Mercado Global de Moderação de Conteúdo: diagnóstico e oportunidades para o Brasil[1]

Gabriel Penna[2]

Emanuella Ribeiro Halfeld Maciel[3]

Thiago Tavares Nunes de Oliveira[4]

Resumo: O artigo realiza uma análise do mercado global de moderação de conteúdo *online,* identificando seus desafios e práticas, através de: i) uma revisão da literatura pertinente, explicando diferentes modelos de negócio e de interação humano-máquina que incidem sobre a disponibilidade de conteúdo online; ii) a apresentação de dados obtidos através de pesquisa exploratória de levantamento de vagas de emprego online, a fim de situar as cidades que possuem força de trabalho dedicada a tarefa e práticas de recrutamento e iii) a análise de *clippings* de notícia de diversos veículos de mídia e de documentários sobre o tema, buscando situar a realidade dos trabalhadores do mercado. Com um foco especial no mercado de moderação em português, o artigo cobre quatro grandes temas: a) localização dos pontos de moderação de conteúdo; b) perfil dos trabalhadores; c) condições de trabalho; e d) a importância do contexto e das habilidades linguísticas.
Palavras-chave: Moderação de Conteúdo; Liberdade de Expressão; Regulamentação de Plataformas.

Abstract: The article analyses the global online content moderation market, identifying its challenges and standard practices, through: i) a review of relevant literature, explaining different business models and human-machine interaction that affect the availability of online content; ii) the presentation of data obtained through an exploratory research of job posts available online, in order to determine the cities in which the content moderation workforce is distributed and their recruitment practices and iii) the analysis of news clippings from different media vehicles, as well as documentaries on the subject, seeking to situate the reality of market workers. Focusing on the Portuguese moderation market, the article covers four major themes: a) the location of content moderation points; b) the profile of agents that moderate content in Portuguese; c) their working conditions; and d) the importance of context and language skills in the market.
Keywords: Content Moderation; Freedom of Expression; Platform Regulation.

---

[1] Trabalho apresentado no V Congresso do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT.DD). Anais disponíveis
[2] Bacharel em Direito pela Universidade Federal de Viçosa e mestrando em Ciência Política pela Universidade Federal de Minas Gerais, gabrielpennaandrade@gmail.com
[3] Bacharel em Direito e Mestranda em Ciência Política pela Universidade Federal de Minas Gerais, Gerente de Projetos da ONG SaferNet Brasil, manu@safernet.org.br
[4] Mestre em Desenvolvimento e Gestão Social pela Universidade Federal da Bahia, Presidente da ONG SaferNet Brasil, thiagotavares@safernet.org.br

Introdução

A moderação de conteúdo produzido por usuários é uma característica essencial da interação digital contemporânea. Seu objetivo principal é garantir um ambiente digital saudável, coibindo atos ilícitos e aplicando as diretrizes da comunidade em questão. Embora seja ligada principalmente às redes sociais, que tem por principal traço ser um repositório de conteúdo produzido primariamente pelo usuário, atualmente a atividade de moderar conteúdo está presente em quase todas as plataformas digitais (Gillespie, 2018, cap. 1) e compreende uma série de diferentes modelos de negócio e fluxos de trabalho que podem ser agrupados sob o rótulo comum de moderação de conteúdo (Roberts, 2016, p. 147).

Quadro 1 – Notícias sobre moderação de conteúdo

Em 2020, a plataforma de exercícios físicos Peloton removeu *hashtags* relacionadas à teoria conspiratória QAnon, popular entre grupos de extrema-direita americana. O objetivo foi impedir a criação de redes extremistas com o uso da plataforma e evitar a disseminação de discursos de ódio e anti-democráticos (Smith, 2020)

Fonte: Elaboração Própria

Uma das maneiras pelas quais o conteúdo inapropriado pode ser localizado é a revisão editorial do que será postado na plataforma (Gillespie, 2018, cap. 4). A partir dessa técnica, o conteúdo será revisado por pessoas ligadas à plataforma em questão antes de ficar disponível ao restante dos usuários. Por ser uma técnica cara e lenta, por exigir revisão manual dos conteúdos, esse modelo é cada vez menos utilizado (Gillespie, 2018).

Um segundo modelo de detecção de conteúdo inapropriado é a denúncia de usuários. A partir desse sistema, conhecido pelo termo *flagging*, qualquer pessoa pode reportar à plataforma conteúdo que julgue incondizente com as diretrizes comunitárias. O grau de detalhamento das denúncias e os fluxos após a denúncia podem variar consideravelmente a depender da plataforma, o que pode gerar dificuldades e ausência de transparência nas decisões de moderação (Gillespie, 2018).

Quadro 2 – Notícias sobre moderação de conteúdo

Em 2022, o trailer da terceira temporada da série *The Boys* foi reportado mais de 20 milhões de vezes por usuários do Twitter, segundo a conta oficial da série (Jacobs, 2022). O motivo seriam as cenas violentas e gráficas mostradas no vídeo. Até o dia 18/04/2022, o conteúdo não havia sido retirado da plataforma, embora tenha sido sinalizado para conteúdo potencialmente sensível.

Fonte: Elaboração Própria

O terceiro modelo de detecção de conteúdo inapropriado funciona a partir de ferramentas de inteligência artificial. O uso de programas de computador permite uma inspeção massificada sobre o conteúdo, permitindo a identificação rápida e derrubada de conteúdos inapropriados ou mesmo perigosos. Por essa razão, as plataformas atualmente usam, em conjunto, técnicas de detecção automática de conteúdo inapropriado, mecanismos de denúncia de usuários e técnicas de revisão humana de conteúdos denunciados, garantindo tanto a rapidez da inteligência artificial quanto a acurácia do julgamento humano. Há, contudo, problemas específicos nesse modelo, como a existência de vieses conservadores nos algoritmos utilizados (Gillespie, 2018, cap. 4)

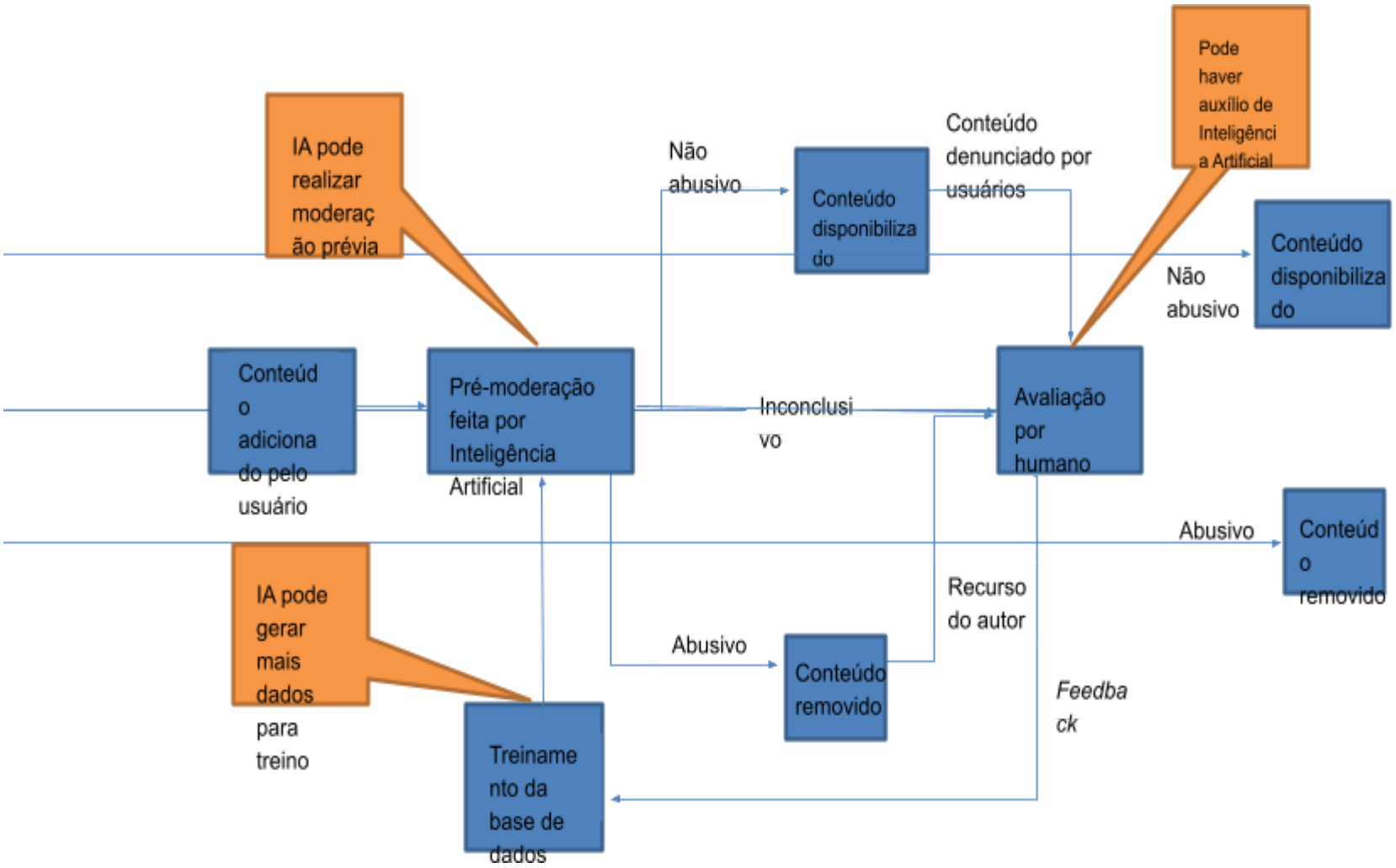
Quadro 3 – Notícias sobre moderação de conteúdo

Em 2019, um estudo do InternetLab revelou que uma ferramenta de análise automatizada de conteúdo classificava os perfis de *drag queens* participantes do *reality show RuPaul's Drag Race* como, em média, mais "tóxicos" do que perfis ligados a supremacistas brancos. A razão foi a alta presença de termos originalmente LGBTfóbicos, mas que foram ressignificados pela comunidade LGBT, como *bitch* ("vadia") e *fag* ("bicha"). Esse exemplo mostra a dificuldade de utilizar indiscriminadamente a inteligência artifical para moderação de conteúdo, dada a dificuldade dos mecanismos de *machine-learning* de entender o contexto social em que a referida frase foi proferida (Gomes; Antonialli; Oliva, 2019)

Fonte: Elaboração Própria

Em geral, os mecanismos de inteligência artificial funcionam como um filtro prévio para a moderação de conteúdo. Logo após o *upload*, a ferramenta analisa o conteúdo, tentando identificar possíveis inadequações. Caso o algoritmo não consiga detectá-los, o conteúdo fica disponível, podendo ser denunciado por usuários. Caso o algoritmo não consiga dar um parecer conclusivo, o conteúdo é enviado para o parecer de um moderador humano. A revisão por moderadores humanos também acontece quando há denúncia por usuário ou, ainda, quando o sistema informatizado considera o conteúdo inapropriado e há recurso por parte do usuário. A decisão do moderador humano é ainda utilizada para fortalecer o sistema informatizado, ao servir como *input* no processo de aprendizagem da máquina (OFCOM, 2019)

Figura 1 – Fluxo de Moderação de Conteúdo



Fonte: OFCOM, 2019, trad. livre

Esse fluxo de trabalho mostra como é importante a atividade humana no processo de moderação de conteúdo, já que a interpretação contextual é dada principalmente por humanos. O trabalho dos moderadores é ainda essencial para criar os bancos de dados que serão utilizados para treinar o algoritmo de detecção automatizada de conteúdo inapropriado (Binns *et al.*, 2017). Esse trabalho de categorização de conteúdo em bancos de dados será chamado, no presente artigo, de "moderação de primeiro nível". Já a moderação de "segundo nível" será aquela que tem por objetivo revisar o parecer dado por um algoritmo de detecção automatizada de conteúdo inapropriado. Por fim, há ainda os times de Trust & Safety, responsáveis pela criação de diretrizes para as comunidades.

Os recursos humanos para a moderação de conteúdo podem ser divididos em três grandes categorias. Na primeira delas, a moderação artesanal, os times são menores e o conteúdo é moderado em uma escala que permite a discussão entre membros da equipe acerca de sua adequação ou não às diretrizes da comunidade. No segundo modelo, o baseado em comunidade, a moderação é feita basicamente por usuários da plataforma que se voluntariam para tal, e há um esquema complexo de hierarquização de usuários em termos de suas prerrogativas de revisão de conteúdo. Por fim, o modelo mais adaptado à dinâmica ágil para moderação de grandes volumes de conteúdo no mundo digital, e por isso mesmo mais utilizado por grandes plataformas, como Google e Meta, é a moderação industrial. Esse modelo é caracterizado pela grande escala e número de usuários das redes que o utilizam, pela operacionalização de regras e pela separação de times de *policy* e *enforcement* nas companhias. Conceitos complexos, como "assédio" e "discurso de ódio" são operacionalizados, de forma a permitir aplicação de regras de forma mais ou menos consistente por times de moderação de conteúdo. Times de operação estão distribuídos pelo mundo, encarregados de realizar a revisão de conteúdos a partir de diretrizes centrais. Tipicamente, as equipes possuem poucos segundos para tomar decisões sobre a adequação ou não do conteúdo e há metas claras quanto ao volume que deve ser moderado. Além disso, são utilizadas técnicas de remoção automatizada por algoritmos (Caplan, 2018).

A "industrialização" do mercado de moderação de conteúdo gerou também sua globalização, com a terceirização de sua realização para empresas baseadas em países do Sul Global, como Filipinas e Índia (Roberts, 2019). O mercado teve um valor estimado de 5,3 bilhões de dólares em 2020 e tem uma estimativa de crescimento de 12,6% até 2026 (Business Wire, 2021).

A globalização do mercado de moderação de conteúdo cria duas grandes preocupações. A primeira delas é com as condições de trabalho, dado que a exportação desse tipo de serviço para países com menor regulamentação das relações trabalhistas gerou um cenário onde trabalhadores de países subdesenvolvidos enfrentam jornadas exaustivas de trabalho e um ambiente insalubre (Roberts, 2019).

Quadro 4 – Notícias sobre moderação de conteúdo

Reportagem da revista Time relatou que um escritório da empresa Sama, em Nairobi (Quênia) era responsável por atividades de moderação de conteúdo em todos os 48 países da África Subsaariana. Enquanto a empresa tem como lema ser uma operação de "Inteligência Artificial Ética", os relatos de ex-moderadores demonstram outro cenário: profissionais que se inscrevem para vagas de *call center* e só descobrem a natureza do trabalho após assinarem um Acordo de Confidencialidade, com alta pressão para realização de decisões em poucos segundos e com a dissolução e demissão de trabalhadores que levantam iniciativas de sindicalização (Perrigo, 2022). Importa ressaltar, ainda, que o continente africano possui uma grande diversidade cultural, política e linguística, tendo entre 1500 e 2000 línguas faladas. Todos esses são fatores de interesse para se pensar a atividade de moderação de conteúdo concentrada em uma única área geográfica. A atividade de moderação de conteúdo dispõe, fundamentalmente, sobre visibilidade de informação - decisões importantes sobre o que pode ou não circular na esfera pública e que possuem efeitos retroalimentam a relação entre real e digital. Uma questão de importância é compreender como lidar com os desafios de compreensão do contexto político-cultural para uma "justa moderação" em cenários de línguas e países sub representados na composição das equipes de moderação.

Fonte: Elaboração própria

Há, também, a preocupação com a própria qualidade da atividade desenvolvida. A moderação tem conteúdo tem impactos nítidos sobre a liberdade de expressão e jornalística, e certas políticas podem restringir seriamente a circulação de certos tipos de conteúdo (Gillespie, 2018, cap. 1). A falta de compreensão acerca do contexto em que determinado conteúdo foi produzido pode gerar a proliferação de conteúdo impróprio ou, ainda, a restrição excessiva de discursos legítimos. Essa dificuldade de compreender corretamente o conteúdo a ser moderado pode ser potencializada em um contexto de globalização do serviço de moderação, dada a possibilidade de moderação de conteúdo de um determinado país em outro com tradições e noções completamente distintas.

David Kaye (2019), antigo Relator Especial das Nações Unidas para a promoção e proteção do direito à liberdade de expressão e opinião, ressalta ainda a importância de descentralizar não só a atividade de moderação de conteúdo, mas também as próprias diretrizes das comunidades, integrando a sociedade civil e ativistas locais aos processos decisórios para definição de quais ações violam as regras da plataforma em questão. Assim, seria possível mitigar um certo déficit democrático gerado pela concentração da

comunicação via *internet* em algumas poucas redes sociais gerenciadas por grandes corporações (Kaye, 2019).

Outro grande risco da moderação de conteúdo é a reprodução de assimetrias de raça, gênero, sexualidades e outras opressões sociais, seja pela ausência de variedade dentro das equipes que realizam moderação, seja pelas próprias regras de Trust and Safety. A rede social Grindr (Hunsberger *et al.*, 2021), por exemplo, publicou um *white paper* em que discutia a necessidade de criar políticas de moderação de conteúdo que fossem inclusivas do ponto de vista de identidade de gênero, sem criar fardos excessivos para pessoas trans e não-binárias, por exemplo.

Quadro 5 – Notícias e retratos da realidade do mercado de moderação de conteúdo

Política, cultura e interpretação: um retrato do racional de decisão possível de uma profissional de moderação de conteúdos

O documentário "The Cleaners", de 2018, mostra a realidade da força de trabalho de moderação de conteúdo nas Filipinas. O retrato é complexo: o documentário demonstra desafios práticos da atividade de moderação, bem como a condição dos trabalhadores, sujeitos à exposição de imagens violentas diariamente e com poucos segundos para realizarem decisões complexas. Em algumas cenas notáveis, o documentário pede que os trabalhadores expliquem o racional que realizariam na moderação de conteúdos que causaram polêmica na mídia ao serem removidos de grandes plataformas, como uma pintura que mostra Donald Trump nu. Abaixo, transcrita a fala:

> Esse é Donald Trump nu. Ele não é um líder forte o bastante para liderar e é por isso que seu pênis foi desenhado pequeno. Ele não é másculo o bastante para lidar a grande tarefa de ser Presidente dos Estados Unidos. Seria deletar. Por quê? Degrada a personalidade de Donald Trump, então deve ser deletado. (The Cleaners, 18'30''-20'12'')

Ilma Gore, autora da pintura, também entrevistada, revelou que sua intenção era questionar padrões de gênero e de poder. Após a viralização da imagem, Ilma teve sua conta banida e o conteúdo removido de circulação. Aqui, importa uma reflexão importante: como a análise de julgamento de um profissional moderador de conteúdo pode ser afetada por sua localização geográfica? Como as percepções internas sobre direitos fundamentais e o que constituem os limites de liberdade de expressão de trabalhadores que crescem e vivenciam diferentes regimes políticos pode afetar seu campo de decisão sobre circulação de conteúdo em nível global? Neste caso, o tempo para decisão é um fator importante, que pode afetar o julgamento ou capacidade de pesquisa para classificação e compreensão do contexto sócio-cultural e político de circulação de conteúdos.

Fonte: Elaboração própria

O presente artigo tem por objetivo fazer um mapeamento do mercado de moderação de conteúdo, realizando um panorama do mercado global e uma análise mais aprofundada do mercado em língua portuguesa.

O texto está dividido em três seções, além desta introdução e de uma conclusão. Na primeira, será descrita a metodologia do trabalho, que envolveu a coleta de informações

sobre o mercado de moderação de conteúdo de fontes diversas, que foram posteriormente classificadas e tratadas. Em segundo lugar, far-se-á um panorama geral do mercado global de moderação de conteúdo, como a localização e os agentes que realizam o serviço. Na terceira seção, o foco será na moderação em português, cobrindo quatro grandes temas: a) localização dos pontos de moderação de conteúdo; b) agentes que realizam moderação de conteúdo em português; c) condições de trabalho; e d) a importância do contexto e das habilidades linguísticas no mercado. Ao longo do texto, é possível encontrar caixas de reflexão que retratam desafios da moderação de conteúdo a partir de casos reais retratados em reportagens e documentários sobre o tema.

1 Metodologia

Há uma dificuldade em localizar o mercado de terceirizadas que realizam prestação de serviços de moderação de conteúdo para grandes plataformas. Esse não é um dado facilmente obtido pelas Centrais de Transparência das plataformas. Além disso, análises do setor realizadas por grandes empresas de consultoria de mercado possuem preço elevado, o que se mostrou um primeiro empecilho para a obtenção dos dados.

O artigo foi construída a partir de três grandes fontes: (i) vagas de trabalho em moderação de conteúdo postadas em sites especializados; (ii) procura por perfis de moderadores de conteúdo na plataforma LinkedIn; e (iii) consulta a reportagens investigativas acerca do mercado de moderação de conteúdo.

As vagas de emprego foram procuradas em duas etapas, realizadas entre os dias 03 de março de 2022 e 29 de março de 2022. Na primeira etapa, foram procuradas vagas de “content moderator” ou de “content moderation” no LinkedIn, Google e nos sites oficiais de empresas consideradas grandes players no mercado de moderação de conteúdo.[5] As vagas encontradas foram fichadas a partir de um código único que as identificavam, sendo extraídos o nome da vaga, a posição presumida dentro de uma escala de moderação de conteúdo, a língua em que haveria a moderação, a empregadora e o local onde seria

[5] A identificação de grandes players de mercado de moderação de conteúdo foi feita através de dados obtidos pela amostra gratuita do relatório “Trust and Safety – Content Moderation Services PEAK Matrix® Assessment 2021”, no qual foi possível identificar 18 empresas divididas entre “Líderes”, “Principais Concorrentes” e “Aspirantes”. A investigação de vagas de trabalho disponíveis nos websites das empresas identificadas auxiliou no refinamento dos dados obtidos.

prestado o trabalho. Nessa fase, foram encontradas 342 vagas distribuídas em 83 cidades de 45 países distintos em todos os continentes, totalizando 83 línguas distintas moderadas.

Após a coleta das vagas, foi construído um índice, denominado de “línguas-empresa” para medir a importância de determinada localidade (cidade ou país) para o mercado global de moderação de conteúdo. O índice é calculado a partir da contagem das linhas sem repetição de uma tabela que mostra as empresas presentes na área e as línguas em que moderam. Assim, se há, por exemplo, duas empresas atuantes na área e duas línguas moderadas, o índice poderá assumir o valor de 4 (se ambas as empresas moderarem as duas línguas encontradas), 3 (se uma das empresas moderar em ambas as línguas e a outra moderar em apenas uma delas) ou mesmo 2 (se cada uma das empresas moderar em apenas uma das línguas distintas).

O índice permite detectar tanto localidades onde há um grande hub multilingue de uma única empresa quanto locais que atraem empresas distintas especializadas, o que permite uma detecção mais ampla de locais de interesse para o mercado global de moderação de conteúdo. Contudo, não é possível inferir a partir do índice qual o tamanho, em volume de conteúdo moderado ou até mesmo em termos de números de funcionários, do polo em questão, ainda que se suponha que haja uma correlação considerável entre o número de empresas atuantes/línguas moderadas e as dimensões dos hubs. Essa fragilidade foi diminuída com o uso das notícias coletadas na fase (iii), que por vezes traziam as informações almejadas.

A segunda parte da coleta de dados consistiu em uma busca mais ampla por vagas de moderação de conteúdo em português. Nesse ponto, foram buscadas palavras-chave em inglês, português, francês e espanhol por vagas de moderação para falantes de português. Nessa fase, encontrou-se 86 vagas em 22 países.

Por ser uma busca mais ampla, que não foi repetida para as outras 82 línguas encontradas na primeira fase, optou-se por não adicionar essas vagas ao banco de dados usado para calcular o índice “línguas-empresa”, dada a possibilidade de existência de vieses que distorceriam os resultados encontrados.

As vagas em português foram fichadas para uma série de categorias distintas. Para essas vagas, além do local e da empresa que prestava os serviços, buscou-se outras informações relativas às condições de trabalho, requisitos e descrição das tarefas desempenhadas. As categorias foram testadas para a sua confiabilidade inter-codificadores.

Durante a segunda fase, procurou-se no LinkedIn por perfis de moderação de conteúdo que realizavam moderação em português. O objetivo dessa fase foi duplo: em primeiro lugar, buscou-se detectar empresas que realizavam moderação de conteúdo em português mas que, pela limitação da procura por vagas de emprego, não possuíam vagas anunciadas; e (ii) coletar informações que não estão disponíveis nas vagas de emprego, como relações comerciais de outsourcing e os mercados alcançados. Nessa fase, que durou até que houvesse a saturação nas informações encontradas, foram coletados 31 perfis.

Por fim, realizou-se uma coleta de reportagens e notícias sobre o mercado de moderação de conteúdo, fichando-as para as categorias descritas em anexo. O objetivo dessa fase foi coletar materiais que não estavam disponíveis a partir das vagas e dos perfis de LinkedIn, como as dinâmicas mais específicas do mercado de moderação.

Essa pesquisa possui algumas limitações. A primeira diz respeito à identificação dos locais onde há moderação de conteúdo e das línguas em que ela se desenvolve. A busca por vagas de emprego em um curto período de coleta acaba por não permitir a detecção de vagas que, por motivos sazonais, não foram anunciadas durante as semanas em que a coleta se deu. Isso porque, via de regra, as vagas são apagadas dos sites de busca após o fim do período de inscrição, e as vagas coletadas se referem, em larga escala, a vagas que se encontravam abertas. Também não é possível detectar, por meio da metodologia apresentada, vagas não disponíveis na internet e em uma das línguas de leitura da equipe de pesquisadores (português, inglês, francês, espanhol e alemão). Em terceiro lugar, as reportagens coletadas demonstram que nem sempre as empregadoras anunciam vagas de moderação de conteúdo como tal, sendo comum apresentá-las como vagas de call center, o que dificulta sua mensuração.

Por fim, a metodologia para detecção de hubs não consegue apreender o tamanho da força de trabalho no local, o que acaba por gerar falsos negativos (locais de importância não são detectados como tais), embora falsos positivos sejam provavelmente raros.

2 Análise das vagas do mercado global de moderação de conteúdo

A análise de vagas mundiais encontrou, conforme já reportado, 342 propostas em 83 cidades de 45 países nas cinco regiões. Em termos percentuais, as vagas se concentraram largamente na Europa (54,6% das ocorrências) e na Ásia (27,15% das ocorrências) e, em

menor escala, nas Américas (12%) e na África (5,5%). A Oceania teve presença residual, com apenas uma vaga encontrada na Austrália.

Do ponto de vista dos países, a média de vagas encontradas por país foi de 7,6 vagas, enquanto a mediana encontrada foi de apenas duas, sugerindo a existência de números extremos dentro da amostra, o que pode ser consequência da existência de uma certa concentração do mercado global. Em termos quantitativos, foram valores discrepantes na amostra[6] (outliers), nessa ordem, a Irlanda (40 vagas), Portugal (37 vagas), Alemanha (30 vagas), Índia (28 vagas), Malásia (28 vagas), Estados Unidos (24 vagas) e Espanha (22 vagas). Os resultados permanecem quando é utilizado o índice de "línguas-empresa" para os países.

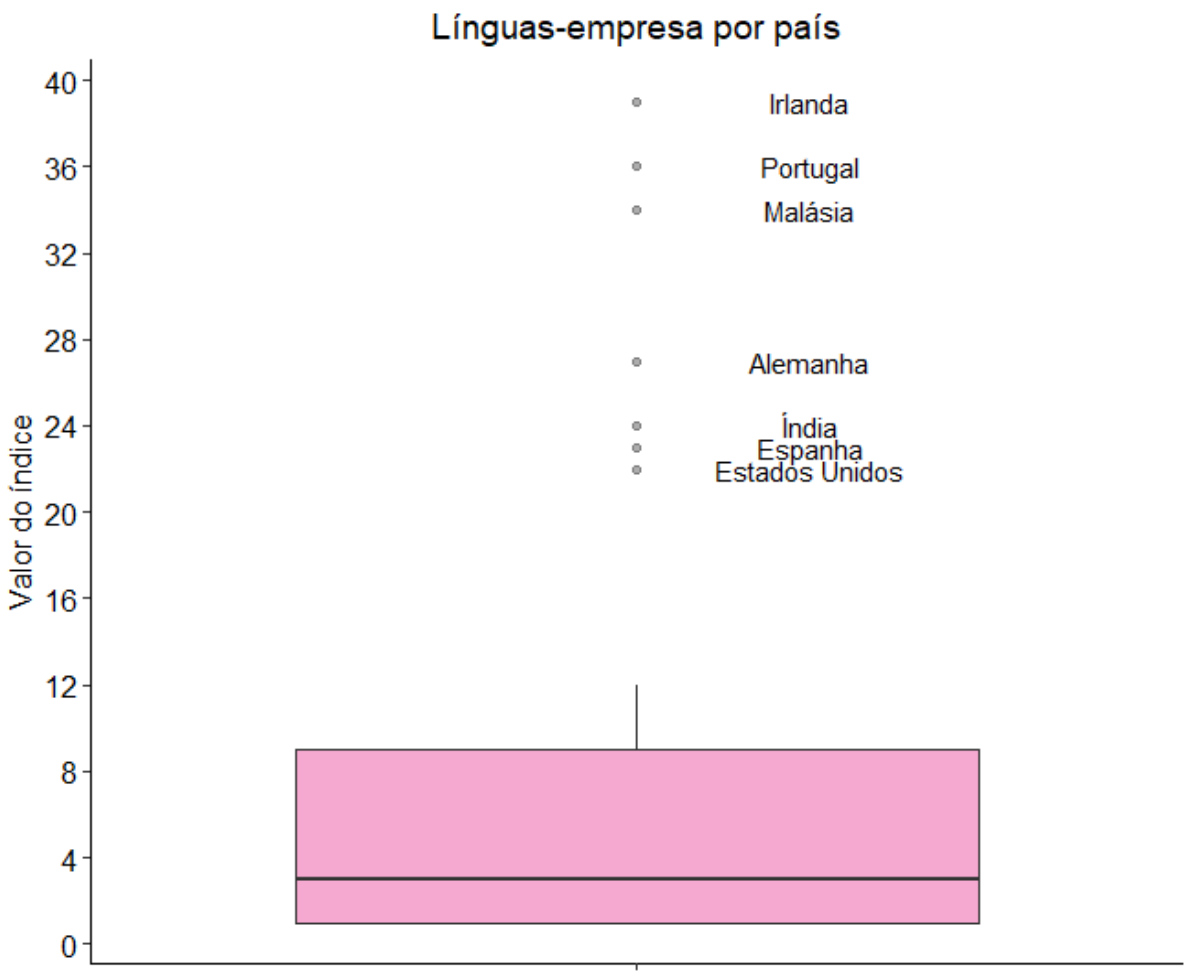
Em certa medida, a distribuição encontrada reflete duas tendências dentro do mercado mundial de moderação de conteúdo. Em primeiro lugar, há uma importância considerável de países europeus, que pode ser causada tanto pelo fácil trânsito interno pelo continente, permitindo a contratação de um volume considerável de imigrantes com conhecimentos linguísticos variados, ou pela própria regulamentação europeia, que tende a ser mais rígida em situações que envolvem a moderação de conteúdo.

Em segundo lugar, há uma grande presença de países asiáticos no mercado, em especial, nas vagas encontradas, da Índia e da Malásia. A ocorrência massiva da Ásia na amostra representa não só a importância do continente mais populoso do mundo em termos de tráfego digital como também o barateamento progressivo da força de trabalho asiática, gerando a migração de postos de trabalho para esses países e a consolidação de um modelo de outsourcing que faz com que a Ásia seja um local privilegiado para o setor de Business Process Outsourcing (BPO).

Outro país asiático que a literatura aponta como importante para o mercado mundial de moderação de conteúdo (Roberts, 2019) é as Filipinas que, embora não tenham pontuado o suficiente no índice de "línguas-empresa" para ser considerado como um valor discrepante, pertence ao terceiro quartil da distribuição, com ao menos sete empresas e seis línguas moderadas em seu território.

Boxplot 1 – Línguas-empresa por país

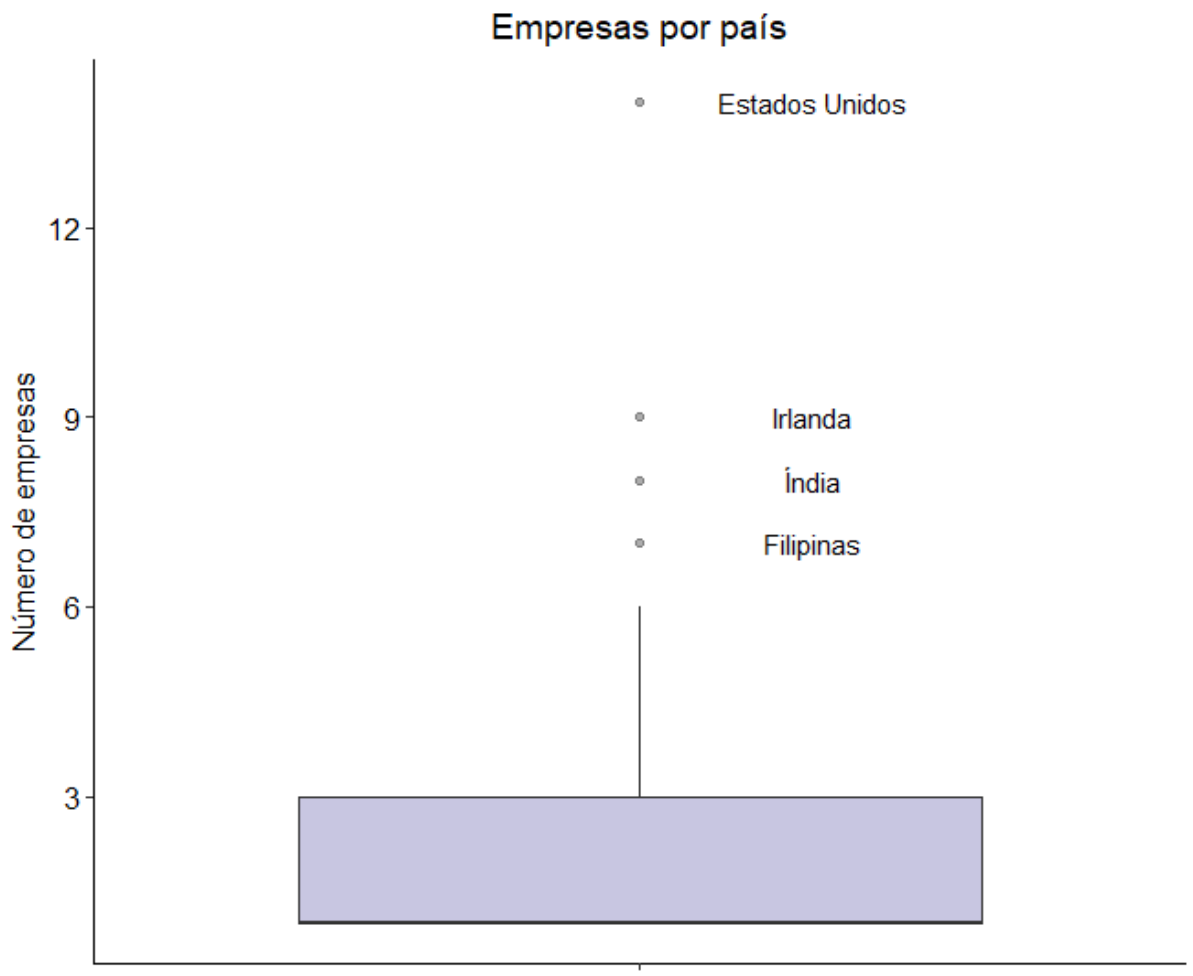
[6] Para a identificação de valores discrepantes superiores, foi usado o critério informal de considerar aqueles maiores que uma vez e meia o intervalo interquartil, contado a partir do valor extremo superior do terceiro quartil. Os valores discrepantes inferiores são aqueles menores que uma vez e meia o intervalo interquartil, contado a partir do valor extremo superior do primeiro quartil.



Fonte: Elaboração própria

Quando se analisa o número de empresas em seu território, contudo, as Filipinas passam a ter um local privilegiado. A média de empresas/país na amostra foi de cerca de 2,5, enquanto a mediana foi de apenas uma. O país é o quarto em número de empresas presentes no território e, junto com Estados Unidos (14 empresas), Irlanda (9 empresas) e Índia (8 empresas), é um valor discrepante da distribuição amostral. Diferente desses outros países, todavia, as empresas presentes nas Filipinas eram exclusivamente terceirizadas, o que sugere ser o país um foco importante no processo global de terceirização da moderação de conteúdo.

Boxplot 2 – Empresas por país

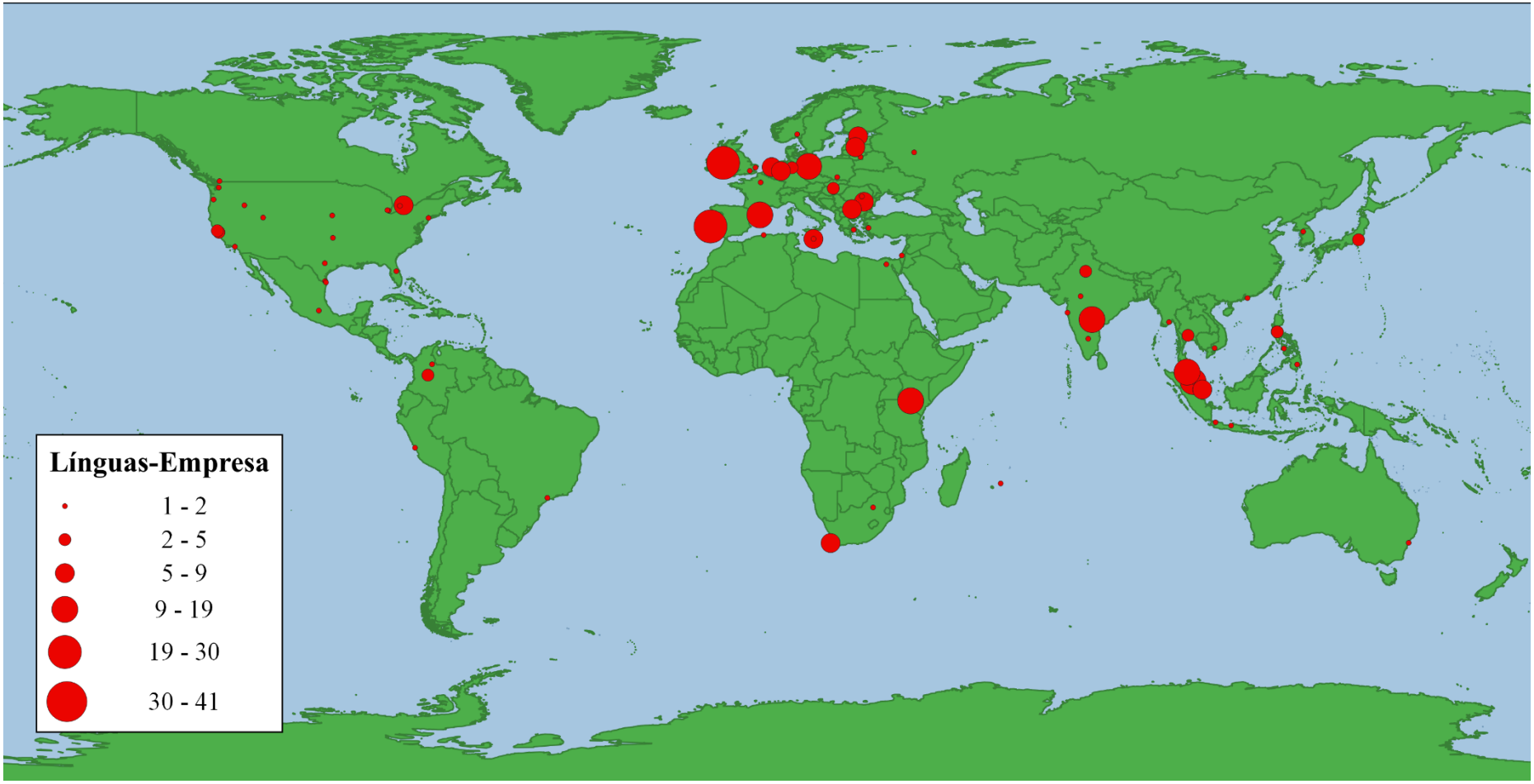


Fonte: Elaboração própria

A pontuação filipina pode ainda ter sido abaixada por uma especificidade do país no mercado de moderação. Uma reportagem do Washington Post mostrou que, diferente de moderadores na Índia ou nos Estados Unidos, a moderação de conteúdo filipina é realizada em postagens de todo o globo e em línguas que são, por vezes, desconhecidas do moderador (Dwoskin; Whalen; Cabato, 2019). A ausência de exigência de conhecimento na língua em que se moderará pode ser responsável por uma diminuição artificial de um dos indicadores que compõem o índice de "línguas-empresa", o de línguas moderadas, o que causa, inevitavelmente, um valor menor no momento da mensuração do índice em questão.

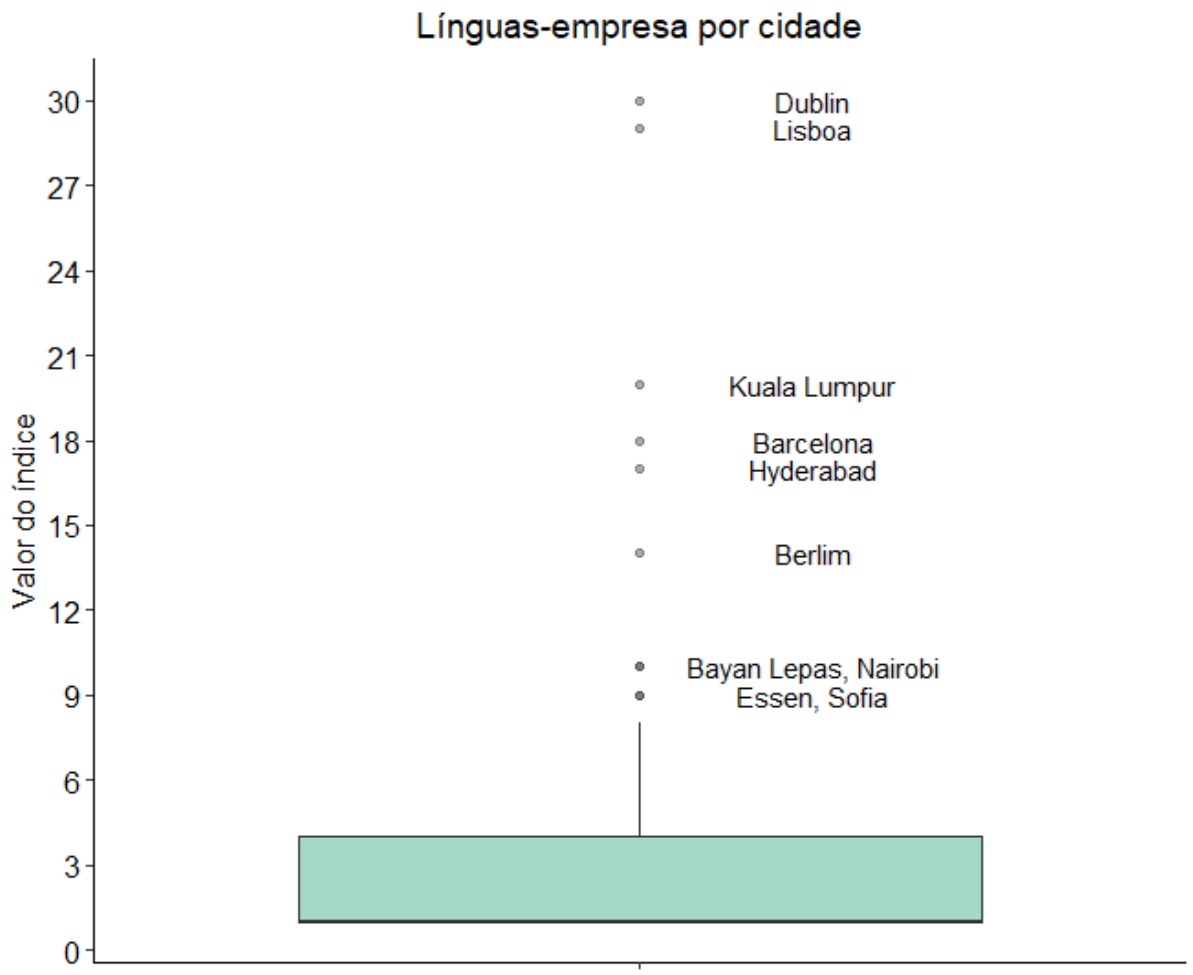
A análise de cidades mostrou uma tendência semelhante à análise dos países. A média de línguas moderadas por cidade é de 3,5 com mediana de uma, o que sugere também a presença de valores exacerbados na amostra. No índice de línguas-empresa, o resultado foi semelhante, com uma amplitude que vai de um a 30, média de 3,9 e mediana também de uma.

Mapa 1 – Localização das cidades onde há moderação de conteúdo, por línguas-empresa



Fonte: Elaboração própria

Boxplot 3 – Línguas-empresa por cidade



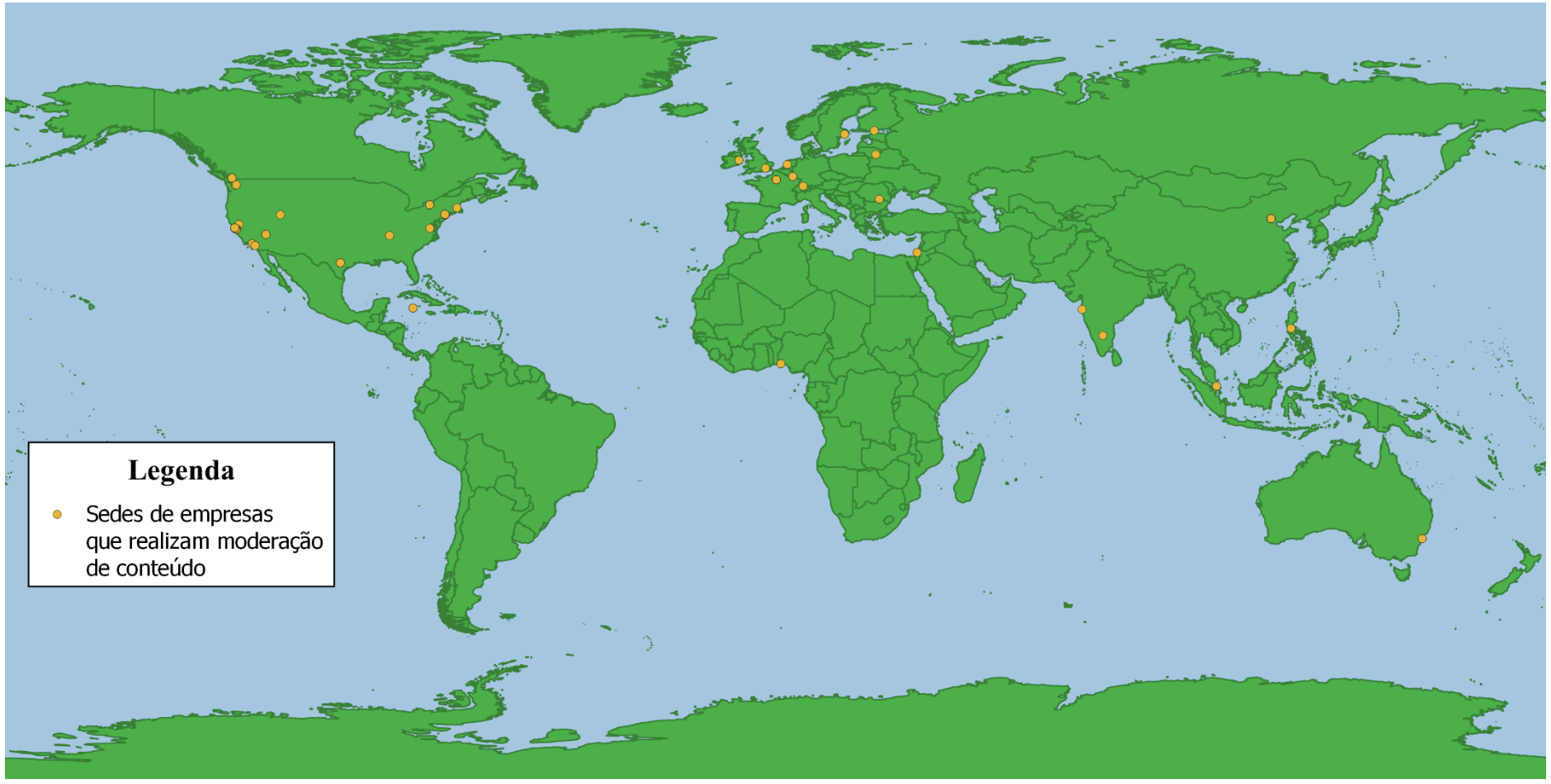
Fonte: Elaboração própria

Foram encontrados os seguintes valores discrepantes no índice de "línguas-empresa": na Irlanda, Dublin (30), na Malásia, Kuala Lumpur (20) e Bayan Lepas (10), em Portugal, Lisboa (29), na Espanha, Barcelona (18), na Índia, Hyderabad (17), na Alemanha, Berlim (14) e Essen (9), no Quênia, Nairobi (10) e, na Bulgária, Sofia (9). Um mapa completo dos pontos de moderação de conteúdo, com indicação de sua pontuação na escala de línguas-empresa, pode ser encontrado acima.

A ausência de cidades americanas como hubs, ainda que este país seja considerado importante no mercado, pode ser explicado pelo enorme espalhamento entre distintas cidades no país. Em média, os países da amostra possuem 1,88 cidade em que há

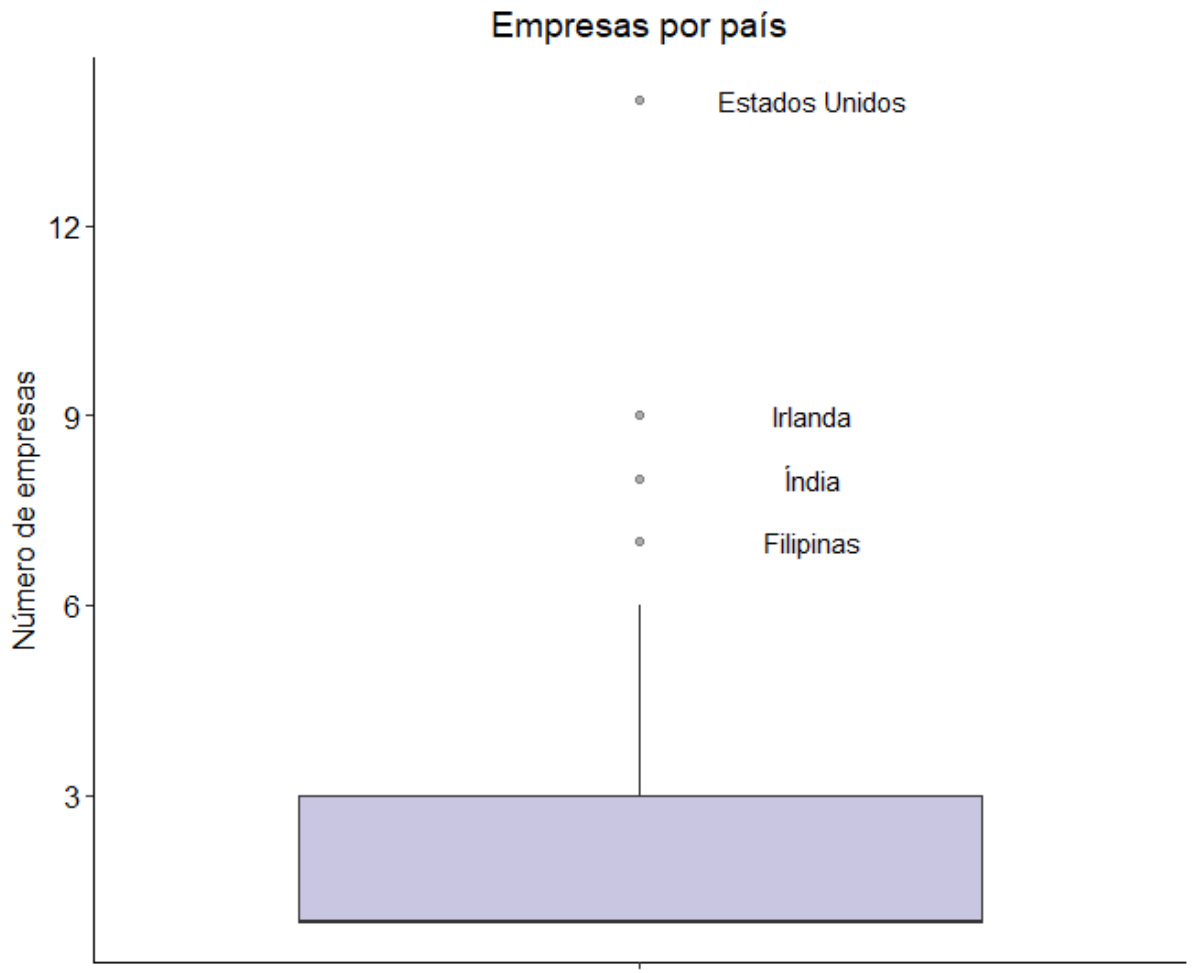
moderação de conteúdo, com a mediana de uma cidade. Já os Estados Unidos possuem ao menos 18 cidades em que há moderação de conteúdo, 13 a mais do que os segundos lugares da amostra, Índia e Filipinas. Esse espalhamento parece ser consequência da centralidade americana na economia digital, o que leva ao surgimento de uma série de empresas em seu território.

Mapa 2 – Sedes das empresas (plataformas e terceirizadas) que realizam moderação de conteúdo



Fonte: Elaboração própria

Boxplot 4 – Empresas por país



Fonte: Elaboração própria

A análise das sedes das empresas que moderam conteúdo, contudo, revela uma conformação totalmente distinta, como pode se ver no mapa 2. Os Estados Unidos possuem o maior número de sedes, com 24, seguido pela China e Cingapura, com três sedes cada. A conformação revela uma concentração em países desenvolvidos, o que confirma a literatura sobre o tema, que aponta para a existência de um mercado composto por grandes conglomerados empresariais que descentralizam o trabalho para o Terceiro Mundo.

Por fim, a análise das línguas moderadas em cada *hub* revela a existência de heterogeneidade entre eles. Embora todos sejam multilíngues, é possível notar que línguas românicas e germânicas, com exceção do inglês, são moderadas, em larga escala, nos *hubs* europeus, enquanto Kuala Lumpur e Bayan Lepas são responsáveis por moderar línguas comuns no Sudeste Asiático e Oceania. Já Nairobi tem vagas especialmente para línguas

africanas e Hyderabad, para línguas faladas especialmente no subcontinente indiano. A exceção está justamente na língua inglesa, que parece ter moderação um pouco mais universal. Essa conformação condiz com o fato de a moderação de conteúdo profissional depender da existência de um exército de reserva de falantes das línguas moderadas, em especial imigrantes, o que justifica a criação de hubs regionais que moderam conteúdo em línguas comuns na região em questão.

Tabela 1 - Relação de Cidades-Hubs e Línguas Moderadas

Hub	Línguas moderadas
Dublin	Alemão, Árabe, Dinamarquês, Espanhol, Espanhol Mexicano, Finlandês, Flamengo, Francês, Hebraico, Indonésio, Inglês, Italiano, Neerlandês, Norueguês, Polonês, Russo, Sueco, Tcheco, Turco, Ucraniano, Vietnamês
Lisboa	Alemão, Árabe, Coreano, Dinamarquês, Espanhol, Finlandês, Francês, Grego, Hebraico, Hindi, Indonésio, Inglês, Italiano, Japonês, Neerlandês, Norueguês, Polonês, Russo, Sueco, Turco, Ucraniano
Kuala Lumpur	Bengali, Burmês, Cambojano, Cantonês, Chinês, Chinês Tradicional, Coreano, Espanhol, Francês, Indonésio, Inglês, Japonês, Malásio, Maori, Tailandês, Tagalog, Vietnamês
Barcelona	Alemão, Dinamarquês, Espanhol, Espanhol Boliviano, Espanhol Colombiano, Espanhol Equatoriano, Espanhol Mexicano, Espanhol Peruano, Finlandês, Francês, Hebraico, Italiano, Neerlandês, Norueguês, Português, Sueco
Hyderabad	Árabe, Assamês, Bengali, Canarês, Cingalês, Hindi, Inglês, Malaialo, Marati, Marvari, Nepalês, Oriá, Pashto, Punjabi, Tamil, Urdu
Berlim	Alemão, Búlgaro, Dinamarquês, Esloveno, Finlandês, Grego, Hebraico, Húngaro, Neerlandês, Norueguês, Sueco, Tcheco, Ucraniano
Bayan Lepas	Burmês, Chinês, Cingalês, Coreano, Hindi, Indonésio, Japonês, Malásio, Tailandês, Vietnamês
Nairobi	Africâner, Amárico, Hausa, Kinyarwanda, Kirundi, Luganda, Oromo, Somali, Swahili, Zulu
Sofia	Armênio, Finlandês, Francês, Húngaro, Inglês, Italiano, Norueguês, Sueco, Turco
Essen	Albanês, Alemão, Árabe, Búlgaro, Curdo, Farsi, Inglês, Romeno, Turco

Fonte: Elaboração própria

Em suma, a pesquisa de vagas permitiu confirmar a prática comercial de descentralizar a moderação de conteúdo para países do Terceiro Mundo e, mesmo em países europeus, naqueles em que a mão-de-obra é mais barata e as taxas de desemprego maiores, como países do Leste Europeu e da Península Ibérica. Países como Irlanda e Estados Unidos são importantes por sua posição no mercado digital e a Alemanha é

relevante, muito provavelmente, por sua legislação rígida, que exige respostas rápidas das plataformas.[7]

O mercado é concentrado em algumas cidades, que moderam bem mais do que a média, seja pela presença de diversas empresas, seja pela presença de equipes multilíngues. Essas cidades, contudo, não são homogêneas, e se especializam, em geral, para a moderação em inglês e outras línguas de interesse regional.

Quadro 6 – Considerações sobre dialetos e linguagem

O espanhol foi uma língua bastante recorrente nos anúncios de vagas: ao menos treze grandes empresas moderam em espanhol. Contudo, somente em duas delas foram encontradas vagas diferenciadas para registros específicos da língua - como espanhol mexicano, peruano ou boliviano. Tendo em vista que a língua espanhola é a língua oficial de dezoito países latinoamericanos, do Estado Livre Associado de Porto Rico, da Espanha e da Guiné Equatorial, tendo também um número considerável de falantes em países como os Estados Unidos (Posner; Sara, 2022), é importante pensar se somente o rótulo de "Espanhol" é suficiente para compreender em profundidade todos esses registros. O uso da língua pode variar bastante dependendo do contexto, e é importante reconhecer a existência de diferentes formas de se falar uma língua tão difundida quanto a língua espanhola. Uma consideração de interesse é o "lunfardo", dialeto utilizado na cidade de Buenos Aires desde o final do Século XIX, que mistura expressões de diversas origens e realiza jogos de palavras e inversão de sílabas. Uma tradução e compreensão concreta dos termos utilizados dificilmente será realizada em ferramentas de tradução, sendo um dialeto que se desenvolveu a partir de vivências culturais específicas da cidade.

Fonte: Elaboração própria

3 Análise das Vagas de Moderação de Conteúdo em Língua Portuguesa

A análise de vagas em língua portuguesa encontrou 86 ofertas localizadas em 39 cidades de 22 países. Os locais de trabalho se distribuíram pelas Américas, Ásia e Europa, não tendo sido encontradas ocorrências no continente africano e na Oceania.

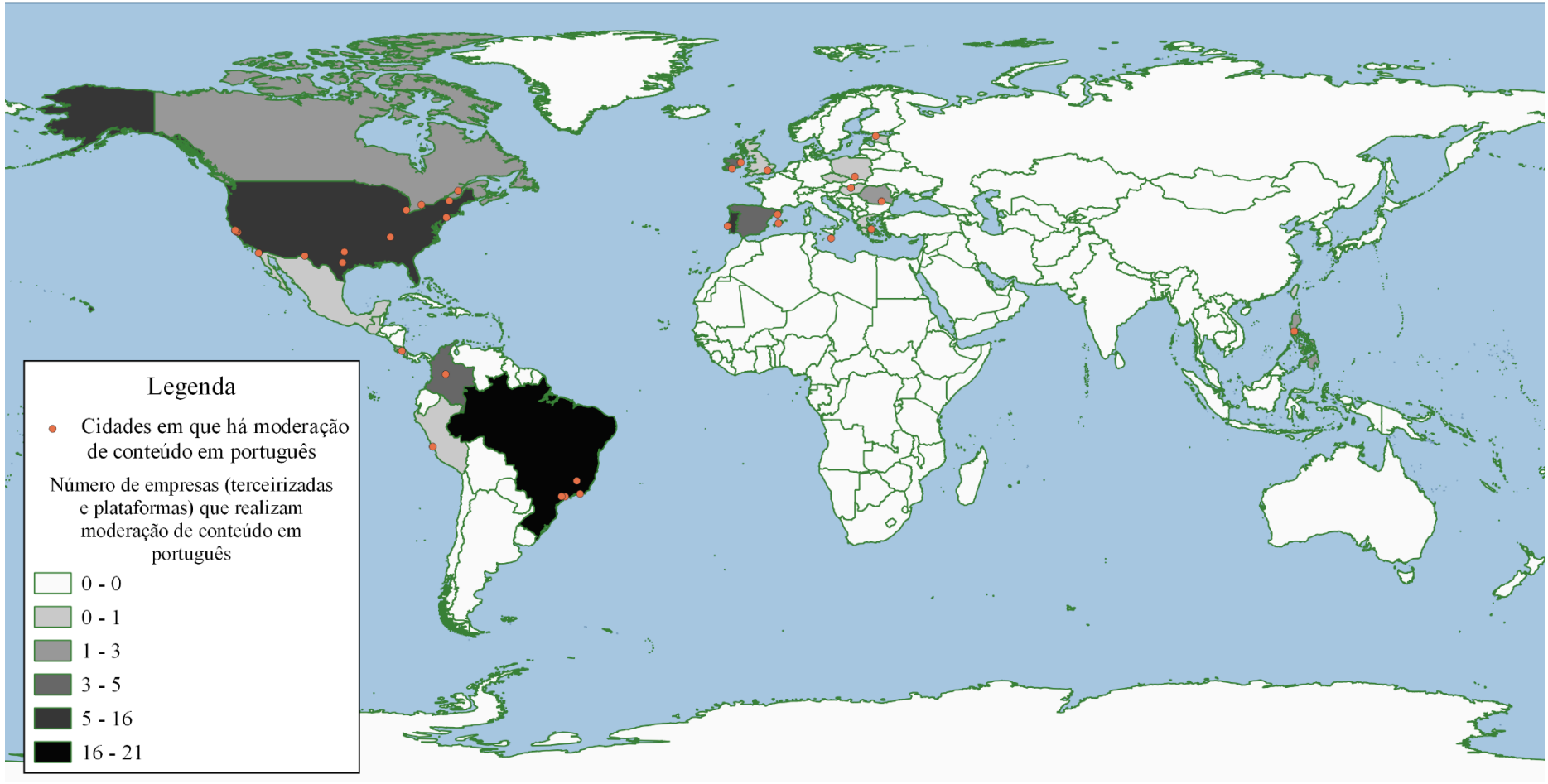
A distribuição de vagas em língua portuguesa seguiu um padrão bastante distinto do encontrado para o mercado de moderação em geral, como é possível perceber nos mapas 3 a 5. Diferente do protagonismo asiático encontrado na primeira fase, no caso da moderação em língua portuguesa, nove em cada dez vagas se encontram nas Américas (50% das ocorrências) e na Europa (40% das ocorrências). Ainda que distinta, essa conformação parece condizer com a conclusão, enunciada na seção anterior, de que os *hubs* de moderação de conteúdo acabam tendo também um aspecto regionalizado. Essa

[7] Em 2016, a Alemanha introduziu o Netzwerkdurchsetzungsgesetz, também conhecido como Network Enforcement Act ou NetzDG. A legislação impõe o prazo de 24h a mídias digitais para remoção ou bloqueio de conteúdos que contenham discurso de ódio, propaganda terrorista e outros tipos de conteúdos manifestamente ilegais. (Center for Democracy and Technology, 2017).

regionalização atrairia o mercado de moderação em português para as Américas, de forma a se aproximar do mercado brasileiro, o maior na língua em questão.

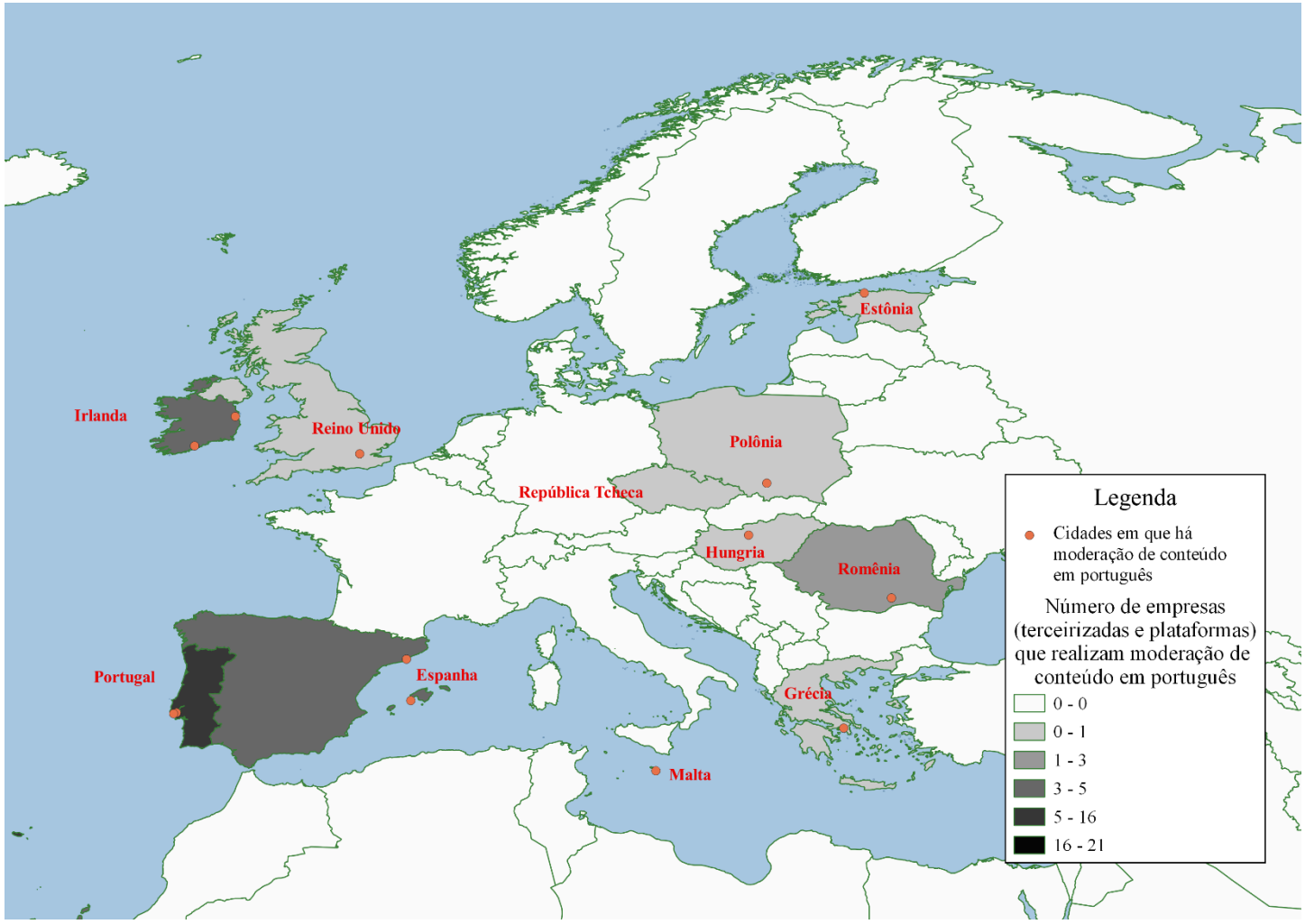
Dos 22 países encontrados em que há moderação de conteúdo em português, apenas dois, Portugal (11,6% da amostra) e Brasil (12,8% da amostra), possuem a língua como oficial. Esses dois Estados correspondem a cerca de 24,4% do mercado encontrado de moderação de conteúdo em português, valor idêntico ao encontrado apenas para os Estados Unidos (24,4%), país que teve o maior número de vagas na amostra. Após os três países, as nações com mais vagas encontradas foram a Espanha (7%) e Irlanda, Grécia e Colômbia (6% cada). A presença da Colômbia e Grécia à frente de países como as Filipinas demonstra como, em termos geográficos, o mercado de moderação de conteúdo em línguas específicas pode ser sensivelmente distinto daquele que é praticado ao redor do mundo.

Mapa 3 – Número de empresas (terceirizadas e plataformas) que realizam moderação de conteúdo em português - Mundo



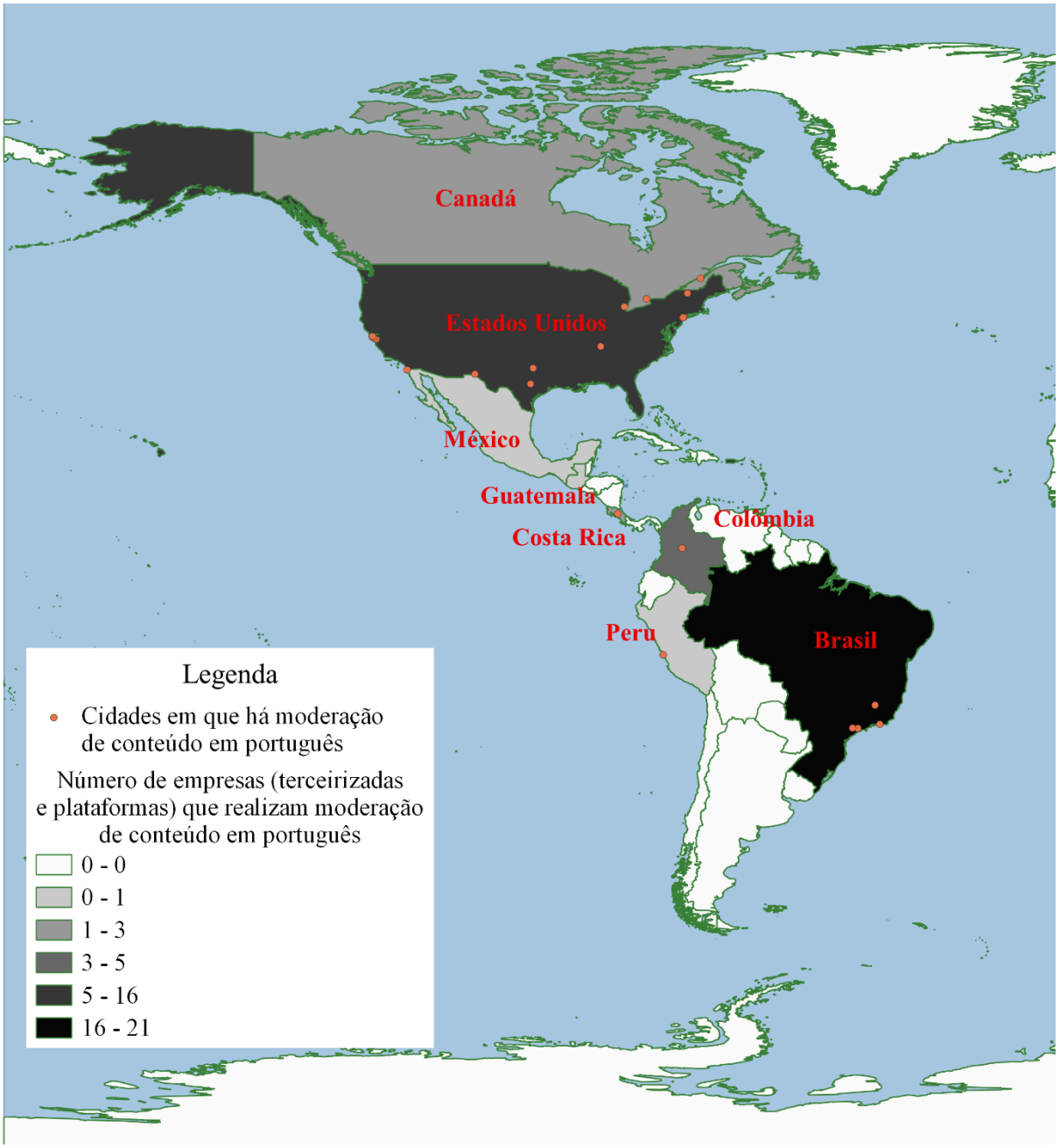
Fonte: Elaboração própria

Mapa 4 - Número de empresas (terceirizadas e plataformas) que realizam moderação de conteúdo em português – Europa



Fonte: Elaboração própria

Mapa 5 - Número de empresas (terceirizadas e plataformas) que realizam moderação de conteúdo em português – Américas



Fonte: Elaboração própria

A ausência de países africanos no mercado de moderação de conteúdo sugere uma distância considerável deste mercado em relação à preocupação com o contexto durante a atividade de moderação. Isso porque a lusofonia africana tem uma quantidade considerável de usuários de internet que, embora não ameacem a hegemonia brasileira no campo, ultrapassam, em muito, o número de usuários de Portugal. A tabela 2, que estima o número de usuários de internet na lusofonia a partir de dados do Banco Mundial, mostra que,

enquanto Portugal seja o segundo país em proporção de usuários de internet (atrás apenas da região chinesa de Macau), sua pequena população faz com que, em termos absolutos, seu número de usuários seja pequeno.

Tabela 2 – Países lusófonos e proporção de usuários de internet

País	Proporção de usuários de internet	População (em milhares)	Estimativa de usuários (em milhares)
Angola	0,36	32.866,27	11.831,86
Brasil	0,739	212.559,41	157.081,4
Cabo Verde	0,619	555,99	344,16
Guiné Equatorial	0,188	1.402,98	263,76
Guiné-Bissau	0,28	1.968,00	551,04
Macau	0,865	649,34	561,68
Moçambique	0,151	31.255,44	4.719,57
Portugal	0,783	10.305,56	8.069,25
São Tomé e Príncipe	0,32	219,16	70,13
Timor-Leste	0,275	1.318,44	362,57

Fonte: Elaboração própria com base nos dados do Banco Mundial (2022a; 2022b).

Em termos comparativos, Angola, onde pouco mais de um terço da população usa a internet, possui mais de três milhões de usuários a mais do que Portugal. Em conjunto, os seis países da lusofonia africana possuem 17,78 milhões de usuários de internet, mais do que o dobro de Portugal.

Portanto, partindo do pressuposto de que a moderação de conteúdo baseada em contexto exige um domínio profundo não só da língua em que se pretende moderar, mas também do registro em questão e da realidade social que cerca aquele uso particular da língua, entende-se que, em situações ideias, haveria não só pontos concentrados de moderação de conteúdo em português no continente africano, mas um espalhamento de forma a apreender toda a diversidade da lusofonia neste continente. A inexistência de vagas

encontradas nesta pesquisa, portanto, se não for resultado de suas limitações, aponta para uma possível inadequação na moderação de conteúdo advindo desses países.

Por fim, ainda em relação à localização dos pontos de moderação de conteúdo em língua portuguesa, a análise de perfis no LinkedIn revelou quatorze novos locais em que há moderação de conteúdo, sendo doze no Brasil, um em Portugal e um nos Estados Unidos. O fato de grande parte desses perfis estarem no Brasil pode ser explicado pela alta taxa de moderação artesanal encontrada nessa fase.[8] Tratam-se, portanto, de locais em que a moderação de conteúdo é menos intensa, o que provavelmente gera também uma menor rotatividade nos postos de trabalho e uma maior dificuldade em se encontrar vagas de trabalho pela metodologia de busca de vagas empregada.

Em relação aos atores que realizam moderação de conteúdo em português, a análise conjunta das vagas e dos perfis no LinkedIn encontrou 48 empresas atuantes em alguma etapa de moderação de conteúdo na língua. Destas, apenas dezessete podem ser consideradas plataformas, sendo o restante empresas que realizam BPO. Essa conformação aponta para o grau de terceirização da atividade de moderação de conteúdo.

Foi encontrada uma concentração grande de plataformas no Brasil (13), seguida pelos Estados Unidos (5). No polo oposto, treze países possuíam moderação de conteúdo em português sendo prestada exclusivamente por terceirizadas, incluindo Portugal (onze terceirizadas encontradas). Em termos brutos, a maior concentração de diferentes empresas prestando moderação de conteúdo em português foi encontrada no Brasil (21) – muito influenciada pela presença de pequenas plataformas – e Estados Unidos (16) – com destaque para grandes companhias do Vale do Silício.

As plataformas que realizam diretamente moderação de conteúdo são bastante diversas entre si. Há, na amostra, tanto empresas que podem ser consideradas como redes sociais de grande porte, como o YouTube e o Tik Tok, ou mesmo de pequeno porte e mais localizadas no mercado brasileiro, como a rede social de relacionamentos Meu Patrocínio. Há, ainda, empresas de comércio eletrônico, como o Mercado Livre e o Sympla, e até mesmo plataformas educacionais interessadas em realizar moderação de comunidades, como a Alura. Essa distribuição demonstra como o mercado de moderação de conteúdo é

[8] As práticas de moderação de conteúdo foram caracterizadas por artesanais através das características da vaga e da plataforma anunciante. Tratam-se, aqui, de empresas que buscam perfis de “Community Managers” para gerenciamento de conteúdos disponíveis em plataformas internas.

diverso e como essa atividade é realizada por inúmeros segmentos digitais com objetivos distintos. Uma proposta de regulamentação sólida para a atividade deve alcançar ambientes cruciais como sites de ofertas e jogos digitais, cuja moderação é central para a proteção de pessoas vulneráveis, como o consumidor e as crianças e adolescentes.

Um outro aspecto analisado pela pesquisa foram as condições de trabalho descritas nas vagas de emprego de moderação de conteúdo em português. Foram descritas características como o turno de trabalho, as horas semanais trabalhadas e se o trabalho exigia conhecimentos em línguas estrangeiras ou experiência.

A primeira grande característica das vagas é trazer poucas informações sobre o posto de trabalho: em somente 15% delas foi possível encontrar remuneração, somente 45% traziam a informação sobre a jornada de trabalho semanal ou se o posto era em período parcial ou integral, em 36% delas não era possível saber se o trabalho era presencial ou remoto e somente 29% traziam a informação sobre qual o turno de trabalho da vaga.

As atribuições das vagas apontam para um meio bastante profissionalizado: em somente 10,7% das vagas, o contratado exerceria atividades de moderação de conteúdo em conjunto com outras de áreas estranhas à moderação; em 83,3% das vagas, os contratados seriam analistas de moderação de conteúdo; em 6%, exerceriam atividades relacionadas à gestão de equipes e elaboração de políticas e estratégias de moderação. Esses dados sugerem que há uma grande massa de analistas, e que cargos de direção são mais escassos, podendo inclusive ser promoções de analistas.

Essa conclusão condiz com os relatórios de transparência informados pelas empresas como parte dos requisitos de cumprimento da NetzDG (ver nota 3). No caso do relatório da Google relativo ao YouTube, por exemplo, foi informado que, entre os setenta e dois funcionários responsáveis por garantir o cumprimento da NetzDG, sessenta e um eram revisores, quatro eram líderes de equipe, cinco eram revisores de qualidade e dois eram instrutores (Google, s.d.).

A alta profissionalização é corroborada pelo fato de que somente 7% das vagas estavam relacionadas à moderação “artesanal”. Isso se deu, ainda, porque as vagas artesanais provavelmente aparecem pouco pelo tamanho reduzido das equipes e baixa rotatividade dos trabalhadores. Uma terceira evidência de profissionalização advém do fato de que somente 8% das vagas eram de trabalho *freelancer*.

Quadro 7 - Notícias sobre Crowdsourcing e Gig Economy

Aonde está a Moderação de 1º Nível? Plataformas de CrowdSourcing, 'GhostWorkers' e a *gig economy* global

Plataformas de crowdsourcing permitem a contratação de serviços em nível global para categorização e análise de dados. Esse é o modelo de negócios de plataformas como o Amazon Mechanical Turk (MTurk), Samasource, CrowdFlower e Microworkers. Prestadores de serviço de quaisquer locais do globo podem passar horas realizando classificação de imagens e conteúdos ofensivos, em uma prática de moderação de conteúdo de 1º nível (Wakefield, 2021). Esses trabalhadores têm sido nomeados pela literatura como "GhostWorkers", devido à invisibilidade do trabalho que realizam com o treinamento de máquinas (Royer, 2021). Através de dois questionários realizados pela Gizmodo sobre a experiência de 1100 trabalhadores na plataforma MTurk, 11% responderam ter realizado tarefas de classificação de conteúdo ofensivo e ilegal, como imagens de decaptação, cenas de suicídio e maus tratos à animais. À época, os trabalhadores diziam não ter orientação ou aviso de que classificariam imagens de violência ou abuso. (Mehrotra, 2020).

Fonte: Elaboração própria

O conhecimento de outra língua é um requisito importante para a maior parte das vagas de moderação de conteúdo: foi um requisito essencial ou desejável em cerca de três quartos delas. O inglês foi a língua mais exigida (55 vagas), seguido de espanhol (7 vagas). Em cerca de 41% das vagas, algum grau de escolaridade era desejável ou requerido: entre essas, 37% exigiam o equivalente ao Ensino Médio completo, 8,5% exigiam ao menos ensino superior em curso e 54% exigia o ensino superior completo. Já a experiência prévia em moderação de conteúdo ou áreas correlatas era exigida em apenas duas em cada cinco vagas. Todo esse cenário aponta para um setor composto por candidatos com escolaridade relativamente alta, porém aberto àqueles que não possuem experiência prévia no setor.

Quadro 8 – A moderação de conteúdo em línguas desconhecidas: reflexões a partir de casos reais

"Eu não falo essa língua" -

O conhecimento da língua em que se pretende moderar um conteúdo é um fator importante, capaz de auxiliar o moderador na identificação de gírias e na melhor compreensão do contexto de um conteúdo moderado. Apesar disso, reportagem investigativa do jornal The Washington Post, já citada anteriormente, demonstrou casos onde profissionais atuantes em empresas terceirizadas nas Filipinas moderavam conteúdos em línguas nas quais não possuíam nenhum conhecimento:

> Apesar de os moderadores dos Estados Unidos e Europa terem relatado ao The Post que realizavam a revisão de conteúdos principalmente nas línguas que falavam, moderadores de conteúdo filipinos, que falavam apenas Tagalog e Inglês, disseram que era comum serem pedidos para revisar conteúdos em **até 10 línguas diferentes.** Eles disseram realizar consultas com um falante fluente, caso houvesse alguém disponível no trabalho, mas caso não, **confiavam no Google Tradutor e Urban Dictionary, ferramentas que às vezes exacerbavam sua confusão e geravam estresse**. (Dwoskin, Whalen & Cabato, 2019)[9]

[9] Trad. própria. Original: *"Although moderators in the United States and Europe told The Post that they were asked to review content primarily in languages that they speak, Filipino moderators who spoke only Tagalog and English said that they were commonly asked to review content in as many as 10 languages. They said they would consult with a fluent speaker if one was available at work, but otherwise relied on Google Translate and Urban Dictionary, tools that sometimes exacerbated their confusion and created stress."*

Fonte: Elaboração própria

Em cerca de 1/5 das vagas, havia a indicação de que o funcionário precisaria de dedicação integral (24/7) ou que os horários não eram definidos, sendo feitos no esquema rotativo. Os horários rotativos eram mencionados em 62,5% das vagas em que o turno de trabalho era mencionado, o que sugere ser uma prática comum no mercado. Entre as vagas em que havia indicação de turno, 50% das vagas eram exclusivamente para o período diurno (5-22h). Em uma a cada seis, o turno noturno era uma possibilidade; em um terço, o turno era exclusivamente noturno.

Quadro 9 – As Condições de Trabalho de Moderadores de Conteúdo: Imprevisibilidade de Turnos

Uma imigrante brasileira na Alemanha que realizava moderação de conteúdo em português relatou ao *The Intercept* como a falta de previsibilidade dos turnos de trabalho afetava a vida desses trabalhadores: "Os turnos não eram fixos, e só descobríamos em qual estaríamos no dia 20 de cada mês. Mensalmente, podíamos pedir até três dias específicos de folga, mas nada nos garantia que teríamos os dias que pedíamos. Tanto isso quanto o turno eram um problema, e diversos colegas tiveram que abandonar cursos de língua por conta da imprevisibilidade da empresa. Como estávamos morando em um país cuja língua não falávamos, aprender alemão não era um luxo, mas uma necessidade" (Ribeiro, 2021, s.p.)

Fonte: Elaboração própria

O trabalho remoto é comum: 45% das vagas em que havia indicação da modalidade de trabalho (presencial ou remoto) são para trabalho remoto; 43%, para trabalho presencial; em 11% delas, o trabalho era temporariamente remoto devido à pandemia de COVID-19.

Por fim, outro ponto importante de análise nas vagas foi o alerta de que o ingressante estaria sujeito ao contato com imagens sensíveis. A importância deste alerta é um consenso, estando presente inclusive nas recomendações do Employee Resilience Guidebook for Handling Child Sexual Abuse Images da The Technology Coalition, que reúne grandes plataformas como Google, Twitter, Meta e Microsoft (The Technology Coalition, 2015). Todavia, somente pouco menos da metade (47,6%) das vagas traziam o referido aviso, o que sugere não ser uma prática muito difundida dentro do mercado.

3.1 Contexto e conhecimento da língua na moderação de conteúdo em português

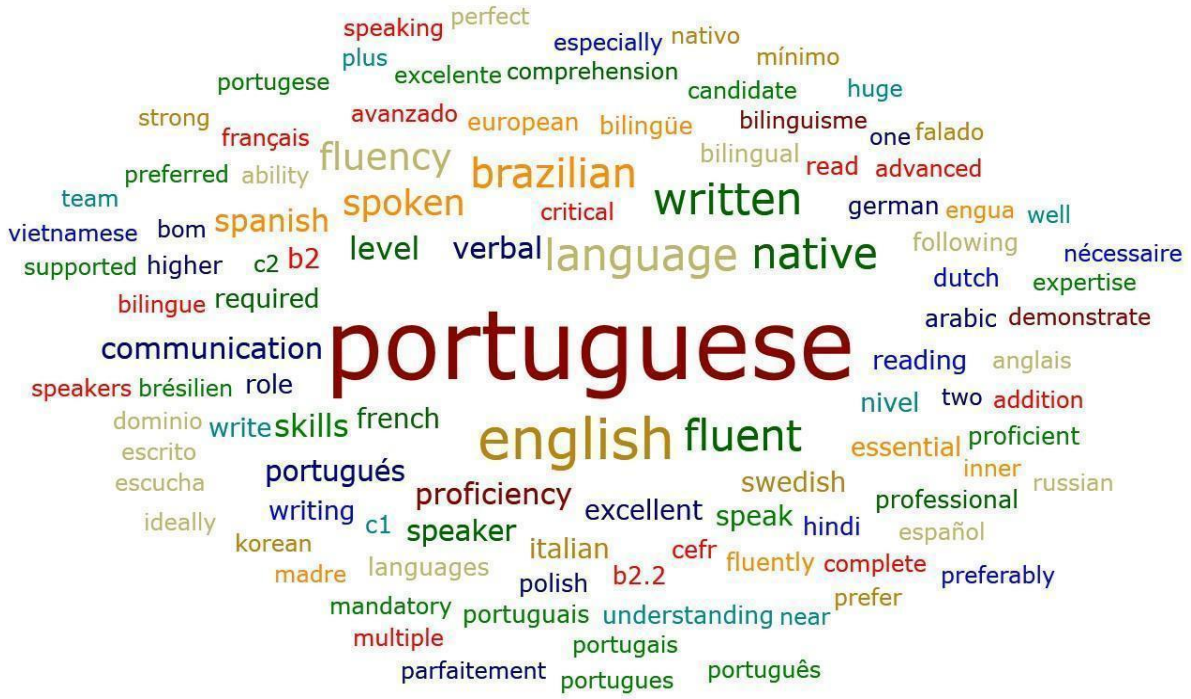
Uma última questão a ser analisada nas vagas é a importância do conhecimento do contexto para a realização da atividade de moderação de conteúdo. Como já descrito acima,

a própria distribuição geográfica dos postos de trabalho sugere não ser essa a principal preocupação dos atores que realizam a atividade comercialmente.

Em 63% das vagas, não havia diferenciação entre português e português brasileiro, sendo exigido apenas “português”; em 12%, havia a diferenciação, mas tanto o português brasileiro quanto o europeu eram aceitos; em 35%, a vaga era específica para o português brasileiro. As vagas em português brasileiro abrangeram Brasil, Portugal e outros cinco países, com destaque para Estados Unidos, Grécia e Espanha.

A definição da exigência de português era comum: em 72 das 86 vagas (83,7%), havia alguma especificação sobre o nível de português requerido. A nuvem de palavras dos requisitos se encontra abaixo.

Figura 1 - Nuvem de palavras dos requisitos linguísticos das vagas encontradas



Fonte: Elaboração própria a partir do software Atlas.ti

A nuvem aponta para a co-ocorrência de várias línguas distintas entre os requisitos, como inglês (English, anglais), espanhol (Spanish, español) e até línguas como o hindi, árabe e coreano. Esse fator se justifica pela profissionalização e globalização do campo, que acaba unindo em um mesmo ambiente de trabalho (e, por vezes, em uma mesma vaga), analistas de diversas línguas distintas.

Os requisitos são, em geral, caracterizados por adjetivos pouco precisos, como “fluente”, “avançado”, “nativo” ou “excelente”. Esse pode ser um fator de dificuldade, dada a vagueza desses termos. Em oito vagas, contudo, o requisito de nível de português estava descrito em termos do Quadro Europeu de Referência para Línguas, o que fornece uma objetividade adicional aos requisitos, aferível inclusive em testes. Entre essas, três vagas exigiam o nível B2 de português, duas exigiam o nível B2.2, duas exigiam C1 e uma exigia o nível C2.

Enquanto a proficiência linguística em português era valorizada nas vagas encontradas, a menção à importância de se conhecer o contexto era sensivelmente mais rara: foi encontrada em somente 23 das 86 vagas (26,7%), o que sugere não ser esse um requisito especialmente almejado pelas empresas. Esse aspecto reforça a ideia de que a moderação de conteúdo comercial em grande escala tende a se preocupar menos com o contexto do conteúdo moderado em relação a outros aspectos como economicidade e eficiência.

Esse achado contradiz o discurso de diversas companhias, que vem apontando a centralidade do contexto para a tomada de decisão sobre moderação de conteúdo. A Meta, no relatório de transparência do Facebook, afirma que

> [t]emos analistas que conhecem e entendem as culturas representadas nas tecnologias da Meta. Por exemplo, hispanofalantes mexicanos, não espanhóis, são contratados para analisar conteúdo do México. É importante que os analistas conheçam o significado específico das palavras, o contexto cultural, as celebridades locais ou as notícias mais recentes para que tenham o contexto necessário sobre a publicação e apliquem nossas políticas corretamente (Meta, 2022, s/p.).

Dada essa assimetria entre o discurso oficial e os achados empíricos, é importante que as empresas sejam transparentes sobre quais as métricas utilizadas para mensurar o conhecimento do analista da realidade a ser analisada, tanto durante a contratação, quanto no dia-a-dia da moderação. Essa consideração é importante sobretudo reconhecendo que o processo de moderação de conteúdo global é centrado em algumas poucas cidades, o que diminui a probabilidade de que o moderador, mesmo sendo um imigrante, esteja a par das complexidades culturais e políticas da região em que ocorrerá a moderação.

Apenas 13,2% das vagas eram de moderação bilíngue: três para moderação em português/inglês, três para moderação em português/espanhol, uma para moderação português/francês, duas para moderação em português/inglês/espanhol e uma para moderação português/espanhol/inglês/francês.

A análise de LinkedIn, contudo, encontrou relatos de outras seis grandes terceirizadas que realizavam moderação bilíngue. Ainda, em ao menos duas delas, moderadores de conteúdo em língua portuguesa atendem, indiscriminadamente, o mercado português e brasileiro. Com isso, o total de terceirizadas que realizam moderação bilíngue passa a ser de, ao menos, doze (38,7%), representando um modelo relativamente comum no mercado de moderação de conteúdo.

A moderação bilíngue ou para mercados diversos não é, por si só, um problema. Contudo, a consideração de que o contexto importa faz com que o trabalho desse moderador envolva o conhecimento profundo de diversas realidades, o que nem sempre é possível do ponto de vista prático. É central, portanto, que as métricas utilizadas para avaliar o desempenho da moderação de conteúdo levem em consideração o conhecimento das distintas realidades e do contexto social a partir do qual o conteúdo a ser moderado foi produzido, evitando cerceamentos excessivos ou mesmo discriminatórios na liberdade de expressão dos usuários.

Em suma, a moderação de conteúdo em língua portuguesa é realizada em diversas situações sociais, nem todas elas na agenda do Legislativo Federal. As condições de trabalho aparentam ser semelhantes àquelas do resto do mundo, embora a vagueza das vagas não permita fazer grandes inferências no tema. A moderação bilíngue ou para vários mercados é uma prática comum, e a importância do contexto parece ser secundária para os atores do setor.

4 Conclusões e Oportunidades de Pesquisas Futuras

A regulamentação da atividade de moderação de conteúdo, ou seja, de detecção, análise e intervenções em conteúdos em circulação no ambiente digital requer um conhecimento das regras utilizadas, do modelo de negócios, das condições do trabalho humano, da acurácia e forma de treinamento de tecnologias de detecção automatizada e dos mecanismos que apoiam a sua realização (Gillespie; Aufderheide, 2020).

Trata-se de um setor econômico em expansão, essencial para o bom funcionamento de redes sociais, mas também utilizada pelos setores de comércio, transporte e logística, jogos, educação, dentre outros. Nesta pesquisa, buscamos identificar a distribuição geográfica do mercado global e português de moderação de conteúdo, identificando hubs estratégicos onde se localizam terceirizadas que prestam esse tipo de serviço.

Os resultados apresentados apontam para uma divergência entre a localização geográfica das empresas prestadoras de serviços e dos países para os quais os serviços são prestados. Como casos notáveis, temos a realização de moderação de conteúdo em 2º nível de registros linguísticos como “Espanhol Boliviano”, “Espanhol Colombiano”, “Espanhol Equatoriano”, “Espanhol Mexicano” e “Espanhol Peruano” localizados geograficamente na Espanha, sem evidências de que os respectivos países na América Latina possuam equipes especializadas na função.

No que diz respeito à língua portuguesa, em que pese uma grande quantidade de vagas encontradas no Brasil, a análise global dos dados parece indicar uma prevalência de atividades de moderação de conteúdo artesanal, com uma distribuição significativa de empresas prestadoras do serviço de moderação em 2º nível em países como Estados Unidos e Portugal.

Nesse sentido, importa uma reflexão sobre a importância do conhecimento do contexto local pelos trabalhadores envolvidos no setor:

Quadro 10 – Desafios de compreensão de contexto para moderação de conteúdo

Quais os desafios de compreensão de contexto para moderação de contexto?
O UOL Tecnologia conversou com 6 dos 800 moderadores terceirizados da CCC (Competence Call Centre), no Brasil. Um dos maiores desafio apontado pelos moderadores entrevistados foi a compreensão de contexto, falas e gírias de outros estados brasileiros: **Pergunta:** Vocês podem dar um exemplo de conteúdo ambíguo que fez vocês errarem? **Resposta:** Às vezes, é questão de região. **Eu sou do interior de Goiás e vejo coisas de forma diferente de quem mora em São Paulo, a vivência é diferente, as expressões são diferentes.** Às vezes, são questões gramaticais, erro ortográfico, falta de vírgula. Você tem que se colocar no lugar da pessoa e entender o que ela quer falar. Isso pode ser super tranquilo ou super difícil. (Uchinaka, 2019).

Fonte: Elaboração própria

A língua, a vivência e a compreensão da realidade é influenciada pelo local onde se vive e pela convivência com diferentes grupos, espaços e ambientes. Na entrevista concedida ao UOL Tecnologia, moderadores brasileiros apontam o desafio da moderação de conteúdos provenientes de outros Estados. Importa ressaltar, entretanto, que essas diferenças podem ser latentes até mesmo em Estados diferentes: o Extremo-Sul da Bahia possui um arcabouço cultural que difere da região do Baixo Médio São Francisco. É também na compreensão cultural de nuances geográficas, apreendidas e aplicadas na análise da

força de trabalho de moderadores de conteúdo, que se pode esperar uma maior acurácia nas decisões realizadas.[10] Por este motivo, essa pesquisa se dedicou a uma compreensão da localização e das condições da força de trabalho que move este setor.

É notável que o Brasil é um país que gera alta demanda de moderação de conteúdos provenientes de usuários(as) brasileiros(as) da rede. Com uma população de 212,6 milhões de habitantes, dos quais mais de 157 milhões possuem algum nível de acesso à internet (Banco Mundial, 2020a; 2020b), o país é o segundo no ranking mundial de maior quantidade de tempo diário online por pessoal. No ranking global de tempo diário utilizando redes sociais, o Brasil ocupa a 3ª posição (KEMP, 2021). O que se percebe, entretanto, é que a demanda de moderação de conteúdo gerado por usuários brasileiros parece ser exportada, não gerando oportunidades significativas de geração de trabalho e de renda para a população nacional.

Cabe retomar o processo de regulamentação do setor de *call-centers*, atividade que possui relevância econômica no mercado de trabalho brasileiro por causa de sua elevada capacidade de geração de empregos (NETO, 2005, p. 145). A entrada em vigor do Código de Defesa do Consumidor, em março de 1991, é um dos fatores explicativos que esclarece o fenômeno, na medida em que a legislação exigiu que empresas se adequassem, disponibilizando Serviços de Atendimento do Consumidor (SAC) e aproximando sua relação com clientes para uma melhor prestação do serviço, coleta de dúvidas, reclamações e sugestões (Neto, 2005, p. 147)[11]. Para além disso, o Decreto nº 6.523/08[12], que regulamentava o SAC, criou uma série de obrigações acessórias quanto à qualidade de atendimento, ao armazenamento de dados e aos prazos de atendimento, obrigações que se converteram em demanda de mão de obra para o mercado em crescimento. O SAC pode ser interpretado como um mecanismo de solução de conflitos e de *accountability* sobre a

[10] A literatura especializada no tema demonstra uma tensão latente que deve ser analisada em propostas de regulamentação do mercado de moderação de conteúdo, qual seja, a dificuldade em se estabelecer regras consistentes e globais para a atividade, sendo sensível quanto aos contextos localizados em que o conteúdo foi produzido (Caplan, 2019, p. 7).

[11] Há, ainda, outros fatores explicativos para a proeminência do mercado de *call-centers* no Brasil, conforme aponta Neto (2005, p. 151), quais sejam: i) oportunidade de expansão competitiva representada pelo serviço e ii) maior disponibilidade de linhas telefônicas no país, configurando um modelo de gestão de relacionamentos entre consumidor-empresa via telefone.

[12] Em 05 de abril de 2022, o instrumento normativo foi revogado pela entrada em vigor do Decreto 11.034/22. A nova regulamentação trás disposições de adaptação do SAC para a melhora na prestação de atendimento durante a pandemia do Covid-19, como previsões de disponibilização de canais alternativos de contato e de ao menos 8h de serviço de atendimento telefônico por humanos, impedindo uma total automatização do atendimento das demandas pelo setor.

prestação de serviços ao consumidor, intermediando relações de forma a garantir maior eficácia do CDC.

Considerando a importância de análises baseadas em contexto para a atividade de moderadores de conteúdo, bem como a posição estratégica do Brasil no consumo de serviços digitais, é possível conceber o mercado de moderação de conteúdos como uma janela de oportunidades para atração de investimentos em nível nacional. Em um contexto de grandes transformações tecnológicas, no qual se projeta a redução de empregos em nível global e a automatização de diversos setores de trabalho, é essencial um pensamento sistêmico capaz de criar novas oportunidades para a população que estejam atreladas à demandas reais para construção de um ecossistema digital mais democrático, participativo e seguro. Apesar disso, não se deve perder o foco de que a atração de investimentos nesse setor deve priorizar a saúde, bem-estar e segurança dos trabalhadores.

Recomendações e propostas de agenda

PARA O ESTADO BRASILEIRO:

1) Analisar a possibilidade de incluir a moderação de conteúdo em um código específico do Cadastro Brasileiro de Ocupações - CBO, permitindo a criação de estatísticas sobre esse mercado;
2) Incluir a moderação de conteúdo como uma preocupação legislativa, de forma a abordar, para além das estruturas de redes sociais, os microssistemas de proteção do consumidor, da criança e do adolescente e do universo de jogos, considerando aspectos como o controle de qualidade dos trabalhos, a garantia de justa remuneração e o respeito à direitos humanos e trabalhistas no ambiente de trabalho;
3) Estruturar uma estratégia de fomento ao mercado nacional de moderação de conteúdo a partir de planos regulatórios;
4) Criar, no âmbito dos órgãos de saúde e segurança do trabalho (Ministério do Trabalho, Ministério Público do Trabalho), forças-tarefas para pensar a legislação aplicável ao trabalho de moderação de conteúdo, analisando a possibilidade de criação de legislação protetiva para a categoria, considerando fatores como jornada e condições de trabalho.

PARA AS EMPRESAS (PLATAFORMAS E TERCEIRIZADAS)

1) Criar uma política de transparência em relação ao processo de moderação de conteúdo, registrando, por exemplo, as empresas terceirizadas prestadores de serviços, os locais onde a moderação acontece, quais línguas são moderadas pela equipe, qual a proveniência dos conteúdos moderados e os requisitos mínimos para contratação de moderadores;
2) Implementar a preocupação com o contexto no processo de recrutamento, treinamento e avaliação do processo de moderação, exigindo um conhecimento sólido da realidade a ser moderada para a realização do trabalho de moderação;
3) Alertar candidatos a vagas de moderação de conteúdo para os riscos inerentes ao trabalho;
4) Fortalecer os serviços de bem-estar para moderadores, diminuindo fontes de estresse como imprevisibilidade de turnos de trabalho, pouca quantidade de tempo para realizar decisões e pressão para atingir KPIs de qualidade;

PARA A ACADEMIA

1) Consolidar uma agenda de pesquisa em temas emergentes de moderação de conteúdo;
2) Criar metodologias inovadoras para entender o mercado e as experiências de moderação de 1º e 2º nível através, por exemplo, de metodologias qualitativas como os grupos focais e realização de questionários (*surveys*).

Referências Bibliográficas

Aggarwal, Manu; Nijhawan, Rhea; Nousher, Aseem (2021). Trust and Safety – Content Moderation Services PEAK Matrix® Assessment 2021. **Everest Group**. Disponível em: https://www2.everestgrp.com/reportaction/EGR-2021-0-R-4262/Marketing. Acesso em 28 abr. 2022.

Banco Mundial (2022a). **Individuals using the internet (% of population)**. Banco Mundial, [s.l.]. Disponível em https://data.worldbank.org/indicator/IT.NET.USER.ZS. Acesso em 1 abr. 2022.

Banco Mundial (2022b). **Population, total**. Banco Mundial, [s.l.]. Disponível em https://data.worldbank.org/indicator/SP.POP.TOTL. Acesso em 1 abr. 2022.

Binns, Reuben; Veale, Michael; Van Kleek, Max; Shadbolt, Nigel Shadbolt (2017). Like Trainer, like Bot? Inheritance of Bias in Algorithmic Content Moderation. *In:* Ciampaglia, Giovanni; Mashhadi, Afra; Yasseri, Taha (eds.).. **Social Informatics:** 9th International Conference (pp. 405-415). Cham: Springer International Publishing.

Business Wire (2021, 16 ago.). Global Content Moderation Solutions Market Report 2021: Market was Valued at $5.3 Bn in 2020. **Business Wire**. Disponível em https://www.businesswire.com/news/home/20210816005282/en/Global-Content-Moderation-Solutions-Market-Report-2021-Market-was-Valued-at-5.3-Bn-in-2020---Exponential-Rise-in-Inappropriate-Content-to-Account-for-a-VAGR-of-12.6-2021-2026---ResearchAndMarkets.com. Acesso em 18 abr. 2022.

Caplan, Robyn (2018, 14 nov.). Content or context moderation? Artisanal, Community-reliant, and Industrial Approaches. **Data & Society**, [s.l.]. Disponível em https://datasociety.net/library/content-or-context-moderation. Acesso em 18 abr. 2022.

Center For Democracy & Technology (2007). **The Netzwerkdurchsetzungsgesetz (NetzDG) Network Enforcement Law**. Washington: Center for Democracy & Technology. Disponível em https://cdt.org/wp-content/uploads/2017/07/NetzDG-Law-Overview.pdf. Acesso em 28 abr. 2022.

Dwoskin, Elizabeth; Whalen, Jeanne; Cabato, Regina (2019, 25 jul.). Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently. **The Washington Post**. Disponível em https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/. Acesso em 18 abr. 2022.

Gillespie, Tarleton (2018). **Custodians of the internet**. New Haven: Yale University Press.

Gomes, Alessandra; Antonialli, Dennys; Oliva, Thiago (2019, 28 jun). Drag queens e Inteligência Artificial: computadores devem decidir o que é 'tóxico' na internet? **InternetLab**. Disponível em

https://internetlab.org.br/pt/noticias/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/. Acesso em 18 abr. 2022.

Google. **Remoções de acordo com a Lei aplicável a redes**. [s.d]. Disponível em https://transparencyreport.google.com/netzdg/youtube?hl=pt_BR. Acesso em 28 jun 2022.

Hunsberger, Alice; Brown, Vanity; Galib, Lily (2021). **Best practices for gender-inclusive content moderation**. Grindr, [s.l.].

Kaye, David (2019). **Speech Police**. Nova York, Columbia Global Reports.

Kemp, Simon (2021, 27 jan.). Digital 2021 Global Overview Report. **We are social**. Disponível em: https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/. Acesso em 9 abr. 2022.

Jacobs, Mira (2022, 19 mar.). The Boys' Twitter Account Reacts to Its Millions of Content Flags. **CBR.com**. Disponível em https://www.cbr.com/the-boys-twitter-millions-content-flags/. Acesso em 18 abr. 2022.

OFCOM (2019). **Use of AI in online content moderation**. [s.l.], Cambridge Consultants.

Mehrotra, Dhruv (2022, 28 jan.). Horror Stories From Inside Amazon's Mechanical Turk. **Gizmodo**. Disponível em: https://gizmodo.com/horror-stories-from-inside-amazons-mechanical-turk-1840878041. Acesso em 29 abr. 2022.

Meta (2022). **As pessoas por trás das equipes de análise da Meta.** [s.l.], Meta. Disponível em https://transparency.fb.com/pt-br/enforcement/detecting-violations/people-behind-our-review-teams/. Acesso em 28 jun. 2022.

NETO, José Borges Da Silva (2005). **Call Centers no Brasil:** um estudo sobre emprego, estratégias e exportações. Dissertação (Mestrado em Economia) – Universidade Federal de Uberlândia/UFU. Uberlândia.

Perrigo, Billy (2022, 17 fev.). Inside Facebook's African Sweatshop. **Time**. Disponível em: https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/. Acesso em 20 abr. 2022.

POSNER, Rebecca And SALA, Marius (2022, 14 fev.). "Spanish language". **Encyclopedia Britannica**. Disponível em https://www.britannica.com/topic/Spanish-language. Acesso em 28 abr. 2022.

Ribeiro, Paulo Victor (2011, 25 jan.). Eles veem execuções, mutilações e suicídio sem suporte psicológico: o relato de uma ex-moderadora do Facebook. **The Intercept**. Disponível em https://theintercept.com/2021/01/25/eles-veem-execucoes-mutilacoes-e-suicidio-sem-supo

rte-psicologico-o-relato-de-uma-ex-moderadora-brasileira-do-facebook/. Acesso em 28 abr. 2022.

Roberts, Sarah (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. *In*: Noble, Safiya; Tynes, Brendesha (eds.). **The intersectional internet:** Race, sex, class and culture online. Nova York, Peter Lang.

Roberts, Sarah (2019). **Behind the Screen:** Content Moderation in the Shadows of Social Media. New Haven, Yale University Press.

Royer, Alexandrine (2021, 9 fev). The urgent need for regulating global ghost work. TechStream. Disponível em: https://www.brookings.edu/techstream/the-urgent-need-for-regulating-global-ghost-work/. Acesso em 29 abr. 2022.

Smith, Adam (2020, 12 out.). Peloton forced to remove QAnon groups as conspiracy theory spreads from social media to workout platforms. **Independent**. Disponível em https://www.independent.co.uk/tech/peloton-qanon-bikes-conspiracy-theory-b985652.html. Acesso em 18 abr. 2022.

The Technology CoalitiON (2015). **Employee resilience guidebook for handling child sexual exploitation images** (version two). [s.l.], The Technology Coalition.

The Cleaners (2018). Direção: Riesewieck, M.; Block, H. Produção de Gebrueder Beetz Filmproduktion, Grifa Filmes & Westdeutscher Rundfunk (WDR). Alemanha. Streaming (88 min).

Uchinaka, Fabiana (2019, 7 jun). No Limite: Facebook abre as portas da moderação de conteúdo para mostrar quem decide o que é certo ou errado na rede. **UOL**, Barcelona. Disponível em https://www.uol.com.br/tilt/reportagens-especiais/como-e-o-centro-de-moderacao-de-conteudo-do-facebook. Acesso em 28 abr. 2022.

Wakefield, Jane (2021, 28 mar.). AI: Ghost workers demand to be seen and heard. **BBC News**. Disponível em: https://www.bbc.com/news/technology-56414491. Acesso em 28 abr. 2022.