Mapeamento do Perfil das Mulheres Brasileiras em Processamento de Linguagem Natural

Helena Caseli[1,7], Evelin Amorim[2,7], Elisa Terumi Rubel Schneider[3,7], Leidiana Iza Andrade Freitas[4,7], Jéssica Rodrigues[5,7], Maria das Graças V. Nunes[6,7]

[1]Universidade Federal de São Carlos (UFSCar) – São Carlos – SP
[2]INESC TEC – Porto – Portugal
[3]Pontifícia Universidade Católica do Paraná (PUCPR) – Curitiba – PR
[4]Universidade Federal do Ceará – Fortaleza – CE
[5]Oxford Internet Institute (OII) – University of Oxford – Oxford – Reino Unido
[6]NILC – ICMC – Universidade de São Paulo (USP) – São Carlos – SP
[7]Brasileiras em Processamento de Linguagem Natural (BPLN)

helenacaseli@ufscar.br, evelin.f.amorim@inesctec.pt, elisa.rubel@pucpr.edu.br, izafreitas@alu.ufc.br, jessica.rodrigues@oii.ox.ac.uk, gracan@icmc.usp.br

Abstract. *Knowing the profile of Brazilian women who work in Natural Language Processing (NLP) is an important step towards the development of policies aimed at increasing inclusion and diversity in this field. This is the first study conducted in Brazil for this purpose. Based on data from public survey, Lattes and Linkedin, we found that the profile is that of a background in computer science or linguistics, working in companies or universities, but with little ethnic diversity and apparent difficulty in balancing professional life and motherhood. Analyzing more specifically the group "Brasileiras em PLN" (Brazilian Women in NLP), we found that we have a significant capacity for publication and supervision, but still a low level of collaboration among our members.*

Resumo. *Conhecer o perfil das mulheres brasileiras que atuam em Processamento de Linguagem Natural (PLN) é um importante passo para o desenvolvimento de políticas e programas que visem aumentar a inclusão e a diversidade nessa área. Este é o primeiro trabalho realizado no Brasil com este fim. A partir de dados coletados via consulta pública, Lattes e Linkedin, notou-se que o perfil é de uma formação em computação ou linguística, atuando em empresas ou universidades, mas com pouca diversidade étnica e aparente dificuldade em conciliar vida profissional e maternidade. Analisando mais especificamente o grupo "Brasileiras em PLN" constatou-se uma expressiva capacidade de publicação e orientação, mas ainda uma baixa colaboração entre nossas integrantes.*

1. Introdução

O Processamento de Linguagem Natural (PLN) é uma área de pesquisa interdisciplinar que envolve principalmente pessoas com formação em Computação e Letras/Linguística. O PLN[1] surgiu como um ramo da Inteligência Artificial (IA) e, segundo [Freitas 2022], tem um lado teórico e um lado aplicado. O lado teórico envolve estudos sobre

[1]Enquanto Processamento de Linguagem (ou Língua) Natural é o nome mais usual no Brasil, também encontramos menções à Linguística Computacional para se referir à área. Em inglês, os termos são *Natural Language Processing* e *Computational Linguistics*, respectivamente.

a formalização da língua, modelos de representação, caracterização de fenômenos linguísticos, etc. O lado aplicado diz respeito à automatização de tarefas que possibilitam a geração de aplicações, como os tradutores automáticos, os corretores gramaticais e, mais recentemente, os assistentes virtuais inteligentes, como a Alexa e a Siri, ou os chatbots, como o ChatGPT.[2]

Na atualidade, o aprendizado de máquina é a estratégia predominante nas pesquisas e produtos do PLN, em especial por meio do uso das redes neurais artificiais profundas (*deep learning*) que "aprendem" a partir de dados textuais presentes em conjuntos de dados denominados *corpus*. A partir de um *corpus* é possível treinar modelos computacionais para realizar tarefas como correção ortográfica [Wang et al. 2020], análise de sentimentos [Zhang et al. 2018], tradução automática [Wu et al. 2016] e geração de texto [Dong et al. 2022]. Embora a parte aplicada seja a que está mais presente no nosso dia-a-dia, o lado teórico é imprescindível para o estudo dos fenômenos linguísticos da maneira apropriada e o desenvolvimento adequado de produtos precisos [Bang et al. 2023].

A área de PLN tradicionalmente abriga linguistas e cientistas da Computação, chamados de linguistas computacionais. Com a demanda cada vez maior em sistemas de PLN usados nos mais variados setores da sociedade, são vários os profissionais que hoje se envolvem com PLN. Ainda que pessoas de diferentes formações se envolvam com a área, é natural que fiquem a cargo dos linguistas computacionais as principais decisões de projeto. A formação desses profissionais enfrenta desafios próprios das áreas multidisciplinares, já que o modelo das universidades brasileiras dificulta a integração de conhecimentos de ciências distintas, como as ditas exatas e humanas.

Já há muito tempo as ciências exatas se caracterizam por atrair estudantes do gênero masculino, enquanto que as humanas, em especial, Letras e Linguística, atraem mais estudantes do gênero feminino. No imaginário popular, meninos são mais racionais ou mesmo capazes de entender e lidar com números e máquinas, enquanto que meninas se dão melhor com leitura, escrita, ensinar e aprender [Hill et al. 2010]. Enquanto esse cenário persistir, é natural que meninas da área de humanas que têm interesse em PLN sintam-se em desvantagem para se juntarem a essa área. Por mais que a contraparte – os informatas – necessite e deseje essa integração, elas parecem convencidas de que jamais vão entender os detalhes de PLN, e com isso deixam de participar em papel de igualdade.

Além dos desafios gerais na área de PLN, que abarcam tanto homens quanto mulheres, há os desafios específicos para uma pessoa que se identifica com o gênero feminino atuar em uma área predominantemente masculina [Hango 2013], seja sua formação em Computação ou Letras. Relatos de sexismo e suas especificidades[3], como o *manterrupting*[4], são constantes. Como apontado por [Mohammad 2020], "gênero" é um conceito complexo, que não possui apenas dois extremos (masculino e feminino) nem está diretamente ligado ao gênero atribuído no nascimento. Contudo, a sociedade ainda defende e prega a dualidade, e os preconceitos associados ao gênero feminino são evidentes em áreas de tecnologia como a Computação e o PLN.

[2]Disponível em: `https://chat.openai.com`. Acesso em: 25 maio 2023.

[3]Disponível em: `https://didthisreallyhappen.net/`. Acesso em: 25 maio 2023

[4]*Manterrupting* é o termo usado para designar a ação de interrupção, por parte de um homem, da fala de uma mulher antes de esperar que ela finalize sua fala.

Por esses e outros motivos, em 2020, um grupo de mulheres brasileiras criou o Brasileiras em PLN[5] do qual nós, as autoras deste artigo, somos as integrantes do Conselho (Board). O BPLN, em Maio/2023, era composto por mais de 140 mulheres brasileiras atuando no PLN. O Brasileiras em PLN foi a primeira iniciativa desse tipo, no Brasil, e define o cenário no qual o levantamento apresentado neste artigo foi realizado.

Neste contexto, este artigo apresenta um levantamento das mulheres, brasileiras, atuantes na área de PLN. Tal levantamento foi realizado de duas maneiras distintas: (1) por meio de uma consulta pública anônima e (2) por meio da coleta de informações públicas nas plataformas Lattes e Linkedin para as integrantes do Brasileiras em PLN.

As questões de pesquisa que embasam nosso trabalho são:

Q1 Qual é o perfil das mulheres brasileiras que atuam no PLN?
Q2 Como é a atuação quantitativa das integrantes do BPLN?
Q3 Quais são os principais temas de pesquisa das integrantes do BPLN?
Q4 Há interação/colaboração entre as integrantes do BPLN?
Q5 Onde estão atuando (academia ou indústria) as integrantes do BPLN?
Q6 Quais são as habilidades (*skills*) que as integrantes do BPLN alegam possuir?

A principal contribuição deste trabalho é trazer o retrato das mulheres brasileiras atuantes no PLN em termos socio-demográficos e quantitativos, de acordo com as informações coletadas de forma anônima ou disponibilizadas publicamente. Como resultados esperados estão: o aumento da visibilidade da atuação das mulheres na área de PLN; a possibilidade de despertar, em futuras pesquisadoras de PLN, o interesse pela área; e a conscientização de organizações para a adoção de iniciativas mais inclusivas.

2. Trabalhos relacionados

A análise de classes de grupos dentro de uma comunidade é um tipo de estudo comum na ciência. Em [Lorens et al. 2020], por exemplo, as autoras analisam a participação de mulheres em comitês de programa de eventos da computação nacionais e internacionais. Como bem pontuado pelas autoras, levantamentos que focam no público brasileiro trazem a possibilidade de entender o problema em análise sob a nossa perspectiva. Entre as conclusões desse trabalho estão a de que as mulheres tendem a concentrar suas participações em eventos nas áreas de sistemas colaborativos, educação, interação humano-computador e inteligência artificial, nessa ordem. As autoras enfatizam, ainda, que os eventos com maior concentração feminina são também eventos interdisciplinares. Tal fato também ocorre na área de PLN, onde a interdisciplinariedade pode fazer com que a quantidade de mulheres seja maior do que em outras áreas da computação [Mohammad 2020, Richter et al. 2023].

Em [Cordeiro et al. 2020], os autores analisam a participação de mulheres em programas de pós-graduação em computação, no Brasil. Entre as conclusões desse estudo estão: a redução da participação das mulheres na docência em programas de pós-graduação em computação, no país, nos últimos 15 anos; e a predominância da faixa etária de 40-69 anos das mulheres atuantes.

Especificamente para a área de PLN, [Mohammad 2020] traz um estudo sobre a participação de mulheres como autoras em artigos disponíveis na ACL Anthology[6], que

[5]Disponível em: `https://brasileiraspln.com/`. Acesso em: 25 maio 2023.
[6]Disponível em: `https://aclanthology.org/`. Acesso em: 25 maio 2023.

é o principal repositório de artigos da *Association for Computational Linguistics* (ACL). Neste estudo foi feita uma análise sobre a quantidade de citações a seus artigos no Google Scholar[7]. A análise de autoria foi realizada considerando artigos publicados entre 1965 e 2019 e, desses, cerca de 29% foram identificados como tendo autoria do gênero feminino: 29,7% (8.532) dos 28.682 autores e 29,2% (10.885) dos 37.297 primeiros autores. O estudo aponta que essa porcentagem não aumentou desde meados de 2000, e que as taxas de mulheres como primeiras autoras foram relativamente maiores em pesquisas com idiomas europeus diferentes do inglês como russo, português, francês e italiano.

O mesmo estudo também traz uma contagem dos bigramas[8] presentes nos títulos dos artigos na tentativa de encontrar os temas nos quais as mulheres são mais ativas como primeiras autoras. Neste nosso trabalho, nós refinamos a lista original de bigramas agrupando aqueles similares (como *dialog systems* e *dialog system*) e descartando os genéricos (como *natural language*) ou indicativos de técnicas (como *machine learning*). Nessa nova compilação dos dados originais de [Mohammad 2020], os temas de pesquisa nos quais as mulheres têm maior participação como primeiras autoras são: recursos linguísticos (*language resources*, 36,5%: 58/159), sumarização (multi) documento (*document summarization* e *multi document*, 36,3%: 91/251), sistemas de diálogo (*dialog system(s)*, 36,2%: 105/290), geração de linguagem (*language generation*, 35,2%: 56/159), entendimento de linguagem (*language understanding*, 31,4%: 38/121), encadeamento textual (*textual entailment*, 31,4%: 33/105) e desambiguação (lexical) de sentido (*word sense* e *sense disambiguation*, 30,9%: 238/769). Em relação a esses bigramas, em [Mohammad 2020] não foi detectada correlação entre a popularidade dos temas e a porcentagem de mulheres como primeiras autoras, indicando que a escolha pelos temas não está diretamente relacionada com o fato de serem *hot-topics*.

Diferente dos trabalhos citados, este não pretende comparar dados levantados para os dois gêneros tradicionais – masculino e feminino – mas sim analisar dados de pessoas que se identificam com o gênero feminino, atuantes ou interessadas na área de PLN, e que responderam a consulta pública ou fazem parte do Brasileiras em PLN.

3. Materiais e métodos

O levantamento apresentado neste artigo foi realizado com base em dados coletados por meio de uma consulta pública anônima e também da coleta de informações públicas nas plataformas Lattes[9] e Linkedin[10] para as integrantes do Brasileiras em PLN.

3.1. Consulta pública

A consulta pública recebeu 92 respostas no período de 13/02/2023 a 01/03/2023. Ela foi projetada para conter duas partes. A primeira era formada por 10 perguntas de cunho sócio-demográfico: faixa etária, identidade de gênero, etnia, estado civil, quantidade de filho(a)s, estado de nascimento, país de residência, escolaridade, situação empregatícia e renda mensal familiar. A última pergunta desta parte era "Você faz parte do grupo Brasileiras em PLN?". Se a pergunta fosse respondida com "Não", a consulta pública era encerrada; caso contrário, a participante era direcionada para a segunda parte da consulta.

[7]Disponível em: `https://scholar.google.com.br/`. Acesso em: 25 maio 2023.
[8]Um bigrama é uma sequência de duas palavras.
[9]Disponível em: `https://lattes.cnpq.br/`. Acesso em: 25 maio 2023.
[10]Disponível em: `https://www.linkedin.com/`. Acesso em: 25 maio 2023.

Na segunda parte da consulta, as perguntas visavam levantar o perfil específico das integrantes do BPLN em relação a: nível de conhecimento em PLN, área de formação, meio pelo qual ficou conhecendo o grupo e ano de ingresso no grupo. Ao final, cada integrante foi convidada a selecionar características/palavras que a definisse ou que representasse sua visão do grupo, bem como se expressar livremente (campo de texto).

3.2. Lattes

A partir da lista de *e-mail* do BPLN, solicitamos que as integrantes preenchessem uma planilha indicando IDs de seus currículos Lattes. Das cerca de 140 integrantes do BPLN, 90 atenderam nosso pedido. Dessas, 42 (46,7%) têm atuação/formação em linguística e 48 (53,3%) têm atuação/formação em computação. Também foi possível identificar 82 CVs Lattes, os quais foram baixados manualmente para serem processados pelo script lucyLattes[11], com adaptações no código feitas por nós.[12] O período de análise foi estipulado entre 2020 (ano de fundação do BPLN) e Fevereiro de 2023 (atual).

3.3. Linkedin

A partir da mesma planilha usada para coleta dos IDs Lattes, solicitamos o preenchimento dos IDs das páginas pessoais do Linkedin. No total, 69 perfis foram identificados e processados por scripts Python gerados por nós.[13] Nessa coleta, notamos que os perfis poderiam estar disponíveis em português ou inglês. Neste último caso, usamos o modelo de linguagem *Helsinki-NLP/opus-mt-en-ROMANCE* disponível de forma gratuita no portal Hugging Face[14] e a biblioteca python `transformers`[15] para que toda a informação textual analisada estivesse em português.

4. Resultados e Discussão

Essa seção traz os resultados dos levantamentos realizados via consulta pública e coleta automática ao Lattes e Linkedin, apresentados para cada uma das questões de pesquisa, nas próximas subseções. Por fim, a seção 4.7 discute algumas limitações deste trabalho.

4.1. Qual é o perfil das mulheres brasileiras que atuam no PLN? (Q1)

Por meio da consulta pública, foi possível traçar um retrato das participantes. Nesse caso, das 92 pessoas que iniciaram a consulta, apenas uma, após ler as instruções, não concordou em seguir adiante. Os dados resultantes da primeira parte da consulta são resumidos na Tabela 1.[16]

[11]Disponível em: `https://github.com/rafatieppo/lucyLattes`. Acesso em: 25 maio 2023.

[12]Disponível em: `https://github.com/brasileiras-pln/levantamento-2023`. Acesso em: 25 maio 2023.

[13]Os scripts foram gerados com base em `https://medium.com/mlearning-ai/how-to-build-a-web-scraper-for-linkedin-6b49b6b6adfc`. Acesso em: 25 maio 2023. Esses scripts também estão disponíveis no github do BPLN.

[14]Disponível em: `https://huggingface.co/Helsinki-NLP/opus-mt-en-ROMANCE`. Acesso em: 25 maio 2023.

[15]Disponível em: `https://pypi.org/project/transformers/`. Acesso em: 25 maio 2023.

[16]Embora a maioria viva atualmente no Brasil, a consulta também recebeu respostas de participantes residentes na África do Sul, Alemanha, Argentina, Colômbia, Espanha, Estados Unidos, França e Portugal.

Tabela 1. Retrato das brasileiras atuantes em PLN

Faixa etária	de 30 a 49 anos	63,7%
Identidade de gênero	Cisgênero	96,7%
Raça	Branca	72,5%
Estado civil	Solteira	42,9%
	Casada/união estável	49,5%
Quantidade de filho(a)s	Sem filho(a)s	68,1%
Nascidas em	São Paulo	30,7%
	Rio de Janeiro	13,6%
	Minas Gerais	12,5%
Residente no	Brasil	87,9%
Nível de escolaridade	Mestrado	31,8%
	Graduação	24,2%
	Doutorado	15,4%
	Pós-doutorado	15,4%
Inserção no mercado de trabalho	Empregada trabalhando 40 horas por semana	62,6%
Renda familiar mensal	entre R$ 5.000,00 e R$ 10.000,00	26,4%
	entre R$ 10.000,00 e R$ 15.000,00	20,9%
	mais de R$ 15.000,00	20,9%

Tabela 2. Retrato das integrantes do BPLN

Nível de conhecimento de PLN	Sólidos conhecimentos	36,2%
	Expert	24,1%
	Conhecimento intermediário	19,0%
	Iniciante	20,7%
Como soube do BPLN	Indicação de amiga(o), colega ou professor(a)	65,5%
	Navegando no Google	13,8%
	Via Linkedin	12,1%
Quando entrou no BPLN	Em 2020	50,0%
	Em 2022	27,6%
	Em 2023	10,3%
	Em 2021	8,6%
Características pessoais apontadas	Estudiosa	75,4%
	Pró-ativa	59,6%
	Resiliente	54,4%

O fato de a consulta ter coletado respostas de uma maioria (72,5%) que se reconhece como branca, desperta a necessidade urgente de maior representatividade étnica na comunidade de PLN. Nesse sentido, políticas de ações afirmativas na área devem ser pensadas pela Ce-PLN[17], a SBC[18] e a Abralin[19], entre outros. Outro fato que vem à tona a partir dos dados é que, embora o estado civil esteja bem balanceado entre as solteiras (42,9%) e as que declararam ter companheiro/as (49,5%), a maioria (68,1%) relata não ter filho/as. Nesse sentindo, vale ressaltar que os impactos da maternidade na carreira da mulher são notórios [Ahmad 2017, Klocker and Drozdzewski 2012]. Segundo [Metz 2005], mães que trabalham possuem a relação entre jornada de trabalho e ascensão gerencial fragilizada, enfrentando barreiras adicionais quando comparadas com mulheres sem filho.

Para a segunda parte da consulta pública, das 91 respondentes, 58 (63,7%) afirma-

[17]Disponível em: `https://sites.google.com/view/ce-pln`. Acesso em: 25 maio 2023.
[18]Disponível em: `https://www.sbc.org.br/`. Acesso em: 25 maio 2023.
[19]Disponível em: `https://www.abralin.org/`. Acesso em: 25 maio 2023.

ram fazer parte do Brasileiras em PLN e, como tal, prosseguiram para as demais questões. As outras participantes, 33 (36,3%) finalizaram sua participação na consulta pública. A Tabela 2 traz um resumo dos dados levantados na segunda parte da consulta pública.

A partir da análise do nível de conhecimento de PLN reportado pelas integrantes do BPLN, notamos a necessidade de engajar integrantes jovens. Uma das estratégias que o grupo pode adotar no sentido de capacitar novas meninas para atuarem na área é promover minicursos de conteúdo introdutório ao PLN, abertos, gratuitos e de ampla divulgação para estudantes dos cursos de graduação em letras/linguística e computação. Em relação ao modo como tomou conhecimento sobre o BPLN, notou-se que o Linkedin[20] se destacou como o melhor canal de alcance de novas integrantes entre as opções adotadas pelo grupo (que incluem o YouTube[21] e o Twitter[22]).

As questões sobre características pessoais permitiam múltipla escolha, enquanto a descrição do BPLN era para ser feita em uma palavra. Nesse último caso, 47 respostas foram obtidas, visto que não era uma pergunta obrigatória. As palavras que mais se repetiram foram: comunidade (4), sororidade (3), potente (3), apoio (2), colaboração (2), interessante (2), motivador (2) e necessário (2). Isso reforça a importância de grupos como este, projetando visibilidade das mulheres em áreas geralmente associadas aos homens e criando um ambiente acolhedor e participativo. Também perguntamos como a participante se definira, em apenas uma palavra. As respostas mais obtidas foram: determinada (4), resiliente (4), curiosa (3), dedicada (3), persistente (3), pesquisadora (2), pró-ativa (2), esforçada (2) e sonhadora (2).

4.2. Como é a atuação quantitativa das integrantes do BPLN? (Q2)

No período de 2020 a 2023, das 82 integrantes do BPLN com CV Lattes identificados, 10 não tinham atualizado seus currículos no período. Assim, a análise apresentada nesta seção leva em consideração as informações contidas nos CV Lattes de 72 integrantes do Brasileiras em PLN.[23] Para essas, constatou-se a participação em 69 projetos de pesquisa e 21 de extensão, 12 livros publicados, 145 artigos publicados em periódicos e 220 em eventos. Em termos de orientações, elas orientaram 19 teses de doutorado, 49 dissertações de mestrado, 53 iniciações científicas e 144 trabalhos de conclusão de curso de graduação e 11 de aperfeiçoamento e especialização. A Figura 1 traz as distribuições das quantidades de publicações em periódicos e eventos, por ano, no período.

Entre os periódicos A1[24] nos quais as integrantes do BPLN publicaram no período de 2020 a 2023 estão: Revista de Estudos da Linguagem (5), Linguamática (4), Domínios de Lingu@gem (3), Future Generation Computer Systems (3) e Artificial Intelligence Review (3). Outros periódicos qualificados que receberam publicações do grupo (3 artigos em cada) foram: Sensors (A2), Journal of Information and Data Management (B1),

[20] `https://www.linkedin.com/company/brasileiras-em-pln/`

[21] `https://www.youtube.com/@brasileiraspln`

[22] `@brasileiraspln`

[23] É importante mencionar que esses dados foram coletados a partir das versões do Lattes disponíveis no dia 27/02/2023.

[24] As classificações Qualis de periódicos foram obtidas por meio de consulta ao site `https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf`, sem selecionar nenhuma "Área de Avaliação" e baixando a planilha gerada como resultado.

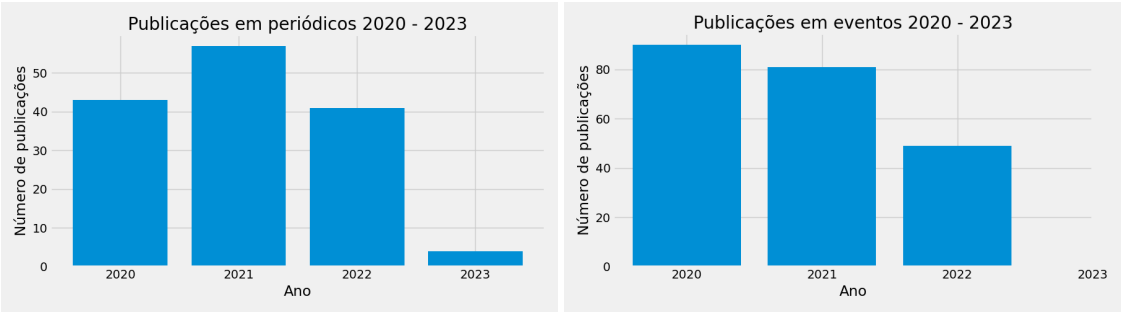

Figura 1. Publicações, por ano, em periódicos (à esquerda) e eventos (à direita)

Revista GTLEX (B2) e Revista Brasileira de Sistemas de Informação (B2).

Entre os eventos[25] nos quais as integrantes do BPLN mais publicaram no período de 2020 a 2023 estão os tradicionais de PLN – STIL (11/B1), PROPOR (7/A4), LREC (5/A1) – e alguns de IA – ENIAC (6/B4), BRACIS (5/A4), ICMLA (4/A2), IJCNN (3/A1), EPIA (3/B2) e FLAIRS (3/A4).[26] Além desses, há publicações em eventos interdisciplinares com: banco de dados (SBBD, 5/A4), redes sociais (BRASNAM, 3/B4), área médica (CBMS, 3/A3), educação (SBIE, 3/A3), sistemas multimídia e web (WebMedia, 3/A4) e mineração de dados (KDMile, 3/B3). A intersecção com as áreas de educação e interação humano-computador corroboram as conclusões de [Lorens et al. 2020] que apontou essas, juntamente com a IA, como três das quatro áreas de maior inserção de mulheres em comitês de programa de eventos da computação.

4.3. Quais são os principais temas de pesquisa das integrantes do BPLN? (Q3)

Para encontrar os temas nos quais essas integrantes do Brasileiras em PLN mais pesquisam realizou-se uma análise dos bigramas presentes nos títulos das publicações. Entre os bigramas mais frequentes estão: *sentiment analysis* (análise de sentimentos), *named entity* (entidade nomeada), *machine learning* (aprendizado de máquina), *terminological accessibility* (acessibilidade terminológica), *fake news*, *question answering*, *information extraction* (extração de informação), *language models* (modelos de linguagem), *deep learning* (aprendizado profundo), *word embeddings*, *hate speech* (discurso de ódio), *open information* e *universal dependencies*.

Contudo, a frequência desses bigramas nos títulos foi de no máximo 5%, indicando uma grande variedade terminológica dos títulos artigos publicados por esse grupo. Notamos, também, que a menção ao idioma investigado (em expressões como *Portuguese language*) é bastante recorrente o que se justifica pelo fato de ser um costume na área indicar o idioma no qual a pesquisa foi realizada quando este é diferente do idioma inglês.

4.4. Há interação/colaboração entre as integrantes do BPLN? (Q4)

Outra informação relevante gerada pelo script a partir dos Lattes é a rede de colaboração entre pesquisadores. A versão original gerava apenas a colaboração em publicações em

[25]Os eventos na computação também são classificados via Qualis, disponíveis em `https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/avaliacao/09012022_RELATORIOQUALISEVENTOS20172020COMPUTACAO.PDF`

[26]Neste caso, são listadas entre parênteses as quantidades de publicações acompanhadas do Qualis atribuído ao evento.

periódicos, nós adaptamos a versão para gerar a colaboração também em publicações em eventos, conforme ilustrado na Figura 2.

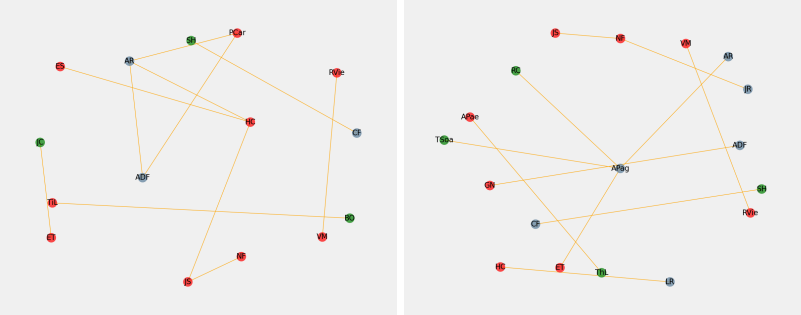

Figura 2. Grafos de colaborações entre as integrantes do Brasileiras em PLN nas publicações em periódicos (à esquerda) e em eventos (à direita)

É possível notar que das 72 integrantes do Brasileiras em PLN com publicações no período de 2020 a 2023, 13 (18%) publicaram conjuntamente em periódicos e 16 (22%) em eventos, indicando que ainda há espaço para aumentar/melhorar as colaborações em pesquisa entre as integrantes do grupo.

4.5. Onde estão atuando (academia ou indústria) as integrantes do BPLN? (Q5)

Considerando que a área de TI continua em expansão[27], levando em conta números de 2022, as pessoas formadas nesta área têm alguma liberdade de optar por um trabalho no mercado ou na academia. Portanto, é natural uma análise da área de atuação profissional das brasileiras especialistas em PLN.

A partir dos perfis de Linkedin coletados, verificamos que 28 perfis, dos 69 analisados, possuíam alguma relação com a academia. Esta relação foi verificada através de um simples casamento de strings utilizando o valor preenchido no campo "Posição Atual" (*Current Position*). Se neste campo o perfil contivesse alguma Universidade ou titulação (Ph.D, ou Ms.C, Professora), então o perfil era considerado acadêmico. Após esta separação automática, foi feita uma verificação manual dos perfis.

Na verificação manual, foi constatado que existem pessoas com titulação que estão ligadas ao mercado também. Dos perfis que declaram posições puramente acadêmicos, a apuração manual detectou apenas 18 perfis. Os perfis restantes eram estudantes de pós-graduação ou pessoas já com a titulação com posições no mercado. Portanto, o grupo possui uma proporção expressiva (73,91%) de representantes da indústria.

4.6. Quais são as habilidades que as integrantes do BPLN alegam possuir? (Q6)

Quando analisamos o grupo acadêmico, as únicas habilidades citadas mais de uma vez, com duas citações cada, são 'Computer Science', 'Machine Learning', 'Inglês', e 'Docência'. Quanto ao grupo não exclusivamente acadêmico, as cinco habilidades mais citadas[28] são 'Aprendizado de máquina' (3), 'Análise de dados' (3), 'Microsoft Office'

[27]Disponível em: `https://g1.globo.com/trabalho-e-carreira/noticia/2022/12/09/vagas-no-setor-de-ti-crescem-34percent-em-10-meses-veja-cargos-em-alta.ghtml`. Acesso em: 25 maio 2023.

[28]O número de citações no grupo estão em parênteses.

(3), 'Linguística' (4), 'Natural Language Processing' (4). Embora não seja possível concluir pela quantidade de pessoas em cada grupo, as mulheres com perfil exclusivamente acadêmico alegam habilidades que podem ser mais úteis na academia no Brasil, como 'Docência'. No grupo não acadêmico encontramos habilidades que podem ser mais úteis na indústria, como 'Microsoft Office'. Entretanto, 'Machine Learning' ('Aprendizado de Máquina') é uma habilidade em comum nos dois grupos.

4.7. Limitações desta pesquisa

Uma das limitações deste levantamento está na forma como os dados foram coletados. O fato de a consulta pública ter sido opcional e anônima não nos permite garantir que todas as integrantes do BPLN responderam à consulta, nem fazer o alinhamento das respostas de uma participante da consulta pública com seus dados públicos nas plataformas consultadas (Lattes e Linkedin). Outra limitação está na quantidade de dados analisados, uma vez que das mais de 140 integrantes do BPLN, foi possível identificar apenas 82 (91%) CVs Lattes e 69 (76,7%) perfis do Linkedin. Além disso, não há garantia de que as informações presentes nessas plataformas estão atualizadas. De fato, notou-se que há CV Lattes que não foram atualizados no período especificado para este levantamento. Por fim, vale ressaltar que a coleta de dados quantitativos foi realizada por meio de scripts automáticos. Embora não tenha sido identificado nenhum erro aparente, tais scripts extraem automaticamente as informações presentes nos arquivos XML dos currículos e podem estar sujeitos a erros de interpretação das informações contidas nos campos desses arquivos.

5. Considerações finais

Embora o número de mulheres nas áreas de ciências, engenharia e tecnologia esteja crescendo, as mulheres ainda estão sub-representadas nas áreas que envolvem exatas, como no PLN. Estudos evidenciam que fatores sociais e ambientais [Wang and Degol 2017, Alam 2022] contribuem para este fenômeno, ao invés da noção errônea de que as mulheres carecem das habilidades necessárias para atuar nessa área. A própria cultura científica pode atuar como uma barreira para as mulheres. Neste cenário, mais estudos, pesquisas e políticas são necessários para reverter a situação e proporcionar às meninas e mulheres que seguem nestas áreas, igualdade nos contextos acadêmico, científico e profissional.

Este trabalho levantou o perfil das mulheres brasileiras que atuam em PLN, apontando a necessidade de investir em políticas e ações para aumentar a diversidade étnica e promover condições que permitam conciliar a vida profissional com a opção pela maternidade. Em relação especificamente às integrantes do Brasileiras em PLN, considerando as 72 integrantes com CV Lattes atualizados no período de 2020-2023, constatou-se uma expressiva capacidade de publicação (média anual de 2 artigos em periódicos e 3 em eventos) e orientação (média anual próxima a 4), mas ainda uma baixa colaboração entre as integrantes (apenas 15% das publicações no período foram conjuntas).

A propagação de grupos que incluem e relacionam as mulheres em áreas correlacionadas habitualmente aos homens, como o Brasileiras em PLN, contribui para aumentar a percepção de que as mulheres também podem (e devem) atuar nas áreas mais masculinizadas, desmistificando estereótipos e promovendo a inclusão e diversidade. Por fim, até onde sabemos, esse é o primeiro trabalho realizado no Brasil para mapear o perfil de mulheres atuando em PLN. Os dados e scripts usados neste trabalho estão disponíveis no GitHub do BPLN: `https://github.com/brasileiras-pln`.

Agradecimentos

Agradecemos a todas as pessoas que participaram da consulta pública e às integrantes do BPLN que fornecerem seus IDs Lattes e Linkedin para possibilitar a coleta de dados para esta pesquisa. Agradecemos também o suporte financeiro de nossas instituições e programas de pós-graduação aos quais estamos vinculadas.

Referências

Ahmad, S. (2017). Family or future in the academy? *Review of Educational Research*, 87(1):204–239.

Alam, A. (2022). Psychological, sociocultural, and biological elucidations for gender gap in stem education: A call for translation of research into evidence-based interventions. In *Alam, A.(2022). Psychological, Sociocultural, and Biological Elucidations for Gender Gap in STEM Education: A Call for Translation of Research into Evidence-Based Interventions. Proceedings of the 2nd International Conference on Sustainability and Equity (ICSE-2021). Atlantis Highlights in Social S*.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Cordeiro, D., Rocha, A., Cassiano, K. K., and da Silva, N. (2020). Representativeness of women in postgraduate programs in computer science in Brazil. In *Anais do XIV Women in Information Technology*, pages 110–119, Porto Alegre, RS, Brasil. SBC.

Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., and Yang, M. (2022). A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38.

Freitas, C. (2022). *Linguística Computacional*. Parábola, São Paulo.

Hango, D. (2013). Gender Differences in Science, Technology, Engineering, Mathematics and Computer Science (STEM) Programs at University. *Insights on Canadian Society*.

Hill, C., Corbett, C., and St. Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics.* American Association of University Women., Washington, D.C.

Klocker, N. and Drozdzewski, D. (2012). Commentary: Career progress relative to opportunity: how many papers is a baby 'worth'?

Lorens, A. L., Botelho, J., Moura, A. F., Duarte, B., and Moro, M. (2020). Participação feminina em comitês de programa de simpósios da computação. In *Anais do XIV Women in Information Technology*, pages 90–99, Porto Alegre, RS, Brasil. SBC.

Metz, I. (2005). Advancing the careers of women with children. *Career Development International*.

Mohammad, S. M. (2020). Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of ACL*, pages 7860–7870, Online. Association for Computational Linguistics.

Richter, A., Yamamoto, J., and Frachtenberg, E. (2023). Why are there so few women in computer systems research? *Computer*, 56(2):101–105.

Wang, M.-T. and Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (stem): Current knowledge, implications for practice, policy, and future directions. *Educational psychology review*, 29:119–140.

Wang, Y., Wang, Y., Liu, J., and Liu, Z. (2020). A comprehensive survey of grammar error correction. *arXiv preprint arXiv:2005.06600*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.