os resultados de maneira dinâmica. O dashboard é dividido em cinco áreas distintas:

1. A primeira área do *dashboard* é destinada ao controle, permitindo que o usuário defina o percentual de acurácia das habilidades apresentadas;
2. A segunda área do *dashboard* é dedicada ao painel que exibe o total de habilidades extraídas dos relatórios;
3. O terceiro painel exibe o total de relatórios utilizados para a extração das habilidades;
4. O quarto painel exibe um *treemap* com todos os relatórios, onde os retângulos têm tamanhos proporcionais à quantidade de habilidades extraídas de cada relatório. Na parte inferior dos retângulos, também é exibida a quantidade de habilidades extraídas de cada relatório;
5. O último painel exibe um *Word Cloud* (nuvem de palavras, em português) que apresenta todas as habilidades diferentes extraídas dos relatórios. O tamanho das palavras no *Word Cloud* representa a incidência de cada habilidade.

Na área cinco, destacam-se as diversas habilidades extraídas dos relatórios técnicos, usando as técnicas apresentadas. Entre essas habilidades, podemos mencionar o pensamento crítico, a resolução de problemas e o *Machine Learning*, que serão fundamentais no mercado de trabalho do futuro.

6 Considerações finais

Neste artigo, realizou-se um estudo com o objetivo de desenvolver um protótipo de solução que faz uso de IA para automatizar a extração das competências do futuro a partir de relatórios técnicos que tratam desse tema. Além disso, foram apresentadas as tecnologias utilizadas na solução. De forma complementar, também foi abordada a evolução dos dados, que são complexos e desestruturados, destacando a importância da IA para realizar a análise desses dados.

Como demonstrado na apresentação desta pesquisa, desenvolvemos um método capaz de extrair as competências do futuro por meio da análise de relatórios relevantes sobre o assunto. Utilizando a técnica de NER em relatórios em inglês, extraímos as entidades-habilidade. Com isso, ao final deste trabalho, podemos afirmar que o modelo criado é capaz de extrair as habilidades do futuro e apresentá-las em um *dashboard*.

Nos próximos anos, testemunharemos o avanço das tecnologias relacionadas ao PLN, o que nos permitirá explorar novas possibilidades para a técnica de NER. São esperados investimentos significativos para o treinamento de novos modelos, o que poderá ampliar a capacidade de extração de entidades e acelerar a realização dessa tarefa. Isso abrirá caminho para avanços ainda maiores na extração de habilidades e informações relevantes a partir de documentos e textos.

Para a construção do método, foram empregados serviços de IA em um modelo de *Cloud Computing*, o que permitiu a extração das entidades habilidades de maneira ágil e mais eficiente do que as técnicas utilizadas anteriormente. Durante o desenvolvimento do trabalho, também observamos que os serviços de IA em ambiente de *Cloud Computing* impulsionam inovações tecnológicas, liberando os usuários da necessidade de desenvolver e treinar algoritmos complexos de IA. Isso possibilita que as empresas se concentrem no

desenvolvimento de seus negócios, enquanto aproveitam os benefícios das soluções de IA prontas para uso.

Além disso, foram considerados outros serviços relacionados ao armazenamento, para manter os relatórios e posteriormente um arquivo com os resultados. Também foi fornecida uma infraestrutura de hardware e software para executar o programa central. Além disso, foi construído e disponibilizado um *dashboard* com os resultados expostos, permitindo que o usuário navegue por esses dados e obtenha *insights* diferentes.

Neste contexto, foram utilizados os seguintes serviços de IA oferecidos no modelo de *Cloud Computing* na solução: *Azure Blob Storage*, *Azure Machine Learning*, *Cognitive Services Text Analytics* e o *Power BI Services*. Esses serviços desempenharam papéis fundamentais no desenvolvimento do método. Além disso, foi adotada uma arquitetura específica para a construção dessa solução, a qual foi fundamental para integrar e utilizar esses serviços de forma eficiente.

Outro aspecto digno de destaque é a comparação entre duas abordagens de extração da entidade-habilidade. Nesse sentido, foram exploradas duas formas distintas de trabalhar com o NER. A primeira abordagem utilizou a biblioteca *spaCy* em *Python*, que apresentou um desempenho inferior em termos de velocidade e um número entidades reconhecidas. Vale ressaltar que, para aumentar o número de entidades reconhecidas, seria necessário ampliar o treinamento, o que resultaria em uma diminuição ainda maior na velocidade de análise do método.

Em seguida, foi apresentada uma segunda abordagem para a extração das habilidades, utilizando o serviço de *Cloud Computing* do *Microsoft Azure*, o *Cognitive Services Text Analytics*. Essa abordagem demonstrou um desempenho superior, com uma capacidade de extração de habilidades igualmente superior, sem a necessidade de treinamento adicional. A seguir, serão apresentados os resultados obtidos com o uso desses dois métodos.

Tabela 1 - Comparativo entre as formas de extração.

Forma de extração	Tempo para extração (s)	Número de habilidades	Treinamento
Biblioteca do Python chamada spaCy	359	*3022*	Necessário treinamento
Cognitive Services Text Analytics	218	11124	Não é necessário treinamento

Baseado nisso, uma nova arquitetura fundada em serviços de IA no modelo de *Cloud Computing* elimina a necessidade de realizar um pré-treinamento semelhante aos documentos reais analisados. Isso resulta em uma redução de tempo significativa na realização da extração, sendo 64,68% mais rápido. Além disso, essa nova arquitetura, juntamente com as modificações realizadas, ampliou em 268,10% o número de habilidades identificadas com o método atual.

Identificamos que a maior contribuição desta pesquisa foi a construção de um protótipo de solução que utiliza a IA para extrair as competências do futuro. Essa solução é capaz de analisar automaticamente relatórios que apresentam as tendências de habilidades. Além disso, nosso protótipo permite a atualização contínua dos relatórios, bastando substituí-los ou adicioná-los ao *dataset* existente. Outra contribuição importante do nosso estudo foi fornecer uma visão abrangente das tecnologias de PNL, auxiliando os usuários e potenciais usuários a reconhecer e avaliar quando utilizar essa técnica. Essa análise detalhada das tecnologias de PNL permite que as organizações obtenham vantagens reais ao adotá-las, proporcionando *insights* valiosos e impulsionando a tomada de decisões fundamentadas.

Referências

AMAZON Web Services. **Amazon Comprehend Documentation**. 2022. Disponível em: https://docs.aws.amazon.com/comprehend/index.html. Acesso em: 15 mar. 2023.

APACHE. **Apache OpenNLP**. 2021. Disponível em: <https://opennlp.apache.org/>. Acesso em: 13 dez. 2021.

AUTOR, David *et al.* **Artificial Intelligence and Work.** 2020. Disponível em: <https://cetic.br/media/docs/publicacoes/6/20210115080952/internet_sectoral_overview_year-12_n_4_artificial_intelligence_and_work.pdf>. Acesso em: 14 mar. 2023.

BAKHSHI, Hasan et al. **THE FUTURE OF SKILLS EMPLOYMENT IN 2030**. 2017. Disponível em: <https://futureskills.pearson.com/research/assets/pdfs/technical-report.pdf>. Acesso em: 18 jan. 2022.

BORDIN, Andréa Sabedra *et al*. Modelo de Descoberta de Conhecimentos e Interesses Baseado em Insumos Textuais Eletrônicos: uma proposta para apoio a gestão do capital humano. **Revista Gestão & Tecnologia**, Pedro Leopoldo - MG, v. 3, n. 13, p. 3-22, 2013.

EXPLOSION. **Industrial-Strength Natural Language Processing.** 2022. Disponível em: <https://spacy.io/>. Acesso em: 15 jan. 2023.

GOOGLE. **Natural Language AI**. 2022. Disponível em: <https://cloud.google.com/natural-language. Acesso em: 11 mar. 2023.

IBM. **Watson Natural Language Understanding**. 2022. Disponível em: https://www.ibm.com/cloud/watson-natural-language-understanding/details. Acesso em: 02 jan. 2023.

LAPRADE, Annette *et al.* **The enterprise guide to closing the skills gap.** 2019. Disponível em: <https://www.ibm.com/downloads/cas/EPYMNBJA>. Acesso em: 23 fev. 2023.

LI, Susan. **Named Entity Recognition with NLTK and SpaCy.** 2021. Disponível em: <https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy8c4a7d88e7da>. Acesso em: 22 dez. 2021.

MARKIEWICZ, Tom; ZHENG, Josh. **Getting Started with Artificial Intelligence**: A Practical Guide to Building Enterprise Applications. 2. ed. Sebastopol, Ca: O'Reilly, 2020. 85 p.

MGI. **Jobs lost, jobs gained:** What the future of work will mean for jobs, skills, and wages. Paris: McKinsey Global Institute, 2017. 160 p. Disponível em: <https://www.mckinsey.com /featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages>. Acesso em: 21 fev. 2023.

MICROSOFT. **What is Azure Cognitive Service for Language?** 2021a. Disponível em: <https://docs.microsoft.com/en-us/azure/cognitive-services/language-service/overview>. Acesso em: 17 maio 2023.

MICROSOFT. **Knowledge Mining Solution Accelerator.** 2021b. Disponível em: https://docs.microsoft.com/en-us/samples/azure-samples/azure-search-knowledge-mining/azure-search-knowledge-mining/. Acesso em: 21 mar. 2023.

NLTK Project. **NLTK**: natural language toolkit. Natural Language Toolkit. 2021. Disponível em: <https://www.nltk.org/>. Acesso em: 12 dez. 2021.

O*NET. **The O*NET® Content Model.** 2022. Disponível em: <https://www.onetcenter.org/content.html>. Acesso em: 15 jan. 2023.

OECD. **OECD Skills Outlook 2021**: learning for life, oecd publishing. Paris: Oecd Publishing, 2021. 229 p. Disponível em: <https://doi.org/10.1787/e11c1c2d-en.> Acesso em: 30 fev. 2023.

PACETE, Luiz Gustavo. **Google lista os desafios da inteligência artificial no Brasil.** 2022. Disponível em: <https://forbes.com.br/forbes-tech/2022/10/google-lista-os-desafios-da-inteligencia-artificial-no-brasil/>. Acesso em: 20 jan. 2023.

PINHEIRO, Breno David Lopes *et al.* A Comparative Analysis of Machine Learning Named Entity Recognition Tools for the Brazilian and European Portuguese Language Variants. **Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021)**, [S.L.], p. 1-12, 29 nov. 2021. Sociedade Brasileira de Computação - SBC. http://dx.doi.org/ 10.5753/eniac.2021.18257.

STANFORD NLP Group. **Overview**: corenlp. CoreNLP. 2021. Disponível em: <https://stanfordnlp.github.io/CoreNLP/>. Acesso em: 12 dez. 2021.

WORLD Economic Forum. **The Future of Jobs.** 2020. Disponível em: <https://www. weforum.org/reports/the-future-of-jobs-report-2020/in-full>. Acesso em: 20 nov. 2020.

ZAGHLOUL, Waleed; TRIMI, Silvana. Developing an innovative entity extraction method for unstructured data. **International Journal Of Quality Innovation**. [S. l.], p. 1-10. 22 maio 2017.