REB
REVISTA DE ESTUDIOS BRASILEÑOS
e-ISSN: 2386-4540
https://doi.org/10.14201/reb20231021151167

Ciências Sociais Computacionais e análise de conteúdo: reflexões a partir da produção latino-americana

Ciencias Sociales Computacionales y Análisis de Contenido: reflexiones a partir de la producción latinoamericana

Social Science Computing and Content Analysis: reflections based on Latin American production

AUTORES

Gleidylucy Oliveira*
gosilva@ufscar.br

Rafael Cardoso Sampaio**
rafael.sampaio@ufpr.br

* Professora adjunta da Universidade Federal de São Carlos (UFSCar, Brasil).

** Professor adjunto do Departamento de Ciência Política da Universidade Federal do Paraná (UFPR, Brasil).

RESUMO:
As Ciências Sociais Computacionais (CSC) têm emergido como um campo híbrido formado pela intersecção das Ciências Sociais e da Ciência da Computação, e que se desenvolve pela ampliação da capacidade de análise dos pesquisadores pelos computadores e pelo exponencial crescimento de dados digitais, além de pesquisas baseadas em simulações computacionais baseadas em agentes. Nesse sentido, diversos temas, áreas e metodologias das humanidades têm sido impactadas. Um dos campos potencializados, nesse sentido, é o estudo de objetos sociais/políticos a partir da comunicação humana com Análise de Conteúdo. Apesar de não ser um método recente, pesquisadores e analistas de conteúdos lidam com dificuldades e limites da pesquisa causados pela subjetividade e replicabilidade e têm visto na automatização por meio de computadores a superação desta questão. Assim, buscamos identificar como se deu a incorporação de uma metodologia tradicional pelas CSC na América Latina buscando escrutinar como os cientistas sociais estão operacionalizando as transições teóricas/epistemológicas nesse campo em formação. Para tanto, fizemos análise cientométrica de artigos publicados por instituições e pesquisadores da região e os dados demonstram uma bibliografia composta de autores mais tradicionais das humanidades, mas com forte incorporação metodológica das técnicas da Ciência da Computação.

RESUMEN:
Las Ciencias Sociales Computacionales (CSC) han emergido como un campo híbrido formado por la intersección de las Ciencias Sociales y las Ciencias de la Computación, y que se desarrolla por la expansión de la capacidad de análisis de los investigadores por el uso de ordenadores y el crecimiento exponencial de los datos digitales, así como de la investigación con simulaciones informáticas basadas en agentes. En este sentido, se han visto afectados diversos temas, áreas y metodologías de las humanidades. Uno de los campos potenciados ha sido el estudio de los objetos sociales/políticos de la comunicación humana con el Análisis de Contenido. Aunque no sea un método reciente, los investigadores y analistas de contenido hacen frente a dificultades y limitaciones derivadas de la subjetividad y la replicabilidad, encontrando en la automatización mediante ordenadores la vía para la superación de estos problemas. Así, buscamos identificar cómo se produjo la incorporación de una metodología tradicional por parte de la CSC en América Latina, escudriñando cómo los científicos sociales operacionalizan las transiciones teórico/epistemológicas en este campo en formación. Realizamos un análisis cienciométrico de artículos publicados por instituciones e investigadores de la región. Los datos ponen en evidencia una bibliografía compuesta por autores de humanidades más tradicionales, pero con una fuerte incorporación metodológica de técnicas de Ciencias de la Computación.

ABSTRACT:
Social Science Computing (SSC) emerged as a hybrid field formed by the intersection of Social and Computer Sciences, and which develops itself through researchers' ability to analyze computers and the exponential growth of digital data's expansion, as well as research based on agent-based computer simulations. In this sense, several themes, areas and methodologies of the humanities have been impacted. In this context, the study of social/political objects based on human communication with Content Analysis is one of the potential fields. Despite not being a recent method, researchers and content analysts deal with research difficulties and limitations caused by subjectivity and replicability of these studies and have seen automation through computers as an overcoming of this issue. Thus, we seek to identify how the incorporation of a traditional methodology by the SSC took place in Latin America, seeking to investigate how social scientists are operationalizing the theoretical-epistemological transitions in this still developing field. For that, we performed a scientometric analysis of articles published by institutions and researchers in the region and the data demonstrate a bibliography composed of more traditional authors from the humanities, but with a strong Computer Science techniques methodological incorporation.

PALAVRAS-CHAVE
Ciências Sociais Computacionais; Análise de Conteúdo; América Latina; *Text as Data*; PLN.

PALABRAS CLAVE
Ciencias Sociales Computacionales; Análisis de Contenido; América Latina; *Text as Data*; PLN.

KEYWORDS
Social Science Computing; Content Analysis; Latin America; Text as Data; PLN.

Recibido:
10/01/2023
Aceptado:
20/06/2023

1. Introdução

As Ciências Sociais Computacionais (CSC) têm emergido e se consolidado como um campo híbrido e ainda em construção. Diferentes autores têm se alternado na tentativa de, a partir de seu avanço e aplicação, definir os seus aspectos mais fundamentais no sentido da concretização deste campo. Entretanto, o que a literatura especializada consegue concordar nesse momento é que este é um espaço científico formado pela intersecção das Ciências Sociais e da Ciência da Computação, trazendo ainda contribuições de conhecimentos e expertises adjacentes, como Estatística, Matemática, Física, Linguística, Engenharia etc. (Conte *et al.*, 2012; Lazer *et al.*, 2020; Cioffi-Revilla, 2017; Edelman, Wolff, Montagne & Bail, 2020; Salganik, 2018). Importante frisar que desde muito tempo as Ciências Sociais se utilizam do conhecimento matemático e estatístico, além do uso de softwares e computadores, por exemplo, para realização de suas pesquisas, em especial, aquelas de abordagem quantitativa. Mas o que ocorre de diferente – e é nisso que a literatura tem se debruçado – é a emergências de objetos, métodos e teorias que antes não eram possíveis de existir ou serem realizados/testados e atualmente podem (Conte *et al.*, 2012; Edelman, Wolff, Montagne, & Bail, 2020). Assim, vale destacar a definição de Lazer *et al.* (2020), que num esforço pioneiro de caracterização, apontavam que a "a ciência social computacional é um campo interdisciplinar que avnaça nas teorias do comportamento humano por meio da aplicação de técnicas computacionais para grandes conjuntos dados oriundos de sites de mídia social, da Internet ou de outros arquivos digitalizados, como como registros administrativos" (Lazer *et al.*, 2020, p. 1060, tradução livre do original em inglês).

Este é outro consenso mínimo das CSC: elas se desenvolvem especialmente, por um lado, pela potencialização da capacidade de análise dos pesquisadores a partir do desenvolvimento de ferramentas computacionais e de programação e, por outro, pelo exponencial crescimento da capacidade de armazenamento e captação de dados no ambiente digital (Salganik, 2018). Acrescenta-se mais recentemente a possibilidade também de pesquisas a partir da modelagem computacional baseada em agentes.

Nesse ambiente, e com novas sociabilidades, as ferramentas tradicionais das Ciências Sociais precisam superar dificuldades de processamento, coleta e análise de dados pelo seu volume, mas também pelos diferentes fenômenos despertados nas interações digitais, resultando não apenas na revisão ou confirmação de teorias, mas no surgimento de outras. Nesse sentido, Lazer *et al.* (2009) apontam que o campo requer a necessidade de formação de cientistas sociais com habilidades computacionais e/ou cientistas da computação com conhecimentos em Ciências Humanas, o que traz desafios institucionais, mas também humanos, em especial, para cientistas sociais já formados. Um deles é a capacidade de integrar conhecimentos e expertises de forma interdisciplinar.

Nesta direção, um dos campos potencializados pelo surgimento e o desenvolvimento das CSC tem sido o estudo de objetos e problemas sociais e políticos a partir da linguagem e da comunicação humana. E um dos métodos mais consolidados nas Ciências Sociais[1] para tratar deste tipo de dado em todo o mundo é a Análise de Conteúdo (AC). Apesar de não ser recente (Bardin, 2008; Krippendorff, 2004), durante muito tempo, esse conjunto de técnicas amplamente difundido enfrentou dificuldades e limites de pesquisa causados pela necessidade de cotejamento da subjetividade, validação, replicabilidade dos resultados, bem como tratamento de um universo de dados amplo. Isso resultou em importantes críticas à cientificidade e à objetividade do método, dado que todo analista também é um leitor "embebido" da linguagem e suas nuances (Krippendorff, 2004; Sampaio & Lycarião, 2021). Além disso, o descompasso entre a capacidade de análise dos pesquisadores e a produção de dados comunicacionais era outro

entrave para o tratamento de *corpus* amplos e para a generalização dos resultados. Diante disso, analistas têm visto o crescimento das CSC e as ferramentas de automatização de análise textual e processamento de linguagem natural como um espaço importante de superação destas dificuldades e potencialização das técnicas da AC.

Dito isto, o objetivo deste artigo foi identificar se e em que medida há a incorporação de uma metodologia tradicional – a AC – pelas CSC, considerando como pressuposto que parte significativa da conformação do campo das CSC atualmente consiste na construção dessas relações entre as práticas já estabelecidas e seu incremento teórico e metodológico a partir da computação. Além disso, queremos trazer um recorte regional entendendo que grande parte desse movimento de virada no campo – e quem sabe de paradigma – vem sendo liderada por universidades e pesquisadores do Norte global. Assim, buscamos escrutinar se e como os cientistas sociais da América Latina/Sul global estão operacionalizando as transições teóricas e epistemológicas provocadas pela consolidação desse campo no que se refere à AC.

Para tanto, algumas questões guiarão nossa análise, como: quais autores são utilizados nos trabalhos para balizar as análises, quais as principais técnicas utilizadas, quais países e instituições mais têm se dedicado a esse movimento e quais temáticas/objetos aparecem mais relacionadas. Importante destacar que, para nós, esse olhar é relevante, porque visa acompanhar o próprio processo de adequação do campo às transformações e desafios trazidos pela CSC, escrutinando a consolidação desta abordagem na região. Assim, entendemos como um primeiro esforço para caracterizar esse movimento de forma a apontar para possíveis questões que poderão ser respondidas em trabalhos futuros sobre o tema.

Assim, realizamos uma análise cientométrica de artigos disponíveis na SciELO, no período de 2012 a 2021, por meio da *Web of Science*, utilizando o software VOSviewer versão 1.6.16. Para considerar o recorte regional, filtramos o país de origem da publicação e dos autores, bem como textos em inglês, espanhol e português, excetuando os países que utilizam essas línguas fora da região[2]. Além disso, utilizamos *strings* relacionadas a um conjunto de descritores ligados às CSC e AC – em especial a Análise de Conteúdo Automatizada – e suas respectivas variações na língua inglesa e na espanhola[3]. A partir desse tratamento, o *corpus* de análise deste trabalho reúne 471 artigos congregando autores de instituições de cinco países: Argentina, Brasil, Chile, Colômbia e México.

Por fim, este trabalho é composto de três partes principais. Na primeira, apresentamos uma revisão na literatura internacional e nacional sobre AC enquanto método e sobre o surgimento, conformação e consolidação das CSC e sua interface com a AC por meio das análises automatizadas e *Text as Data* (logo, não consideramos *Text as Data* como sinônimo de AC, mas como um dos tipos de técnicas que podem ser utilizadas para AC). A segunda parte é composta por metodologia, descrição dos dados e resultados, seguidos das primeiras impressões das análises. E, na terceira parte, fazemos as considerações finais apontados para alguns achados que consideramos relevantes, como a forte presença de trabalhos de instituições brasileiras, bem como que a incorporação da AC nas CSC vem no sentido contrário – ou seja, da incorporação das técnicas desenvolvidas pela Ciências da Computação e *Text as Data* por analistas de conteúdo. Não fica claro pelos dados coletados e pela análise empregada que pressupostos e condições da AC orientam a escolha e uso das técnicas identificadas. Além disso, há uma forte presença de técnicas de análise textual ou *Text as Data*[4] sem isso se converter, necessariamente, em uso da literatura de análise textual. Esta pode até ter sido utilizada, mas não aparece como bibliografia significativa ou correlacionada estatisticamente.

Por fim, entendemos que este trabalho vem coadunar com os objetivos desse dossiê por trazer um olhar amplo sobre a combinação de Ciências Sociais e Política e Computação olhando para dentro do próprio campo científico e para cientistas sociais latino-americanos, buscando não apenas fazer uma reflexão sobre as técnicas, mas também teórica-epistemológica sobre as CSC, AC e apresentar dados e reflexões ainda não estabelecidas nos trabalhos do campo.

2. Linguagem, política e análise de conteúdo

A comunicação, a política e as sociabilidades andam *pari passu*. Por meio das trocas simbólicas e comunicativas – verbal, escrita ou visual – os atores interagem e os processos sociais se desenrolam. É por isso que, desde muito tempo, a comunicação e seus conteúdos despertam o interesse de investigadores das Ciências Sociais e produzem esforços de diferentes áreas no sentido de sua coleta, organização e análise (Bourdieu, 1989; Habermas, 1984; Lasswell, 1978; Lipman, 2008; Sartori, 1998).

Um dos métodos mais importantes – pela sua difusão e tempo de desenvolvimento – é, sem dúvida, a Análise de Conteúdo (AC). Como apontam diferentes autores (Bardin, 2008; Krippendorff, 2004; Neuendorf, 2017), há registros de uso da AC há séculos, mas apenas após 1920 é que esta se desenvolve a partir da análise das comunicações de guerra e ampliação dos *media* ao redor do mundo, promovendo o aumento massivo da circulação das informações (Sampaio & Lycarião, 2021). Sua aplicabilidade se estabeleceu a partir da necessidade de inferir os conteúdos de diferentes tipos de comunicação – verbal e não verbal – a partir de uma hermenêutica controlada e procedimentos replicáveis. Assim, "a análise de conteúdo é um método de pesquisa observacional utilizado para avaliar sistematicamente o conteúdo simbólico de todas as formas de comunicação registrada" (Kolbe & Burnett, 1991, p. 243, tradução da autoria).

Nesse sentido, um primeiro ponto fundamental de toda AC é o seu caráter procedimental controlado, onde etapas e condições devem ser atendidas para permitir o controle da interpretação e subjetividade dos pesquisadores[5]. Entretanto, mesmo com este aspecto de "etapas e procedimentos controlados", o método e as diferentes técnicas que encerram não passaram ilesos de problemas quanto à objetividade, replicabilidade dos resultados ou validade das pesquisas. Nesse sentido, Krippendorff destaca que diferentes problemas podem "enviesar" os resultados de uma análise de conteúdo, como

> os analistas de conteúdo podem discordar das interpretações de um texto. As instruções de codificação podem não ser claras. As definições das categorias podem ser ambíguas ou não parecerem aplicáveis ao que se supõe descrever. Os codificadores podem ficar cansados, tornarem-se desatentos a detalhes importantes ou possuir diferentes inclinações. Dados não confiáveis podem levar a resultados de pesquisa incorretos (Krippendorff, 2004, p. 1, tradução da autoria).

É ainda a partir da constatação das dificuldades para a execução de uma AC com alto grau de confiabilidade que Grimmer e Stewart (2013) apontam que toda análise de conteúdo seria "incorreta" e precisaria de processos de validações e correções sucessivos para mitigar suas imprecisões. Entretanto, apesar de complexo, esse trabalho não é impossível e, nesse sentido, diferentes esforços foram e vêm sendo desenvolvidos no sentido de atender às condições de cientificidade, objetividade e validade da AC, desde processos coletivos de refinamento até testes estatísticos sobre os resultados do trabalho dos codificadores (Lima, 2013; Sampaio & Lycarião, 2021). Estes testes mediriam assim o grau de coerência entre os resultados encontrados e objetivos investigados, bem como a estabilidade desses resultados ao longo do tempo e de diferentes codificadores, e ainda o grau de precisão (*accuracy*) das classificações realizadas[6].

Outro caminho que tem surgido nesse sentido é o desenvolvimento de técnicas de análise automatizada de conteúdo em que, por meio do uso de softwares e linguagens computacionais, analistas e pesquisadores têm reduzido o trabalho de codificação e tratamento humano dos dados, contornando em alguma medida a arbitrariedade de parâmetros. Na Figura 1, Grimmer e Stewart (2013) apontam uma classificação destes métodos de acordo com os objetivos da análise, bem como a presença ou ausência de supervisão humana do processo de tratamento computacional.

Importante destacar que este conjunto de técnicas, oriundos do desenvolvimento de ferramentas pela Ciência da Computação, também tem sido nomeado como o campo do *Text as Data* ou AT (Benoit, 2020; Izumi & Moreira, 2018; Moreira, Pires, & Medeiros, 2022). No texto que citamos acima de Grimmer e Stewart (2013), os autores utilizam os termos análise textual e análise de conteúdo automatizada como sinônimos[7], entretanto, esse ponto parece ainda não ser consenso, dado que em trabalhos mais recentes

essa semelhança acaba sendo substituída apenas pela nomeação do campo como "*Text as Data*" (Benoit, 2020).

Para além dessa discussão – que acreditamos que ainda levará um tempo a se resolver e dependerá, em certa medida, da própria construção do campo e do trabalho cotidiano dos cientistas sociais no uso de técnicas da computação para observar e analisar nosso objetos – *Text as Data* não pode ser confundido completamente com Análise de Conteúdo Automatizada, dado que a Análise de Conteúdo não se debruça apenas sobre *corpus* textuais, dando conta de um conjunto mais amplo de signos comunicacionais ou mesmo expressões não linguísticas, como gestos, comportamentos, imagens, sintomas patológicos etc. Isto porque estes códigos também carregam conteúdos, tornando-se passível de inferência e técnicas para identificá-los.

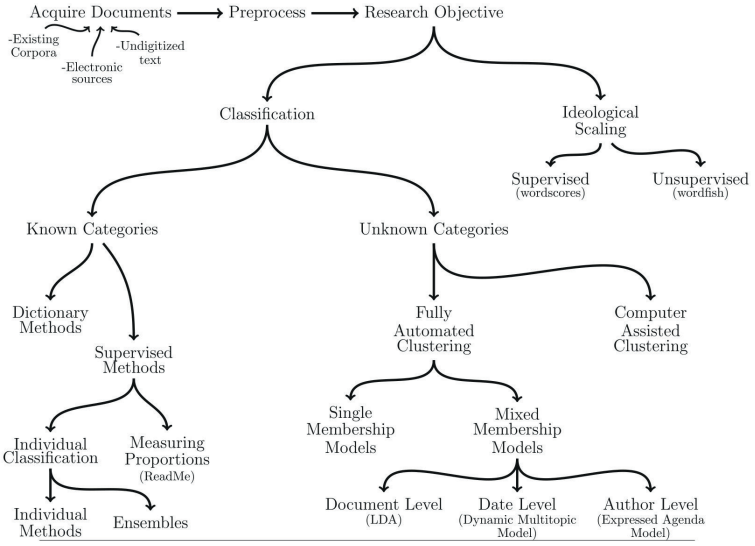

Figura 1. Classificação dos Métodos de Análise Textual Automatizado de acordo com Grimmer e Stewart (2013). Fonte: Grimmer e Stewart (2013, p. 2).

Quadro 1. Domínios possíveis da AC

CÓDIGO E SUPORTE	UMA PESSOA	COMUNICAÇÃO DUAL	GRUPO RESTRITO	COMUNICAÇÃO DE MASSA
Linguístico				
Escrito	Agendas, pensamentos, conjecturas, diários	Cartas, respostas a questionários e teste projetivos, trabalhos escolares	Todas as comunicações escritas trocadas dentro de um grupo	Jornais, livros, anúncios publicitários, cartazes, literatura, textos jurídicos, panfletos etc.
Oral	Delírios do doente mental, sonhos	Entrevistas e conversações de qualquer espécie	Discussões, entrevistas em grupo, conversações de qualquer espécie em grupo	Discursos, exposições orais, palestras, televisão, rádio, publicidade
Icônico				
Sinais, grafismos, imagens, fotografias, filmes etc.	Rabiscos mais ou menos automáticos, sonhos	Respostas a testes projetivos, comunicação entre duas pessoas por meio de imagens	Toda comunicação icônica em um pequeno grupo	Sinais de trânsito, cartazes, cinema, publicidade, televisão
Outros códigos semióticos				
Tudo o que não é linguístico e pode ser portador de significações (ex.: notas musicais, código olfativo, objetos diversos, comportamento etc.)	Manifestações de doenças mentais, posturas, tiques, gestos, dança, coleções de objetos	Comunicação não verbal com destino a outrem (postura, gestos, distância espacial, manifestações emocionais, objetos cotidianos, vestuário, alojamento etc.), comportamentos diversos como ritos e regras de cortesia		Meio físico e simbólico: sinalização urbana, monumentos, arte, mitos, estereótipos, instituições, elementos da cultura

Fonte: elaboração própria a partir de Bardin (2008).

Além disso, a Análise Textual ou "*Text as Data*" usa como unidade de registro e de análise a mesma dimensão: a palavra ou o conjunto de caracteres ordenados de um texto, como a frase (Izumi & Moreira, 2018; Moreira, Pires, & Medeiros, 2022). O principal pressuposto da Análise Textual é que os conteúdos das trocas comunicativas podem ser encontrados por meio de análises matemáticas sobre esse conjunto lexical, que passam a ser tratadas estatisticamente por meio da contagem de frequência e testes multivariados, sendo analisadas sozinhas e/ou em contexto. Ou seja, mesmo em análises de redes semânticas, clusters ou de tópicos (Blei, Ng, & Jordan, 2003; Camargo & Justo, 2013; Cúrcio, 2006; Lebart & Salem, 1994), em que os termos de um conjunto de textos são analisados em relação com as outras palavras estatisticamente relevantes para inferir os sentidos compartilhados de forma mais ampla e contextual, a unidade é a palavra[8].

Já na AC, as unidades de análise podem variar de tamanho e dimensão. Pensando especificamente em AC de textos, estas podem ser a palavra e a frase, mas também, o parágrafo, a *quasi-sentence*, entre outros. Assim, não precisa ir muito além para perceber que a análise de conteúdos textuais a partir das palavras é um tipo de abordagem que coaduna perfeitamente com a AC. Contudo, a análise de conteúdo não é sinônimo apenas de análise textual (AT) ou *text as data*. Assim, AT é, no nosso entendimento, um *subset* (não exclusivo) da AC, em especial, da Análise de Conteúdo Automatizada.

Por outro lado, é importante ainda acrescentar que a Análise de Conteúdo Automatizada pode variar desde o uso de técnicas computacionais interessadas em organizar e classificar os dados comunicacionais até outros modelos que buscam prever e propor comunicações com o uso massivo de aprendizagem de máquina (*machine learning*)[9] e seus desdobramentos[10]. Assim, esse campo é composto de um espectro de objetivos e técnicas científicas de diferentes graus de sofisticação, que variam de uma maior interferência e ação do pesquisador sobre o trabalho com os dados para uma menor presença humana e tarefas cada vez mais automatizadas pelos computadores, que aprendem sobre a linguagem humana à medida que trabalham com os dados e entregam *outputs* cada vez mais robustos[11]. Entretanto, isso não resulta numa menor relevância do cientista social.

> Os métodos de aprendizado de máquina são idealmente adequados para auxiliar os pesquisadores a aproveitarem ao máximo a abundância de dados. No entanto, é importante ressaltar que as ferramentas de aprendizado de máquina são apenas instrumentos e não uma nova e mágica solução que resolve os problemas de longa data enfrentados pelos cientistas sociais. O aprendizado de máquina obterá os melhores resultados quando aplicado de maneira apropriada aos problemas de pesquisa. (...) Enquanto nos concentramos no trabalho empírico, a abundância de dados e as novas aplicações tornaram a teoria – e a teoria formal em particular – em algo ainda mais importante para as ciências sociais (Grimmer, Roberts e Stewart, p. 414, tradução livre do original em inglês).

Fato é que a apropriação rápida e massiva de técnicas de computação para o processamento de dados textuais tem gerado transformações ainda nas tarefas dos cientistas em geral envolvidos na sua realização (Grimmer, Roberts, & Stewart, 2022), quer seja um cientista da computação que trabalha com Processamento de Linguagem Natural (PLN) aplicada aos textos ou, no nosso caso, os cientistas sociais. Nesse sentido, cabe apontar que

> embora os cientistas da computação geralmente (mas não exclusivamente!) estejam interessados em recuperação de informações, sistemas de recomendação e tarefas linguísticas de referência, uma comunidade diferente está interessada em usar "texto como dados" para aprender sobre fenômenos previamente estudados, como nas áreas de ciências sociais, literatura e história (Grimmer, Roberts, & Stewart, 2022, p. 4, tradução livre do original em inglês).

Essas transformações apontadas pelos autores são resultado do movimento de um campo em formação, como o das CSC, mas também do processo de incorporação e desenvolvimento de métodos sobre perspectivas de pesquisa mais consolidadas. Com isso, e instigados por essas reflexões, este artigo buscou investigar se e como a AC vem sendo incorporada pelas CSC (ou vice-versa), olhando de forma mais específica para a análise textual e automatizada, dada a incorporação das técnicas e o forte desenvolvimento da computação e do PLN. Neste sentido, na seção a seguir apontamos como realizamos a análise que orienta este artigo, bem como apresentamos os resultados preliminares da pesquisa.

3. AC e CSC na produção da América Latina

Para realizar os objetivos deste artigo, buscamos verificar diferentes termos e expressões que pudessem remeter a qualquer tipo de análise de conteúdo e/ou de análise textual automatizada ou computadorizada em conjunto com CSC nas três línguas de interesse: português, espanhol e inglês.

Um primeiro desafio da pesquisa era justamente acessar bases que fossem representativas da produção latino-americana e que também apresentassem uma boa organização dos metadados para o uso de softwares cientométricos. Acabamos decidindo pela base SciELO, que é aberta e abrangente em termos dessa produção, apesar de sua óbvia concentração da produção brasileira[12]. Para tanto, fizemos a extração através da coleção SciELO na *Web of Science*.

Outro desafio foi o fato de termos um número muito maior de publicações sobre análise de conteúdo (Sampaio, Lycarião, Codato, Marioto, Bittencourt, Nichols, & Sanchez, 2022). A referência a Bardin e outros autores da área tenderia a se mostrar muito grande para ser incluída no *corpus*, então preferimos não utilizar o termo. Realizamos a seguinte *string* de busca com termos em português e espanhol (ou inglês, quando notamos se tratar de termo ainda muito utilizado pela literatura):

> ("análise de conteúdo computadorizado" OR "análise de conteúdo automática" OR "análise de conteúdo automatizada" OR "Ciências Sociais Computacional" OR "Ciências Sociais Computacionais" OR "Text as Data" OR "Texto como dado" OR "Text Mining" OR "mineração de texto" OR "análise semântica" OR "redes semânticas" OR "análise textual" OR "análise léxica" OR "natural language processing" OR "processamento natural de linguagem" OR "análise de sentimento" OR "análisis de contenido computarizado" OR "análisis de contenido automatizado" OR "Ciencias Sociales Computacional" OR "Ciencias Sociales Computacionales" OR "Texto como Datos" OR "minería de textos" OR "análisis semântico" OR "redes semânticas" OR "análisis textual" OR "procesamiento natural del lenguaje" OR "sentiment analysis" OR "análisis de los sentimientos" OR "análisis léxico") NOT ("análise textual discursiva" AND "análisis textual discursivo")[13].

A busca foi realizada em 16 de agosto de 2022 e os resultados iniciais apontam 889 referências. Posteriormente, excluímos a produção de 2022 (incompleta) e de todas as coleções da SciELO que não fossem de países latino-americanos, nomeadamente Portugal, Espanha e África do Sul. Também filtramos apenas artigos de pesquisa e de revisão e apenas textos em português e espanhol. Finalmente, feitos todos esses filtros, também retiramos pesquisadores com filiações de outros países de fora do continente. Todos esses passos foram para nos limitarmos à autores e revistas latino-americanas. Após tais filtros e exclusões, chegamos a um total de 471 referências. Esses dados foram analisados no Excel e no SPSS para avaliar frequência e no VOSviewer versão 1.6.16 para análises bibliométricas (Eck & Waltman, 2010).

3.1. Resultados

Dentro do esperado, os textos sobre os assuntos de análises automatizadas de textos vêm crescendo ao longo do tempo. Em 2002, eram apenas três artigos sobre os vários termos pesquisados, enquanto em 2021 foram 48. Observando o Gráfico 1, a inflexão ocorre claramente a partir de 2010, quando o número de estudos passa a ser duas a três vezes maior que nos anos anteriores.

Os dados mostram que grande parte da produção coletada é brasileira. Pode ser que, de fato, o Brasil seja um produtor naturalmente superior de textos acadêmicos na área, dado o tamanho da população. Porém, é também possível que a SciELO ainda não tenha a mesma importância para o restante dos países da região como possui para a ciência brasileira. De toda sorte, 219 trabalhos foram oriundos do Brasil, 57 do Chile, 49 da Colômbia, 46 do México, 22 da Argentina e outros 22 de Cuba.

O mesmo padrão se repete nas afiliações dos autores, dos quais 12 das principais 21 instituições de pesquisa são brasileiras. Em segundo lugar, o Chile contou com quatro universidades, e, em seguida, a Colômbia com três.

Ao fazermos as análises de cocitação[14] de referências, notamos que não há ligações o suficiente para termos uma rede, o que evidencia que as diferentes literaturas não possuem tantas conexões. Optamos, então, por fazer a rede de cocitação de autores, na qual todas as obras dos mesmos autores são aglutinadas.

Os resultados são apresentados na Figura 3. O maior cluster, vermelho, é formado por onze referências. As principais brasileiras são Bardin e seu manual de análise de conteúdo, textos de Camargo e Justo sobre o uso do software *Iramuteq* para análises textuais léxicas, além de algumas referências ao trabalho de Valentim na ciência da informação sobre organização de bancos de dados e inteligência competitiva. Uma outra parte da referências deste cluster está em espanhol e tratam de análises de redes semânticas, como Valdez, Vera-Noriega e Figueroa, sendo a maioria abaixo dos anos 2000 (mais antigas). Ademais, temos autores mais teóricos sobre representações sociais, como Jodelet, Bourdieu e Moscovici.

O segundo maior cluster, verde, é formado por nove referências. Seria o cluster mais linguístico de nosso *corpus*, apresentando autores como Chomsky, Parodí, Lakoff, Halliday, Jurafsky e Kintsch, porém também vemos discussões sobre análises semânticas latentes nos trabalhos de Landauer, Venegas e Deerwester, sendo boa parte das referências também antigas.

O terceiro cluster, azul, é formado por seis referências e faz alusão a diferentes vertentes da análise de discurso, notadamente no uso de Fairclough, Foucault, Maldavsky e Van Dijk, além de referências à Psicanálise e Sociologia das sociedades modernas, citando Bauman e Freud, além de Maldavsky novamente.

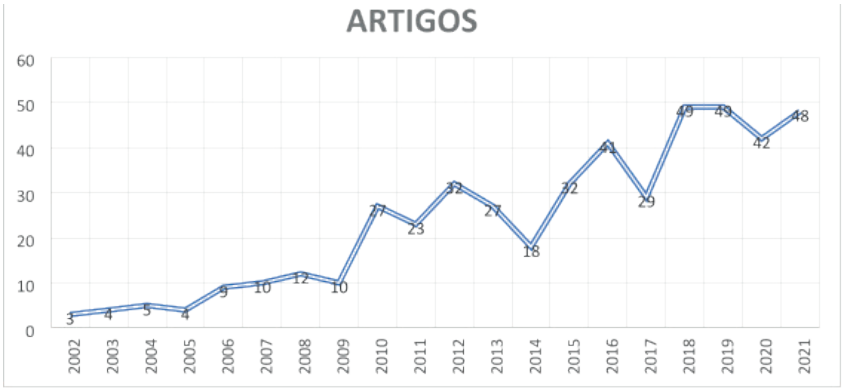

Gráfico 1. Número de artigos por ano. Fonte: elaborado a partir de dados de Web of Science.

Tabela 1. Afiliações dos autores

AFILIAÇÕES	ARTIGOS
Universidad Nacional Autónoma de México (México)	23
Universidade Federal de Santa Catarina (Brasil)	20
Universidad de La Frontera (Chile)	18
Universidade de São Paulo (Brasil)	18
Universidade Federal do Rio Grande do Sul (Brasil)	13
Universidad Nacional de Colombia (Colômbia)	12
Universidade de Brasília (Brasil)	11
Universidad de Buenos Aires (Argentina)	10
Pontificia Universidad Catolica de Valparaiso (Chile)	9
Universidad de Concepcion (Chile)	9
Universidade Estadual de Londrina (Brasil)	9
Universidade do Estado do Rio de Janeiro (Brasil)	8
Universidad del Valle (Colômbia)	7
Universidade Federal de Minas Gerais (Brasil)	7
Universidade Federal do Rio de Janeiro (Brasil)	7
Universidade Federal Fluminense (Brasil)	7
Universidad Catolica del Maule (Chile)	6
Universidad de Antioquia (Colômbia)	6
Universidade Federal de Santa Maria (Brasil)	6
Universidade Federal de São Carlos (Brasil)	6
Universidade Federal do Rio Grande (Brasil)	6

Fonte: elaboração própria a partir de dados de Web of Science através de extração no VOSViewer.

O menor cluster, amarelo, apresenta quatro referências. As referências parecem estar entre a Educação (alfabetização, formação da linguagem para crianças etc.), com Paulo Freire, Delizoicov e a formação da linguagem, que apresenta Vygotsky e Moraes Roque, que é um teórico tanto da análise de conteúdo quanto da análise textual discursiva.

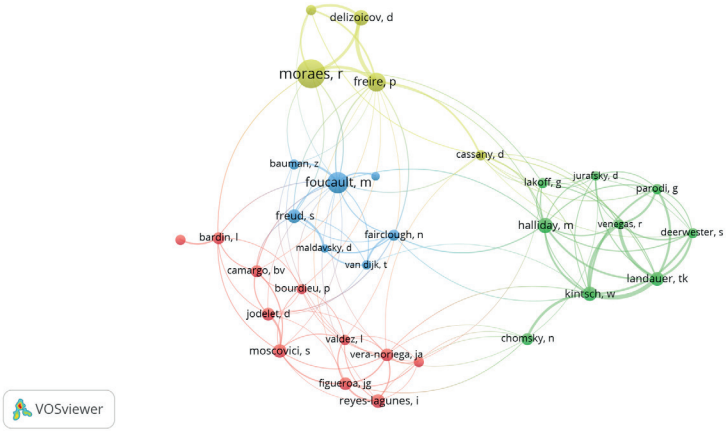

Figura 3. Grafo de cocitações de autores (10 ou mais citações).Fonte: elaboração própria a partir do VOSViewer. Nota: Para termos de replicabilidade, *type of analysis*: co-*citation*, *unity of analysis*: *cited authors*, *counting method*: *full counting*, ao menos 10 citações por autor, método de normalização *LinLog/Modurality*, *attraction*: 2, *repulsion*: -1 e *weights*: *citations*.

Posteriormente, buscamos verificar a partir dos termos que os autores mais utilizaram em seus resumos se eles trataram mais de questões técnicas ou teóricas. Novamente, fizemos uso do VOSviewer para gerar uma rede de coocorrência de palavras.

Desta vez, temos três grupos. O maior, vermelho, é formado por 30 termos, e parece ser o mais próximo do objetivo de nosso estudo. Dentro do que esperávamos, vemos os nomes das técnicas de processamento automático, como *natural language processing*, *text mining*, *semantic analysis*, *text analysis*, além de referências a prováveis objetos de estudo (*author*, *document*, *term*, *theme*, *topic*, *user*, *word*), questões metodológicas (*case*, *corpus*, *model*, *set*, *toolk*, *type*, *user*, *view*) e também elementos mais teóricos (*discourse*, *field*, *issue*, *language*, *perspective*, *representation*).

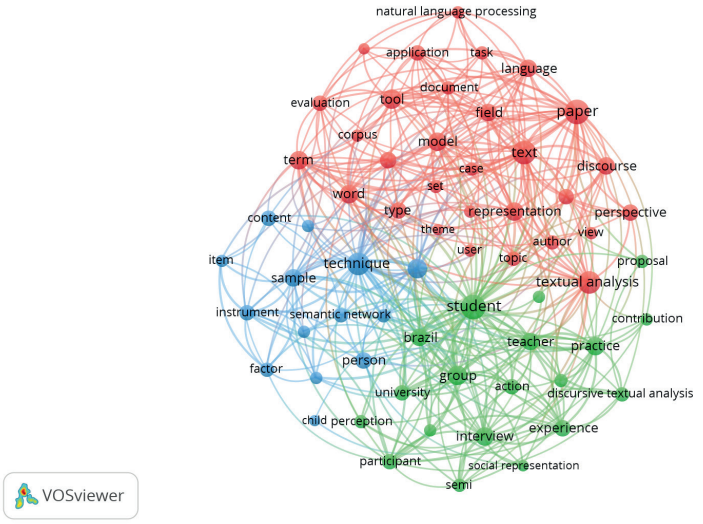

Figura 4. Grafo de coocorrência de palavras extraídas dos resumos (20 ocorrências ou mais). Fonte: elaboração própria a partir do VOSviewer. Nota: Para termos de replicabilidade, mapa através de dados textuais, *Abstract fiel*, *ignore structered abstract labels* e *copyright statements*, *binary counting*, coocorrências mínimas de um termo: 20, o que gerou 105 termos, e 63 foram considerados os mais relevantes. O método de normalização escolhido foi o *LinLog/Modurality*, *attraction*: 2, *repulsion*: -1, *weights*: *occurrences*.

O segundo grupo, verde, parece estar mais próximo de pesquisas qualitativas, uma vez que apresenta o nome de técnicas, como *semi[structered]interview*, *group*, *questionnaire*, *discursive textual analysis*, além de elementos que técnicas qualitativas se importam mais, como *perception*, *social representation*, *experience, practice*. Como a palavra "Brazil" é central, é bem provável que sejam estudos brasileiros e que muitos se concentrem em Educação, uma vez que *university*, *teacher* e *student* são outros termos de destaque.

Além disso, buscamos aprofundar a análise da produção a partir de mais dois testes. Considerando que as palavra-chave são elementos sintetizadores importantes dos conceitos e partes principais do trabalho acadêmico, geramos um grafo de coocorrência destas. Para dar melhor inteligibilidade e leitura, consideramos todas as palavras que aparecessem mais de 5 vezes no corpus. O grafo segue abaixo.

Aqui podemos ver cinco clusters. O principal, vermelho, parece se encontrar entre a enfermagem e a educação, tendo a análise de conteúdo como sua técnica principal, mas provavelmente numa perspectiva mais qualitativa. O segundo cluster, verde, é dedicado a estudos da educação e algo entre representações sociais e estudos de gênero, não apresentando de forma explícita o uso das técnicas. Daí em diante, os clusters parecem mais conectados às técnicas de análise. O cluster azul parece ser mais conectado a análises semânticas, apresentando os termos mineração de dados, representação de conhecimento, processamento de linguagem natural, redes semânticas e semântica. Por sua vez, o cluster amarelo parece mais conectado a análises mais léxicas, apresentando os termos corpus linguísticos, mineração de opiniões e análise de sentimentos, apresentando redes sociais e mais especificamente o Twitter como objetos de análise. Interessantemente, o termo "análise textual" em si ficou próximo de análise de discurso, o que pode evidenciar que essa metodologia e sua teoria servem como base de análises textuais.

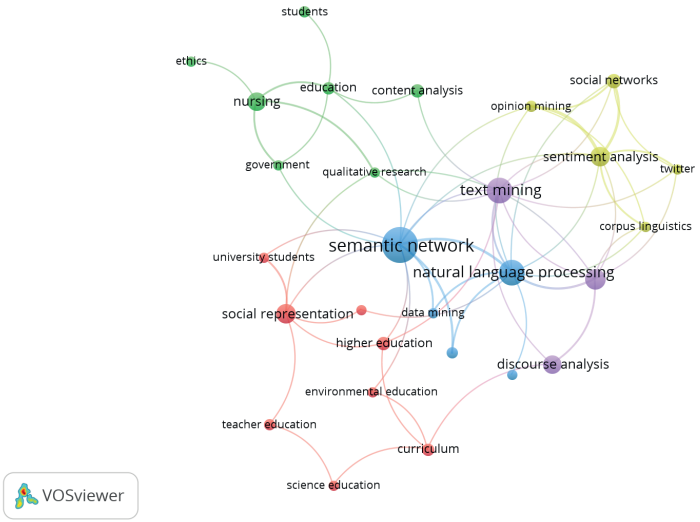

Figura 5. Grafo de coocorrência de palavras-chave (cinco ocorrências ou mais). Fonte: elaboração própria a partir do VOSViewer. Nota: Para termos de replicabilidade, *type of analysis*: *co-occurrence unity of analysis*: *author keywords*, *counting method*: *full counting*, ao menos 5 coocorrências, gerando 31 termos, estando 28 conectados entre si e apresentados no grafo acima, método de normalização *LinLog/Modurality*, *attraction*: 3, *repulsion*: -1 e *weights*: *occurrences*.

Tabela 2. Clusters de palavras-chave (5 ocorrências ou mais)

CLUSTER 1 (VERMELHO)	CLUSTER 2 (VERDE)	CLUSTER 3 (AZUL)	CLUSTER 4 (AMARELO)	CLUSTER 5 (ROXO)
content analysis	*curriculum*	*data mining*	*Corpus linguistics*	*Discourse analysis*
Education	*environmental education*	*knowledge representation*	*Opinion mining*	*Text analysis*
Ethics	*gender*	*natural language processing*	*Sentiment anaylsis*	
Government	*higher education*	*semantic network*	*Social networks*	
Nursing	*science education*	*Semantics*	*Twitter*	
qualitative reseach	*social representation*			
Students	*teacher education*			
text mining	*university students*			

Fonte: elaboração própria a partir do VOSViewer.

Tabela 3. termos de maior destaque na análise de coocorrência de palavras dos resumos

TERMO	OCORRÊNCIA	SCORE DE RELEVÂNCIA DO VOSVIEWER
natural language processing	26	44,033
Item	27	32,569
Semi	29	27,725
text mining	22	24,132
discursive textual analysis	33	23,449
Task	30	21,777
Interview	59	18,639
Instrument	41	17,466
Language	50	16,524
natural semantic network	25	16,273
social representation	23	16,271
Participant	39	15,231
Document	31	13,723
Application	38	12,561
Paper	97	12,202
Perception	29	11,865
Practice	56	10,963
Child	22	10,745
Questionnaire	28	10,483
Contribution	27	10,418
Fator	37	1,671
Experience	47	1,407
Tool	62	1,12
Text	91	0,9138
Evaluation	39	0,902
Teacher	53	0,8669
Set	30	0,8563
semantic analysis	48	0,8545
Field	51	0,8449
Sample	51	0,8331
Stage	28	0,8057
Case	29	0,769
Corpus	28	0,7042
Question	32	0,6949

TERMO	OCORRÊNCIA	SCORE DE RELEVÂNCIA DO VOSVIEWER
University	*41*	*0,6943*
Action	*38*	*0,65*
Author	*34*	*0,6252*
Evidence	*26*	*0,6185*
View	*26*	*0,574*
Discourse	*52*	*0,5525*
Issue	*40*	*0,5438*
Chile	*30*	*0,5299*
Term	*59*	*0,5212*
Brazil	*51*	*0,5016*
Model	*55*	*0,4998*
Group	*62*	*0,4962*
User	*25*	*0,4909*
textual analysis	*85*	*0,4776*
Proposal	*32*	*0,4244*
Type	*44*	*0,4191*
Topic	*37*	*0,4109*
Investigation	*25*	*0,4087*
Perspective	*44*	*0,3978*
Addition	*25*	*0,3719*
Person	*43*	*0,361*
Representation	*56*	*0,346*
Contente	*39*	*0,3422*
Theme	*24*	*0,3396*
Word	*58*	*0,3212*
Student	*102*	*0,3186*
semantic network	*35*	*0,3147*
Meaning	*66*	*0,2821*
Technique	*88*	*0,2181*

Fonte: elaboração própria a partir do VOSViewer.

O VOSViewer também mostra os termos considerados mais relevantes no *corpus* analisado[15]. Este foi o último ponto que analisamos. Dentro do esperado, os termos com maior pontuação tendem a ser aqueles ligados aos nomes das técnicas, como *natural language processing*, *text mining*, *discursive textual analysis*, *natural semantic network*.

4. Considerações finais

Ao considerarmos esta última tabela (Tabela 3) em conjunto com as Figuras 3 e 4, vemos que há uma forte presença de palavras sobre as técnicas, com predominância da nomenclatura das técnicas da Computação. A palavra *Content Analysis* só apareceu na Figura 3, das palavras-chaves. Em todos os outros testes não aparece como relevante. Já termos ligados à computação e análise computacional da linguagem – como "*Natural Language Processing*", "*text mining*" e *discursive textual analysis*, entre outros. O que isso pode demonstrar? uma forte presença das técnicas da computação nesses trabalhos, demonstrando assim a forte influência desta sobre um método que nasce no campo das humanidades - a análise de conteúdo. Dito de outra forma: ao considerarmos o campo das CSC e este método em específico, vemos que este vem se construindo pela forte incorporação da computação pelos cientistas sociais.

Retomando à questão que guiou nossa investigação, a incorporação do método da AC nas CSC vem muito mais pelo avanço e uso das técnicas computadorizadas. Entretanto, quando olhamos para a Figura 3 – dos autores mais citados e cocitados – as referências bibliográficas continuam predominantemente do campo das humanas, não da computação. Bardin, Foucault, Bourdieu entre outros são os autores de destaque. Assim, não se sustenta a percepção de que o movimento ocorre apenas da computação para as Ciências Sociais. Mas, no âmbito teórico e epistemológico dos trabalhos, a matriz ainda é fortemente orientada pela literatura das Ciências Sociais. Com isso, podemos apontar, neste levantamento preliminar que se, de um lado, no âmbito das técnicas é uma forte presença e contribuição da computação, no âmbito da orientação epistemológica, os autores ainda partem das referências mais clássicas da AC, demonstrando assim uma combinação de ambos os campos, formando um espaço híbrido.

Isso pode vir ao encontro do que Grimmer e Stewart (2013) apontam de que nenhuma análise automatizada prescinde completamente da ação do pesquisador. Mesmo na classificação completamente organizada pelo computador, a inferência, ou seja, a interpretação dos dados precisa do cientista social e de sua reflexão sobre eles. Neste sentido, é necessário avançar na investigação de como a hermenêutica controlada proposta pela AC aparece combinada com a análise qualitativa e automatizada, reforçando ainda mais o caráter híbrido da construção das CSC.

Por fim, entendemos que outros cruzamentos podem ser feitos a partir desses dados iniciais para pensar pelo menos dois novos eixos de investigação: sobre como os temas se relacionam com os países da região e ao longo do tempo, para verificar como a intersecção entre Ciências Sociais e Computação vem tendo entrada por diferentes objetos em países distintos. E, por fim, uma análise mais pormenorizada sobre as/os autoras/es que são alvo dessa investigação compreendendo que essas transformações mobilizam gerações e expertises distintas de pesquisadoras(es). Sabemos que esses são passos futuros a serem dados e que o artigo acima já aponta a fertilidade da reflexão sobre a incorporação da computação pelas Ciências Sociais latino-americanas.

NOTAS

[1] Estamos considerando aqui uma concepção ampla de Ciências Sociais como sinônimo da área de Humanidades, incluindo assim todas as disciplinas e áreas de formação que utilizam o comportamento humano, sua interação e relações como objeto, incluindo assim Educação, Comunicações, Administração, Filosofia, Economia entre outras. Em alguns casos, mesmo formações da área de ciências da vida utilizam Análise de Conteúdo em suas pesquisas de forma intensa (Sampaio, Lycarião, Codato, Marioto, Bittencourt, Nichols, & Sanchez, 2022).

[2] Como Portugal e Espanha, por exemplo.

[3] Na seção "metodologia" deste artigo apresentamos a *string* e descritores mais detalhadamente.

[4] Aqui utilizaremos Análise Textual (ou sua sigla AT) como sinônimo de *Text as Data* e técnicas computacionais de análise textual, e vice-versa.

[5] Neuendorf (2017), por exemplo, apresenta nove etapas para a realização da AC. Sampaio & Lycarião (2021) ainda dividem em 12 etapas a serem seguidas. As condições para uma AC ser considerada científica são validade (interna e externa), confiabilidade (dos codificadores) e replicabilidade da pesquisa como um todo.

[6] Para saber mais sobre esses testes, ver Hayes e Krippendorf (2007); Feng (2014), bem como as revisões feitas por Lima (2013) e Sampaio & Lycarião (2021).

[7] Mais recentemente, em 2022, os autores lançaram um livro intitulado *Text as Data*, juntamente com a pesquisadora Margaret E. Roberts, onde eles adotam apenas o termo "*text as data*" para definir um campo onde tem ocorrido a apropriação cada vez mais massiva e rápida de ferramentas computacionais para a análise do comportamento humano e complementam que este é "*a fast-moving field and the state of the art can easily change within the space of six months*" (Grimmer, Roberts & Stewart, 2022).

[8] Um exemplo disso é a *Latent Dirichlet Allocation* (LDA, Blei, Ng, & Jordan, 2003), técnica utilizada para a identificação de tópicos a partir de um conjunto de palavras relacionadas baseada na decomposição de textos no plano vetorial. Ela é um dos mais comuns algoritmos para identificação de tópicos guiada por dois princípios fundamentais: a) todo documento é um conjunto de tópicos/temas; b) todo conjunto de temas é composto de um conjunto de palavras. Blei, Ng e Jordan (2003) demonstram que todo tópico é, de fato, uma variável latente multinomial, ou seja, representa a probabilidade de uma distribuição específica de um conjunto de palavras. E isso é calculado a partir de distribuições multinomiais e testes de correlação entre as palavras que compõem os conjuntos.

[9] Consideramos *Machine Learning* aqui de forma ampla, como o conjunto de técnicas que melhoram os sistemas e atividades computacionais por meio da aprendizagem. E, como aponta Zhou, as máquinas aprendem a partir dos dados: *In computer systems, experience exists in the form of data, and the main task of machine learning is to develop learning algorithms that build models from data* (2021, p. 2).

[10] Redes neurais e *deep learning* são alguns exemplos nesse sentido.

[11] Como é o caso da aplicação ChatGPT, entre outras, que tem gerado grande discussão em torno da produção de *corpus* textuais, autoria, desinformação ou, até mesmo, trabalho acadêmico, dada sua capacidade de fazer classificações sobre *corpus* de textos.

[12] Apesar de enviesar parte dos resultados, a escolha pela SciELO pode ser justificada em ao menos duas questões: 1) não há uma base realmente significativa da produção da América Latina. Grandes indexadores como WoS e Scopus não contemplam a maior parte de nossos periódicos e indexadores locais, como o Latindex e Redalyc, também não podem ser vistos como representativos da produção brasileira e, 2) a maior parte dos indexadores locais não apresenta um bom conjunto de metadados, que permitam seu estudo através de softwares bibliométricos, como o VOSViewer. Para uma discussão sobre o tema, ver Tennant (2020).

[13] Análise textual discursiva é uma técnica de pesquisa que se encontra entre a análise de conteúdo e a de discurso, portanto tendendo a utilizar codificações manuais (Moraes & Galiazzi, 2006), portanto foi importante retirar a expressão completa, mantendo apenas a "análise textual".

[14] Conforme Grácio (2020), a cocitação identifica a ligação ou semelhança de dois documentos citados pela frequência de ocorrência conjunta em uma lista de referências posterior de autores citantes. Portanto, se dois autores ou documentos são citados juntos, em pesquisa posterior, há proximidade conceitual, temática ou metodológica entre os citados na avaliação do autor citante, sendo provável que estejam relacionados em termos de conteúdo.

REFERÊNCIAS BIBLIOGRÁFICAS

Bardin, L. (2008). *Análise de conteúdo*. Lisboa: Edições 70.

Benoit, K. (2020). Text as Data: an overview. In L. Curini, & R. Franzese. *The SAGE Handbook of Research Methods in Political Science and IR*. Londres: Sage Publications.

Blei, D. M., Ng., A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, 993-1022.

Bourdieu, P. (1989). *O poder simbólico*. Lisboa/Rio de Janeiro: Difel/Bertrand Brasil.

Camargo, B. V., & Justo, A. M. (2013). *Tutorial para uso do software IRaMuTeQ* (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires). Florianópolis: Laboratório de Psicologia Social da Comunicação e Cognição – UFSC.

Cioffi-Revilla, C. (2017). *Introduction to Computational Social Science: Principles and Applications*. Londres: Springer.

Conte, R. *et al*. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214, 325-346.

Cúrcio, V. R. (2006). Estudos estatísticos de textos literários. *Revista Texto Digital*, *2*(2), 9-28.

Eck, N. J. van, & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

Edelman, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology*, 46, 61-81.

Feng, G. (2014). Intercoder reliability indices: disuse, misuse, and abuse. *Quality & Quantity*, 48, 1803-1815.

Grácio, M. C. C. (2020). *Análises relacionais de citação para a identificação de domínios científicos*. São Paulo: Cultura Acadêmica.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267-297.

Grimmer, J; Roberts, & M. Stewart, B.M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 24, 395-419

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data*. Princenton: Princenton University Press.

Habermas, J. (1984). *A mudança estrutural da esfera pública*. Rio de Janeiro: Tempo Universitário.

Hayes, A. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77-89.

Izumi, M., & Moreira, D. (2018). O texto como dado: desafios e oportunidades para as ciências sociais. *BIB - Revista Brasileira de Informação Bibliográfica em Ciências Sociais*, 86, 138-174.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: an examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research*, *18*(2), 243-250.

Krippendorff, K. (2004). *Content Analysis: an introduction to its methodology*. SAGE Publications.

Lasswell, H. (1978). A estrutura e a função da comunicação na sociedade. In G. Cohn (Org.). *Comunicação e indústria cultural*. São Paulo: Cia Editora Nacional.

Lazer, D. *et al*. (2009). *Computational Social Science. Science*, *323*(5915), 721-723.

Lazer, D. *et al*. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060-1062.

Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Lima, J. Á. (2013). Por uma Análise de Conteúdo mais fiável. *Revista Portuguesa de Pedagogia*, *47*(1), 7-29.

Lipman, W. (2008). *Opinião pública*. Petrópolis: Vozes.

Moreira, D., Pires, A., & Medeiros, M. A. (2022). Do 'texto como texto' ao 'texto como dado': o potencial das pesquisas em Relações Internacionais. *Revista de Sociologia e Política*, *30*, 1-29.

Moraes, R., & Galiazzi, M. C. (2006). Análise textual discursiva: processo reconstrutivo de múltiplas faces. *Ciência & Educação (Bauru)*, 12, 117-128.

Neuendorf, K. A. (2017). *The Content Analysis: guidebook*. Thousand Oaks: Sage Publications.

Salganik, M. (2018). *Bit by bit: social research in digital age*. Nova Jersey: Princeton University Press.

Sampaio, R. C., & Lycarião, D. (2021). *Análise de Conteúdo Categorial: manual de aplicação*. Brasília, DF: ENAP. Recuperado em 4 de janeiro de 2023, de https://repositorio.enap.gov.br/bitstream/1/6542/1/Analise_de_conteudo_categorial_final.pdf.

Sampaio, R. C., Lycarião, D., Codato, A. N., Marioto, D. J. F., Bittencourt, M., Nichols, B. W., & Sanchez, C. S. (2022). Mapeamento e reflexões sobre o uso da análise de conteúdo na SciELO-Brasil (2002-2019). *New Trends in Qualitative Research*, *15*, e747-e747.

Sartori, G. (1998). *Homo Videns: la sociedad teledirigida*. Rio de Janeiro: Taurus.

Tennant, J. P. (2020). Web of Science and Scopus are not global databases of knowledge. *European Science Editing*, *46*, e51987.

Zhou, ZH. (2021). *Machine Learning*. Springer Nature Singapure.