Jornalismo automatizado na prática: o uso de geração de linguagem natural para cobertura eleitoral

Automated journalism in practice: the use of natural language generation for electoral coverage

Periodismo automatizado en la práctica: el uso de la generación de lenguaje natural para la cobertura electoral

—

Stefanie Carlan da SILVEIRA
Brasil
Universidade Federal de Santa Catarina
stefanie.silveira@ufsc.br

Matheus Costa NUNES
Brasil
Universidade Federal de Santa Catarina
matheuscnxavier@gmail.com

Chasqui. Revista Latinoamericana de Comunicación
N.º 154, diciembre 2023 - marzo 2024 (Sección Monográfico, pp. 193-210)
ISSN 1390-1079 / e-ISSN 1390-924X
Ecuador: CIESPAL
Recibido: 03-10-2023 / Aprobado: 21-12-2023

Resumen
O jornalismo automatizado é o foco desta pesquisa, que busca revisitar os principais conceitos ao redor da utilização da geração de linguagem natural para a produção de notícias. Argumenta-se que esta forma de produzir notícias une técnicas de apuração jornalísticas à linguagens de programação, processamento de bases de dados, recursos estatísticos e valores tradicionais da profissão. Empiricamente, a pesquisa analisa o projeto de cobertura automatizada das eleições municipais de 2020, conduzido pelo portal G1, no Brasil. Ao final, expõe-se que velocidade e o ganho de escala justificam o advento do jornalismo automatizado na atualidade. Conclui-se que é o volume de dados e seu tratamento que habilita a automação de notícias.
Palavras-chave: textos automatizados; geração de linguagem natural; datificação; mediação algorítmica; jornalismo digital.

Abstract
Automated journalism is the focus of this research, which seeks to revisit the main concepts surrounding the practice that uses natural language generation to produce news. It is argued that this form of journalism combines investigation techniques with programming languages, database processing, statistical resources and traditional values of the profession. Empirically, the research analyzes the automated coverage initiative of 2020's municipal elections, conducted by the G1 portal, in Brazil. In the end, it is stated that speed and gains in scale justify the emergence of automated journalism. Furthermore, it is the volume of data and its processing that enables such automation of news.
Keywords: automated texts; natural language generation; datafication; algorithmic mediation; digital journalism.

Resumen
El periodismo automatizado es el tema de esta investigación, que pretende revisar los principales conceptos en torno al uso de la generación de lenguaje natural para la producción de noticias. Se argumenta que esta forma de producir noticias combina técnicas de investigación periodística con lenguajes de programación, procesamiento de bases de datos, recursos estadísticos y valores tradicionales de la profesión. Empíricamente, la investigación analiza el proyecto de cobertura automatizada de las elecciones municipales de 2020, realizado por el portal G1, en Brasil. Al final, se afirma que la velocidad y los avances en escala justifican hoy el periodismo automatizado. Sin embargo, es el volumen de datos y su tratamiento lo que permite automatizar las noticias.
Palabras clave: textos automatizados; generación de lenguaje natural; datificación; mediación algorítmica; periodismo digital.

Introdução

O jornalismo se insere dentro de um cenário de inovação transversal, onde novas tecnologias criam impacto multissetorial que vão além da própria mídia. Entre os vários desdobramentos provocados pelas mídias digitais, como o jornalismo multimídia, o jornalismo imersivo (Deuze, 2004) e o jornalismo de dados, uma nova ramificação aparece em função do crescimento da geração de linguagem natural: o jornalismo automatizado.

A fim de compreender as características deste tema e discuti-lo, o conceito de jornalismo automatizado é analisado, neste trabalho, como um desdobramento do jornalismo de dados, dentro das ramificações existentes no jornalismo digital. A justificativa para tal escolha ocorre em função de ambas as formas de produção e veiculação de notícias unirem técnicas de apuração a linguagens de programação, processamento de bancos de dados, recursos estatísticos e formatos clássicos de visualização da informação.

Para aproximar a revisão teórica de uma análise empírica, optou-se por analisar o caso da cobertura feita pelo portal brasileiro G1 nas eleições municipais de 2020. O projeto foi identificado como sendo pioneiro no Brasil em aplicar a geração de linguagem natural para uma cobertura de escopo nacional. A iniciativa reportou o resultado das eleições em mais de 5.568 municípios brasileiros em menos de 24 horas, durante o primeiro e o segundo turno.

Para compreender a automação no jornalismo a partir deste caso, esta pesquisa realizou entrevistas em profundidade (Duarte, 2005) com os integrantes da equipe responsável pelo projeto. Além disso, um corpus composto de 2.966 mil notícias publicadas no primeiro turno das eleições foi submetido a uma análise de similitude (Marchand e Ratinaud, 2012) a fim de identificar padrões entre os textos. Ambas as metodologias foram aplicadas com o objetivo de responder a seguinte problemática: de que forma operou o jornalismo automatizado do G1 para que uma equipe multidisciplinar de programadores, jornalistas e revisores colaborasse e interagisse com ferramentas digitais publicando mais de cinco mil notas de jornalismo político em um único dia?

Marco teórico

Para Salaverría (2015), os últimos 30 anos marcaram o jornalismo com a explosão do jornalismo digital. Essa forma de jornalismo tem no *software* sua ferramenta principal. Mecanismos de buscas auxiliam na publicação, sistemas de gestão de conteúdo (CMS) executam quase todo o processo de publicação, plataformas de redes sociais passam a ser não só ambientes de veiculação de notícias, mas objetos de apuração jornalística. A atuação do indivíduo-jornalista no meio digital foi afetada por novas mediações, novas competências e formas de agregar valor à notícia. Como Deuze (2004) aponta, as competências centrais dos profissionais foram alteradas.

Esse processo de ruptura, experimentação e teste dos potenciais das novas mídias no jornalismo foi fartamente documentado por pesquisadores na primeira década do século XXI. Entre as diversas tendências propiciadas pela ida do jornalismo para o ambiente digital, Santos (2016) aponta o jornalismo automatizado (*automated journalism*) como uma aplicação possível, mesmo que ainda em estágio de prova de conceito na primeira década do século XXI. A ampla disseminação do jornalismo automatizado seria viável na medida que jornalistas acumulassem a destreza e familiaridade com instrumentos computacionais, desenvolvendo uma *inteligência híbrida.*

O emprego de recursos como a programação para o processamento de grandes volumes por jornalistas constitui um movimento amplo de introjetar a lógica computacional para dentro de outras profissões (Lima Junior, 2011). Esse movimento obriga jornalistas a cruzarem campos do conhecimento das ciências ditas sociais com as comumente chamadas de "ciências duras". Linguagens de programação como *Python, SQL (Structured Query Language)* e *R* passam a ser usadas dentro de redações para encontrar elementos narrativos no ciberespaço.

Uma das novas competências necessárias ao profissional de jornalismo diz respeito à necessidade de lidar com grandes volumes de informações na *web*. O jornalista de outrora se preocupava em apurar e coletar informações, para depois interpretá-las e reportá-las. Tal profissional que atuava em um contexto de escassez, passa a encarar um cenário de abundância (Meyer, 2002). Essa mudança de paradigma fez com que de todos os tratamentos a serem dados à informação, processá-la seja o mais importante.

O jornalismo automatizado vai, portanto, acumular uma obrigação já presente no jornalismo de dados - significar largos conjuntos de informação *online* - mas acrescido do desafio de aumentar a escala da quantidade de notícias. Ambas estas práticas de reportagem emergem em um cenário de convergência como uma amálgama do jornalismo digital, ciência de dados, cultura *web*, recursos computacionais, e valores norteadores da profissão jornalística (Bradshaw, 2014; Gray, 2014; Mayer, 2002; Coddington, 2015).

Segundo Gray et al (2014. p. 4), o dado encontrado na *web* passa a servir para "delimitar a forma de uma história", enquanto, simultaneamente, o jornalista emprega sua habilidade tradicional de contextualizar e dar utilidade social a esse dado com o seu repertório cultural. A datificação no caso do jornalismo é impactada por uma contumaz celeridade e pressa em publicar os eventos reportados.

O *discurso da velocidade* apontado por Örnebring (2010) é uma propensão histórica da profissão em tornar a produção de notícias mais rápida e em maior escala. O caminho encontrado por algumas empresas midiáticas para atingir este ideal produtivo foi atribuir ao *software* uma atividade que antes era exclusiva ao homem: a escrita. Denominado por alguns autores de jornalismo automatizado (Graefe, 2016), repórter *robot* (Carlson, 2016), ou jornalismo algorítmico (Dörr,

2015), uma recém-surgida maneira de produzir notícias desponta com a característica da 'autonomia' para a criação de conteúdo jornalístico.

O conceito de jornalismo automatizado

Para a automação da redação de notícias funcionar, é preciso mobilizar processamento computacional, recursos de análise de dados, algoritmos de geração de linguagem natural e sistemas de gestão de conteúdo (*Content Management System*) para as publicações. Segundo Coddington (2015), essa combinação de recursos tem se tornado cada vez mais recorrente ao longo da última década.

O jornalismo automatizado é definido por Carlson (2016, p. 417, tradução nossa[1]) como "processos algorítmicos que convertem dados em narrativas textuais jornalísticas com intervenção humana limitada, ou nenhuma, para além da programação inicial". Portanto, segundo o autor, esta forma de jornalismo seria caracterizada por uma participação humana no momento antes da redação e publicação dos textos, se atendo somente à programação do algoritmo que produz as narrativas. Já a definição de Graefe (2016) adiciona um pouco mais de detalhe ao explicar que o processo de automação age sobre a 'compilação, análise, criação e publicação das notícias'.

> Jornalismo automatizado refere ao processo de usar *software* ou algoritmo para automaticamente gerar histórias noticiosas sem a intervenção humana - depois, é claro, da programação inicial do algoritmo. Logo, uma vez que o algoritmo é desenvolvido, permite automatizar cada etapa do processo de produção jornalística, desde a compilação e análise de dados, até de fato a criação e publicação das notícias. (Graefe, 2016, p. 14, tradução nossa).[2]

De forma similar, o conceito de *robô jornalista* também coloca destaque na autonomia de algoritmos para a escrita de textos jornalísticos. Cunhado em pesquisas acadêmicas desde 2010, o *robô jornalista* ou *repórter* é um conceito que já esteve em voga (Clerwall, 2014; Latar, 2015, 2015; Levy, 2012; Rutkin, 2014), mas hoje é criticado como uma conceituação banal que apela para a imagem de autômatos antropomórficos escrevendo textos jornalísticos (Linden, 2017).

Um dos primeiros pesquisadores a voltar sua atenção para a automação no jornalismo, Van Dalen (2012) nomeia o fenômeno de "notícias escritas por máquinas". "Algoritmos podem automaticamente gerar notícias com base em

1 Algorithmic processes that convert data into narrative news texts with limited to no human intervention beyond the initial programming.

2 Automated journalism refers to the process of using software or algorithms to automatically generate news stories without human intervention—after the initial programming of the algorithm, of course. Thus, once the algorithm is developed, it allows for automating each step of the news production process, from the collection and analysis of data, to the actual creation and publication of news.

informação estatística e um conjunto de *stock phrases*, sem a interferência de humanos jornalistas" (p. 648, tradução nossa)[3]. Outra autora que desloca sua atenção do processo para o produto é Carreira (2017) com o conceito de notícias automatizadas. Segundo sua dissertação, a escolha do termo se dá "porque do ponto de vista de produto ou gênero jornalístico, somente a notícia pode ser automatizada até o momento" (p. 105). A definição de Carreira dá foco à 'notícia' por concluir que somente esse tipo de conteúdo jornalístico pode ser automatizado, excluindo gêneros opinativos e reportagens. Isso ocorre, pois "as notícias são possíveis de serem automatizadas porque elas produzem uma informação primária sobre um evento concreto e objetivo e porque seguem o passo a passo do lide jornalístico" (Carreira, 2017, p. 121).

De maneira complementar, Caswell e Dorr (2018) explicam que de fato o jornalismo automatizado na prática se atém a textos descritivos e relativamente simples, ou seja, à notícia. No entanto, essa característica não se dá por conta de deficiências nos *softwares* de Geração de Linguagem Natural, que têm uma capacidade comprovada de produzir narrativas mais complexas, mas sim por uma escassez de conjuntos de dados estruturados. Para se realizar uma cobertura guiada-por-eventos[4], com um maior grau de complexidade, seriam necessários bancos de dados igualmente mais complexos, com tipos de informações que hoje não são usuais.

Nicholas Dörr (2015) inicia suas pesquisas sobre a área cunhando o conceito de *jornalismo algorítmico* e o atrelando especificamente à produção de notícias com o emprego de *softwares* de Geração de Linguagem Natural. Ou seja, o emprego da tecnologia de Geração de Linguagem Natural é centralizada como o método do jornalismo automatizado. A definição de Dorr detalha o processo em três etapas (*input*, *throughput* e *output*) o processo de automação do jornalismo, com o uso de terminologias próprias da Teoria da Computação.

> Jornalismo Algorítmico é definido como o processo (semi)-automatizado de GLN pela seleção eletrônica de dados a partir de bases de dados privadas ou públicas (input), a atribuição de relevância pré-selecionada ou não-selecionada a característica dos dados, o processamento de base de dados relevantes para estruturas semânticas (throughput), e a publicação do texto final em plataformas online e offline com um certo alcance (output). É produzido dentro ou fora de um ambiente editorial com diretrizes e valores do jornalismo profissional que atendem a padrões de topicalidade, periodicidade, publicidade, e universalidade, estabelecendo, portanto, uma esfera pública. (Nicholas Dörr, 2015, p. 702, tradução nossa,)[5]

3 Algorithms can now automatically generate news stories on the basis of statistical information and a set of stock phrases, without interference from human journalists.

4 O termo original cunhado por Caswell e Dorr (2018) é "event-driven narratives".

5 "Algorithmic Journalism is defined as the (semi)-automated process of NLG by the selection of electronic data from private or public databases (input), the assignment of relevance of pre-selected or non-selected data characteristics, the processing and structuring of the relevant datasets to a semantic structure (throughput), and the publishing of the final text on an online or offline platform with a certain reach

As três etapas mencionadas por Dörr descrevem o processo de automação, colocando enfase na fase intermediária: o *throughput*. Também é notável como a conceituação de Dörr propõem a não exclusão da participação de jornalistas no processo, abrindo margem para identificar dinâmicas de interação entre esses profissionais e algoritmos de Geração de Linguagem Natural. Fatores como a atribuição de relevância aos dados, a possibilidade de o processo ser '(semi)-automatizado', a citação das bases de dados (públicas ou privadas) e do ambiente editorial, assim como os valores e diretrizes do jornalismo profissional, apontam para uma dinâmica de *assemblage* entre diferentes atores. A relevância da constante interferência entre algoritmo, programadores, valores profissionais jornalísticos, padrões estéticos da notícia e assim por diante, configuram uma dinâmica também mencionada pela cibernética: o ciclo de *feedbacks*.

A presente pesquisa adota uma definição de jornalismo automatizado enquanto prática que inclua tanto os processos computacionais, quanto a ação humana. Logo, considera-se que o jornalismo automatizado é o processo (semi)-automatizado de Geração de Linguagem Natural pela seleção pré-programada de informações disponíveis em bases de dados públicas, ou privadas (*input*), a atribuição de relevância a característica dos dados, seguida do processamento em estruturas semânticas (*throughput*) e a publicação dos textos na forma de notícia (*output*). O processo ocorre dentro ou fora do ambiente editorial, com o assunto da cobertura sendo especificado pelos aspectos da base de dados (*input*). Os valores éticos e padrões estéticos do jornalismo são transferidos para o algoritmo de Geração de Linguagem Natural (*throughput*), numa dinâmica de modificação relacional mútua que afeta tanto os atores (algoritmos e humanos), quanto o produto final (texto).

As aplicações da Geração de Linguagem Natural consistem no "processo de renderizar pensamento em linguagem" (p. 121, tradução nossa[6]). A Geração de Linguagem Natural trabalha com um fluxo informacional que vai do conteúdo para a forma, da intenção para o texto. O 'gerador' equivale a uma pessoa com algo a dizer, só que na forma de um programa computacional. Este 'algo a dizer' será transferido de um programador para o algoritmo. O *software* parte de uma intenção comunicativa para depois determinar o que será escrito, selecionando palavras e recursos retóricos, todos pré-adequados a uma gramática. Por meio da formatação das palavras no texto redigido, o programa estabelece a prosódia do discurso. Segundo Reiter (2012), todo esse processo se divide em três etapas: a) Planejamento; b) Microplanejamento; e, c) Realização.

O primeiro estágio consiste em decidir qual informação comunicar (determinação de conteúdo), assim como a organização da informação

(output). It is produced inside or outside an editorial office or environment along professional journalistic guidelines and values that meet the criteria of topicality, periodicity, publicity, and universality, and thus establishes a public sphere. The technology of NLG is furthermore identified as the central technical innovation that enables Algorithmic Journalism."

6 Trecho original: Natural language generation (NLG) is the process by which thought is rendered into language.

(estruturação do documento). O segundo, define como as informações serão expressas no texto gerado. Por último, a realização é a efetivação de um texto concreto, baseado nas informações selecionadas no planejamento e moduladas pelas escolhas linguísticas do microplanejamento (Reiter, 2012).

Metodologia

O G1[7] é um portal de notícias brasileiro criado pelo Grupo Globo, debaixo da liderança da Central Globo de Jornalismo. O site existe desde o início dos anos 2000 e, apesar de *online*. disponibiliza notícias de outras organizações jornalísticas do grupo como a TV Globo, GloboNews, rádio CBN, Jornais O Globo, entre outros. Em 2020, o portal realizou a cobertura dos resultados das eleições municipais de 5.568 municípios brasileiros em menos de 24 horas, com o auxílio de *softwares* voltados para a redação/publicação de notícias. Os textos traziam dados diversos sobre os prefeitos e vereadores vitoriosos, como o nome, o partido, o número de votos recebidos, a idade, o estado civil, o grau de instrução, a profissão e o patrimônio declarado. O anúncio do portal sobre o projeto permite ao leitor saber que a base de dados aberta do Tribunal Superior Eleitoral foi usada de insumo para a redação das notícias.

A fim de atingir o objetivo proposto, esta pesquisa realizou entrevistas em profundidade com os profissionais responsáveis pelo projeto: Felipe Grandin, Tiago Reis, Hector Iankovski e Rafael Muniz. A fim de aproximar os relatos e compará-los, as entrevistas seguiram um modelo semi-aberto. A ideia é que se parta de poucos questionamentos abrangentes derivados do problema de pesquisa, "mas que cada questionamento possa ser afunilado no qual perguntas gerais vão dando origem a específicas" (Duarte, 2005, p. 64).

Em paralelo, 2.966 mil notícias publicadas foram coletadas e submetidas a uma análise de similitude (Marchand e Ratinaud, 2012) para medir o grau de similaridade lexical e semântica dos textos produzidos no projeto do G1. Originalmente, a metodologia proposta por Marchand e Rautinaud (2012) emprega o software IRAMUTEQ, porém, a análise do conteúdo com módulos em Python dispõe do mesmo potencial para observar recorrências lexicais e padrões de textos, principalmente em grandes volumes de dados. Portanto, para executar a análise de similitude, duas etapas se fizeram necessárias: a coleta e a identificação de padrões textuais.

Para realizar a coleta de textos foi empregado o módulo *Beaultiful Soup* de raspagem automatizada com Python. Em sequência, para identificar padrões de similaridade entre as estruturas textuais foi empregado, também em Python, o módulo de expressões regulares (*regex*). Essa função permite medir quantitativamente a recorrência de expressões, ordens de palavras e extensão do texto. Com uma análise quantitativa da recorrência de palavras, de natureza

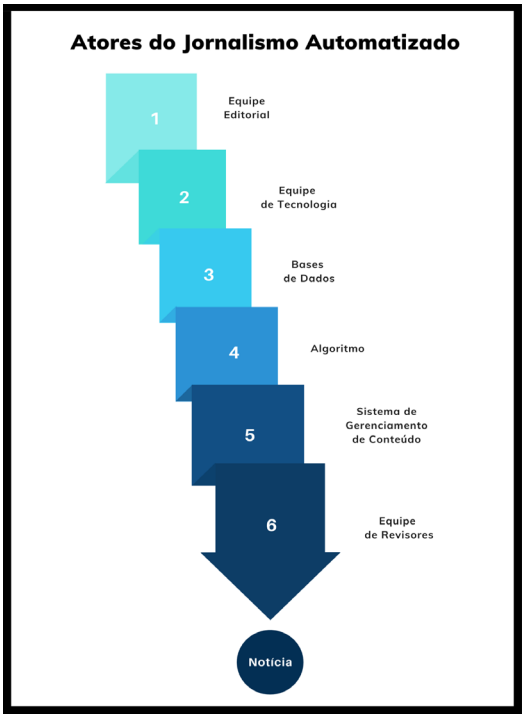
7 https://g1.globo.com/

lexical, é possível observar a capacidade da Geração de Linguagem Natural de produzir textos com diferentes informações - mesmo que as reportagens sigam um número limitado de modelos (*template*) para redação automatizada.

Discussão e resultados

A partir da fundamentação teórica, alinhada às entrevistas em profundidade com os profissionais do G1, foi feita a identificação das principais etapas desta iniciativa de automação de notícias. Os relatos dos executores do projeto permitiram listar seis atores fundamentais para a geração automática das notícias. São eles a (1) equipe editorial, (2) equipe de tecnologia, (3) as bases de dados, (4) o algoritmo, (5) o sistema de gerenciamento de conteúdo[8] e a (6) equipe de revisores. Cada um desses seis atores cumpriu um papel chave na elaboração e publicação dos textos criados pelo G1 para cobertura das eleições de 2020. Para descrever o funcionamento dessa colaboração em rede, parte-se para uma apresentação dos entrevistados na coleta de dados deste estágio da pesquisa.

Figura 1. Atores no Jornalismo Automatizado presentes no caso do G1



Nota. Esta imagem foi produzida para esta pesquisa

8 Chamados em inglês de *Content Management System (CMS)*, são softwares que permitem o armazenamento, gerenciamento e publicação de páginas web. Existem vários tipos no mercado, sendo comuns que empresas de mídia desenvolvam seus próprios sistemas.

O coordenador do projeto foi o editor da equipe de jornalismo de dados do G1, Tiago Reis. O jornalista liderava em 2020 o 'núcleo de dados, *fact-checking* e projetos especiais' do portal, composto por uma equipe de quatro pessoas. Dentro dessa equipe, somente duas pessoas se envolveram diretamente no projeto eleitoral, ele e o então repórter Felipe Grandin (1). Ambos trabalharam lado a lado em um primeiro momento, participando de reuniões quinzenais com a equipe de ciência de dados e desenvolvendo uma "definição do *template*".

Tiago Reis e Felipe Grandin discutiram o que entraria no texto, ou seja, os dados. Essas informações, provindas das (3) bases de dados eram familiares a ambos, que estavam habituados a se debruçar sobre fontes oficiais para produzir reportagens. Duas bases de dados foram escolhidas por sua robustez técnica, confiabilidade, noticiabilidade e formatação homogênea[9]. As duas são administradas pela mesma instituição: o Tribunal Superior Eleitoral (TSE). A primeira é chamada oficialmente de "Divulgação de Candidaturas e Contas Eleitorais", referida de forma despojada pelos entrevistados como "Divulgacand". Ela é descrita pelo TSE como uma plataforma que "apresenta informações detalhadas sobre todos os candidatos que pediram registro à Justiça Eleitoral e sobre as suas contas eleitorais e as dos partidos políticos". A segunda base é o "Resultados", uma coleção de arquivos que podem ser acessados por recorte de ano, região, pleito e tipos de cargos.

As informações identificadas por Reis e Grandin como dotadas de "critérios de noticiabilidade", entraram no *template*. As duas bases precisaram ser "ingestadas", de acordo com o jargão da computação, para as informações alimentarem o texto. Os responsáveis por fazer essa mediação foram os membros da (2) equipe de ciência de dados. Hector Iankovski, cientista de dados, e Rafael Muniz, engenheiro de dados, ambos integravam a equipe de tecnologia do Grupo Globo. O editor Tiago Reis define a atuação de Iankowski como "a pessoa que ficou responsável por pensar nessa parte de Processamento de Linguagem Natural". Já Muniz é descrito como o responsável pela arquitetura dos dados.

A partir das entrevistas realizadas, percebeu-se que a atuação da (2) equipe de tecnologia foi o elo central entre todos os atores envolvidos no projeto de automação de notícias na cobertura eleitoral do G1 em 2020. Tanto os relatos dos jornalistas, quanto os dos programadores, colocam o trabalho de desenvolvimento do algoritmo como a ligação entre os primeiros e os últimos agentes. A partir da (2) equipe de tecnologia, o (4) algoritmo foi elaborado, assim como monitorado. O seu *output* é o que mobiliza a ação do (5) sistema de gerenciamento de conteúdo, que por sua vez distribui a notícia entre a (6) equipe de revisores.

Primeiramente, a (1) equipe editorial selecionou as (3) bases de dados para depois criar modelos de texto, chamados de *template.* Das diversas formas de

9 Bases de dados podem trazer informações com uma formatação constante, seguindo um padrão. Ou podem trazer inúmeros escritos de formas, diversas, pontuações diversas. Bases má formatadas são chamadas por analistas de "sujas".

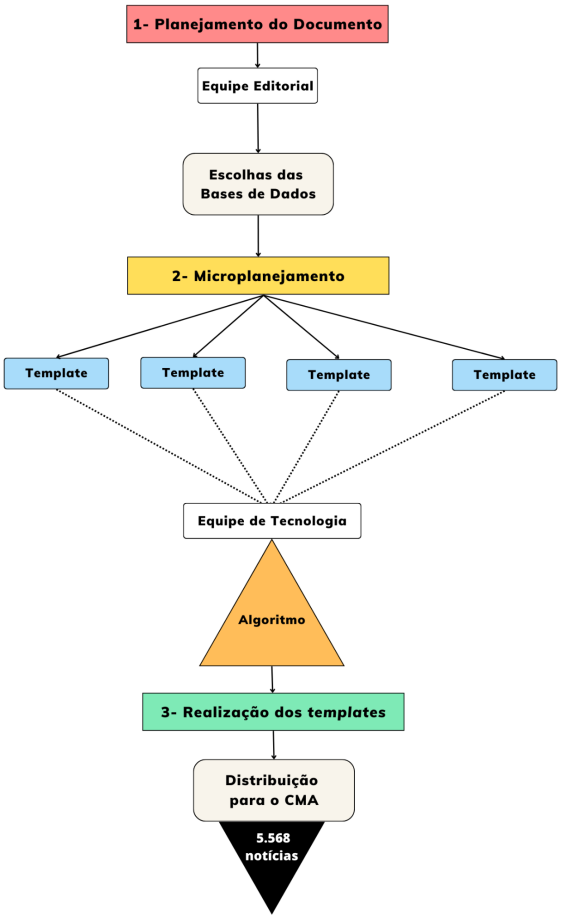
Geração de Linguagem Natural, o uso de expressões-chaves é um dos caminhos mais simples de aplicar esta tecnologia, segundo Dong *et al* (2022). Com este método é possível realizar texto-para-texto, ou dado-para-texto como forma de expansão de texto. Ou seja, o resultado final gerado será um texto mais extenso do que os dados de *input.* Dentro deste *template* são feitas inclusões (dados ou outras palavras) que, por sua vez, podem passar por conjugações, substituições de expressões por sinônimos e formatações. O *template* expressa, portanto, as escolhas linguísticas que o algoritmo GLN terá de executar para produzir o resultado desejado.

Num segundo momento, Iankowski e Muniz tiveram de criar uma base de dados interna, dentro da ferramenta *BigQuery* do *Google Cloud Service*, contendo as informações de Resultados e Divulgacand, conforme apontado pela equipe editorial. A base interna precisava fazer requisições automáticas aos repositórios do TSE para se manter sempre atualizada. Isso foi feito mediante uma outra ferramenta chamada *Functions* programada em *Python.* Depois, a equipe de tecnologia partiu para programação do algoritmo de GNL.

A tarefa consistiu na redação de um script em *Python* que avaliasse os dados enviados em tempo real pela apuração do TSE, consultasse os dados salvos sobre o pleito e os candidatos, a fim de escolher entre um *template*, ou outro. Simultaneamente, o algoritmo precisava fazer microdecisões quanto à conjugação, flexão de gênero e número.

O código em *Python* é a forma em que o algoritmo está codificado. Já as *queries* são consultas feitas ao banco de dados de forma automatizada. A forma como o *Big Querry* e o algoritmo em *Python* se relacionavam era por meio da lógica do "disparo de eventos". A ideia é que um acontecimento dentro da linha de produção da notícia seja o gatilho para um próximo acontecimento, compondo assim a automação. O que conecta as partes do *conjunto técnico* (Simondon, 1988) é o "disparo de eventos". Ou seja, uma "rotina" de códigos que provocam a ativação da base de dados, do algoritmo e da realização dos templates, seguida pela distribuição dos arquivos dentro de um CMS.

Figura 2. As etapas de Geração de Linguagem Natural no projeto do G1



Nota. A figura foi elaborada para esta pesquisa, baseado-se em Diakapoulos (2019), Reiter (2012) e McDonald (2010)

Passado os preparativos do algoritmo de GLN, a equipe de tecnologia sabia que não haveria tempo hábil de fazer mudanças no dia da eleição. Para que tudo estivesse pronto e verificado, era preciso conduzir testes previamente várias vezes. "O caminho não pode ter um único gargalo para a publicação. A gente estava bastante preocupado com essa questão do desempenho de performance do nosso *pipeline* de execução. [...] A gente fez isso para que não tivesse um gargalo, ou um acúmulo de eventos e isso acarretasse um *delay*" (Iankoviski, 2023). Ou seja, uma demora em uma requisição (*querry*) em um banco de dados, acarreta no atraso da geração de um texto, o que represa a geração de outros textos em um efeito cascata.

Finalizadas as testagens, restou à equipe fazer a integração com o sistema de gerenciamento de conteúdo (5). O arquivo em formato "json" era o *output* final da Geração de Linguagem Natural. Um arquivo foi gerado para cada cidade do país, contendo título, subtítulo e corpo da notícia. Junto dos elementos textuais jornalísticos, o arquivo também continha uma série de *tags* que ajudavam na sua distribuição dentro do sistema de gerenciamento de conteúdo. Portanto, os 5.558 documentos em formato "json" foram o *output* de todo o *pipeline*, assim como o resultado dos esforços de Hector Iankovski e Rafael Muniz. Restou à dupla o trabalho de monitorar a geração dos arquivos no dia do pleito, junto da equipe editorial, conforme as urnas foram apuradas e os dados transmitidos.

O esforço conjunto da equipe composta com jornalistas e profissionais de TI conquistou a façanha de publicar uma notícia para cada cidade do Brasil em um único dia. Não sem deixar marcas nos textos publicados. Observar as notícias publicadas pelo G1 permite materializar as dinâmicas explicadas pela equipe, fazendo o caminho contrário do que uma automação jornalística teve que fazer, tal qual uma engenharia reversa da notícia.

Observou-se no caso do G1 nas Eleições Municipais de 2020 uma forma de automatizar notícias, a partir de um processo algorítmico por Regras & Heurística (Vajjala *et al*, 2020), que aplicou o método de expressões-chave e *templates* de Geração de Linguagem Natural (Dong *et al*, 2022). Mais especificamente, essa geração seguiu a via do texto-para-texto e do dado-para-texto, como forma de expansão de conteúdo.

A análise de similitude das 2.966 notícias confirma muitas das afirmações dos entrevistados. Todos os dados de porcentagem de votos recebidos pelos candidatos são provindos da base de dados do TSE. O cruzamento dos valores da base de 'Resultados' com os votos contidos no lide dos textos retornou uma similaridade de 100%[10]. Ou seja, para as notícias que cobriam as prefeituras que tiveram o pleito definido no primeiro turno, a correspondência entre os valores foi total. No entanto, nem todas as notícias do *corpus* declaram vitória em primeiro turno, havendo textos que tratam de outros casos. Desta forma, a análise do conteúdo demonstra a realização de cinco *template*s definidos por Tiago Reis e Felipe Grandin.

Há majoritariamente quatro *templates* que conformam quase toda a cobertura, acrescido de um para um caso excepcional. O primeiro *template*, também o mais frequente, é para vitória em primeiro turno. O segundo, de disputa eleitoral prorrogada para o segundo turno. Terceiro, para candidatura sub judice, aguardando decisão judicial. Quarto, para candidatura impugnada. A exceção é um *template* para o caso de empate técnico.

10 Essa similaridade não se dá necessariamente nas casas decimais. Pois os valores publicados na base do TSE passaram por uma contabilidade total dos votos. As notícias publicadas no G1 foram feitas a partir do momento em que o candidato estava matematicamente eleito, ou seja, essa diferença pode se refletir nas casas decimais.

Quadro 1. Relação de templates no projeto de jornalismo automatizado do G1

Tipo	Quantidade	Porcentagem
Vitória em 1º Turno	2.860	96,4%
Disputa em 2º Turno	32	1,1%
Sub Judice	39	1,3%
Candidatura Impugnada	34	1,1%
Empate	1	0,03%

Nota. Tabela feita para esta pesquisa

Como é possível ver na tabela acima, 96,4% das notícias seguem à risca o padrão do *template* para **Vitória em 1º turno**. Há no mínimo dois motivos para isso. Primeiro, a imensa maioria das cidades brasileiras concorrem nas eleições pelo sistema proporcional. Segundo, somente uma pequena parte das cidades brasileiras estão habilitadas pelo próprio TSE a concorrer no sistema majoritário, onde o candidato deve obter 50% dos votos mais 1 para levar o pleito. Essa definição é feita para seguir os artigos 28 e 29, inciso II, e 77, da Constituição de 1988. Tal marco legal também explica a incidência do segundo *template* **Disputa em 2º Turno**, onde somente 32 notícias derivam da sua realização. No ano de 2020 haviam 95 cidades habilitadas a terem 2º turno, entre as quais 57 seguiram de fato para o próximo pleito. Uma fração um pouco maior que a metade foi capturada pelo *corpus*.

Já na terceira e quarta linha da tabela nota-se a realização do *template* **Sub Judice** e **Candidatura Impugnada** que correspondem a 1,3% e 1,1% dos dos casos, respectivamente. Nessas notícias, encontra-se já no título um padrão que faz a consideração de aguardo de decisão judicial, apesar da contagem de votos ser favorável a um candidato. "Edezio Bastos, do DEM, tem maioria dos votos em Brejolândia, mas candidatura está sub judice" (G1, 2020). No caso das **Candidaturas Impugnadas,** há uma consideração no último parágrafo de que o candidato teve seu recurso negado pelo TSE. O padrão se repete para todas as cidades em que o pleito não pode ser concluído por questões legais, contendo o nome do mais votado, seguido da ressalva específica para aquele caso.

O quinto e último *template* realizado foi o de **Empate.** A chance de acontecer um empate exato em uma eleição por sistema majoritário é mínima, mas há precedentes em quase todos os pleitos municipais. Por essa razão a equipe editorial e a equipe de tecnologia precisaram prever a realização deste template para o caso da base de **Resultados** apresentasse valores idênticos para dois candidatos, no mesmo município. Isso aconteceu em Kaloré (PA) em 2020, onde os candidatos Edmilson (PL) e Ritinha (PSD) tiveram ambos 1.186 votos.

Uma das questões recorrentes apresentadas pelos entrevistados sobre a árvore de decisão, foi a conjugação de gênero. O algoritmo deveria ser capaz de identificar qual candidato(a) identificava-se como homem ou mulher, a fim de

conjugar as palavras da notícia com a devida ortografia. A análise de similitude das notícias contidas no *corpus* verificou que isso de fato ocorreu.

Considerações Finais

O termo 'velocidade' foi um dos mais cunhados, tanto dentro da teoria quanto no relato dos entrevistados. O argumento é de que a busca por ganho de escala e agilidade na produção de informação é o que justifica a automação de notícias (Örnebrig, 2010; Latzer, 2016). Para Felipe Grandin e Tiago Reis, o emprego de algoritmos de redação era a única forma de viabilizar a cobertura eleitoral de 5.568 municípios brasileiros.

Para Lazter *et al.* (2016), os algoritmos na indústria midiática são formas de agregar valor tanto em âmbito individual, quanto corporativo e social. Entre os benefícios da automação estão a redução de custos transacionais, customização de serviços e aumento de performance. Entretanto, o principal benefício econômico da automação é o ganho de escala. O que se buscava nos primeiros estágios da Revolução Industrial, se mantém verdade no século XXI. A busca por velocidade e quantidade na produção de notícias é, na opinião do autor, o fator determinante para automatizar a produção de conteúdo jornalístico.

A velocidade com que a automatização avança e ganha espaço dentro das empresas noticiosas torna difícil inscrever todos os contornos do jornalismo automatizado, porém, algumas conclusões podem ser tiradas a partir da literatura revisada e do estudo de caso do G1. Inicialmente, sabe-se que é possível transferir parâmetros estéticos de um jornalista humano para uma máquina. O lide é a principal técnica de escrita dos jornalistas. Dominar essa forma de escrita é um exercício perene da profissão, atravessada pelo *ethos* jornalístico de objetividade e impessoalidade. O algoritmo é, por definição, um passo-a-passo codificado na forma de *software* (Latzer, 2016). O que o torna competente em replicar esse passo-a-passo de forma que o resultado final é indistinguível daquele escrito por um humano (Carreira, 2017).

Já na realização dos *templates*, há casos em que as notícias foram complementadas por atores humanos. A redação automatizada prevê, portanto, que produtos jornalísticos sejam gerados por sistemas de Geração de Linguagem Natural, e depois elaborados por jornalistas de carne e osso. Essa cooperação entre atores humanos e não-humanos (Latour, 2012) é uma das premissas fundamentais da teoria da *assemblage* (Delanda, 2016), que se expressa nas notícias a partir dessa redação mista, feita camada a camada por diferentes agentes. Fica evidente na iniciativa do G1, como na prática o processo é (semi) automatizado, por ser orquestrado entre pessoas e algoritmos.

No exercício da definição de *templates*, tarefa realizada pela equipe editorial, percebe-se ainda um segundo tipo de transferência. O algoritmo reproduz as experiências pregressas de Tiago Reis e Felipe Grandin, na medida em que replica seus conhecimentos, técnicas de redação e noções de noticiabilidade.

A tecnologia acoplada ao repórter acaba por absorver rotinas do ofício e executá-las (Gumbrecht, 2010). Esse processo é uma via de mão dupla. Quando a equipe editorial se refere à automação como um potencializador do trabalho jornalístico, é unicamente porque parte da inteligência dos repórteres é transferida para a máquina, mesmo que de maneira episódica e específica.

Outro ponto relevante da definição dos *templates* surge na necessidade da cobertura ser previsível, para então ser automatizável. O ato de preparar uma quantidade limitada de modelos de textos para certos cenários exige que os dados sejam conformados dentro desse número limitado de modelos. Segundo Groover (1980), o aspecto antecipatório, ou preditivo, é uma marca da automação industrial. Um fabricante deve ter uma quantidade específica de moldes, processos e insumos para dar vazão à sua produção.

Por último, podemos atestar a importância das bases de dados para a automação de notícias. Conforme ficou explícito nas marcas de apuração deixadas nas notícias, não haveria notícias com informações confiáveis se não houvesse bancos de dados que as organizassem e as transmitissem. Essa constatação vai ao encontro das teorias de autores como Diakopoulos (2019), Nicholas Dörr (2015) e Van Dalen (2012) que colocam o jornalismo automatizado como um desdobramento do jornalismo de dados. O paradigma partilhado por ambos os campos é o mesmo: há um volume avassalador de dados que precisa ser significado na forma de notícia.

Se a velocidade e o ganho de escala justificam o jornalismo automatizado, é o volume de dados que o habilita. Essa é a principal conclusão desta pesquisa. O algoritmo de redação de texto existe, antes de mais nada, para lidar com uma quantidade massiva de informações sobre um único evento. No caso das eleições municipais, 39 mil candidatos e 5.568 municípios compõem a complexidade de um único acontecimento político.

Referências

Carlson, M. (2016). Automated journalism: A posthuman future for digital news? In: The Routledge companion to digital journalism studies. Routledge.

Coddington, M. (2015). Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. Digital Journalism, 3(3), 331-348.

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. Journalism Practice, 8(5), 519-531.

Deuze, M. (2004). What is multimedia journalism? Journalism Studies, 5(2), 139-152.

Diakopoulos, N. (2019). Automating the news. EUA: Harvard University Press.

Duarte, J. (2005). Entrevista em profundidade. Métodos e técnicas de pesquisa em comunicação. São Paulo: Atlas, 1, 62-83.

Dong, C., et al. (2022). A survey of natural language generation. ACM Computing Surveys, 55(8), 1-38.

Dörr, K. N. (2015). Mapping the field of algorithmic journalism. Digital journalism. 4:6, 700-722, DOI: 10.1080/21670811.2015.1096748

Graefe, A. (2016). Guide to automated journalism. Disponível em: https://www.cjr.org/tow_center_reports/guide_to_automated_journalism.php. Acesso em: 16 de agosto de 2022.

Gray, J., Chambers, L., & Bounegru, L. (2012). The data journalism handbook: how journalists can use data to improve the news. EUA: O'Reilly Media, Inc.

Groover, M. P. (2002). Automation, Production Systems, and Computer-integrated Manufacturing 2nd ed. Assembly Automation. EUA: Pearson.

Latar, N. L. (2015). The robot journalist in the age of social physics: The end of human journalism? In: G. Einav (ed.), The New World of Transitioned Media. Switzerland: Springer, Cham, 65-80.

Latour, B. (2005). Reagregando o Social: uma introdução à teoria Ator-Rede. Salvador – Bauru: EDUFBA - EDUSC.

Latzer, M., et al. (2016). The economics of algorithmic selection on the Internet. In: J. M. Bauer & M. Latzer (eds.). Handbook on the Economics of the Internet. UK: Edward Elgar Publishing, 395-425.

Levy, S. (2012). The rise of the robot reporter. Wired, 20(5), 132-139.

Linden, C. G. (2017). Decades of Automation in the Newsroom. Digital Journalism, 5(2), 123–140.

Marchand, P., & Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT, 2012, 687-699.

McDonald, D. D. (2010). Natural Language Generation. In: N. Indurkhya & F. J. Damerau (Eds.), Handbook of Natural Language Processing (Vol. 2, pp. 121-144). EUA: Routledge.

Meyer, P. (2002). Precision Journalism. 4th ed. Lanham, MD: Rowman and Littlefield.

Örnebring, H. (2010). Technology and journalism-as-labour: Historical perspectives. Journalism, 11(1), 57-74.

Reiter, E., Sripada, S. G., & Robertson, R. (2003). Acquiring correct knowledge for natural language generation. Journal of Artificial Intelligence Research, 18, 491-516.

Rutkin, A. (2014). Rise of robot reporters: When software writes the news. New Cientist, 221(2962), p. 22.

Santos, M. C. (2016). Narrativas Automatizadas e a Geração de Textos Jornalísticos: A estrutura de organização do lide traduzida em código. Brazilian Journalism Research, 12(1), 160-185.

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). Practical natural language processing: a comprehensive guide to building real-world NLP systems. EUA: O'Reilly Media.

Van Dalen, A. (2012). The algorithms behind the headlines. How machine written news redefines the core skills of human journalists. Journalism Practice, 6(56), 648–658.