XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT 8 – Informação e Tecnologia

ADERÊNCIA LEXICAL A DADOS PUBLICADOS PARA PRODUTORES RURAIS

LEXICAL ADHERENCE TO PUBLISHED DATA FOR RURAL PRODUCERS

Jacquelin Teresa Camperos-Reyes – Universidade Federal do Pará (UFPA)
Ricardo César Gonçalves Sant'Ana – Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp) Campus de Marília

Modalidade: Trabalho Completo

Resumo: Esta pesquisa insere-se no contexto da disparidade entre a linguagem presente em dados governamentais sobre agricultura e desenvolvimento rural e a que é utilizada em informações direcionadas a pequenos e médios produtores rurais no Brasil. O estudo utilizou o modelo de "Aderência Lexical", que compara o vocabulário entre diferentes fontes de dados, buscando medir a similaridade e, consequentemente, a potencial inteligibilidade das informações. Para testar o modelo, foram analisados dados do site de dados abertos do governo brasileiro (dados.gov.br) e notícias da Confederação da Agricultura e Pecuária do Brasil (CNA) direcionadas aos produtores. A pesquisa identificou alta aderência lexical (80.48% de similaridade) apenas na categoria "Crédito". As demais categorias de necessidades informacionais (Mercado, Tratos Culturais, Direitos e Oportunidades), apesar de presentes nas notícias da CNA, não foram encontradas nos conjuntos de dados governamentais. Conclui-se que a Aderência Lexical é um instrumento válido para aproximar a linguagem dos dados às necessidades dos usuários, e para evidenciar a necessidade da integração do contexto do usuário ao contexto dos dados, e assim ampliar o aproveitamento das informações. A pesquisa contribui para a Ciência da Informação ao demonstrar a importância de considerar os fluxos informacionais junto às necessidades dos usuários, para o desenvolvimento de soluções, especialmente no contexto de pequenos e médios produtores rurais.

Palavras-chave: aderência lexical; acesso a dados; necessidades informacionais; agricultura familiar.

Abstract: This research is et within the context of the disparity between the language found in government data on agriculture and rural development and that used in information aimed at small and medium-sized rural producers in Brazil. The study used the "Lexical Adherence" model, which compares the vocabulary between different data sources, aiming to measure similarity and, consequently, the potential intelligibility of the information. To test the model, data from the Brazilian government's open data portal (dados.gov.br) and news from the Brazilian Confederation of Agriculture and Livestock (CNA) directed towards producers were analyzed. The research identified high lexical adherence (80.48% similarity) only in the "Credit" category. The other categories of

informational needs (Market, Cultural Practices, Rights, and Opportunities), although present in CNA's news, were not found in the government data sets. It is concluded that Lexical Adherence is a valid tool for aligning the language of the data with user needs, highlighting the necessity of integrating user context with data context to enhance the usefulness of the information. The research contributes to Information Science by demonstrating the importance of considering information flows and user needs in the development of solutions, especially in the context of small and medium-sized rural producers.
keywords: lexical adherence; data access; informational needs; family farming.

1 INTRODUÇÃO

A abertura de dados governamentais se manifesta em uma ampla gama de opções de recursos disponíveis. O poder público, diante da dinâmica do acesso à informação, publica conjuntos de dados, que ao vir de todas as esferas orgânicas, implica em uma riqueza pelos assuntos tratados. Os usuários dos dados, com as suas caraterísticas estruturais e contextuais, têm cada vez mais opções de utilização de dados disponibilizados.

Considerando a Barreto (1994, p. 1), ao postular que "a importância que a informação assumiu na atualidade pós-industrial, recoloca para o pensamento questões sobre a sua natureza, seu conceito e os benefícios que pode trazer ao indivíduo e no seu relacionamento com o mundo em que vive", parte-se da premissa de que ao procurar o aproveitamento de dados disponíveis, é necessário que eles estejam publicados com características que favoreçam a sua interpretação.

Diante disso, particularmente no contexto de benefícios das relações entre atores dos fluxos de informação, este estudo localiza-se no âmbito da disponibilização de dados e considera aspectos do acesso a esse volume de recursos. Entende-se que esse volume versus os benefícios obtidos por práticas que o usam, manifesta-se em magnitudes diferentes, densidades de dados que devem ser aproximadas, onde, se bem uma integração plena entre as densidades dos dados seja utópica, ações que contribuam com a conciliação de realidades e o fenômeno informacional observando demandas dos sujeitos alvos, concebem-se como necessárias.

A pesquisa estima a potência do uso de dados e informações por produtores, enquadrada em suas demandas informacionais, com foco na interface em que esses dados são apresentados pelo detentor governo, uma vez que, efetuada a transdução[1] de conteúdos

[1] Sant'Ana (2019) conceitua transdução como o processo de transformação de um tipo de sinal em outro, aplicável a distintas energias; o processo, discernido no âmbito desta área de estudo, foi intitulado como

da coleta à recuperação (Sant'Ana, 2019), considera-se determinante verificar a proximidade entre esses dados disponíveis e os sujeitos alvos.

O objetivo é verificar a viabilidade de um modelo que reconheça a Aderência Lexical a dados publicados pelo detentor governo, usando necessidades informacionais de produtores rurais como marco de análise.

Entende-se Aderência Lexical como uma ferramenta de análise que possibilite mensurar a similaridade vocabular entre duas fontes de dados, o que permitiria obter um primeiro indício da potencial interpretação de dados publicados por um detentor, e passíveis de serem coletados por sujeitos alvo em determinado contexto. Terá um viés lexical observando, como recomendado por Gray *et al*. (2009), aspectos que uma vez tornados acessíveis quanto dados oficiais, determinem o seu uso com base em um formato adequado para os usuários.

Como elementos do escopo da pesquisa, adverte-se sobre a decisão de utilizar, para a prova do modelo, uma fonte de dados online sobre produtores rurais no lugar de dados coletados diretamente com esse grupo de usuários. Do lado governo, optou-se por conjuntos de dados publicados pelo governo do Brasil, orientados a produtores vinculados com a Economia Solidária. A análise das unidades foca no nível lexical atentando-se à similitude entre conjuntos de palavras, observando que outras opções como colocação e relevância fazem parte de desdobramentos da pesquisa. As instruções computacionais utilizadas não são parte do escopo deste artigo devido a limitações de espaço.

Embora o estudo reconheça a importância de linguagens documentárias e de indexação no domínio da Ciência da Informação, a opção por analisar a linguagem natural no âmbito desta pesquisa justifica-se por sua aproximação com a realidade comunicacional dos agricultores familiares. A utilização de termos e expressões presentes no cotidiano desse público, como observado em site de notícias para o público-alvo, permite uma análise mais próxima de suas necessidades informacionais reais, transpondo as barreiras da linguagem técnica e formal frequentemente presente em sistemas de recuperação de informação tradicionais. A proposta não visa construir um sistema de recuperação da informação, mas

transdução informacional, e implica no acompanhamento de processos de persistência, medições, atributos de representações dos conteúdos e suas interfaces, perfiladas como próximas às características dos sujeitos alvo.

sim analisar a "Aderência Lexical" como um primeiro passo para identificar e minimizar as disparidades na comunicação entre o governo e os agricultores familiares. A identificação da discrepância, por si só, já oferece um diagnóstico relevante que pode subsidiar, em etapas posteriores, o desenvolvimento de soluções para aproximar a linguagem dos dados governamentais às necessidades informacionais específicas desse público, contribuindo para uma comunicação mais eficaz e para o desenvolvimento rural sustentável.

2 DETERMINAÇÃO DOS ELEMENTOS PARA MONTAGEM DO MODELO

Segundo o contexto em que dados são disponibilizados, e as necessidades dos sujeitos que os detém ou quem os usará, terão acontecido ações conforme o fim esperado nessa disponibilização, seja localizar, armazenar, preservar etc. Assim, na compreensão de que dados inacessíveis podem perder sua relevância e incorrer em uma perda de esforços de quem os produziu, estima-se necessário que eles sejam atingidos com análises de elementos relacionados com seu acesso, e dessa forma, realmente servir aos sujeitos que deles precisam.

Na perspectiva dos sujeitos-salvo, entende-se que demandam por esses recursos para fundamentar processos estratégicos como tomadas de decisões, por isso, considerar caraterísticas da recuperação pode implicar no melhor uso dos dados acessados.

Entende-se que o acesso à informação implica na melhora da qualidade de vida dos indivíduos, ao quebrar com exclusões sistêmicas que os restringem em suas opções de transformar o ambiente ao qual pertencem (Carvalho, 2010; Ruediger, 2006), de tal maneira que a possibilidade da abertura dos dados de governo, veio como início para a construção de uma cidadania digital (Carvalho, 2010), agindo em sincronia com o pensamento de Saracevic (1996) quanto à busca do equilíbrio da equação homem-tecnologia, visando-o a partir do homem e não da tecnologia em si (Camperos-Reyes, 2018).

No contexto de países em desenvolvimento, se uma opção que se encontra ao interior de cada governo, como os dados governamentais, pode abrir as portas à competitividade e ao desenvolvimento, isto dentro de uma conjuntura tecnológica favorável pela abrangência e escopo global, dita opção poderia estar direcionada a áreas que de fato já se encontram contribuindo no corpus econômico e social de cada país. É o caso do domínio da agricultura e do desenvolvimento rural.

O acesso a dados e informações pode ser determinante para um desenvolvimento integral, influenciando diretamente aspectos socioeconômicos de uma sociedade. Partindo da afirmação de que as Tecnologias de Informação e Comunicação se configuram como agentes de desenvolvimento em áreas rurais (Ferrás Sexto; García; Pose, 2011) e que elas adquirem o caráter de usáveis dependendo dos atributos de acesso a elas, considera-se importante que conjuntos de dados que estejam sendo publicados pelo governo para grupos de produtores rurais, realmente atinjam esses sujeitos alvo. Embora o pressuposto de que tal disponibilização possa estar impregnada de ruídos, acredita-se na relevância de discutir oportunidades para que os dados possam ser aproveitados adequadamente pelos sujeitos.

Esse estudo o empenhou-se em aproximar a comunidade acadêmica à conjuntura que envolve ativos informacionais, na sua disponibilização e nas condições de acesso, no domínio da agricultura e desenvolvimento rural, uma vez que "a agricultura é o setor que produz mais emprego no mundo, fornecendo a forma de vida de 40% da população mundial e é a maior fonte de ingressos e trabalho nos lares pobres rurais" (Organização das Nações Unidas, 2023).

Esse setor agiu com resiliência durante o período da pandemia; foi tal que alavancou a queda dos demais setores econômicos, somente superada pelo auge inefável e necessário das TIC. Na América Latina e o Caribe, 11 de 16 países experimentaram aumento no Produto Interno Bruto (PIB) do setor agropecuário (Comissão Econômica para a América Latina e o Caribe, 2021)

Nesse âmbito, o sujeito informacional, "pequeno produtor", constitui um grupo destacado no país, observa-se que no Brasil, do total de estabelecimentos agropecuários (5.073.324), o 67% correspondem a agricultores familiares (Instituto Brasileiro de Geografia e Estatística, 2021).

Deve considerar-se que, diante a relevância no acesso a ativos informacionais, o ecossistema não pode estar completo focando apenas na tecnologia e nos dados em si. Nossa estrutura tecnológica, pequena como asseverado por Davenport (1998), porém marcante no que diz diante das vantagens competitivas que trazem para aqueles grupos sociais que sabem fazer uso dela, poderia ser observada como pivô na busca do aprimoramento de um Estado, sem desconsiderar que o desafio é fazer que exista equilíbrio na ecologia, percebido mediante o aprimoramento das condições de vida palpável nos grupos sociais atingidos, estabelecendo pontes entre a ciência e o desenvolvimento de uma

região ou de um país.

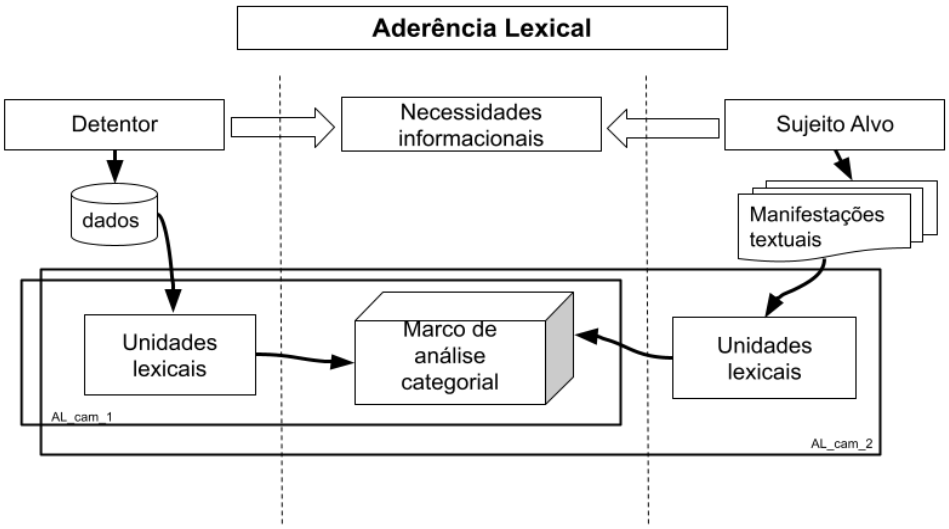
Como o estudo tenciona encontrar explicações acerca de condições de acesso baseadas em eventos regulares nas fontes de dados escolhidas, compreende-se como necessário usar um marco de análise para tal investigação, integrando assim o uso de categorias de necessidades informacionais dos usuários consumidores dos dados. Visando a proposta do ciclo informacional enunciada por Barreto (1998), informação - conhecimento - desenvolvimento – informação, acredita-se que, na medida em que dados publicados pelo governo estejam relacionados com as necessidades informacionais dos sujeitos alvo, existe uma maior possibilidade de que eles sejam usufruídos.

3 PROPOSTA DE MODELO DE ANÁLISE DE ADERÊNCIA LEXICAL

A lente que foi usada para observar um aspecto no acesso a dados é aplicada para analisar a proximidade entre as unidades lexicais geradas, propondo um elemento para propiciar a interpretação de fluxos de informação, prioridade da Ciência da Informação manifesta pelo Borko (1968), e pretendendo um aprimoramento da experiência humana ao participar de fluxos informacionais disponíveis (Camperos-Reyes *et al*., 2020).

O modelo para análise da Aderência Lexical se alicerça na tríade: dados do detentor + necessidades informacionais + sujeito alvo, que em este estudo está instanciada no contexto de dados de governo, e necessidades informacionais de pequenos e médios produtores, que de forma generalizada pode verificar-se conforme a Figura 1. Ela é aplicável em duas camadas: 1) para verificar a proximidade entre dados publicados para o sujeito alvo e as suas necessidades informacionais; 2) para verificar a similaridade vocabular entre duas fontes de dados que mantêm necessidades informacionais em comum.

Figura 1 - Aderência lexical entre duas fontes de dados



Fonte: elaborado pelos autores

É uma proposta que se espera gerar um diagnóstico sobre a proximidade entre as fontes escolhidas, que não pode ser observado apenas pelo volume das unidades lexicais extraídas, e sim ao considerar uma visão holística, no caso deste estudo, pela perspectiva outorgada pelas categorias de necessidades informacionais.

3.1 Instância da proposta para avaliação nas fontes escolhidas

Para a prova da proposta foram triangulados procedimentos metodológicos, a saber, Pesquisa Documental, Revisão Sistemática de Literatura, Mineração Textual e Análise de Similaridade. A interseção formada na aplicação dos procedimentos metodológicos remete à referência heurística da Ciência da Informação (Gómez, 2003), que possibilita o descobrimento de fatos durante o percurso da pesquisa.

O foco da Pesquisa Documental foram fontes de dados e informações do governo do Brasil, dados.gov.br. Foi considerada a documentação para o acesso e uso de conjuntos de dados, buscando estruturas de aproximação para o entendimento dos dados publicados.

A Revisão Sistemática de Literatura (RSL) junto a meta-análise, atentou à produção científica entregando necessidades informacionais no acesso a dados na agricultura (Camperos-Reyes, 2023).

Para a coleta dos dados do lado do produtor rural, unidades lexicais contidas em notícias para eles, foi determinado realizá-la a partir de uma organização que os agrega, a

Confederação da Agricultura e Pecuária do Brasil (CNA), pois considerou-se factível atingir indícios que caracterizam a pequenos e médios produtores quando eles estão agrupados.

A linguagem usada nas comunicações, por ser geradas no entorno de agrupações de produtores rurais, é uma linguagem direcionada para alcançar esse público-alvo, portanto, presumem-se escritas com um teor adequado às caraterísticas dos grupos de produtores.

Entende-se que a linguagem natural ao ser um recurso mediador entre cenários da compreensão humana, é uma linguagem "eficaz na tradução dos fenômenos em pensamento porque é constituída por um elemento primordial para as associações lógica, psicológica, ideológica e interpretativa: o signo" (Martines; Moreira; Almeida, 2022, p. 33)

Abordar documentos escritos na linguagem natural com ferramentas tecnológicas propicia amplas possibilidades de análise. Em um primeiro momento, pelas capacidades *per se* da linguagem natural como sistema de signos expressivo e predileto na comunicação (Korn; Huss; Cumbers, 1988), de outro lado, técnicas informáticas permitem revelar informações em grandes corpos textuais, que aproveitando das capacidades da linguagem natural, viabilizam a obtenção de inferências conforme interesses de pesquisa em um contexto determinado.

Assim sendo, a coleta das comunicações foi realizada mediante Mineração de Textos, implementando algoritmos na linguagem R. Determinou-se usar essa técnica pois permite extrair informação das fontes mediante identificação de padrões, tendências, ou índices, atendendo interesses de estudos que abordam grandes coleções de textos (Feldman; Sanger, 2006). Ao extrair essas informações e relacioná-las com outros elementos discernidos, conformam-se novos feitos e geram-se inferências, tudo isto a partir do processamento de grande corpus textuais, originalmente não codificados ou estruturados (Hearst, 2003; Kao; Poteet, 2007).

Em relação à coleta de dados do lado do governo, foi realizada extraindo rótulos dos recursos de conjuntos de dados e as descrições registradas pelos publicadores, de tal forma que ambas as fontes forneceram dados manifestados na linguagem natural.

Dessa forma, as duas fontes escolhidas provêm de mediadores da informação que estão fora de manifestações explícitas produzidas por profissionais da Ciência da Informação.

Observaram-se conjuntos de dados, ou *datasets* recuperados mediante o uso de descritores "Pequeno produtor", "Desenvolvimento rural", "Associacao agricultura", "Cooperativa agricultura". Realizou-se extração manual dos rótulos e descrições justificando-

se necessário discriminar a disponibilidade de cada recurso recuperado na busca.

Devido a que a forma de apresentação de informações e os pressupostos em relação ao seu acesso estão mais relacionados com atributos de quem os disponibiliza do que de quem fará uso deles, foi determinado o uso dessa fonte de dados do detentor governo, de modo a ser comparada com os elementos lexicais do lado produtor rural.

As descrições e a rotulagem dos conjuntos de dados foram usadas em razão de que eles são referências tangíveis ao conteúdo a que têm acesso os usuários. Esses elementos manifestam características dos dados e outorgam uma aproximação das possibilidades de uso daqueles recursos. Conforme com Santos (2013), essas formas de representação são criadas na perspectiva de oferecer condições favoráveis no processo de acesso a recursos informacionais.

Portanto, a prova do modelo foi realizada com dados do tipo não estruturados, dados extraídos do site com informações para produtores rurais, e dados estruturados, recuperados dos rótulos e das descrições de conjuntos de dados. A mineração central de dados das fontes, independentemente da natureza na origem, foi realizada com procedimentos padronizados.

Para a análise da Aderência Lexical, em primeira instância foram extraídas as unidades lexicais, palavras das notícias publicadas e dos conjuntos de dados. Logo após, de forma automatizada, foi descoberta a proximidade das duas fontes com as categorias de necessidades informacionais para agir como marco estruturante da análise.

Foram usadas as bibliotecas da linguagem R '*tidytext'*, '*tm*' e '*philentropy'*; nestas, as funções *'findAssocs'* e *'jaccard'* fizeram trabalho conjunto para a classificação das notícias e conjuntos de dados conforme as necessidades informacionais.

A função '*findAssocs*', permitiu descobrir quais palavras aparecem com maior frequência quando encontrado o nome da categoria de necessidade informacional. Com os resultados da aplicação da função, a lista de palavras foi analisada na ordem decrescente do índice de coocorrência, que na função apresenta-se com valores entre 0 e 1, sendo que valores próximos a 1 indicam maior coocorrência. Como critério de escolha das palavras foram consideradas apenas palavras do tipo substantivo. Com as palavras identificadas foi construído um vetor para cada categoria de necessidade informacional e, mediante a função

'jaccard', foi calculado com qual categoria, cada *dataset* e notícia, tinha proximidade, produzindo assim uma classificação por categorias.

Uma vez classificadas tanto as notícias quanto os conjuntos de dados, foi calculado o índice de similaridade entre as palavras usadas pelo governo nos conjuntos de dados e as notícias publicadas para grupos de produtores, mediante a função *'jaccard'*. Essa função entrega um índice com valor decimal que está entre 0 e 1, sendo que um valor próximo de zero indica baixa similaridade entre os textos comparados.

Conforme a estrutura adotada para os dados durante o pré-processamento, *jaccard* foi observada como a função mais pertinente para calcular a similitude entre os conjuntos de palavras, de tipo assimétrico que foram conformados: conjuntos de palavras advindos das comunicações para produtores, conjuntos de palavras advindos de rótulos e descrições de datasets, e conjuntos de palavras que descrevem necessidades informacionais de produtores.

Na mineração textual existem diversas funções que calculam similaridade entre conjuntos de dados; a escolha da função depende tanto dos fins da análise quanto da estrutura em que se encontram os dados. Há funções direcionadas a dados simétricos, assimétricos, segundo a natureza, dados binários, categóricos, numéricos, ou para combinações de tipos de dados, ainda podendo ser orientadas para análises de dissimilaridade em dados de essas e outras naturezas. Han, Kanker e Pei (2012) indicam *jaccard* como uma opção apropriada para analisar conjuntos de dados como os resultantes da fase de pré-processamento deste estudo. Ainda, Huang (2008) indica que, não objetivando análises de significado nos dados coletados, *jaccard* é uma das opções adequadas.

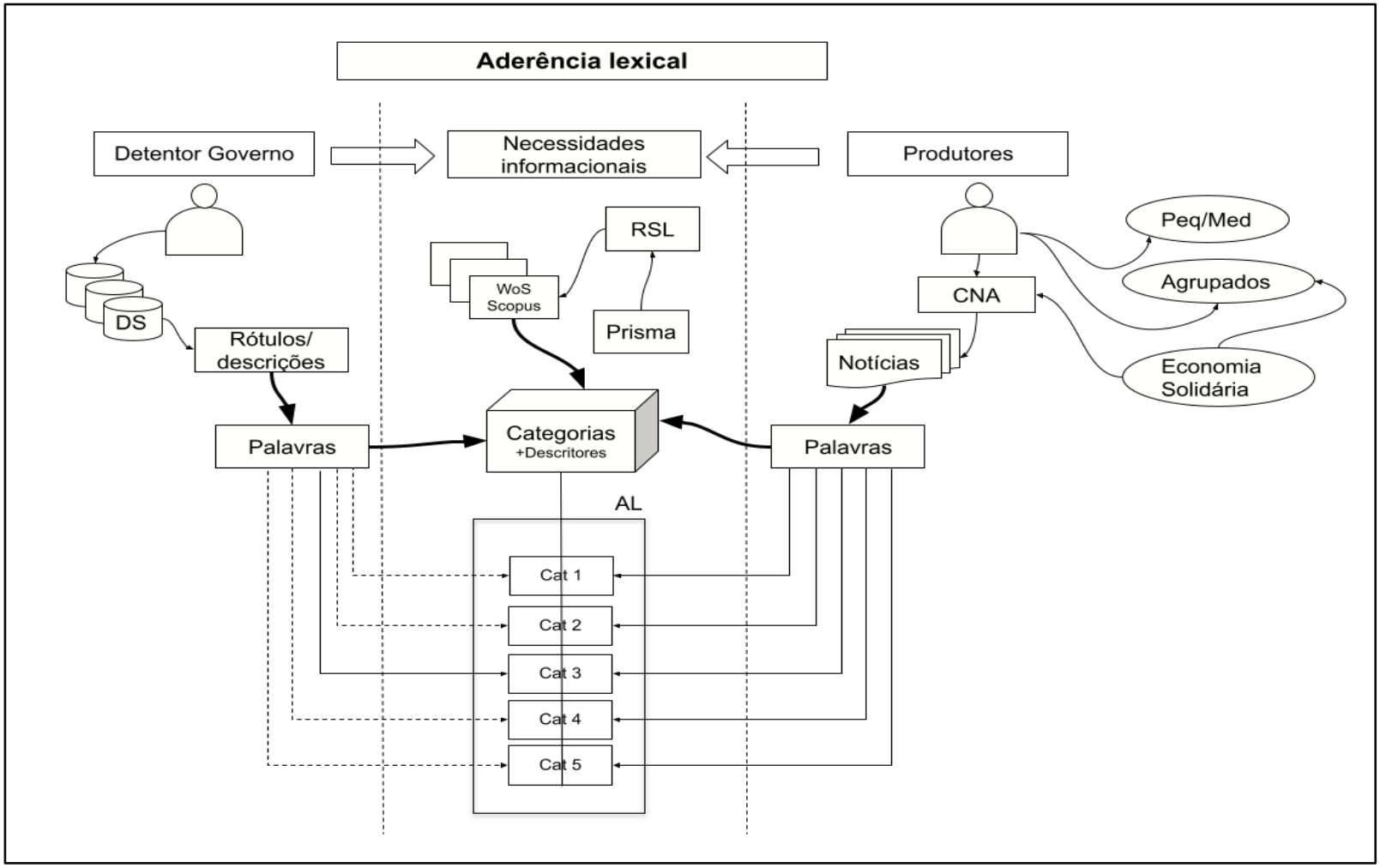
À diferença da análise morfológica, que também considera palavras isoladas em um texto a fim de identificar classes gramaticais, este estudo individualiza as palavras com o intuito de encontrar a proximidade entre o léxico usado nas comunicações orientadas a grupos de produtores rurais e dados publicados por um governo, análise intermediado com um marco estruturante de necessidades informacionais.

Assim, o reconhecimento de indícios de aderência lexical a dados publicados pelo governo do Brasil pôde ser visto em dois níveis, proximidade dos dados de governo com categorias de necessidades informacionais para pequenos e médios produtores, e

proximidade entre unidades lexicais extraídas de dados publicados pelo governo e comunicações publicadas para pequenos e médios produtores. A Figura 2 apresenta uma especialização da proposta considerando as particularidades das duas fontes de dados escolhidas para a prova do modelo.

O estudo de Camperos-Reyes (2023), determinou, mediante RSL, que as categorias de necessidades informacionais mais relevantes para pequenos e médios produtores são na ordem, Mercado, Tratos Culturais, Crédito, Direitos e Oportunidades. Considerando essas categorias para analisar as duas fontes de dados mediante o índice de similaridade, se encontrou, do lado das notícias mineradas, que todas foram abordadas na ordem descendente Mercado, Oportunidades, Crédito, Tratos Culturais e Direitos; do lado dos dados de governo, unicamente foi tratada a categoria Crédito.

Figura 2 – Modelo para análise da Aderência Lexical



Fonte: elaborado pelos autores

A prova do modelo proposto apontou que tanto para a comunidade científica como para uma instituição relacionada com produtores, destacam assuntos categorizados como Mercado, tais como preços, tendências de consumo, publicidade, canais de comercialização etc., enquanto não houve um conjunto de dados de governo que os tratasse. Foi também o caso dos assuntos categorizados como Oportunidades, onde ressaltaram aspectos de associativismo, cooperativismo, negócios, empreendedorismo, e desenvolvimento de carreira, que também não foram identificados na amostra do estudo do lado detentor governo.

Portanto, ao analisar os dados extraídos índices de similaridade no nível lexical, se encontrou que houve proximidade unicamente na categoria Crédito. A Aderência Lexical entre as fontes de dados nessa categoria, aponta que 80.48% das unidades lexicais usadas nos dados de governo, foram similarmente encontradas nas notícias mineradas do site da CNA. Desde o ponto de vista da similaridade vocabular entre as duas fontes, é entendida como favorável em um primeiro nível na busca de inteligibilidade entre estas, como indicado por Sant'Ana (2018), quando prescreve que as formas de publicação dos dados devem contribuir com as possibilidades de interpretação pelos usuários.

É importante frisar a existência de unidades lexicais que foram identificadas unicamente nos conjuntos de dados, que correspondem a siglas usadas no contexto de políticas públicas sobre atributos de programas de governo (BSM, SIATER, SIMOG, UPFS)[2], as quais podem manifestar a necessidade de melhor aproximação por parte dos sujeitos visando o seu potencial uso.

Os resultados assinalam uma oportunidade de melhora na pertinência dos dados que estão sendo disponibilizados pelo governo com as necessidades informacionais dos sujeitos a quem se orientam, apenas na categoria "Crédito" houve possibilidade de análise. As demais categorias de necessidades informacionais (Mercado, Tratos Culturais, Direitos e Oportunidades), apesar de presentes nas notícias da CNA, não foram encontradas nos conjuntos de dados governamentais. É possível afirmar que entre as fontes escolhidas existe

[2] Brasil sem miséria (BSM); Sistema informatizado de ATER (SIATER); Sistema de monitoramento e Gestão da Secretaria Especial de Agricultura Familiar e do Desenvolvimento Agrário (SEAD) (SIMOG); Unidades produtivas familiares (UPFS).

uma comunicação insuficiente ao não tratar assuntos relevantes pelos usuários que eventualmente irão usufruir dos dados publicados.

Outro aspecto a ressaltar foi a quantidade de dados disponíveis do lado detentor governo, magnitude que aponta para novas estratégias metodológicas que permitam coletar dados em outros contextos, resultando em uma maior quantidade e diversificando manifestações da origem dos dados a analisar.

A análise de Aderência Lexical proposta, age como um mecanismo de mensuração da integração do contexto de usuário ao contexto de dados (Santos; Sant'Ana, 2019), usando as unidades lexicais como pontos de comparação e, portanto, é uma forma de interpretação de fluxos informacionais, matéria de estudo da Ciência da Informação conforme Borko (1968); é um resultado que contribui na busca de soluções a situações permeadas pelo uso das TIC no contexto do acesso a dados e a sua potencial interpretação, como auxílio para o direcionamento e tomada de decisões, neste caso, por pequenos e médios produtores.

4 CONSIDERAÇÕES FINAIS

Wersing (1993) propõe somar na freada ao desmoronamento do universo do conhecimento, entendido neste estudo como um incentivo para desenvolver estruturas que contribuam para entender as possibilidades de uso de dados publicados por determinados atores e para determinados usuários, deixando imperar uma sensibilização com aspectos técnicos e sociais do acesso a dados.

Essa proposta de análise desenvolveu um roteiro para abordar elementos teóricos e técnicos que permitiram considerar elementos conceituais, contextos particulares à realidade local, e ferramentas tecnológicas que possibilitaram a culminação do objetivo proposto, revelando que a Aderência Lexical é um instrumento útil para identificar proximidade entre fontes de dados e necessidades informacionais dos sujeitos.

Além da escolha de outras fontes que tencionam a publicação de dados para os sujeitos alvo, e não sobre eles, o modelo admite ampliar o espectro dos dados do detentor, como conjuntos de dados publicados por instituições e não por assuntos, o que pode entregar uma maior quantidade de unidades para serem analisadas, aprofundando nas opções técnicas para análise de aderência lexical, até com trabalho interdisciplinar com áreas como linguística, comunicação e computação, o que pode evoluir para um inter-modelo construído em conjunto com áreas interessadas.

Potencialidades das ferramentas tecnológicas para mineração de dados, particularmente, mineração textual, admitem refletir sobre estudos a serem desenvolvidos com base tanto nos resultados aqui obtidos, quanto sobre novas camadas que possam chegar a atingir um nível da aderência semântica aos dados publicados.

Opções como análise de colocação - contexto em que as palavras aparecem - e da relevância - importância da palavra no contexto específico – se identificam como aplicação futura da proposta no sentido de extrair unidades significativas não unicamente por similitude entre conjuntos e frequências absolutas, e sim de proeminência, assim como a caracterização de assuntos mediante técnicas como *Topic Modelling*, para modelagem temática de corpus.

Estas alternativas mencionadas configuram formas de extrair inferências a partir de fontes de dados dos sujeitos alvo, porém, é necessário verificar se fornecem indícios acerca do usufruto de dados publicados para eles. Novas propostas precisam coletar direta ou indiretamente, amostras que evidenciem o aproveitamento dos dados publicados para eles.

Aponta-se que a busca por indícios das possibilidades de interpretação de dados por produtores levou à decisão de usar manifestações indiretas dessas possibilidades, como é o caso das notícias mineradas do site da CNA, o que naturalmente apresenta vieses técnicos, mas, sobretudo, com uma implementação à distância do próprio sujeito informacional. Abre-se a oportunidade de propor novos projetos na busca de uma visão do que o sujeito manifesta diretamente.

Em suma, os resultados desta pesquisa e estudos futuros, são ainda oportunidades de reflexão sobre a avaliação da proficuidade do volume de dados disponibilizados para sujeitos alvos como os pequenos e médios produtores, setor certamente relevante para países como o Brasil, onde um dos interesses particulares é a potencialidade de uso outorgada pelo tratamento descritivo dado a esses recursos informacionais. Estudos realizados em áreas representativas do desenvolvimento de um país contribuem no robustecimento da Ciência da Informação, gerando contributos que tenham a possibilidade de auxiliar os planos estratégicos em níveis de detentores de dados de governo.

REFERÊNCIAS

BARRETO, A. de A. A questão da informação. **São Paulo em perspectiva**, v. 8, n. 4, p. 3-8, 1994.

BORKO, H. *Information Science: What Is It?* ***American Documentation***, v. 19, n. 1, p. 3–5, 1968.

CAMPEROS-REYES, J. T. **Metadados nas instruções de governos para publicadores de dados**. 2018. Dissertação. (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2018. Disponível em: http://hdl.handle.net/11449/152838. Acesso em: 25 jun. 2024.

CAMPEROS-REYES, J.T.; VECHIATO, F. L.; VIDOTTI, S. A. B. G.; SANTANA, R. C. G. Encontrabilidad de la información en sites que promueven Datos Abiertos. ***Palabra Clave (La Plata)***., v.10, p.e109, 2020. Disponível em: https://www.palabraclave.fahce.unlp.edu.ar/article/view/PCe109. Acesso em: 24 jun. 2024

CAMPEROS-REYES, J. T. **Aderência Lexical a dados publicados para produtores rurais**. 2023. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2023.Disponível em: http://hdl.handle.net/11449/242749. Acesso em: 25 de jun. 2024.

CARVALHO, A. M. G. de. **Apropriação da informação: um olhar sobre as políticas públicas sociais de inclusão digital**. 2010. Tese (Doutorado) – Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010. Disponível em: http://hdl.handle.net/11449/103358.

COMISSÃO ECONÔMICA PARA A AMÉRICA LATINA E O CARIBE. ***La paradoja de la recuperación en América Latina y el Caribe*** *- Crecimiento con persistentes problemas estructurales: desigualdad, pobreza, poca inversión y baja productividad*. 2021. Disponível em: https://repositorio.cepal.org/bitstream/handle/11362/47043/5/S2100379_es.pdf.

DAVENPORT, T. H. **Ecologia da informação**: porque só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998.

FELDMAN, R.; SANGER, J. 2006. **The text mining handbook**: advanced approaches in analyzing unstructured data. Cambridge University Press, 2006. Disponível em: https://dl.icdst.org/pdfs/files/25a6d982ee80e1db7a4ebf7eeca4e0ec.pdf.

FERRÁS SEXTO, C.; GARCÍA, Y.; POSE, M. New ways to buy and sell: an information management web system for the commercialization of agricultural products from family farms without intermediaries. *In*: CRUZ-CUNHA, M. M.; MOREIRA, F. (Ed.). **Handbook of research on mobility and computing**: evolving technologies and ubiquitous impacts. Hershey: Information Science Reference, 2011. p. 1182–1198.

GÓMEZ, M. N. G de. As relações entre ciência, Estado e sociedade: um domínio de visibilidade para as questões da informação. **Ciência da Informação**, v. 32, p. 60-76, 2003. Disponível em: https://www.scielo.br/j/ci/a/KQV7G77RcxK7F6cCH96pVDb/?lang=pt.

GRAY, J. *et al*. **Unlocking the potential of aid information**. 2009. Disponível em: https://web.archive.org/web/20130122095149/http://www.unlockingaid.info/wp-content/uploads/2010/02/UnlockingAidInformation.pdf. Acesso em: 27 jun. 2024.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3 ed. Waltham: Morgan Kaufmann. 2012.

HEARST, M. What is text mining. **SIMS**, UC Berkeley, v. 5, 2003. Disponível em: https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf.

HUANG, A. Similarity measures for text document clustering. In: PROCEEDINGS OF THE SIXTH NEW ZEALAND COMPUTER SCIENCE RESEARCH STUDENT CONFERENCE (NZCSRSC2008), 6., 2008, Christchurch, New Zealand. **Anais** [...] Christchurch, 2008. p. 9-56.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. 2021. Disponível em: https:// www.ibge.gov.br. Acesso em 27 jun. 2024.

KAO, A.; POTEET, S. R. (Ed.). **Natural language processing and text mining**. USA: Springer, 2007.

KORN, J.; HUSS, F.; CUMBERS, J. D. Natural language for modelling situations. In: **IEE Colloquium on Natural Language Understanding**. London: IET, 1988.

MARTINES, A. R.; MOREIRA, W.; ALMEIDA, C. C. de. Do signo ao tesauro: contribuições de três correntes da linguagem. **Ciência da Informação**, v. 51, n. 1, 2022. Disponível em: https://revista.ibict.br/ciinf/article/view/5543. Acesso em 27 jun. 2024.

RUEDIGER, M. A. Perspectivas da governança na era da informação: estado e sociedade civil. In: MARTINS, P. E. M.; PIERANTI, O. P. (org.) **Estado e gestão pública**: visões do Brasil contemporâneo. Rio de Janeiro: FGV, 2006.

SANT'ANA, R. C. G. Transdução Informacional: impactos do controle sobre os dados. In: MARTÍNEZ-ÁVILA, D; SOUZA, E.A.; GONZALEZ, M. E. Q. (org.) **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução? Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2019. p.117-128. ISBN 978-85-7249-054-2.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A ALIMENTAÇÃO E A AGRICULTURA. 2023. Disponível em: https://www.fao.org/brasil/pt/. Acesso em: 27 jun. 2024.

SANTOS, P. L. V. A. C. Catalogação, formas de representação e construções mentais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 6, n. 1, 2013. Disponível em: http://hdl.handle.net/20.500.11959/brapci/119476. Acesso em: 27 jun. 2024.

SANTOS, P. L. V. da C.; SANT'ANA, R. C. G. Camadas de representação de dados e suas especificidades no cenário científico. In: DIAS, G. A.; OLIVEIRA, B. M. J. F. de. **Dados científicos**: perspectivas e desafios. João Pessoa: Editora UFPB, 2019. p. 53-66.

SARACEVIC, T. Ciência da informação: origem, evolução e relações**. Perspectivas em ciência da informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22308. Acesso em: 27 jun. 2024.

WERSIG, G. Information science: the study of postmodern knowledge usage. **Information processing & management**, v. 29, n. 2, p. 229-239, 1993. Disponível em: https://www. sciencedirect.com/science/article/abs/pii/030645739390006Y. Acesso em: 27 jun. 2024.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Código de Financiamento 001.