A construção e análise do *corpus* no estudo de fenômenos em plataformas: limites e interseções entre a semiótica, a etnografia digital e a ciência de dados

Building and analyzing the corpus in the study of platform phenomena: limits and intersections between semiotics, digital ethnography, and data science

Santiago Videla
svidela@sociales.uba.ar

Licenciado em Ciências da Comunicação (UBA). Docente de Semiótica das Midiatizações (UBA-FSOC).

Resumo

Este trabalho propõe discutir o problema da construção do corpus e da captação dos dados de análise no estudo de fenômenos relacionados às plataformas, em particular aqueles vinculados a aspectos da vida musical. Com esse objetivo, propõe-se percorrer os limites e as interseções entre a semiótica das mediações, a etnografia digital e a ciência de dados.
Palavras-chave: midiatização, sociosemiótica, metodología, plataformas.

Abstract

This paper aims to discuss the problem of corpus construction and data collection for analysis in the study of platform-related phenomena, particularly those connected to aspects of musical life. To this end, it proposes to explore the boundaries and intersections between the semiotics of mediatization, digital ethnography, and data science.
Keywords: mediatization, sociosemiotics, methodology, platforms.

Introdução

Este trabalho propõe-se abordar os desafios de construir um corpus que funcione como uma amostra representativa quando nos deparamos com um universo como o das plataformas mediáticas, difícil de definir como tal em razão de sua persistente variabilidade. No mesmo sentido, procuraremos estabelecer os desafios que implica delimitar os aspectos/propriedades do objeto de pesquisa, de modo que possam ser processados por sistemas automatizados e produzir resultados significativos na análise de fenômenos sociais midiatizados.

Para isso, percorreremos os avanços das ciências de dados, das etnografias de redes e dos estudos culturais e semióticos, a fim de estabelecer suas diferenças, bem como um marco de acordos comuns para esse tipo de investigação. Trata-se, portanto, de ajustar um processo de diálogo em torno de um conjunto de objetos que, até o momento, apresentam dificuldades para o analista.

Desde a nossa perspectiva, aquilo que ocorre nas plataformas constitui, em princípio, uma midiatização. Ou seja, um tipo de sistema de intercâmbio que mobiliza questões tecnológicas, restrições discursivas e práticas sociais associadas (Fernández, 2023).

Assim, por exemplo, o que ocorre em um comentário de um post no Instagram da cantora argentina Lali Espósito será, em parte, resultado das restrições e possibilidades técnicas determinadas pelo Instagram (ali é possível utilizar GIFs, mas no YouTube não, por exemplo), em parte dos gêneros e estilos em circulação e, em parte, das decisões dos nativos desses contextos. Em Videla (2023), mostramos como um post de colaboração publicitária de uma artista musical romena (Inna) é respondido tanto com mensagens esperáveis — como expressões de afeto ou emojis (como os "foguinho") — quanto com outro tipo de mensagem, construída a partir da combinação desses elementos com títulos de suas canções e trechos de suas letras, em uma estrutura sofisticada que pode ser pensada como equivalente aos caligramas de Apollinaire. Também mostramos ali um post de uma selfie diante do espelho da cantora argentina Lali Espósito que recebia respostas diretamente vinculadas a ela, e outras como comele la boca a Rosalía (um terceiro não presente na foto, mas mencionado em referência a uma recente passagem da artista pela Espanha, local de residência da citada) ou Es más fácil sacarle el puesto a Roncaglia que responda Lali (em referência à permanência de um jogador de futebol do Club Atlético Boca Juniors na posição).

É certo que aquilo que ocorre tanto no post quanto no comentário condensa, em alguma medida, práticas prévias e até

DOI: https://doi.org/10.4013/qt.2025.132406 — Vol. 13, nº 24, janeiro-dezembro/2025

pré-midiatizadas da relação músico–fã. A declaração de amor sempre existiu: antes das midiatizações, ocorria à beira do palco, e hoje se esboça em um aplicativo na forma de um "foguinho" ou de um coração. O mesmo vale para a presença, nos comentários, do repertório proverbial das interações entre amigos (como o comentário futebolístico), que, ao transitar para o espaço público massivo, parece produzir efeitos de cumplicidade com terceiros não necessariamente conhecidos por seu autor. Sabemos também que o desenvolvimento de algum tipo de prática artística visual como oferenda ao ídolo já estava presente nos primórdios da vida dos músicos populares (Videla, 2009). Por isso, o primeiro desafio consiste em compreender o grau de novidade que essas práticas apresentam em função da série social que mobilizam no sistema de intercâmbio. Em seguida, trata-se de aproximar-se da determinação do peso que essas práticas assumem no conjunto das interações do músico e da plataforma. Daí que, quando encontrarmos comentários que aparentemente pareçam dissociados da proposta do post, eles não devam ser tratados, a princípio, como casos isolados.

Em Videla (2025), insistíamos também que essas aparições, talvez raras ou pouco frequentes nas redes de outros artistas, deveriam ser compreendidas como parte das restrições do sistema de intercâmbio. Isso porque, por definição, elas integram esse sistema e devem poder ser recuperadas como tais, considerando-se, além disso, sua incidência no conjunto das interações. Assim, o primeiro desafio consiste em compreender o volume e a frequência desses casos.

O olhar da ciência de dados

A ciência de dados busca modelos "que descrevam padrões e comportamentos a partir dos dados com o objetivo de tomar decisões ou realizar previsões" (García *et al.*, 2018, p. 5). Trata-se de um campo de enorme crescimento, resultante do volume gigantesco de informação produzido pelo fato de vivermos em plataformas desde o início deste século, na etapa que a literatura denomina Big Data. Sua incidência abrange "numerosos grupos de pesquisa de diferentes áreas (computação, estatística, matemática, engenharia etc.) que trabalham na proposição de novos algoritmos, técnicas de computação e infraestruturas para a captura, o armazenamento e o processamento de dados" (García *et al.*, 2018, p. 6).

Tradicionalmente, nos manuais de metodologia, atribui-se aos estudos qualitativos a análise de poucos casos, porém em profundidade, enquanto a análise de muitos casos ou de universos mais amplos é associada aos métodos quantitativos (Sabino, 1992; Sautu, 2005; Marradi *et al.*, 2018; Hernández Sampieri *et al.*, 2005; Corbetta, 2007). Para Sabino, os métodos qualitativos "tentam recuperar para a análise parte dessa complexidade do sujeito e de seus modos de ser e de agir no meio que o circunda. O íntimo, o subjetivo, por definição dificilmente quantificáveis, constituem, portanto, o terreno no qual se movem os métodos qualitativos" (Sabino, 1992, p. 65). No entanto, essa busca pela compreensão subjetiva dos textos entra em tensão com a escala intrínseca do fenômeno que nos ocupa nas plataformas: o comentário aparece entre milhares de comentários e o post entre milhares de posts. Estamos, necessariamente, diante de algo que tensiona o que é próprio do estilo individual com o estilo esperado nas plataformas (Videla, 2023).

Por isso, consideramos complementar recorrer à distinção proposta por Lévi-Strauss entre análises estatísticas e mecânicas. Para o autor, "um modelo cujos elementos constitutivos se encontram na mesma escala dos fenômenos será denominado 'modelo mecânico', e 'modelo estatístico' aquele cujos elementos se encontram em uma escala diferente" (Lévi-Strauss, 1987, p. 325). O objeto que nos ocupa implica uma escala que excede a posição do observador e, portanto, a abordagem retoma aspectos estatísticos. A partir daí, não se deve perder de vista que a natureza do que será analisado consiste em um conjunto de palavras, imagens e sons. A própria observação de palavras já constituiu um problema para a tradição da content analysis, que, há mais de cinco décadas, trabalha com o uso de ferramentas computacionais para o processamento da informação, enfatizando que, se não se atentar para aquilo que denominamos sistema de intercâmbio, "frequentemente, tais abordagens não passam de contar palavras sem considerar seus significados" (Krippendorff, 1980, p. 87).

Para García *et al.* (2018, p. 14), Big Data é "o conjunto de dados cujo tamanho supera consideravelmente a capacidade de captura, armazenamento, gestão e análise dos softwares convencionais de bases de dados". No entanto, a definição pressupõe também "a variedade do conteúdo e a velocidade com que os dados são gerados, armazenados e analisados. Essas dimensões são as '3V' com as quais a empresa Gartner descreveu o Big Data, isto é, volume, velocidade e variedade dos dados".

Atualmente, grande parte dos avanços nessa área tem em comum a problematização da busca por eficiência no processamento desse volume crescente de dados, entendidos como gerados pela vida fortemente digitalizada da sociedade. Nessa linha, propõe-se a passagem do modelo de Big Data para o de Fast Data (Sánchez Piccardi *et al.*, 2021). Já não basta processar muito; agora, exige-se velocidade. A análise de fluxos permanentes e ilimitados de dados conduz à adoção de soluções como o empilhamento de softwares de leitura de bases de dados. Trabalhos como a tese de Fajardo (2023) discutem a efetividade de diferentes alternativas, como Apache Spark e Apache Flink, ambas permitindo ao analista acessar volumes incalculáveis de informação.

A essa informação, nessa corrente, dá-se o nome de dado. Para esses autores, um dado corresponde à coleta de "um conjunto de fatos (uma base de dados, BD), e os padrões são expressões que descrevem um subconjunto dos dados (um modelo aplicável a esse subconjunto)" (García et al., 2018, p. 30). No entanto, a natureza da captura dos dados, sua ponderação e sua representabilidade não fazem parte das problematizações centrais. Araoz e Cellone (2025) percorrem a discussão sobre a noção de dado nas diversas ciências sociais, enfatizando a ideia de que os dados são construções do analista, e não algo dado de antemão. Nesse sentido, a noção de dado à qual aderimos aproxima-se mais da noção de padrões proposta por García et al.

Recordemos que, como antecipado anteriormente, nosso objeto de estudo (o post em uma rede, a interface do home banking, um aplicativo mobile, um site) é uma midiatização. Como tal, apresenta propriedades significantes em sua superfície discursiva, aquilo que Verón (1987) define como marca. Quando, nas ciências de dados, se fala em dados, para nós trata-se de marcas. O objetivo do analista é converter essas

marcas em indícios do processo produtivo. Assim, seguindo Fernández (2023), um dado é uma marca posta em relação com suas condições produtivas, quando inscrita na análise em vínculo com o processo produtivo de produção de sentido. Um dado é, portanto, algo colocado em relação com algo. Fernández denomina esse movimento de passagem do grillado à matriz.

Isso não supõe, em hipótese alguma, uma crítica às ciências de dados e a seus resultados. Pelo contrário, trata-se do primeiro passo de um ajuste terminológico que permita estabelecer pontes de trabalho.

As etnografias digitais

A definição de midiatização com a qual trabalhamos entende que estamos diante de um sistema de intercâmbio de sentido que coloca em funcionamento três séries do social: a dos dispositivos técnicos, a dos discursos e a das práticas sociais. Há mais de cinquenta anos, Christian Metz já enfatizava, em sua definição do grande regime do significante cinematográfico, a incidência desses três níveis (como recupera Fernández, 1994). O que constitui o cinema é, portanto, o cinematógrafo (o tecnológico), a seleção genérico-estilística e uma arquitetura particular das salas, bem como um modo específico de aproximação a elas. Além disso, no caso do cinema, da televisão ou mesmo do rádio, contamos com o auxílio de mais de sessenta anos de literatura descrevendo o fenômeno. Nessas tradições, as referências ao que fazem os nativos são numerosas e apresentam relativo acordo entre os pesquisadores. Já no caso das plataformas, em parte devido à novidade e em parte às limitações materiais das análises, sabemos pouco sobre o que as pessoas fazem com seus aplicativos. Escreve-se bastante sobre padrões de comportamento (percentuais de uso do like, de compartilhamentos, de visualizações), mas pouco sobre o significado disso e sobre seus efeitos discursivos.

Por exemplo, uma parte significativa da literatura sobre narrativas transmídia pressupõe a existência de um usuário típico, de caráter universal, baseado em suposições aceitas pelo conjunto dos autores, que consome relatos de forma ordenada e vive uma vida digital sem que haja, aparentemente, um contraste empírico que o comprove (como se observa nos trabalhos de Irigaray, 2021, 2022; Alabadejo, 2017; Lovato, 2018; Pratten, 2011). Sem que isso seja entendido como uma crítica a essa posição — da qual reconhecemos a produtividade —, destacamos que o que é próprio desses espaços dedicados ao estudo do comunicacional é a pressuposição de um conjunto de acordos sobre as práticas dos nativos, que se baseia, em parte, na própria experiência dos autores e, em parte, na repetição de cadeias de citações às quais se atribui valor de dado.

Isso nos obriga a revalorizar áreas de trabalho como a representada pelos textos de Pink et al. (2019), que se situam em um extremo oposto. Inseridos no que denominam etnografia digital, esses autores explicam tratar-se de uma disciplina que "esboça uma aproximação ao trabalho etnográfico no mundo atual, convidando os pesquisadores a considerar como vivemos e pesquisamos em um ambiente digital, material e sensorial" (Pink et al., 2019, p. 17). Ou seja, o foco está na compreensão do que o nativo faz com essa plataforma. Certamente, não se trata de um campo homogêneo. Aronica mostra que "a etnografia realizada na internet é conhecida por diversos termos: ciberetnografia (Escobar, 1994); etnografia do ciberespaço (Hakken, 1999); etnografia virtual (Hine, 2000); antropologia dos meios (Ardévol; Vayreda, 2002); etnografia mediada ou etnografia de/em/através dos meios (Beaulieu, 2004); etnografia do digital (Estalella, 2006); autoetnografia (Espinosa, 2007); netnografia (Turpo, 2008); etnografia da internet (Ardévol; Estalella; Domínguez, 2008) e etnografia digital" (2019, p. 30). Ainda assim, identifica-se certo consenso em que "consiste em descrições detalhadas de ambientes, acontecimentos, pessoas, interações e condutas observáveis. Incorpora aquilo que os participantes comentam, suas experiências, atitudes, crenças, pensamentos e reflexões tal como são expressos por eles, e não como são percebidos pelo pesquisador" (Aronica, 2019, p. 29).

As interações em plataformas costumam ser descritas como parte de um conjunto de condutas. Para Pink et al. (2019, p. 138, *tradução nossa*)

> *A experiência cotidiana sugere que essa ideia pode lançar luz sobre a dinâmica social presente em plataformas sociais tão distintas quanto listas de e-mail, Twitter, Weibo, Facebook e WhatsApp. A terminologia e a sintaxe variam de uma plataforma para outra ('fio', 'hashtag', 'tendências', 'comentários', 'chat' etc.), mas todos esses espaços organizam as conversações por meio de uma série específica de entradas conectadas, isto é, por meio de fios*[1].

Essa abordagem não se pergunta (nem tem por que fazê-lo) sobre a significação dos likes ou se comentar em uma bolha de comentários do Facebook equivale a fazê-lo após um fio no X (ex-Twitter). Em um trabalho anterior, Theviot (2014) analisa uma série de conversações no Facebook em torno da militância política de determinados usuários. O estudo é altamente detalhado e reconstrói aspectos do sistema de intercâmbio entre o mural e os comentários, embora não se detenha na descrição da superfície discursiva. Ou seja, os fenômenos de significação tendem a ser relegados na análise, e propostas como a de semiodata, de Raimondo Anselmino (2024), adquirem um valor particular ao buscar sua recuperação.

Seguindo Fernández, não é menor a importância de prestar atenção ao discursivo. Em especial, ao que o autor denomina megustos diferentes, definidos como "uma espécie de efeito lateral, micro dentro do micro, que, no entanto, faz parte do que há de mais profundo na compreensão interacional do Facebook como rede de contatos" (2018, p. 108, *tradução nossa*)[2]. A mesma ação de dar like a uma selfie em um espaço público insere-se em diferentes sistemas de intercâmbio, conforme exista entre quem curte e quem é curtido uma relação de chefe–empregado, de amizade ou de conquistador–conquistado. Sabemos também que, em última instância, um fio não é o mesmo que um hashtag ou uma tendência: estes pressupõem fenômenos de significação distintos, derivados dos dispositivos nos quais se inscrevem.

Nesses trabalhos, surge também o problema da escala de observação. Ela costuma envolver dezenas de casos e, às vezes, centenas de posts ou comentários em redes. Mesmo

[1] Original em espanhol.
[2] Original em espanhol.

considerando a riqueza inerente à capacidade do pesquisador, não é o mesmo eleger como objeto um artista que publica três posts por dia e recebe cerca de 3.000 comentários, do que um artista que publica uma vez a cada quinze dias e recebe cerca de 50 comentários. O valor dos achados para avançar em conclusões de nível meso é muito diferente e envolve riscos consideráveis (na linha do efeito túnel advertido por Paolo Fabbri, 1999).

Parte da semiótica e dos estudos culturais segue essa direção: estudos de caso único ou de poucos casos sobre acontecimentos em plataformas (como se pode observar em AAVV, 2023). Sem desconsiderar a riqueza desses aportes e sua validade metodológica, o volume e a variedade de conteúdo produzido em plataformas como TikTok ou Instagram também exigem um olhar que, sem perder de vista a importância central das observações em nível micro, não deixe de atender à sua inevitável escalabilidade. Insistimos, contudo, que esses aportes e perspectivas que viemos descrevendo são fundamentais para a construção do objeto e nos obrigam a um diálogo articulador.

A proposta sociosemiótica

Como já afirmamos, nosso objeto de pesquisa é o sistema de intercâmbio de sentido midiatizado. Ao analisá-lo, levantaremos marcas da textura do dispositivo técnico, aspectos dos formatos discursivos e formularemos hipóteses sobre práticas de uso a partir das possibilidades e restrições materiais. Sempre levando em conta a série histórica que torna possível cada um dos elementos que nos ocupam.

Em relação ao analista dos meios tradicionais do século passado, estamos em desvantagem. Já não nos deparamos com um objeto relativamente estável como aquele com o qual eles trabalhavam. Verón (1985) propunha, para o estudo dos meios gráficos, a adoção de períodos longos (dois anos), nos quais a conjuntura afetasse a amostra de maneira relativamente homogênea, apoiando-se na premissa — permita-se aqui a simplificação — de que o jornal ou a revista permanecem mais ou menos iguais ao longo do tempo. Em contrapartida, a vida do músico que faz posts ou do usuário que comenta se transforma entre os 16 e os 18 anos, entre os 19 e os 21, ou entre os 26 e os 28. Ele se apaixona, entra em conflito, entristece-se, muda sua equipe de trabalho, fracassa, obtém sucesso... e tudo isso afeta, em alguma medida, o modo como constrói sua comunicação. Além disso, por exemplo, o Facebook (Meta) altera a cada dois anos o ambiente visual e aquilo que pode ou não ser feito em todas as suas plataformas, estabelecendo novos possíveis, restrições e negociações por parte dos usuários no momento de comunicar.

Sem que isso implique negar a existência de traços que permanecem ao longo do tempo, a variabilidade individual em todo o sistema é um fator que não deve ser menosprezado. Não porque deva ser considerada no nível do caso isolado, mas sim como elemento ordenador da metodologia de coleta.

Em Uma mecânica metodológica, José Luis Fernández (2023, p. 154, *tradução nossa*) afirma que:

> *Não se pode selecionar um corpus útil e representativo para os objetivos da análise proposta sem levar em conta o contexto sociocultural no qual se desenvolvem os intercâmbios e o estado da arte existente sobre o tema. Se esses planos já são conhecidos, deverão ser sintetizados, de maneira mais ou menos exaustiva, no momento de decidir quais materiais serão efetivamente analisados e quais — o que é muito importante — ficarão de fora*[3].

Trata-se de uma seleção capaz de dar conta de que a unidade de análise é o sistema de intercâmbio midiatizado, uma vez que essa abordagem "permite deixar de lado [...] tanto as grandes configurações macro das midiatizações e suas relações com o social quanto a esperança paroquial da especialização" (Fernández, 2021, p. 196, *tradução nossa*)[4].

A dívida dessa proposta — se é que se pode atribuí-la ao autor — reside na resolução procedimental. Não se apresenta um caso exemplar de corpus que oriente o leitor. Por isso, antes de avançar, devem ficar claros os limites dentro dos quais perseguimos o objetivo da pesquisa:

- O olhar analítico deve ser micro, com expectativas de formalizações em nível meso (Fernández, 2021). Isso implica, necessariamente, a realização de descrições exaustivas de casos;
- O objeto tensiona as categorias tradicionais do qualitativo e do quantitativo. Ou seja, o olhar micro não obriga, sem desconhecer a escala do fenômeno, a manter exclusivamente uma perspectiva qualitativa.

Na etnografia digital, o objeto é descrito em relação ao que os nativos fazem. Nós entendemos que um olhar sociosemiótico parte do sentido produzido (Verón, 1987). É a materialidade que deve ser colocada em relação com suas condições produtivas. Borelli et al. (2024, p. 249) retomam Verón para afirmar que "o princípio da estrutura interna de um corpus parte da escolha em função de certa homogeneidade", recordando que, para ele, "todo texto 'é um objeto heterogêneo, suscetível de múltiplas leituras, situado no cruzamento de uma pluralidade de 'causalidades' diferentes".

Em Videla (2023), havíamos proposto que o mais próximo de uma amostra representativa é, em primeiro lugar, assumir uma limitação: dispõe-se apenas do equivalente a uma fotografia de um momento determinado. Qualquer generalização que pressuponha validade ao longo do tempo é, necessariamente, questionável.

Para avançar, deve-se escolher um conjunto de hashtags, de posts ou de contas (à maneira de conglomerados) suficientemente amplos e representativos do segmento a ser estudado e do conjunto no qual se inscrevem, selecionando um corpus de unidades por algum sistema de aleatoriedade que represente, da forma mais confiável possível, uma instantânea de um momento do sistema de intercâmbio analisado. Para diversos autores, esse tipo de seleção por conglomerados não garante a todos os elementos do universo a mesma probabilidade de aparecer (Marradi et al., 2018). Contudo, o acesso à totalidade sem uma restrição desse tipo é ainda mais problemático, sobretudo quando não existe, no momento da publicação deste artigo, uma API que capture tudo de uma rede, nem um "todo" que possa ser considerado representativo para além do momento da captura.

[3] Original em espanhol.

[4] Original em espanhol.

Considerando que a chave é evitar o viés fundacional em qualquer avanço científico, propomo-nos a dialogar com outras equipes que vêm desenvolvendo esforços nessa direção. Uma delas é a mencionada por Raimondo Anselmino, que propõe uma articulação entre a semiótica e as análises computacionais, denominada semiodata. Nessa perspectiva, "integra-se o estudo empírico da colocação em discurso a partir de um ponto de vista sociosemiótico (Verón, 1998) com o emprego de diversas ferramentas e métodos computacionais, bem como com a análise univariada e multivariada de dados e metadados" (Raimondo Anselmino et al., 2024, p. 194, *tradução nossa*)[5].

Seu esforço concentra-se em obter, por meio de procedimentos computacionais, o acesso ao maior número possível de unidades de análise, apesar das restrições impostas pelas plataformas. O material é reunido em bases de dados e, "por sua vez, o módulo de extração, transformação e carga de dados do Pentaho permitiu sistematizar as planilhas de cálculo de ambos os corpus em uma base de dados relacional. A partir da base de dados MySQL construída com o gerenciador MySQL WorkBench, realizou-se uma análise exploratória e descritiva que acompanhou a observação semiótica convencional" (Raimondo Anselmino et al., 2024, p. 211, *tradução nossa*)[6]. Embora os primeiros resultados apresentados possam parecer, a um olhar desatento, semelhantes aos oferecidos por soluções existentes desde 2010, como o Hootsuite (número de cliques, número de visualizações, quantidade de posts com *links*, entre outros), há na proposta um esforço consistente para avançar na complexidade estilística dos usuários e das publicações analisadas, esforço que deve ser valorizado e considerado.

O enfoque enfrenta dificuldades equivalentes àquelas que aqui pretendemos superar. A elevada qualidade da análise da informação, por vezes, entra em tensão com a massividade dos dados e com a origem não semiótica das APIs e de seus desenvolvedores. As informações que elas fornecem obrigam a uma leitura fora de contexto e expõem constantemente o analista ao risco da content analysis ao qual nos referíamos ao citar Krippendorff. Ou, na mesma linha, seguindo a proposta de Fernández sobre os megustos, tampouco seria possível saber o que significa esse número — seja qual for — de interações.

Um ponto de partida equivalente encontra-se no trabalho de Borelli et al. (2024), que, ao estudar fenômenos de circulação de sentido, explicam que inicialmente analisam "as métricas relacionadas à árvore máxima dos títulos e, posteriormente, as dos textos, bem como aplicamos a classificação hierárquica descendente apenas nos textos" (p. 249), o que lhes permite trabalhar, em seguida, com grafos arbóreos de relações semânticas. Dessa forma, tornam evidentes eixos de circulação discursiva que iluminam o peso de determinadas configurações de sentido, as quais de modo algum teriam sido compreendidas ou consideradas sem essa intervenção.

Raimondo et al. (2024) também realizam o esforço de articular o engaging com o uso de emojis nos posts, dando conta de estilos de consumo dos usuários. Os autores afirmam que "assim como ocorre com os demais recursos desse tipo, os emojis funcionam como marcas de certos mecanismos significantes atuantes nas propriedades específicas do discurso informativo plataformizado" (Raimondo et al., 2024, p. 212, *tradução nossa*)[7]. Essa afirmação nos conduz a questões complexas acerca de seu funcionamento na instância de reconhecimento. Tomemos, por exemplo, um simples emoji de um rosto com uma lágrima na caixa de comentários. Sem entrar na distinção de uso por segmento etário (para menores de 25 anos, tende a indicar empatia com o outro; para maiores, tende a expressar o sentimento próprio, e sua repetição indicaria maior intensidade), esse emoji indica que o post entristeceu o comentarista (algo como "vi isso e me deixou triste")? Trata-se de uma manifestação de emoção genérica diante do post? Indica um choro moderado por comoção? Ou expressa empatia, acompanhando o sentimento do outro ("vi seu post e compartilho o que você sente")? Esse problema já foi enfrentado em Videla (2023) e, até o presente, não dispomos de outra proposta senão assinalá-lo como um aspecto a ser considerado, insistindo em que um dos caminhos possíveis de solução é avançar na relevância dos estudos estilísticos entre os usuários.

Na proposta de semiodata, evidencia-se tanto o potencial quanto os desafios envolvidos na geração de descritores. Em Videla (2023), por exemplo, uma parte significativa do trabalho evidenciou a dificuldade de nomear muitos dos vídeos do TikTok em termos genérico-estilísticos e temáticos. Ali apresentamos uma classificação que, embora relativamente ordenadora do que ocorria no TikTok, mostrou-se deficiente em razão do elevado nível de matizes e flexibilizações conceituais mobilizadas. Em etapas posteriores, optamos por falar, em vez de gêneros, em formatos, conforme propõe Fernández (2023). Isso se deve ao fato de que a noção de gênero remete à história da literatura ocidental dos últimos dois mil anos, enquanto a ideia de formato convoca metadiscursos menos rígidos, ainda que relativamente estáveis — embora tampouco resolva plenamente a variedade e a diversidade do que emerge nas plataformas. Assim como enfrentamos dificuldades para diferenciar a presença do humor como formato genérico do humor como matiz estilístico, Raimondo et al. (2024, p. 211) acabam por encontrar a categoria "conteúdo multimídia" para designar algo que "compreende publicações com conteúdo visual ou audiovisual que costumam versar sobre temáticas consideradas 'brandas' (espetáculos, entretenimento, interesse humano)". O objeto obriga o pesquisador a adotar soluções que anulam as diferenças entre uma temática (espetáculos), um tipo discursivo (entretenimento), uma classificação branda (interesse humano) e um aspecto do dispositivo técnico (conteúdo audiovisual).

Esforços como os que descrevemos chamam nossa atenção para aquilo que deve ser o centro de nossas discussões: a definição de acordos na busca pelos aspectos a analisar do objeto e acordos na passagem de marca a pegada. Acordos sobre como os convertemos em atributos/variáveis.

Sobre grades e matrizes

Seguindo Fernández (2023), é necessário atender tanto às propriedades do sistema de troca quanto às do objeto de estudo. Se este se compõe de posts, posso iniciar a pesquisa sem considerar seus comentários. Sua vida discursiva é, em princípio, independente e se relaciona mais com a plataforma e

[5] Original em espanhol.
[6] Original em espanhol.
[7] Original em espanhol.

com outros posts, que, estes sim, devem ser considerados e levantados. Em contrapartida, se a pergunta de pesquisa se orienta para os comentários, torna-se impossível estudá-los sem uma análise do post e da plataforma. Do mesmo modo, não é possível deixar de incluir na amostra posts equivalentes e seus comentários. Caso contrário, corre-se o risco de afirmar que jogos visuais com emojis e palavras ou comentários provenientes do repertório popular estão presentes em todos os posts de músicos e fãs, algo que é, em princípio, falso. Há músicos cujos posts não recebem esse tipo de comentário. Há variações entre artistas, entre plataformas e entre posts de um mesmo artista.

Fernández (2023) propõe que o primeiro passo necessário consiste em realizar a descrição do objeto/corpus. Trata-se de algo tão simples quanto sofisticado e arriscado: aproximar-se do objeto, observá-lo e anotar o que ele tem. A vantagem é que aquilo sobre o que o pesquisador falará será efetivamente o objeto, e não um preconceito oriundo de uma conversa de bar. O risco é o de gerar categorias banais ou denominações pomposas e desnecessárias. A partir dessa descrição, a proposta é desenvolver grades de análise que são, por sua vez, para esse autor, a primeira análise do objeto.

A matriz de dados é geralmente considerada uma ferramenta básica nas ciências sociais. Trata-se de um instrumento "formado pelo cruzamento de um feixe de vetores paralelos verticais e um feixe de vetores paralelos horizontais" (Marradi et al., 2018). Dependendo da tradição de pesquisa, os casos estarão nas linhas ou nas colunas. A matriz é uma construção do analista.

O grillado proposto por Fernández constitui o passo prévio. Aquilo que um post possui deve ser traduzido em um conjunto de aspectos que permitam sua descrição. Para Lazarsfeld, estamos diante de aspectos ou dimensões cuja nomeação (o nome na linha ou na coluna) é um processo literário (1973) e que deve ser decomposto em indicadores.

"Somente a partir desse ponto é possível começar a discutir o tratamento dos dados a partir de diferentes pontos de vista, e o grillado começa a se converter em uma matriz de dados" (Fernández, 2023, p. 211). Estamos diante da passagem de um aspecto — que para nós é a marca ou pegada do processo produtivo — ao momento em que ele é convertido, ao ser nomeado, em sua presença na grade. Trata-se de um passo inevitável do trabalho analítico, que deve ser realizado manualmente. Para isso, e para facilitar a visualização humana da análise, é mais favorável que os casos estejam dispostos em colunas. As bases de dados, em geral, organizam os casos em linhas. Por isso, com o simples uso da função de transposição, pode-se gerar uma alternativa para o pesquisador humano, novamente reversível mediante o uso da mesma função.

As dificuldades dos indicadores

Se o olhar do analista incide sobre o post — por exemplo, um videoclipe musical —, aquilo que ocorre na tela constituirá aspectos de textura. O ritmo poderá, por exemplo, ser pré-codificado na grade como agitado, calmo ou intenso.

Aquilo que, para o pesquisador humano, será resolvido por meio de um acordo simples e com alto grau de consenso — mensurável em estudos de validação — não é tão simples ao interagir com algoritmos básicos ou de *deep learning*. O software que determinará o que é "calmo" deve considerar o BPM? A quantidade de cortes de plano por minuto? Se trabalhássemos com um transformer que converta vídeo em texto, talvez a quantidade de cenas descritas por unidade de tempo pudesse funcionar como uma unidade de medida do ritmo. Aquilo que é óbvio para o observador humano é difícil de traduzir e padronizar.

O problema se torna ainda mais complexo quando se trata de processar categorias mais elaboradas, como a de motivo (Fernández, 2023). Isso porque estamos diante de processos de metaforização (Lazarsfeld, 1973) de aspectos do objeto. Em um projeto com colegas do Brasil, de Mar del Plata, de Torino e da Polônia, estamos trabalhando a análise de parágrafos produzidos por uma IA. Tomemos como exemplo um dos parágrafos analisados:

> *Após muitas discussões — e algumas ameaças cruzadas — decidiram dividir o prêmio e usar uma parte para montar juntos uma consultoria. Contra todas as expectativas, descobriram que suas diferenças, bem canalizadas, se complementavam. Martina cuidava da estratégia e dos processos, enquanto Diego se ocupava das relações com os clientes. Não se tornaram amigos, mas aprenderam a se respeitar, e a empresa prosperou mais do que qualquer um dos dois teria imaginado.*

A partir de nossas experiências pessoais, compreendemos rapidamente que aí está presente a figura/motivo do empreendedor: investiramum montante de dinheiro e criaram uma consultoria que eles mesmos administravam.

Concordaremos com o leitor que empresário se distingue de empreendedor pelos montantes de investimento. No entanto, criadores de aplicativos que alcançam avaliações milionárias e partem de aportes familiares com vários zeros também são considerados empreendedores. Outro critério possível é a quantidade de empregados: poucos empregados caracterizariam um empreendedor. Mas é difícil diferenciar um empreendedor de uma empresa familiar, que poderia muito bem ser o caso de uma consultoria formada por dois amigos. O mesmo problema surgiria se insistíssemos no nível de faturamento. O que significa "prosperou mais do que imaginaram"? Ainda que a definição de indicadores seja uma tarefa própria da pesquisa social desde quase suas origens, não é irrelevante observar que, diante desse tipo de objeto — assumimos que esse parágrafo produzido por IA coloca os mesmos desafios que qualquer comentário humano em plataformas —, a dificuldade se amplia.

Em Videla (2025a), demonstramos que esse caminho voltado à automatização da análise é compartilhado por diversas equipes em diferentes universidades. Trata-se de um esforço alinhado à busca por algoritmos capazes de resolver a classificação do corpus, com as dificuldades que mostramos que isso implica. Nesse trabalho, contudo, apresentamos um avanço que combina a grade analítica com o uso de grandes modelos de linguagem (*Large Language Models* – LLM).

O trabalho do analista, em estudos de caso como o que citamos, consiste em estabelecer manualmente as variáveis qualitativas (atributos) da grade, em relação aos saberes prévios e à sua formação. A primeira coluna ao lado do atributo escolhido (seguindo o exemplo, o motivo do empreendedor individual) será destinada à sua definição (aparece no texto a construção de um personagem que decide empreender uma atividade comercial, de forma individual, como se partisse do

zero). Pode-se repetir uma linha com o motivo do empreendedor de uma pequena empresa familiar ou do empreendedor tecnológico, tantos atributos quantos a primeira leitura de uma amostra permitir identificar ao analista.

Em Videla (2025a), mostramos como um LLM é capaz de compreender essa definição e devolver automaticamente, por meio de um código em Google App Script, o valor 1 quando esse motivo aparece no texto analisado, ou deixar o campo em branco quando não aparece — de modo análogo ao que faria um analista humano. Embora a equipe de pesquisa ainda se encontre em etapas iniciais, os primeiros resultados são alentadores nesse sentido.

Discussão

Ao longo deste trabalho, analisamos o desafio que envolve a construção de um corpus representativo para o estudo de fenômenos em plataformas midiáticas, particularmente no campo musical. O trabalho analítico tensiona a capacidade do pesquisador de capturar a quase totalidade do universo do objeto com as limitações materiais de sua abordagem. Por isso, insiste-se na importância de não perder de vista as metodologias estatísticas que têm dominado o campo até o presente, ao mesmo tempo em que se propõe um enfoque ancorado nos aprendizados da sociosemiótica das midiatizações.

Desse modo, para construir uma amostra representativa, torna-se necessário considerar o sistema de troca de sentidos como unidade de análise. Não apenas o post em si, mas também seus comentários e a plataforma em que se produz, já que essas interações não são independentes entre si.

A proposta de grillado de Fernández (2023) apresenta-se, assim, não apenas como uma referência epistemológica, mas como uma ferramenta metodológica capaz de articular-se com o futuro da pesquisa mediada por inteligência artificial, sempre em diálogo com campos científicos afins, como a ciência de dados e a etnografia digital.

Entendemos, portanto, que a solução para o problema da construção do corpus e de sua análise não é individual. Ela exige, além da articulação com as engenharias de dados, a produção de consensos de pesquisa. O trabalho isolado de diferentes equipes orientadas a objetivos equivalentes, sem acordos terminológicos prévios, é pouco produtivo e economicamente ineficiente. Não se trata apenas de negociar com um cientista da computação ou com um etnógrafo uma linha de uma grade analítica, nem de proclamar a centralidade de um campo de estudos em detrimento de outro. Trata-se, sobretudo, de manter abertos espaços de diálogo.

Esforços como os das equipes multidisciplinares coordenadas por Natalia Raimondo na UNR, da Rede de Pesquisadores em Plataformas (REDINPLA) e da Rede de Pesquisadores em Midiatizações (Midiaticom) apontam nessa direção, e este trabalho se insere nesse percurso coletivo de discussão e construção.

Referências

ALBALADEJO ORTEGA, S. *LEGO brick learning*: hacia un modelo de alfabetización transmediática a través del storytelling. 2017. Tese (Doutorado) – Universidad Católica San Antonio de Murcia, Murcia (Espanha), 2017.

ARAOZ V.; CELLONE, F. Restricciones y posibilidades de la investigación con datos digitales. In: PORTO LOPEZ, P. (org.). Intervenciones semióticas: focalizar, transformar, expandir: *actas* del 11° Congreso Argentino de Semiótica. Buenos Aires: Libros de Crítica, Área Transdepartamental de Crítica de Artes, 2025.

ARONICA, S. La etnografía digital: descripción de un caso de aplicación para el análisis de interacciones virtuales. In: *VI Simposio Argentino sobre Tecnología y Sociedad (STS 2019)* – JAIIO 48. Salta, 2019. p. 28–39.

BORELLI, V.; FRIGO, D.; ROMERO, L. M. Circulação de sentidos em textos noticiosos sobre mortes pela pandemia no Brasil. *MATRIZes*, São Paulo, v. 18, n. 1, p. 239–263, 2024. DOI: https://doi.org/10.11606/issn.1982-8160.v18i1p239-263.

CETOCCHI, C.; SZNAIDER, B. *Mediatizaciones en pandemia*. Buenos Aires: Editorial Carrera Comunicación UBA, 2023.

CORBETTA, P. *Metodología y técnicas de investigación social*. Madrid: McGraw-Hill, 2007.

FABBRI, P. La caja de los eslabones que faltan. In: FABBRI, P. *El giro semiótico*. Barcelona: Gedisa, 1999.

FAJARDO, H. *Estudio comparativo entre Apache Spark y Apache Flink en el procesamiento de streaming en entornos Big Data*. 2023. Trabalho de Conclusão de Curso (Especialização em Inteligência de Dados Orientada a Big Data) – Universidad Nacional de La Plata, 2023.

FERNÁNDEZ, J. L. *Los lenguajes de la radio*. Buenos Aires: Atuel, 2023.

FERNÁNDEZ, J. L. *Una mecánica metodológica*. Buenos Aires: La Crujía, 2023.

FERNÁNDEZ, J. L. *Vidas mediáticas*. Buenos Aires: La Crujía, 2023.

GARCÍA, et al. *Ciencia de datos*: técnicas analíticas y aprendizaje estadístico. Tarragona: Altaria, 2018.

HERNÁNDEZ SAMPIERI, R. et al. *Metodología de la investigación*. México: McGraw-Hill, 2005.

IRIGARAY, F. De los conceptos de espacio, territorio y lugar al de posterritorio: territorialidad expandida en el ecosistema urbano. In: *Transmedia storytelling*. Buenos Aires: RIA Editorial, 2021.

KRIPPENDORFF, K. Validity in content analysis. In: MOCHMANN, E. (ed.). *Computerstrategien für die Kommunikationsanalyse*. Frankfurt: Campus, 1980. p. 69–112. Disponível em: http://repository.upenn.edu/asc_papers/291. Acesso em: 29 nov. 2025.

LAZARSFELD, P. De los conceptos a los índices empíricos. *Metodología de las ciencias sociales*, v. 1, p. 35–46, 1973.

LÉVI-STRAUSS, C. *Antropología estructural*. Buenos Aires: Paidós, 1987.

LOVATO, A. The transmedia script for nonfictional narratives. In: *Exploring transmedia journalism in the digital age*. Hershey: IGI Global Scientific Publishing, 2018. p. 235–252.

MARRADI, A.; ARCHENTI, N.; PIOVANI, J. I. *Manual de metodología de las ciencias sociales*. Buenos Aires: Siglo Veintiuno, 2018.

PINK, S. et al. *Etnografía digital*. Madrid: Ediciones Morata, 2019.

PRATTEN, H. *Getting started with transmedia storytelling*. Londres: Robert Pratten, 2011.

RAIMONDO ANSELMINO, N. et al. El engagement de La Nación y Clarín en Facebook. *Zer*, Bilbao, v. 29, n. 57, p. 191–220, 2024. DOI: https://doi.org/10.1387/zer.26698.

SABINO, C. A. *El proceso de investigación*. Buenos Aires: Panamericana Editorial, 1992.

SÁNCHEZ PICCARDI, M. L.; PALOMO, L. E. Del big data al fast data: enfoques modernos de streaming de datos para el procesamiento de datos masivos en tiempo real. *Difusiones*, v. 21, n. 21, p. 38–58, 2021. Disponível em: http://ediciones.ucse.edu.ar/ojsucse/index.php/difusiones/article/view/401. Acesso em: 29 nov. 2025.

SAUTU, R. et al. Recomendaciones para la redacción del marco teórico, los objetivos y la propuesta metodológica de proyectos de investigación. In: *Manual de metodología*. Buenos Aires: CLACSO, 2005.

THEVIOT, A. Devenir « ami » avec 4500 enquêtés: les enjeux éthiques de l'analyse d'interfaces semi-privées. *Tic & Société*, v. 7, n. 2, 2014. Disponível em: http://journals.openedition.org/ticetsociete/1608. DOI: https://doi.org/10.4000/ticetsociete.1608. Acesso em: 29 nov. 2025.

VERÓN, E. *La mediatización*. Buenos Aires: Universidad de Buenos Aires, Facultad de Filosofía y Letras, Cursos y Conferencias, 1986.

VERÓN, E. *La semiosis social*: fragmentos de una teoría de la discursividad. Barcelona: Gedisa, 1987.

VIDELA, S. La investigación en plataformas: entre la construcción de la especificidad y la apertura interdisciplinaria. Apresentação em: *Coloquio CIM*. Universidad Nacional de Rosario, 13 nov. 2025. 2025a.

VIDELA, S. Sistemas de intercambio en plataformas entre músicos y fans. In: VIDELA, S.; ROES DALMOLIN, A. (org.). *Investigar en plataformas mediáticas desde América Latina*: un estado de situación y una herramienta para el diálogo. Santa Maria: Editorial Universidad Federal de Santa María, 2025b. p. 39–70.

VIDELA, S. TikTok en confinamiento: análisis de los efectos de la pandemia de COVID-19 en las formas de producción de contenido. In: CETOCCHI, C.; SZNAIDER, B. *Mediatizaciones en pandemia*. Buenos Aires: Editorial Carrera Comunicación UBA, 2023.

Artigo submetido em 20/07/2025
Aceito em 18/12/2025