

NID NLP: plataforma de análise textual para pesquisa aplicada em Comunicação

NID NLP: a text analysis platform for applied research in Communication

Márcio Carneiro dos Santos¹

Universidade Federal do Maranhão (UFMA) – São Luís/MA

marcio.carneiro@ufma.br

 <https://orcid.org/0000-0002-0663-1300>

RESUMO: O artigo apresenta o NID NLP, plataforma de análise textual desenvolvida para apoiar pesquisas em Comunicação com *corpora*² extensos. O artefato integra preparação de dados, análise lexical, reconhecimento de entidades, modelagem de tópicos, redes, similaridade e exportação em um fluxo modular. A pesquisa é aplicada e adota desenvolvimento iterativo. A contribuição central está na integração de métodos consolidados com rastreabilidade, preservação do texto original e retorno aos documentos, sem delegar ao sistema a interpretação científica.

Palavras-chave: Comunicação; Análise textual; Métodos digitais; Processamento de Linguagem Natural; Pesquisa aplicada.

Abstract: This article presents NID NLP, a text analysis platform designed to support Communication research with large *corpora*. The artifact integrates data preparation, lexical analysis, named-entity recognition, topic modeling, networks, similarity, and export in a modular workflow. The study follows an applied, iterative development approach. Its main contribution is the integration of established methods with traceability, preservation of original texts, and systematic return to source documents, without delegating scientific interpretation to the system.

Keywords: Communication; Text analysis; Digital methods; Natural Language Processing; Applied research.

¹ Pesquisador chefe do Núcleo de Inteligência de Dados da UFMA - NID. Trabalho desenvolvido dentro de projeto financiado pela FAPEMA e CNPq. Email: marcio.carneiro@ufma.br

² **Corpora:** plural de *corpus*. Em pesquisa textual, refere-se a conjuntos organizados de textos ou documentos reunidos para análise.

INTRODUÇÃO

A digitalização das práticas comunicacionais alterou também as condições de pesquisa no campo da Comunicação. Notícias, postagens em redes sociais, comentários, transcrições, documentos institucionais e resultados de mecanismos de busca podem formar *corpora* com centenas ou milhares de registros. Nessas situações, procedimentos exclusivamente manuais tornam-se onerosos e, em alguns casos, insuficientes para reconhecer recorrências, comparar subconjuntos ou localizar relações dispersas no material. O problema não elimina a necessidade de leitura, contextualização e teoria; ao contrário, amplia a importância de decisões metodológicas sobre coleta, preparação, análise e validação.

A aproximação entre Comunicação e métodos computacionais antecede a expansão recente da inteligência artificial generativa. Perspectivas como *text as data*, métodos digitais e *Computational Communication Science* demonstraram que procedimentos automatizados podem ampliar a escala de observação dos textos, desde que seus resultados não sejam confundidos com interpretações autossuficientes. Grimmer e Stewart (2013) ressaltam que modelos de análise textual são representações incompletas da linguagem e exigem validação substantiva. Van Atteveldt e Peng (2018), por sua vez, associam a expansão da pesquisa computacional em Comunicação ao crescimento simultâneo da disponibilidade de dados e da capacidade de processamento, sem desconsiderar problemas de validade e mensuração.

Uma barreira persistente está na distância entre as perguntas de pesquisa e os conhecimentos técnicos tradicionalmente necessários para operar bibliotecas, preparar ambientes computacionais e adaptar códigos. Essa dificuldade atravessa uma trajetória de pesquisa aplicada que, em diferentes momentos, produziu experimentos de coleta de notícias, análise de redes sociais, exploração textual, automação jornalística e visualização de dados. O aprendizado acumulado indicou que o desafio não era apenas disponibilizar algoritmos, mas organizá-los em percursos compreensíveis para

pesquisadores que trabalham com textos sem necessariamente possuir formação em programação (Santos, 2016; Santos, 2018; Santos, 2019).

É nesse contexto que foi desenvolvido o NID NLP³, plataforma que reúne em uma única interface procedimentos antes distribuídos em *scripts* e ferramentas independentes. Sua navegação acompanha nove etapas: Entrada, Preparação, Visão Geral, Análise Textual, Entidades, Tópicos, Redes, Similaridade e Exportação. A organização foi concebida para aproximar recursos computacionais de questões recorrentes na investigação: Quais termos predominam? Onde aparecem? Quem é mencionado? Que atores surgem associados? Que estruturas temáticas atravessam o corpus? Quais documentos apresentam maior proximidade? E em quais textos os resultados podem ser novamente verificados?

A ênfase no retorno aos documentos ganha relevância diante da popularização dos grandes modelos de linguagem. Interfaces conversacionais reduziram a barreira técnica para tarefas de síntese, classificação e exploração, mas parte das operações internas pode permanecer pouco visível ao usuário. Métodos computacionais tradicionais também dependem de modelos, parâmetros e escolhas de preparação e estão sujeitos a erro; sua vantagem, em determinadas situações, reside na possibilidade de tornar explícitas as operações que ligam corpus, transformação e resultado. O NID NLP, portanto, não é proposto como alternativa excludente à IA generativa, mas como infraestrutura complementar para situações em que rastreabilidade e inspeção das evidências são requisitos importantes.

A questão que orienta o artigo é: como integrar métodos consolidados de análise textual em uma plataforma acessível a pesquisadores da Comunicação, preservando a rastreabilidade das operações e a possibilidade de retorno aos documentos de origem? O objetivo é apresentar o processo de desenvolvimento do NID NLP, descrever sua arquitetura funcional e discutir suas contribuições e limitações como artefato de pesquisa aplicada.

³ **NLP (Natural Language Processing)**: em português, **Processamento de Linguagem Natural (PLN)**. É a subárea da inteligência artificial e da computação dedicada ao desenvolvimento de métodos capazes de analisar, interpretar e processar textos e outras formas de linguagem humana.

1 MÉTODOS COMPUTACIONAIS, ESCALA E RASTREABILIDADE

Textos sempre ocuparam posição central na pesquisa em Comunicação, mas a digitalização acrescentou uma propriedade operacional: além de objetos simbólicos, podem ser tratados como dados passíveis de organização e processamento. Quando o corpus cresce, métodos computacionais permitem observar frequências, distribuições, entidades, associações e proximidades que dificilmente seriam identificadas por leitura linear de todo o material. Essa mudança de escala não substitui abordagens qualitativas. Ela cria um movimento entre níveis de observação: medidas agregadas ajudam a localizar padrões, e os documentos associados a esses padrões podem ser novamente examinados de forma contextual.

A noção de *distant reading*, formulada por Moretti (2007), é útil para compreender esse deslocamento. Gráficos, mapas e outras representações abstratas tornam visíveis estruturas que não aparecem com facilidade na leitura de unidades isoladas. Para a Comunicação, o interesse dessa perspectiva está menos em opor leitura próxima e distante do que em conectá-las. Um resultado computacional pode funcionar como mecanismo de seleção e navegação: aponta ocorrências, grupos ou relações que merecem leitura posterior. Essa circulação também converge com a proposta dos métodos digitais de considerar as propriedades específicas dos objetos produzidos em ambientes conectados (Rogers, 2013).

Os métodos incorporados a esse tipo de investigação não respondem à mesma pergunta. Frequências e n-gramas permitem observar recorrências lexicais; concordâncias recolocam termos em contexto; reconhecimento de entidades organiza referências a pessoas, instituições e lugares; modelagem de tópicos procura padrões de coocorrência distribuídos em coleções documentais; redes transformam ocorrências em estruturas relacionais; medidas de similaridade aproximam documentos a partir de uma representação escolhida. A utilidade de cada técnica depende, portanto, do problema investigado e das unidades consideradas. Um mesmo corpus pode produzir leituras distintas porque a seleção de documentos, o pré-processamento e os parâmetros alteram a representação computacional disponível para análise.

Essa diversidade torna inadequada a ideia de uma versão única e definitivamente “limpa” do texto. Remover pontuação, uniformizar caixa, excluir *stopwords*⁴ ou lematizar⁵ pode ser produtivo em determinadas operações lexicais, mas prejudicial em tarefas que dependem de estrutura sintática, capitalização ou contexto. A preservação do documento original e a produção de representações derivadas para finalidades específicas constituem, assim, decisões metodológicas e não apenas procedimentos técnicos. No NID NLP, essa separação funciona como base para recuperar o texto que antecede qualquer transformação e para registrar que os resultados são dependentes da forma de representação adotada.

A literatura recente sobre *text as data* reforça a necessidade de validação. Birkenmaier, Lechner e Wagner (2024) mostram que medidas derivadas computacionalmente exigem procedimentos explícitos de validação, especialmente quando passam a representar conceitos substantivos. Em vez de tratar a saída de um algoritmo como evidência suficiente, torna-se necessário documentar escolhas de corpus, transformações textuais, parâmetros e critérios de interpretação. Na Comunicação Computacional, Van Atteveldt *et al.* (2019) defendem ainda fluxos reutilizáveis, nos quais dados e código possam ser documentados como parte do processo científico.

A emergência da IA generativa modifica esse cenário, mas não elimina tais exigências. Sistemas conversacionais permitem interagir com textos por meio de linguagem natural e são úteis em tarefas de exploração, síntese e apoio à produção. Entretanto, quando a investigação exige saber quais unidades foram contabilizadas, como uma relação foi construída ou quais documentos sustentam determinado padrão, procedimentos estruturados mantêm valor metodológico. A distinção não é entre técnicas “corretas” e “incorretas”, mas entre diferentes formas de produzir e verificar evidências.

⁴ **Stop words:** palavras muito frequentes em um idioma, como artigos, preposições e algumas conjunções, que podem ser removidas em determinadas análises por contribuírem pouco para a diferenciação temática ou lexical dos textos

⁵ **Lema / lematização:** lema é a forma básica de uma palavra, usada como referência para suas variações flexionadas. Lematização é o processo computacional de reduzir essas diferentes formas ao respectivo lema, como “falou”, “falando” e “falaram” → “falar”.

Métodos de NLP também erram, sobretudo em domínios específicos ou em línguas com menor disponibilidade de recursos, mas seus estágios podem ser isolados, parametrizados e inspecionados.

A rastreabilidade pode ser entendida aqui em quatro níveis complementares. O primeiro é documental: cada ocorrência deve permanecer associada ao texto de origem. O segundo é transformacional: deve ser possível distinguir o texto recebido das versões usadas por cada técnica. O terceiro é paramétrico: escolhas que afetam o resultado precisam ser explicitáveis e, quando possível, reproduzíveis. O quarto é interpretativo: visualizações e medidas devem conduzir novamente aos casos que lhes dão sustentação. Esses níveis não resolvem o problema da validade, mas criam condições para que erros, decisões e divergências possam ser examinados em vez de permanecerem ocultos numa resposta final.

Essa perspectiva orienta o NID NLP: automação não é entendida como interpretação científica. Frequências, entidades, tópicos, redes e similaridades são representações parciais do corpus. O sistema procura tornar essas representações navegáveis e vinculadas aos textos que as originaram, deslocando o papel do computador para localizar, calcular, relacionar e visualizar, enquanto a atribuição de significado permanece vinculada à pergunta de pesquisa, ao referencial teórico e ao contexto.

2 PERCURSO METODOLÓGICO E DESENVOLVIMENTO DO ARTEFATO

O trabalho caracteriza-se como pesquisa aplicada com desenvolvimento iterativo de artefato. A plataforma deriva de um projeto de maior duração voltado à criação de protocolos e ferramentas para análise de grandes volumes de texto em Comunicação e áreas correlatas. A finalidade não é propor novos algoritmos de Processamento de Linguagem Natural, mas combinar métodos consolidados em uma solução destinada a um problema concreto: permitir que pesquisadores explorem *corpora* extensos sem

depende, para cada etapa, da edição de código ou da montagem de ambientes computacionais distintos.

O público de referência é formado por pesquisadores que têm o texto como matriz relevante de seus objetos e que podem trabalhar tanto com um documento isolado quanto com coleções estruturadas. Por isso, a especificação considerou formas de entrada comuns em rotinas de pesquisa: texto colado, arquivos TXT e planilhas CSV/XLSX. Nos dados tabulares, a lógica adotada é preservar as demais colunas como metadados, porque informações como data, fonte, autoria, categoria ou grupo podem ser necessárias para filtragens e comparações posteriores. Essa decisão evita que a preparação para NLP empobreça o corpus ao descartar informações coletadas originalmente.

A opção pelo desenvolvimento de artefato aproxima-se da Design Science, na qual a produção e avaliação de soluções para problemas relevantes também pode constituir atividade de pesquisa (Hevner *et al.*, 2004). No caso aqui descrito, experiências anteriores com ferramentas específicas revelaram duas limitações recorrentes: fragmentação do fluxo de trabalho e dependência de conhecimento técnico para integrar coleta, preparação, análise e visualização. O NID NLP surge como tentativa de transformar esse aprendizado disperso em infraestrutura modular.

A especificação foi orientada por quatro requisitos: reduzir a barreira operacional; aceitar formatos frequentes na pesquisa com textos; preservar o conteúdo original e seus metadados; e permitir retorno sistemático às unidades que sustentam os resultados. O desenvolvimento ocorreu incrementalmente, recuperando procedimentos já experimentados em Python, atualizando bibliotecas e reorganizando-os em uma aplicação web. Sistemas de IA generativa foram utilizados como apoio à prototipação e à implementação de requisitos, mas não para definir autonomamente a lógica metodológica do sistema. A seleção das técnicas e a avaliação das saídas permaneceram orientadas pelas necessidades do projeto.

Outro critério foi priorizar o processamento de textos em português. Isso envolve preservar acentos e caracteres Unicode, utilizar recursos linguísticos adequados ao idioma e evitar que decisões herdadas de exemplos em inglês sejam aplicadas

automaticamente ao corpus. A opção é relevante porque parte das ferramentas disponíveis em NLP foi historicamente desenvolvida e avaliada com maior intensidade para a língua inglesa. A interface simplifica o acesso, mas não elimina essa assimetria: resultados de entidades, lematização ou outras análises linguísticas continuam dependentes da qualidade dos modelos empregados e precisam ser verificados pelo pesquisador.

A plataforma foi estruturada de modo modular, separando interface e funções analíticas. O percurso geral pode ser sintetizado como: objeto de pesquisa → constituição do corpus → preparação → exploração computacional → visualização → retorno aos documentos → interpretação. A sequência é orientadora, não obrigatória; cada investigação pode utilizar apenas os módulos compatíveis com sua pergunta.

A decisão de separar o texto original das versões de processamento também atravessa a arquitetura. O sistema considera que uma tarefa de frequência pode requerer normalização distinta daquela utilizada no reconhecimento de entidades, enquanto a modelagem de tópicos necessita de uma representação específica dos documentos. A separação entre essas camadas reduz a possibilidade de aplicar transformações agressivas ao corpus inteiro e cria uma referência comum para auditoria. Metodologicamente, o texto original funciona como ponto de retorno sempre que uma medida agregada exige interpretação contextual.

A avaliação realizada nesta fase é predominantemente funcional. Diferentes *corpora* foram usados para verificar se os módulos operavam de forma integrada, se produziam saídas exploráveis e se permitiam recuperar os documentos associados aos resultados. Essa etapa não equivale a uma avaliação sistemática de usabilidade nem à validação de cada algoritmo incorporado. Testes com pesquisadores do público-alvo, coleta de dificuldades e novos ciclos de aperfeiçoamento constituem etapa posterior do projeto.

Entre os testes de integração foram utilizados conjuntos de conteúdos jornalísticos e ambientais, além de materiais empregados em experiências anteriores de análise textual. Um dos *corpora* recentes reuniu documentos relacionados à Amazônia obtidos

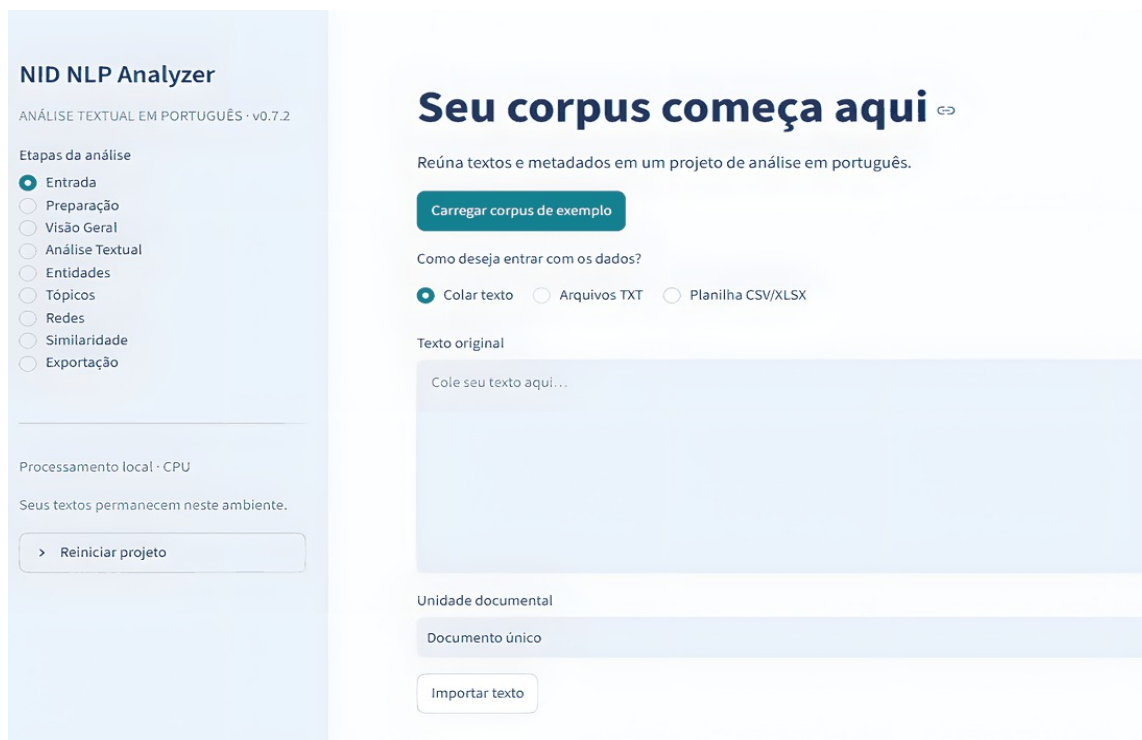
em resultados de busca na web e foi processado para verificar, no mesmo fluxo, frequência, dispersão de termos, entidades, tópicos e redes de coocorrência. Os resultados tiveram função exclusivamente demonstrativa: o objetivo não era produzir conclusões substantivas sobre o corpus, mas testar continuidade entre módulos, consistência das saídas e capacidade de retorno aos documentos.

3 ESTRUTURA E FUNCIONALIDADES DO NID NLP

A interface traduz o processamento computacional em etapas próximas ao fluxo de uma investigação, evitando que o usuário seja inicialmente confrontado com nomes de bibliotecas ou algoritmos. A tela inicial permite colar texto, carregar arquivos TXT ou importar planilhas CSV/XLSX; a navegação lateral apresenta os nove módulos do sistema. A Figura 1 registra essa organização.

A ordem dos módulos sugere uma progressão entre constituição do corpus, diagnóstico, exploração e saída, mas não representa um protocolo obrigatório. O pesquisador pode iniciar por uma leitura geral, restringir-se às análises lexicais ou combinar operações mais complexas. A interface procura traduzir capacidades técnicas em perguntas investigativas: localizar recorrências, identificar atores, observar associações, reconhecer estruturas temáticas, comparar documentos e exportar evidências para outras etapas da pesquisa.

Figura 1 – Tela inicial e etapas de análise do NID NLP



Fonte: Elaborado pelo autor

Tabela 1 – Princípios de projeto do NID NLP

Princípio	Decisão de implementação	Finalidade metodológica
Preservação	Manter o texto original e criar representações derivadas conforme a tarefa.	Evitar que uma única limpeza altere indiscriminadamente todas as análises.
Rastreabilidade	Associar resultados a documentos, trechos ou unidades que os originaram.	Permitir verificação e retorno ao contexto.
Modularidade	Separar etapas e permitir uso independente dos módulos.	Subordinar a técnica à pergunta de pesquisa.
Acessibilidade	Oferecer interface web sem exigir edição de código pelo usuário.	Reduzir barreira operacional sem automatizar a interpretação.

Fonte: Elaborado pelo autor

Tabela 2 – Síntese das funcionalidades da plataforma

Etapas	Função principal	Saídas e vínculo com o corpus
Entrada	Constituir o corpus por texto, TXT, CSV ou XLSX.	Documentos e metadados preservados para as etapas seguintes.
Preparação	Normalizar e gerar representações adequadas a cada técnica.	Texto original permanece disponível para auditoria e contexto.
Visão Geral	Diagnosticar dimensões e características básicas do corpus.	Medidas descritivas apoiam decisões antes de análises mais complexas.
Análise Textual	Explorar frequência, concordância, dispersão e n-gramas.	Listas e gráficos podem ser confrontados com ocorrências nos textos.
Entidades	Identificar pessoas, organizações, lugares e outras categorias reconhecidas pelo modelo.	Entidades permanecem ligadas a documentos e trechos de ocorrência.
Tópicos	Explorar padrões de coocorrência temática em <i>corpora</i> maiores.	Termos e documentos representativos apoiam interpretação dos tópicos.
Redes	Construir relações de coocorrência e explorar estrutura relacional.	Nós e arestas podem ser examinados com seus contextos textuais.
Similaridade	Comparar proximidade entre documentos ou grupos.	Resultados orientam seleção de casos para leitura e comparação.
Exportação	Retirar tabelas e resultados para outras etapas da pesquisa.	Favorece reuso em codificação manual, estatística ou análise de redes.

Fonte: Elaborado pelo autor

O princípio transversal é a preservação do texto original. Transformações necessárias a tarefas distintas são mantidas separadas, porque procedimentos como lematização, remoção de *stopwords* ou normalização podem ser adequados a uma análise e inadequados a outra. Assim, o documento recebido permanece como referência para auditoria, concordância, contexto de entidades e recuperação de trechos. A rastreabilidade não garante correção automática, mas permite identificar onde um resultado foi produzido e quais decisões de processamento participaram de sua construção.

A modularidade também evita transformar o percurso em sequência rígida. Um estudo pode combinar apenas frequência e concordância; outro, entidades e redes; um terceiro, modelagem de tópicos e similaridade. A função da interface é reduzir complexidade operacional, não substituir escolhas metodológicas. Em todas as etapas, a pergunta de pesquisa deve anteceder a escolha do recurso computacional.

Nas análises lexicais, a combinação entre medida e contexto é particularmente importante. Uma lista de frequência oferece uma visão inicial do vocabulário, mas termos recorrentes podem assumir sentidos diferentes. A concordância permite recuperar a palavra ou expressão acompanhada de seu entorno textual; a dispersão mostra sua distribuição no corpus; n-gramas deslocam a observação de palavras isoladas para sequências recorrentes. Em conjunto, essas operações funcionam como instrumentos exploratórios e podem orientar recortes posteriores, sem converter frequência em relevância substantiva de forma automática.

O mesmo cuidado vale para entidades e tópicos. O reconhecimento de entidades organiza menções a pessoas, organizações, lugares e outras categorias previstas pelo modelo linguístico, mas classificações equivocadas precisam ser conferidas no texto. Na modelagem de tópicos, grupos de termos não recebem significado por decisão algorítmica: sua interpretação depende dos termos de maior peso e dos documentos que melhor representam cada distribuição. A plataforma procura facilitar esse retorno para que a rotulagem temática continue sendo uma decisão analítica do pesquisador.

Redes e similaridade acrescentam uma dimensão relacional. Nas redes de coocorrência, duas entidades ou termos podem ser conectados quando aparecem na mesma unidade definida para a análise; pesos e medidas estruturais ajudam a localizar associações recorrentes, posições centrais e possíveis comunidades. A visualização, porém, não autoriza inferir causalidade ou vínculo social fora do que a coocorrência representa. De forma equivalente, similaridade indica proximidade segundo uma representação textual adotada; não significa identidade de sentido. Essas saídas são mais produtivas quando tratadas como mecanismos de seleção de casos e formulação de perguntas.

A visualização funciona como camada de exploração. Gráficos de frequência e dispersão, representações de tópicos e redes de ocorrência permitem alterar filtros, selecionar elementos e observar diferentes perspectivas sobre o conjunto. Particularmente nas redes, a seleção de um nó pode ser associada a seus vizinhos, medidas e contextos textuais. O gráfico deixa de ser apenas produto final e passa a operar como interface de navegação do corpus.

A exportação encerra o fluxo sem aprisionar a análise na aplicação. Tabelas de entidades podem seguir para codificação manual; redes podem ser tratadas em software especializado; documentos representativos de tópicos podem alimentar análise qualitativa; resultados tabulares podem ser reutilizados em planilhas ou programas estatísticos. Essa continuidade é coerente com pesquisas multimétodo, frequentes na Comunicação.

As mesmas funções podem compor percursos distintos conforme o objeto. Em estudos jornalísticos, entidades e concordâncias podem auxiliar a localizar fontes, atores e enquadramentos a serem examinados posteriormente; em comunicação organizacional, frequências e similaridade podem apoiar a comparação de documentos institucionais; em corpora de discursos ou postagens, tópicos e redes podem orientar a seleção de conjuntos relevantes para leitura qualitativa. Esses exemplos indicam possibilidades de uso e não equivalem a protocolos prontos: a definição das unidades, dos parâmetros e das inferências continua dependente do desenho de cada pesquisa.

4. CONTRIBUIÇÕES, LIMITES E RELAÇÃO COM A IA GENERATIVA

A contribuição do NID NLP não está na novidade isolada das técnicas incorporadas. Frequência lexical, concordância, reconhecimento de entidades, modelagem de tópicos, medidas de similaridade e análise de redes possuem tradição e implementações consolidadas. A contribuição do artefato está na integração dessas possibilidades em um ambiente orientado ao pesquisador da Comunicação e na articulação entre medida agregada e retorno ao contexto textual.

Essa integração possui consequência metodológica. Uma entidade recorrente pode ser localizada, examinada em rede e conferida nos trechos em que ocorre; um tópico pode ser observado a partir de seus termos e dos documentos que mais contribuem para ele; uma proximidade entre textos pode ser usada para selecionar casos para leitura detalhada. A plataforma procura, portanto, diminuir a distância entre cálculo e contexto, favorecendo um movimento contínuo entre exploração quantitativa e interpretação qualitativa.

Também há uma contribuição relacionada à interdisciplinaridade. A incorporação de métodos computacionais à Comunicação costuma ocorrer em zonas de contato com Ciência da Computação, Ciência de Dados, Estatística e Design. A plataforma não elimina a necessidade desse diálogo, mas desloca parte do esforço técnico para uma infraestrutura reutilizável. Isso permite que o pesquisador concentre maior atenção na construção do corpus, na escolha do procedimento e na interpretação dos resultados, mantendo visíveis as decisões que precisam ser justificadas metodologicamente.

A rastreabilidade constitui o princípio mais importante desse desenho. Frequências permanecem relacionadas às unidades do corpus; entidades, às suas ocorrências; relações em rede, às unidades de coocorrência; tópicos, aos termos e documentos que participam de sua formação. Isso não elimina erros. Um modelo linguístico pode classificar incorretamente uma entidade; um tópico pode ser pouco interpretável;

decisões de pré-processamento podem alterar resultados. A vantagem é tornar essas operações separáveis e discutíveis, em vez de apresentar uma interpretação acabada como única saída do sistema.

Nesse ponto, a relação com a IA generativa é de complementaridade. Modelos generativos são especialmente úteis para síntese, interação exploratória, apoio à escrita, formulação de hipóteses e desenvolvimento de aplicações. Procedimentos estruturados podem ser preferíveis quando a investigação exige mensuração reproduzível, parametrização ou identificação das unidades que sustentam cada resultado. O próprio NID NLP foi acelerado por sistemas generativos de programação, enquanto seu núcleo analítico preserva técnicas tradicionais de NLP. Diferentes recursos, portanto, podem ocupar funções distintas no mesmo processo de pesquisa.

Há limitações importantes. O desempenho das funções depende dos modelos e bibliotecas utilizados, que podem apresentar erros e vieses. A interface não corrige automaticamente problemas de qualidade do corpus, como duplicatas, falhas de extração, mistura de idiomas ou vieses de coleta. Além disso, a avaliação realizada até o momento é funcional: demonstra integração dos módulos, mas não mede de forma sistemática usabilidade, aprendizagem ou adoção por pesquisadores. Esses aspectos definem a próxima etapa de avaliação do artefato.

A acessibilidade da interface também não deve ser confundida com neutralidade metodológica. Ao esconder detalhes de programação, uma aplicação pode facilitar o uso, mas as decisões sobre tokenização, normalização, unidade de ocorrência, número de tópicos ou medida de similaridade continuam produzindo efeitos. Por isso, a simplificação adequada não é a que elimina parâmetros relevantes, e sim a que os apresenta de forma compreensível e os registra para posterior consulta. Essa tensão entre facilidade de uso e transparência constitui um dos pontos que precisarão ser examinados empiricamente nos testes com usuários.

Como produto de pesquisa aplicada, a plataforma deve ser compreendida como versão em desenvolvimento. Parte do conhecimento metodológico encontra-se materializada na própria arquitetura: separação das representações textuais,

organização por etapas, associação entre resultados e documentos e possibilidade de exportação. O artefato não substitui o artigo científico, mas constitui resultado adicional, passível de uso, avaliação, modificação e transferência.

A avaliação futura precisa considerar não apenas se os módulos funcionam, mas se a interface torna compreensíveis as consequências das escolhas realizadas. Uma ferramenta acessível pode criar falsa sensação de simplicidade quando oculta parâmetros ou transforma resultados probabilísticos em respostas definitivas. Por isso, usabilidade e transparência devem ser tratadas de forma conjunta: reduzir cliques e código é desejável, desde que o sistema continue oferecendo informação suficiente para que o usuário reconheça limites, ajuste parâmetros e retorne aos dados.

Essa agenda de avaliação⁶ também permitirá observar em que medida a plataforma efetivamente reduz barreiras para pesquisadores não programadores e quais operações continuam exigindo mediação técnica ou formação adicional. O caráter iterativo do projeto decorre justamente dessa possibilidade de incorporar o uso real como fonte de requisitos para novas versões.

5. CONSIDERAÇÕES FINAIS

O NID NLP foi desenvolvido para responder a uma dificuldade recorrente da pesquisa em Comunicação: trabalhar com *corpora* cuja escala dificulta procedimentos exclusivamente manuais sem exigir que o pesquisador se torne programador para utilizar métodos computacionais. A plataforma integra preparação de dados, análise lexical, entidades, tópicos, redes, similaridade, visualização e exportação em um ambiente modular.

O argumento central do trabalho é que a expansão da IA generativa não torna obsoletos os métodos tradicionais de análise textual. Sistemas generativos ampliam

⁶ O acesso ao programa de testes das ferramentas do NID se dá através da adesão das instituições interessadas ao PDIOrg - Programa de Desenvolvimento Inovador das Organizações desenvolvido no Programa de Pós-graduação em Comunicação - Modalidade Profissional - PPGCOMPro da UFMA.

possibilidades de interação e síntese; métodos estruturados mantêm relevância quando a investigação precisa explicitar as operações que ligam corpus e resultado. A escolha deve depender da pergunta formulada e do tipo de evidência necessário.

Nesse sentido, a contribuição do NID NLP é menos algorítmica do que metodológica e integradora. O sistema procura transformar técnicas consolidadas em instrumentos acessíveis, preservando o retorno aos documentos e a responsabilidade interpretativa do pesquisador. Ao reduzir a complexidade operacional sem ocultar o percurso analítico, cria uma camada intermediária entre a leitura manual de grandes *corpora* e a delegação integral da análise a sistemas automatizados.

As próximas etapas incluem avaliação sistemática com pesquisadores, identificação de dificuldades de uso e novos ciclos de aperfeiçoamento. O desenvolvimento do artefato permanece, portanto, aberto. Seu princípio orientador é estável: ampliar as formas de observar objetos comunicacionais digitais sem separar os resultados computacionais dos textos, contextos e decisões que lhes dão origem.

REFERÊNCIAS

BIRKENMAIER, Lukas; LECHNER, Clemens M.; WAGNER, Claudia. *The search for solid ground in text as data: a systematic review of validation practices and practical recommendations for validation*. **Communication Methods and Measures**, v. 18, n. 3, p. 249-277, 2024. DOI: 10.1080/19312458.2023.2285765. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/19312458.2023.2285765>. Acesso em: 28 ago. 2026.

GRIMMER, Justin; STEWART, Brandon M. *Text as data: the promise and pitfalls of automatic content analysis methods for political texts*. **Political Analysis**, v. 21, n. 3, p. 267-297, 2013. DOI: 10.1093/pan/mps028. Disponível em: <https://web.stanford.edu/~jgrimmer/tad2.pdf>. Acesso em: 28 ago. 2026.

HEVNER, Alan R.; MARCH, Salvatore T.; PARK, Jinsoo; RAM, Sudha. *Design science in information systems research*. **MIS Quarterly**, v. 28, n. 1, p. 75-105, 2004. DOI: 10.2307/25148625. Disponível em: https://wise.vub.ac.be/sites/default/files/thesis_info/design_science.pdf. Acesso em: 24 ago. 2026.

MORETTI, Franco. **Graphs, Maps, Trees: Abstract Models for Literary History**. London: Verso, 2007.

ROGERS, Richard. **Digital Methods**. Cambridge, MA: MIT Press, 2013.

SANTOS, Márcio Carneiro dos. Narrativas automatizadas e a geração de textos jornalísticos: a estrutura de organização do lead traduzida em código. **Brazilian Journalism Research**, v. 12, n. 1, p. 160-185, 2016.

DOI:10.25200/BJR.v12n1.2016.757. Disponível em:

<https://bjr.sbpjor.org.br/bjr/article/view/757>. Acesso em: 11 jul. 2026.

SANTOS, Márcio Carneiro dos. Pesquisa aplicada em Comunicação: o estranhamento da interdisciplinaridade que nos assombra. **Comunicação & Inovação**, São Caetano do Sul, v. 19, n. 41, p. 18-33, 2018. DOI: 10.13037/ci.vol19n41.5469. Disponível em:

https://seer.uscs.edu.br/index.php/revista_comunicacao_inovacao/article/download/5469/2550/17877. Acesso em: 11 jul. 2026.

SANTOS, Márcio Carneiro dos. A datificação de um campo de conhecimento: como algoritmos, números e abordagens quantitativas estão mudando a comunicação.

Organicom, São Paulo, v. 16, n. 31, p. 145-157, 2019. DOI: 10.11606/issn.2238-2593.organicom.2019.161444. Disponível em:

https://revistas.usp.br/organicom/pt_BR/article/view/161444. Acesso em: 11 jul. 2026.

SANTOS, Márcio Carneiro dos. Pesquisa aplicada, métodos digitais e big data journalism: o conceito de distant reader aplicado à pesquisa na Comunicação. In: MARTINEZ, Monica; SILVA, Marcos Paulo da; STORCH, Laura (org.). **Pesquisa em Jornalismo**: dos conflitos em pauta aos conflitos do campo. Brasília: SBPJor, 2019. p. 46-62.

VAN ATTEVELDT, Wouter; PENG, Tai-Quan. *When communication meets computation: opportunities, challenges, and pitfalls in computational communication science*.

Communication Methods and Measures, v. 12, n. 2-3, p. 81-92, 2018. DOI:

10.1080/19312458.2018.1458084. Disponível em:

<https://www.tandfonline.com/doi/full/10.1080/19312458.2018.1458084>. Acesso em: 18 ago. 2026.

VAN ATTEVELDT, Wouter; STRYCHARZ, Joanna; TRILLING, Damian; WELBERS, Kasper. *Toward open computational communication science: a practical road map for reusable data and code*. **International Journal of Communication**, v. 13, p. 3935-3954, 2019. Disponível em: <https://ijoc.org/index.php/ijoc/article/view/10631>. Acesso em: 20 ago. 2026.