SBPJor – Associação Brasileira de Pesquisadores em Jornalismo
18° Encontro Nacional de Pesquisadores em Jornalismo
3 a 6 de Novembro de 2020

REPÓRTER-ROBÔ: entre conceitos e práticas do jornalismo

Manoella Fortes Fiebig[1]
Claudia Irene de Quadros[2]
Universidade Federal do Paraná

Resumo: O estudo reflete sobre abordagens e pesquisas referentes ao conceito de jornalismo de automação, tendo como objetivo principal compreender como as aplicações de processamento de linguagem natural funcionam. Apresentamos um guia com três aplicações diferentes para os programas de automação de notícias (*news automation software*): 1) via funções, 2) processamento de linguagem natural e 3) processamento de linguagem natural por módulos (em *pipeline*). Dessa maneira, esperamos demonstrar a evolução da arquitetura de softwares e suas respectivas possibilidades. O estudo ainda traz exemplos de desenvolvedores de programas de produção automática de notícias, discutindo as possibilidades de um repórter-robô capaz de realizar, além da sistematização dos dados do *lead*, a contextualização das narrativas jornalísticas.

Palavras-chave: jornalismo de automação, natural language generation, software, repórter-robô.

1. A produção jornalística automatizada e seus impactos

As máquinas têm substituído os jornalistas, como mostra o artigo "*The Rise of the Robot Reporter*"[3], de Jaclyn Peiser (2019), no *The New York Times.* A jornalista aponta que algumas empresas têm adotado procedimentos parciais ou totais de automação de conteúdo, como a *Bloomberg News* e a *Associated Press* (A.P.). A A.P, por exemplo, automatiza editorias inteiras com notícias sobre a bolsa de valores, os relatórios

[1] Manoella Fortes Fiebig. Doutoranda em Comunicação pela Universidade Federal do Paraná. Mestre em Comunicação pela Universidade Federal do Paraná. Integra o Grupo de Pesquisa COM XXI. Bolsista Capes. Orcid: https://orcid.org/0000-0003-0073-8622. E-mail: manoellaff@gmail.com
[2] Claudia Irene de Quadros: Pós-doutora em Comunicação pela Universidade Pompeu Fabra, docente do Programa de Pós-Graduação em Comunicação da Universidade Federal do Paraná (UFPR). Integra o Grupo de Pesquisa COM XXI e a Rede de Pesquisa JORTEC . Orcid: http://orcid. org/0000-0003-1322-8971. E-mail: clauquadros@gmail.com.
[3] "The Rise of the Robot Reporter", by Jaclyn Peiser. Disponível em: <http://bit.ly/NYTimesJP>

financeiros, os resultados esportivos e os relatórios de previsão do tempo. Para isso, fez uma parceria com a *Automated Insights* que desenvolve programas de processamento de linguagem natural.

Em todo o mundo, empresas como a *Automated Insights* desenvolvem sistemas para criar textos de maneira totalmente automatizada, com base nos dados[4] coletados na web, transformando-os em arquivos contextualizados. Eles também adotam o formato estilístico de texto jornalístico para seguir a estrutura do lead e o manual da empresa. Nos estudos de jornalismo, há vários termos adotados para a automatização de notícias (STORCH; FEIL, 2019), como "repórter-robô", "jornalismo automatizado" e "jornalismo algorítmico". Mais adiante discutiremos cada um deles.

Neste artigo, optamos por usar o termo repórter-robô para demonstrar possíveis impactos nas práticas jornalísticas. Jaclyn Peiser (2019) mostra que empresas de desenvolvimento de software, como a Cloudera, esperam que ferramentas de inteligência artificial se tornem cada vez mais capazes de aprofundar a contextualização das notícias. Essas empresas garantem que os postos de trabalho dos jornalistas serão mantidos, afirmando que a automatização possibilita que esses profissionais invistam mais tempo no esforço criativo e investigativo. Milosavljevic & Vobic (2019) argumentam que a automação está cada vez mais comum no jornalismo. Para eles há implicações complexas, mas não ameaças. A literatura científica do jornalismo digital, no entanto, mostra que na evolução tecnológica há reduções de empregos nas redações jornalísticas (LARANGEIRA; QUADROS, 2007).

Nesse sentido, é importante resgatar pesquisas sobre essas transformações no jornalismo para acompanhar a evolução de conceitos e de contextos de determinado período. O jornalismo guiado por dados (TRÄSEL, 2014) tem sido adotado por jornalistas na produção de suas reportagens há bastante tempo. Para Milosavljevic & Vobic (2019), a tecnologia de automatização de notícias é percebida como uma ferramenta sem "olfato jornalístico". Eles entendem que este faro, intrínseco ao repórter humano, o torna tão relevante mesmo coexistindo com a operação de sistemas autônomos nas redações.

[4] Neste trabalho, dados são representações em código das informações presentes na web. São uma espécie de matéria-prima das informações que após tratamento adequado podem ser encontradas em forma de texto, códigos, números, tabelas, áudios, vídeos etc.

A produção autônoma de conteúdo jornalístico por meio de softwares guiados pelo processamento da linguagem natural (Sirén-Heikel et.al, 2020) e pela web semântica (Lammel & Mielniczuk, 2012) – conceitos apresentados mais adiante – é uma realidade em diversos países e instituições jornalísticas, que encontram em empresas de desenvolvimento de software parceiras de negócio rumo à automação do jornalismo. Nos Estados Unidos, a *Associated Press*, a *YahooNews!* e a *ProPublica* são exemplos de empresas que utilizam recursos algorítmicos para a produção de textos jornalísticos. O mesmo acontece do outro lado do atlântico, com relatos de empresas na Finlândia, Suécia e Inglaterra que também usufruem da tecnologia para criar textos jornalísticos automatizados (MILOSAVLJEVIC & VOBIC, 2019; SIRÉN-HEIKEL, 2020). Logo, a automação do jornalismo figura como uma tendência que cresce simultaneamente em diversos pontos do globo, impactando o cotidiano de salas de redação, nos processos jornalísticos e na administração de empresas com características locais, nacionais e internacionais.

Aqui o nosso objetivo é reunir autores e conceituações sobre jornalismo de automação para continuar uma incursão teórico-epistemológica já iniciada pelo Rede JorTec, da SBPJor. Defendemos que as discussões de conceitos são de extrema importância para o desenvolvimento da ciência. Na revista *Digital Journalism*, os editores Scott A. Eldridge II et al.(2019) mostram a necessidade de verificar o que já foi produzido sobre o assunto, buscando um equilíbrio entre o passado e o presente. Os referidos editores discutem especificamente sobre a consolidação do conceito de jornalismo digital, embora alguns autores, como pontuam, adotem outras terminologias.

Ao discutirmos conceitos já adotados para o jornalismo automatizado, pretendemos analisar diferentes perspectivas e seus possíveis impactos. Por isso, apresentamos, em primeiro lugar, uma sistematização do tema com o aporte de pesquisadores nacionais e internacionais para compreender as evoluções tanto das práticas jornalísticas rumo à automação, mas também do pensamento científico que envolve este processo. Além disso, mostramos uma espécie de mapeamento do campo do jornalismo de automação para compreender o impacto desta nova modalidade que adentra as redações. Configura-se, portanto, como um estudo preambular e teórico sobre o estado da arte do jornalismo de automação e suas diversas conceituações e concepções ao longo

dos últimos anos. No contexto brasileiro, elaboramos um quadro teórico para demonstrar a evolução do pensamento científico acerca do fenômeno para, em seguida, introduzir também alguns conceitos provenientes de pesquisadores internacionais sobre o assunto.

Em um segundo momento, discutimos algumas possibilidades de estrutura lógica para o funcionamento destes softwares de geração de discurso. Desta forma, buscamos tensionar conceitos de jornalismo de automação com estudos em ciência da computação para apresentar conceitos de processamento de linguagem natural em pipeline, web semântica e ontologias. A intenção é orientar o leitor num percurso teórico rumo à compreensão do funcionamento desses sistemas.

Para tanto, são apresentadas três formas de configuração de softwares de automação que criam textos jornalísticos. O primeiro está relacionado ao conceito de função, apresentado por Arce (2009), no qual o lead é reconstruído com base em funções e os dados são organizados dentro desta estrutura pré-definida. Em seguida, o conceito de processamento natural de linguagem, do inglês *natural language generation* (NLG), é introduzido no texto para demonstrar duas possibilidades distintas de arquitetura de software, que podem aprimorar a entrega de textos relacionais e contextualizados. Além disso, vale ressaltar as importantes contribuições da web semântica e das ontologias (Lammel e Mielniczuk, 2012) no sentido de possibilitar a criação de narrativas com mais recursos contextuais. Por fim, o estudo apresenta exemplos de empresas que desenvolvem softwares que realizam a produção automática de notícias. Com o objetivo de discutir a viabilidade e as limitações de um repórter-robô o texto também procura incentivar a reflexão sobre os papéis desempenhados por humanos e máquinas na automação de notícias.

2. A evolução conceitual do Jornalismo de Automação

O jornalismo de automação é um tema abordado há décadas no contexto acadêmico e ficcional internacional. No contexto acadêmico brasileiro, destacamos alguns pesquisadores, como Tacyana Arce (2009), Walter Teixeira Lima Jr (2012; 2015), Mendonça (2016) e Magalhães (2017) que procuram discutir questões teórico-epistemológicas e empíricas sobre o assunto. Em 2009, Arce abordou o conceito de "lead automatizado", uma forma de "aceleração do processo de produção jornalística e simplificação do processo de tratamento da informação" (ARCE, 2009, p. 01).

A autora pontua que o lead, por ser uma estrutura fechada e com diretrizes claras, poderia ser produzido de forma automática por meio de um software. As perguntas básicas do lead (Quem? Onde? Quando? Por que? Para quê?), quando organizadas e sistematizadas num software de geração de textos automático, seriam funções (no sentido matemático do termo) que figurariam como uma estrutura lógica dentro do sistema. Para a autora, criar variáveis para cada função (pergunta do lead), seria a saída para criar combinações estruturantes de uma notícia. Desse modo, cada pergunta do lead (L=) seria uma função diferente (f) dentro do sistema, e as respostas para cada função seriam variáveis, de acordo com cada notícia (f=v1+v2). O trabalho do software seria montar, rapidamente, o primeiro parágrafo de uma notícia (L= (f=v1+v2) + (f1=v1+v2), estruturando funções e variáveis de forma coerente. O tempo de produção de uma notícia seria reduzido, mas este modelo de aplicação para softwares de automação de notícias é bastante criticado por ser limitado em suas atribuições.

Vinicius de Sousa Mendonça (2016), que trata esse fenômeno de "o jornalismo sem repórter", observa que os softwares já alcançaram um nível de aperfeiçoamento que leitores comuns não percebem a diferença entre textos produzidos por repórteres ou por programas de automação. Essa evolução de softwares, segundo Mendonça (2012), está diretamente relacionada ao desenvolvimento da NLG - um nicho de pesquisas em inteligência artificial, proveniente da Ciência da Computação. No Gmail, por exemplo, quando o usuário inicia um texto suas frases podem ser completadas automaticamente pelo sistema de processamento de linguagem. Esse sistema ainda é capaz de identificar falas mais recorrentes com os contatos de cada usuário por meio de *machine learning*, relacionado ao processamento de linguagem natural aplicado ao envio de e-mails. Esse dispositivo fica cada vez mais refinando com o passar do tempo, identificando as frases mais utilizadas por cada usuário e criando conversações coerentes.

3. Notícias automatizadas: dados que se transformam em narrativas

Alguns termos do jornalismo automatizado precisam ser discutidos. Antes, contudo, é necessário destacar a importância das reflexões já desenvolvidas sobre o conceito do jornalismo digital (MIELNICZUK,2004 ;SCOTT A. ELDRIDGE II ET AL, 2019) e do jornalismo de dados (BARBOSA, 2007; LIMA JR, 2012; TRASEL, 2014).

Neste artigo, voltamos o olhar para o conceito de jornalismo de dados - uma vez que a automação tem como princípio básico a utilização de grandes quantidades de dados, tratados e sistematizados em softwares para a criação de discursos. No quadro 1, mostramos como pesquisadores de jornalismo brasileiros conceituavam a base de dados até chegar ao uso de jornalismo de dados. Na primeira década dos anos 2000 , houve uma forte influência da obra de Lev Manovich (2001) para falar das possibilidades das bases de dados. Na década seguinte, encontramos mais pesquisas empíricas da realidade brasileira.

QUADRO 1. CONCEITOS DO USO BASE DE DADOS NO JORNALISMO

PESQUISADOR(A)	CONCEITO	ANO
Elias Machado	Machado procura compreender a base de dados no jornalismo digital como uma cultura da sociedade das redes, elencando três funções: "1) de formato para a estruturação da informação, 2) de suporte para modelos de narrativa multimídia e 3) de memória para conteúdos publicados"(p.301).	2004
Claudia Quadros	Quadros observa a base de dados como produtora de conhecimento do jornalismo para além dos repositórios, chamando a atenção para a necessidade de repensar o seu uso nas práticas jornalísticas também com o desenvolvimento da web semântica.	2004/2005
Carla Schwingel	Schwingel defende a consolidação das bases de dados complexas no jornalismo digital de quarta geração (iniciando no período da sua publicação), explorado por meio de ferramentas automatizadas para apurar, editar e veicular informações na produção de produtos jornalísticos. Na sua tese, investiga os Sistemas de Gerenciamento de Conteúdo (SGCs).	2005/2008
Suzana Barbosa	Barbosa compreende a base de dados como um aspecto-chave do jornalismo digital. Para ela, o Jornalismo Digital em Base de dados (JDBD) permite criar, manter, atualizar, disponibiliza e circular produtos jornalísticos dinâmicos.	2007
Walter Teixeira Lima Jr.	Demonstra a necessidade do jornalista compreender o pensamento computacional para cruzar dados e utilizar as bases de dados de forma dinâmica, resgatando a evolução histórica do *data journalism*. Para ele, o profissional deve "adquirir habilidades técnicas/tecnológicas que proporcionem transformar-se em data jornalista."(p.221).	2012
Daniela Bertocchi	Define o conceito de antenarrativa jornalística (dados e metadados) em que o jornalista seleciona e cadastra dados em sistema de narrativas (software) responsável por organizá-los e publicá-los. Também adota o termo Jornalismo Digital em Base de dados (JDBD) em sua tese de doutorado.	2013
Marcelo Ruschel Träsel	Na tese de doutorado Träsel define o jornalismo guiado por dados (JGD) por práticas profissionais que usam a base de dados "como principal fonte de informação para produção de notícias." (p. 106).	2014
André Rosa de Oliveira	Aborda o conceito de Jornalismo Estruturado, com base em dados. Segundo o autor "a produção, a hierarquização e a classificação de notícias no ambiente Web envolvem não apenas variáveis humanas, mas também (e cada vez mais) computacionais. Algoritmos e sistemas culminam com a automatização de processos e produtos" (p. 20) jornalísticos.	2016
Frederico S. M. de Carvalho	Explora os Sistemas de Gerenciamento de Conteúdo (SGCs) para ambientes jornalísticos. Busca compreender como os sistemas funcionam nas redações automatizadas. No doutorado, em andamento, o pesquisador busca compreender as aplicações de inteligência artificial e machine learning nos SGCs.	2019

Fonte: as autoras a partir de teses e artigos científicos, 2020.

A proposta aqui não é esgotar definições já realizadas pelos pesquisadores brasileiros, mas destacar a evolução de terminologias a partir de novos contextos tecnológicos e do mercado jornalístico. Antes de continuar a reflexão sobre a automação do jornalismo, definimos alguns conceitos importantes à luz de novos contextos.

QUADRO 2. Conceitos

TERMOS	CONCEITOS
Jornalismo Digital	Jornalismo que emprega tecnologia digital para tratar os dados em forma de bits. A terminologia tem se consolidado nos estudos da área por ser mais abrangente do que "jornalismo on-line", "webjornalismo" ou "ciberjornalismo".
Jornalismo de Dados	É o jornalismo guiado por dados processados por máquinas. As bases de dados permitem o desenvolvimento de um modelo dinâmico para criar, gerir, atualizar e circular produtos jornalísticos.
Jornalismo Algorítmico	São as notícias geradas por programas que produzem milhares de matérias simultâneas, com base em grandes conjuntos de dados e sua configuração é realizada via algoritmos (*rule based*).
Jornalismo de Automação	Corresponde a jornalismo produzido por software com a utilização de processamento de linguagem natural (NLG) associado às ontologias e web semântica para criar contextualização entre as narrativas.
Repórter-Robô	O software de NLG passa a desenvolver características de autonomia na coleta, produção e circulação de notícias. Utiliza *machine learning* em *pipeline* para aprimorar seus processos.

Fonte: As autoras a partir da revisão bibliográfica deste estudo, 2020.

A automação de discursos jornalísticos emergiu como uma tecnologia capaz de reconfigurar as relações dentro das redações e transformar o paradigma jornalístico (CHARRON & BONVILLE, 2016) com o uso de dados para a construção de narrativas noticiosas. A tecnologia também mudou o cenário da mídia global e afetou a indústria, a economia, a política, a cultura e a sociedade. (SIRÉN-HEIKEL, 2020)

Na indústria jornalística, por exemplo, observamos a reconfiguração dos processos produtivos, a crise no modelo de negócio e a necessidade de o jornalista aprender novos conhecimentos. A automação de processos dentro do fazer jornalístico e o uso de dados transformou a maneira como o conteúdo é produzido com o uso dos repórteres-robô.

Esse jornalismo de automação é considerado por Lindén (2018) como a "combinação entre algoritmos, ciências sociais, processos matemáticos e sistemas para a produção de notícias" (LINDÉN, 2018, p. 08). Ele pontua que o jornalismo procura acompanhar as evoluções tecnológicas, tal como ocorreu, na década de 90, com a troca das máquinas de escrever pelos computadores nas redações. Ele compara o

funcionamento desse processo, realizado por robôs produtores de notícias, com o próprio fazer jornalístico. Basicamente, o dever do desenvolvedor de software é configurar os seguintes comandos no sistema:

> Procure uma informação que atenda a um conjunto predeterminado de regras de produção de notícias, tais como curiosidade, relevância e impacto, procurar três fontes independentes e reconhecidas para a contextualização e comentários, produza um artigo de 200 palavras redigido conforme o manual de estilo da redação e o submeta a um editor ou diretamente ao público. (LINDÉN, 2018, p. 09)

Essas instruções básicas na rotina de repórteres humanos podem ser empregadas em programas de computador. O jornalismo de automação é baseado em regras que permitem introduzir (imputando[5]) padrões no sistema produtor de notícias. Por isso, Lindén (2019) também entende a necessidade de discutir os algoritmos e seus impactos no jornalismo.

Desde a criação do grupo de pesquisa sobre Inteligência Artificial, em 2008, Lindén e sua equipe desenvolvem diversos estudos sobre jornalismo de automação. Membros do grupo (Sirén-Heikel, et.al. 2020) discutiram recentemente alguns conceitos ao explorar possibilidades de automação no jornalismo. Na opinião dos pesquisadores, os conceitos de "*automated journalism*" e "*news automation*" estão relacionados ao uso do processamento de linguagem natural em sistemas que se baseiam em duas formas de *input*[6] (entrada) de dados e *setup* (configuração) distintas para a produção de notícias:

QUADRO 3. Modelos de *input e setup* dos softwares em NLG:

RULE BASED	MACHINE LEARNING
Os sistemas com configuração de *input e setup* via *rule based* obedecem a determinadas regras/padrões pré-estabelecidos no momento do desenvolvimento da ferramenta para a produção de conteúdo.	Corresponde aos sistemas que têm a capacidade de se retroalimentar automaticamente criando uma curva de aprendizado à medida em que é utilizado. Os dados podem ou não corresponder à regras específicas neste formato.

Fonte: Sirén-Heikel, et.al. (2020)

Em alguns casos, os sistemas podem se utilizar destas duas aplicações em seu desenvolvimento, cada uma fornecendo bases para uma etapa distinta do processamento

[5] Imputando é o neologismo derivado da palavra *input* (entrada, em inglês) utilizado por desenvolvedores de software e significa introduzir algo (um dado, uma regra, uma variável) dentro do sistema.
[6] A expressão "input" significa entrada.

de dados. Embora haja a distinção conceitual entre as duas, ambas podem gerar notícias automaticamente, tendo como base um algoritmo regido pela NLG e pela aplicação de web semântica.

A web semântica procura agrupar dados disponíveis na web (textos, imagens, sons, tabelas, gráficos, relatórios e códigos), atribuindo-lhes significados legíveis para máquinas e humanos. Segundo Lammel e Mielniczuk (2012) "basicamente o que a web semântica realiza é a identificação dos significados presentes na rede"(LAMMEL & MIELNICZUK, 2012, p. 182). Por sintetizar o conceito de ontologia - a busca pela "essência" e o significado real dos dados disponíveis na internet-, a web semântica permite que buscas e associações de informações ocorram de forma autônoma.

O conceito de ontologia concede uma resposta positiva ao questionamento de Sirén-Heikel et.al (2020), quando os autores refletem sobre limitações do jornalismo de automação. Para eles, a principal dificuldade enfrentada pelos sistemas de computador menos sofisticados seria a incapacidade de conceder um significado contextual aos acontecimentos relatados.

Quando um software usa recursos da web semântica associados à NLG no momento de responder algumas questões de um acontecimento, como o "por que", é possível preencher por significados atrelados aos termos (palavras-chave, acontecimentos passados, conceitos, nomes de pessoas etc.), com informações recuperadas das base de dados da web (sites, imagens, registros oficiais governamentais, notícias similares etc). Deste modo, a web semântica seria uma resposta possível ao problema da falta de relação entre o simples "lead automatizado" (ARCE, 2009) e a contextualização do fato relatado (LAMMEL & MIELNICZUCK, 2012).

Ainda que relevante para sua época, o conceito matemático de função e variáveis para a criação do "lead automatizado" não é mais adequado ao quadro de desenvolvimento de softwares atual, pois está limitado ao exercício de uma função (pergunta) e suas variáveis (respostas), num fluxo de vai e vem entre os dados. A formatação de um texto jornalístico pode ser mais completa, relacional e aprofundada com a NLG.

Neste artigo discutimos possibilidades abertas pela NLG para compreender como funcionam estruturas de software por meio de *pipeline* (DEVYATKIN et.al., 2019, p.02).

Este modelo estrutural pode figurar como um fluxograma do tratamento dos dados dentro de um software de automação jornalística descentralizado, percorrendo diversos caminhos na construção de textos. Por estruturar uma maior quantidade de dados simultâneos na produção de discursos, consegue recuperar informações e relacioná-las ao contexto de cada notícia.

4. Avançando na compreensão da NLG aplicada ao jornalismo

Os estudos de processamento de linguagem natural, ou simplesmente NLG, surgem com avanços das pesquisas sobre Inteligência Artificial. A NLG tem a capacidade de transformar um conjunto de dados não linguísticos - disponíveis principalmente na internet - em narrativas coerentes. Muitos destes dados, como aponta Ribeiro (2019), precisam de um conhecimento especializado para sua interpretação. E a NLG permite que textos sejam gerados de forma automática a partir de imagens, vídeos, gifs, áudios, gráficos, cálculos, textos, documentos etc. Desse modo, os dados podem ser transformados em uma linguagem compreensível para os leitores de um portal jornalístico. Assim representações gráficas de mapas meteorológicos, como exemplifica Ribeiro (2019), geram de forma automática a previsão do tempo.

Devyatkin et.al. (2019) acreditam que essa é uma das tarefas mais importantes de NLG: identificar e converter dados em estruturas gráficas para transformá-las em texto. Essas estruturas ontológicas ou semânticas, como pontuam os autores, conseguem transformar gráficos *RDF*[7], que combinam RDF-triplos, com elementos com referência cruzada em um texto coerente. Desse modo, a NGL processa os dados, reproduz seus significados, combinando-os com dados de diferentes formatos para criar relações entre eles para gerar um único texto.

Para Devyatkin et.al. (2019) existem seis passos fundamentais para que um software de NLG funcione de forma adequada: a determinação do conteúdo, estruturação do texto, agregação de frases, a lexicalização, seleção de expressões e, finalmente, a

[7] RDF - Resource Description Framework - é um modelo padrão para intercâmbio de dados na Web. O RDF permite que dados estruturados e semiestruturados sejam misturados, expostos e compartilhados em diferentes aplicações.

realização da sentença. A seguir (TAB. 1), explicitamos quais são as tarefas executadas pelo sistema em cada uma destas etapas:

TABELA 1
Passos Fundamentais de NLG aplicada ao jornalismo

ETAPAS	PROCEDIMENTOS
Determinação do conteúdo	Entrada do dado no sistema. O sistema identifica a natureza do dado e deve selecionar quais dados serão utilizados para executar aquela função (texto).
Estruturação do texto	O sistema identifica quais dados devem aparecer primeiro, realizando as melhores combinações. Também determina a ordem que as informações serão apresentadas no texto.
Agregação de frases	O sistema decide quais informações serão apresentadas em cada frase.
Lexicalização	Fase na qual o sistema opta por determinadas combinações de vocabulário para tornar o texto compreensível.
Seleção de expressões de referência	Consiste em determinar a maneira apropriada de se referir aos conceitos e objetos contemplados no plano do documento para evitar ambiguidade.
Realização da sentença	O sistema combina as frases para criar textos estruturados.

Fonte: Devyatkin et.al., 2019; Ribeiro 2019.

Todos estes passos podem ser realizados em sequência ou separadamente. Existe uma combinação de regras para lidar com essas diferentes etapas (DEVYATKIN ET.AL, 2019). Essas regras podem ser criadas de forma manual ou por modelos de aprendizado de máquina. Dependendo do objetivo comunicativo imputado no sistema, é possível chegar a um texto ou a um áudio. Para isso, esse sistema denominado de *rule based* utiliza uma arquitetura fixa para que os dados percorram sequencialmente todos os processos desde sua entrada no sistema até a saída. Segundo os autores, algumas destas arquiteturas de NLG permitem a criação de estruturas completas de software de forma não supervisionada e, ainda assim, atingem um nível de qualidade relativamente alto nos textos gerados. Na proposição de Devyatkin et.al. (2019), o software de processamento de linguagem natural atua em sequência (FIG. 1):

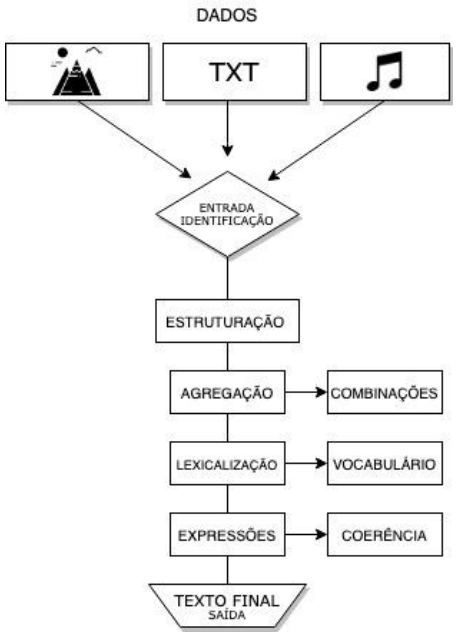


FIGURA 1 – Modelo de estrutura de NLG via *rule based*.
Fonte: Devyatkin et.al., 2019; Ribeiro 2019.

Esta estrutura básica da figura 1, no entanto, a NLG pode ser aperfeiçoada com o uso de uma cadeia de processos (*pipeline*). De acordo com Ribeiro (2019),

> Um modo alternativo de pensar sobre isso é em termos de uma rede ponderada de múltiplas camadas, onde a geração equivale a um melhor primeiro percurso: em qualquer etapa i, o classificador Ci produz a saída mais provável, que leva à próxima etapa Ci + 1 ao longo do caminho mais provável. Essa generalização está conceitualmente relacionada à visão da NLG em termos de políticas na Aprendizagem por Reforço, que define um percurso através de sequências de estados que podem ser organizados hierarquicamente (RIBEIRO, 2019, p. 20).

Neste modelo, a estrutura do sistema não é estática, como o da Figura 1, e o software pode oferecer diversos caminhos possíveis (etapas) que são percorridos de acordo com as necessidades dos dados. Num sistema de NLG em *pipeline* a cadeia de processos é descentralizada e realizada em camadas determinadas não pelo sistema, mas sim pela própria necessidade do dado que será tratado.

O uso de uma cadeia de processos (*pipeline)* possibilita a criação de novas etapas no processamento dos dados. Por exemplo, entre as etapas de entrada e de estruturação de dados é possível inserir uma etapa de relevância, na qual o sistema confere se os dados selecionados são importantes para aquele discurso. Entre a etapa de seleção de expressões

e o texto final, pode-se inserir as etapas de conferência e validação, verificando assim se todas informações estruturadas são verdadeiras.

Num sistema completo, a etapa de publicação seria o último item, uma vez que próprio sistema pode disparar a notícia para um portal de empresa jornalística sem a interferência de um editor. Assim, o software seria capaz de oferecer uma solução completa de produção de textos jornalísticos: desde a entrada do dado no sistema, tratamento, combinação, descarte, estruturação, validação e publicação. Na tabela 2, apresentamos uma possibilidade (das muitas) de arquitetura de software para a produção de notícias utilizando NLG em modo *pipeline* para exemplificar seu funcionamento:

TABELA 2
Passos Fundamentais de NLG em modo *pipeline* aplicada ao jornalismo

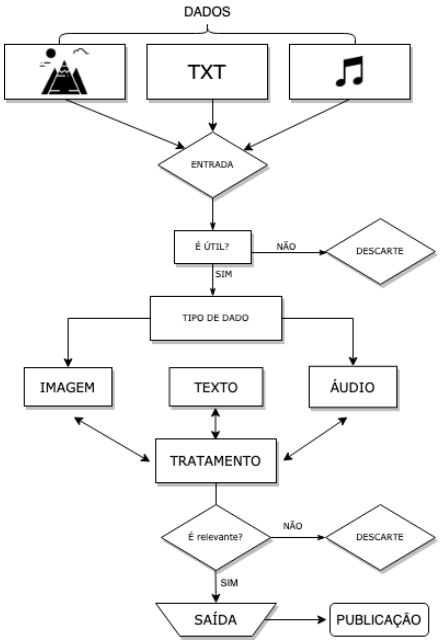
ETAPA	PROCEDIMENTO
Dados	Em seu sentido informacional, um dado é todo aquele arquivo de informação que é armazenado que pode ser coletado na web.
Entrada	É por onde os dados entram no software para serem trabalhados em *natural language generation*.
Validação do Dado	O sistema deve validar se aquele dado contém as informações necessárias para gerar novos arquivos;
Reconhecimento do Dado	O sistema irá reconhecer o formato do dado e poderá, inclusive, transformá-lo em outro formato. Por exemplo: um dado em imagem poderá ser lido e transformado em texto;
Tratamento do Dado	É a etapa em que o dado vai ser transformado em notícia, esta é a etapa de produção textual, conferência da coesão do texto, estruturação do texto e lexicalização, que utiliza as ontologias e a web semântica para gerar conexões e combinações entre frases, acontecimentos, palavras etc.
Validação da Informação gerada	O sistema deverá conferir a validade das informações produzidas. Trata-se da etapa de autenticação da matéria jornalística, função geralmente desempenhada por um jornalista editor nas empresas tradicionais.
Saída	O texto jornalístico está pronto para ser publicado. Neste momento, o sistema pode disparar a matéria para o editor do jornal, ou publicá-la automaticamente no portal do jornal

Fonte: as autoras.

Os sistemas de NLG, de acordo com Ribeiro (2019), "podem ser distinguidos em dois tipos, dependendo dos seus dados de entrada, Text-to-Text, que converte texto para texto e Data-to-Text, que converte dados para texto" (RIBEIRO, 2019, p. 06). Na figura 2, o esquema de entrada de dados é variado, com um sistema capaz de converter imagens ou áudios para texto (*data-to-text*). Neste modelo data-to-text os dados de entrada podem variar muito e são os chamados dados não-linguísticos, como números,

imagens, gráficos, áudios, relatórios, tabelas etc. Na figura 2, demonstramos como funcionaria a NLG no módulo *pipeline*:

FIGURA 2 – Modelo de estrutura de NLG em *pipeline.*



Fonte: as autoras

Por ser alimentado pela web semântica e por ontologias, os sistemas que utilizam o processamento natural de linguagem em *pipeline* ainda são capazes de criar desde textos simples (lead automatizado) até textos mais complexos, como os persuasivos[8]. Ribeiro (2019, p. 07-08) cita vários tipos de textos gerados pela NLG: informativos, simplificados, persuasivos, sistemas de diálogo, explicações, recomendações etc. Os textos informativos são os produzidos a partir de dados concretos, como a previsão do tempo ou o resultado de uma partida de futebol.

Com os avanços em desenvolvimento de software e com o apoio do processamento natural de linguagem, da web semântica e do processamento em camadas (*pipeline*), os sistemas podem ir além da automatização do lead de uma notícia, eles conseguem criar textos relacionais, embasados e verificados. A seguir, destacamos três

[8] Vasco Ribeiro (2019) menciona a capacidade destes softwares de criar diversos formatos textuais, incluindo aí texto com vieses e objetivos distintos, contribuindo inclusive para o fomento à desinformação.

iniciativas que utilizam a NLG que já apresentaram resultados considerados satisfatórios no mercado jornalístico:

- **Automated Insights:**

A *Associated Press* utiliza sistemas de produção de notícias com base em NLG desde 2014 com a inserção de robôs que informam sobre relatórios financeiros. No site da *Automated Insights*, empresa de desenvolvimento do software da A.P., há uma informação sobre o aumento significativo de textos gerados por NLG. O software feito para *Associated Press,* o *Wordsmith*, é capaz de gerar textos com o mesmo estilo de redação da A.P. O *Yahoo!Sports,* outra empresa jornalística, também produz relatórios esportivos personalizados no jogo *Yahoo! Fantasy* com *Wordsmith*. Ele gera notícias para centenas de jogadores e times envolvidos no jogo.

- **Opportunity Gap:**

A *ProPublica*, organização sem fins lucrativos de Nova Iorque, criou um aplicativo para ler e contextualizar dados dos direitos civis do Departamento de Educação dos Estados Unidos. O aplicativo, denominado de *Opportunity Gap*[9], produziu mais de 52.000 matérias em três meses sobre a qualidade das escolas americanas, gerando relatórios com as características socioeducativas de cada região dos EUA.

- **Arria NLG:**

Essa é uma das empresas mais populares de desenvolvimento de software de NLG do mundo. Em 2019, a *BBC News* contratou a *Arria NLG* para auxiliar na cobertura jornalística das eleições no Reino Unido, com o objetivo de fornecer aos seus assinantes relatórios completos com estatísticas em tempo real sobre as campanhas. Em dez horas a BBC publicou 689 textos sobre os resultados dos 690 distritos eleitorais do país, num total de 100 mil palavras. (EXAME, 2019, on-line). Soo Hutton, Engenheiro de Software da *BBC News Labs*, relata que o software combina processamento de dados, geração de histórias e aprovação editorial em um processo simples de apenas um clique. O software funciona com a coleta de dados brutos, gerando automaticamente narrativas jornalísticas

[9] ProPublica explica como utilizar o aplicativo: https://www.propublica.org/article/how-you-can-use-our-opportunity-gap-project-in-your-reporting

com base em modelos projetados pelos jornalistas da *BBC*, conforme o guia de redação e estilo da empresa.

As três empresas mantêm em comum o fato de tratarem centenas de dados transformando-os em narrativas jornalísticas contextualizadas e de interesse público. Esses softwares podem ser considerados "repórteres-robôs", pois podem centralizar as funções de um jornalista: recebem a pauta, vão atrás de fontes, estruturam o texto, verificam a veracidade das informações, criam conexões com eventos passados, produzem os textos (falados ou escritos) e, ainda, os publicam automaticamente em poucos segundos.

A partir desse do processo produtivo automatizado surgem muitas questões sobre o futuro dos jornalistas, como: Em que medida os " repórteres- robôs" podem criar textos tão ou mais refinados que os repórteres de redação? Quais os limites entre softwares e humanos? Quais são os efeitos sobre o jornalista diante deste cenário do jornalismo automatizado?

5. Considerações e caminhos a percorrer

Por meio de um levantamento teórico sobre o conceito de automação no jornalismo, tentamos demonstrar três possíveis aplicações para o desenvolvimento de sistemas de automação de discursos para o jornalismo. Ao compreender como um software de automação de discursos jornalísticos funciona, inferimos que os repórteres-robôs possuem capacidades potenciais para tornar obsoleta a figura de um jornalista de redação.

O jornalismo de automação tem sido utilizado por várias redações, como apontamos ao longo do artigo. Longe de querer criar uma imagem determinística sobre os seus avanços em detrimento de recursos humanos dentro das redações, evidenciamos a necessidade desse fenômeno ser pesquisado. Neste artigo, procuramos mostrar como esses softwares de automação funcionam no âmbito do jornalismo, refletindo sobre as potencialidades da NLG no campo e os possíveis impactos.

Se antes a automação estava restrita ao lead automatizado para responder questões básicas do jornalismo, hoje com a NLG, combinando ontologias e web semântica, os textos podem ser relacionais, contextualizados e profundos. Como destaca Ribeiro

(2019), é possível que estes sistemas reforcem a desinformação no âmbito das redes sociais digitais. Estes softwares são capazes de criar textos intencionais e persuasivos. Por isso, resgatamos a clássica discussão levantada pelo filósofo Piérre Lévy (1999). Tecnologias não são boas, nem más. A adjetivação fica por conta do usuário. Por isso, também é necessário compreender como essas tecnologias estão sendo usadas por empresas e instituições dentro de um contexto político e econômico adjacente aos avanços do jornalismo de automação.

Cabe ressaltar, ainda, que o levantamento teórico apresentado neste estudo faz parte do estado da arte sobre jornalismo de automação, figurando como uma fase preliminar de uma pesquisa de doutorado. Nesta discussão é iminente necessidade de jornalistas (e universidades) aprimorarem seus conhecimentos em programação, linguagens e desenvolvimento de softwares, pois, ninguém mais capacitado para criar e gerenciar um "repórter-robô" dentro de uma redação do que um repórter-humano, dotado de vivências, valores, escopo teórico e técnicas empíricas. Assim, mais do que pensar sobre a completa obsolescência da profissão de jornalista neste cenário, é necessário refletir sobre quais os papéis que os jornalistas poderão desempenhar daqui para frente. Desenvolvedores de software pode ser um deles.

Referências

ARCE, T. **O Lead Automatizado:** Uma Possibilidade de Tratamento da Informação para o Jornalismo Impresso Diário. Revista e-xacta, V. 2, N. 3 (2009).

Appelgren, E., & Linden, C-G. (2020). Data Journalism as a Service: Digital Native Data Journalism Expertise and Product Development. **Media and Communication**, 2020.

BARBOSA, S. J**ornalismo Digital em Base de Dados (JDBD):** Um paradigma para produtos jornalísticos digitais dinâmicos. Tese de doutorado do PPG em Comunicação e Culturas Contemporâneas, Salvador, UFBA, 2007.

BERTOCCHI, D. **Dos dados aos formatos**: um modelo teórico para o design do sistema narrativo no jornalismo digital. São Paulo, USP, 2013.

CARVALHO, F. S. M**. Sistemas de Conteúdo por jornalistas**: um modelo adequado às demandas no jornalismo contemporâneo. Dissertação de mestrado do PPG em Jornalismo, Florianópolis, UFSC, 2019

COSTA, C. R. **O futuro do trabalho do Jornalista é o digital.** Líbero, v. 43, p. 43-54, 2019.

LINDEN, Carl-Gustav. Algoritmos para Jornalismo: o futuro da produção de notícias. Líbero, v. 41, p.05-27, 2018.

DEVYATKIN, D.; et. al. **Genetic algorithm-based sentence packaging in natural language text generation.** In: IOP Conf. Ser.: Mater. Sci. Eng., 2019.

ELDRIDGE II, S. A; HESS, K.; TANDOC JR, E. Digital Journalism (Studies) – Defining the Field. **Digital Journalism,** v 7, número 3, 315-317, 2019.

EXAME. **Tecnologia de geração de linguagem natural da Arria expande cobertura eleitoral da BBC no Reino Unido.** Revista Exame. 17 de dezembro de 2019.

LAMMEL, I.; MIELNICZUK, L. **Aplicação da Web Semântica no jornalismo.** Estudos de Jornalismo e Mídia, vol. 9 nº 1, 2012.

Lindén, T. C-G. Algoritmos para Jornalismo: o futuro da produção de notícias. São Paulo: Cásper Libero **Líbero**, 2018.

Linden, T. C-G. (Ed.), Tuulonen, H. E. (Ed.), Bäck, A., Diakopoulos, N., Granroth-Wilding, M., Haapanen, L., ... Toivonen, H. **News Automation**: The rewards, risks and realities of 'machine journalism'. Frankfurt: World Association of Newspapers and News Publishers, WAN-IFRA, 2019.

LARANGEIRA, A.; QUADROS, C.I. A**ssim Caminha o Jornalismo do Século XXI**: do digital ao neo-analógico. Anais do XVI Encontro da Compós, Curitiba: UTP, 2007.

LÉVY, P.. **Cibercultura**. Tradução de Carlos Irineu da Costa. São Paulo: Editora 34, 1999.

LIMA JÚNIOR, W. T.. **Big Data, Jornalismo Computacional e Data Journalism:** estrutura, pensamento e prática profissional na Web de dados. In: Estudos em Comunicação. Dezembro de 2012. V. 12, p.207-222.

_____. **Projeto Rede JorTec:** produção colaborativa de pesquisa visando à experimentação e criação de inovações tecnológicas digitais. Comunicação & Sociedade, v. 37, p. 47, 2015.

MAGALHÃES, D. L. **Precisão, rapidez e robôs:** um panorama atual do Jornalismo algorítmico. In: Revista Temática. Ano XIII, n. 08 - 2017.

MANOVICH, L. **The language of new media**. Cambridge. MIT. 2001.

MENDONÇA, V. S. **Notícias Geradas por Software:** O Jornalismo Sem Repórter. 2016.

MIELNICZUK, L.. Sistematizando alguns conhecimentos sobre jornalismo na web. In: PALACIOS, Marcos; MACHADO, Elias.(Org.). **Modelos de jornalismo digital**. Salvador: Calandra, 2003, p. 37-54.

MILOSAVLJEVIC, M.; VOBIC, I. **Our task is to desmitify fears:** analysing newsroom management of automation in journalism. In: Journalism, 1:1-19, 2019.

RIBEIRO, V. F. **Jornalista-Robot:** produção automática de conteúdos de texto como apoio ao jornalismo desportivo. 2019

SIRÉN-Heikel, S.; LEPPANEN, L.; LINDÉN, C-G; BACK, A. **Unboxing news automation:** Exploring the imagined affordances of automation in news journalism. In: Nordic Journal of Media Studies. 1:47-66, 2020.

TRÄSEL, M. R. **Entrevistando planilhas** : estudo das crenças e do ethos de um grupo de profissionais de jornalismo guiado por dados no Brasil. Tese (Doutorado em Comunicação Social) - PUCRS, Porto Alegre, 2014.

OLIVEIRA, A. R. DE. **Metadados como atributos da informação estruturada em bases de dados jornalísticas na web**. Tese de doutorado do PPG em Comunicação Social da Universidade Metodista. São Bernardo, 2016.

PEISER, J. **The Rise of the Robot Reporter.** The New York Times. 05 de fevereiro de 2019.

QUADROS, C. I. Base de dados: a memória extensiva do jornalismo. **Em Questão**, Porto Alegre: UFRGS, 2005.

SCHWINGEL, C. Jornalismo Digital de Quarta Geração: a emergência de sistemas automatizados para o processo de produção industrial no Jornalismo Digital. In: Compós, 2005, Niterói: UFF, Compós, 2005.

_____. **Sistemas de produção de conteúdos no ciberjornalismo:** a composição e a arquitetura da informação no desenvolvimento de produtos jornalísticos. Tese de doutorado defendida no PPG Comunicação e Culturas Contemporâneas, Salvador, UFBA, 2008.