UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM JORNALISMO

Matheus Costa Nunes

ALGORITMOS E HUMANOS, PARCEIROS DE REDAÇÃO: A cobertura do portal G1 nas eleições municipais de 2020

FLORIANÓPOLIS - SC
2023

Matheus Costa Nunes

ALGORITMOS E HUMANOS, PARCEIROS DE REDAÇÃO: A cobertura do portal G1 nas eleições municipais de 2020

Dissertação apresentada ao Programa de Pós-Graduação em Jornalismo, do Centro de Comunicação e Expressão, da Universidade Federal de Santa Catarina, na Linha de Pesquisa 2, como requisito para obtenção do título de Mestre.

Orientador(a): Profa. Dra. Stefanie Carlan da Silveira

FLORIANÓPOLIS - SC
2023

NUNES, Matheus Costa
ALGORITMOS E HUMANOS, PARCEIROS DE REDAÇÃO : A cobertura do portal G1 nas eleições municipais de 2020 / Matheus Costa NUNES ; orientadora, Dra. Stefanie Carlan da Silveira, 2023.
216 p.

Dissertação (mestrado profissional) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Jornalismo, Florianópolis, 2023.

Inclui referências.

1. Jornalismo. 2. jornalismo automatizado. 3. textos automatizados. 4. datificação. 5. mediação algorítmica. I. da Silveira, Dra. Stefanie Carlan. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Jornalismo. III. Título.

Matheus Costa Nunes

ALGORITMOS E HUMANOS, PARCEIROS DE REDAÇÃO: A cobertura do portal G1 nas eleições municipais de 2020

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 23 de junho de 2023, pela banca examinadora composta pelos seguintes membros:

Prof. Dr. Márcio Carneiro Santos,
Universidade Federal do Maranhão

Profa. Dra Rita Paulino,
Universidade Federal de Santa

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Jornalismo pelo Programa de Pós-Graduação.

Coordenação do Programa de Pós-Graduação

Prof.(a), Dr.(a)
Orientador(a) Stefanie Carlan

Florianópolis, 2023

A cada instante, milhões de acontecimentos se sucedem, ou não; a cada instante, variáveis se transformam em dados, o virtual se torna presente, e é assim que, a cada instante, o mundo se apresenta num estado diferente. Em sua escala diminuta, o que quer que um escritor venha a compor, está sempre fadado a fazer esse tipo de trabalho: como tudo pode acontecer, cabe a ele decidir que uma coisa aconteça, e não outra. (CARRÈR, 2016, p.75)

AGRADECIMENTOS

Gostaria de agradecer antes de mais nada aos meus pais, que sempre me estimularam a estudar, aprender e enfrentar novos desafios. Devo boa parte dos meus projetos mirabolantes à criação dada por Claudionor Moura Nunes Junior e Wanira Tabatta Wanderley Costa.

À minha vó, Irineide da Costa, que me transmitiu o hábito da leitura. Sem esse costume, escrever uma dissertação não teria sido tão gratificante.

Agradeço ao corpo docente das diversas instituições de ensino que passei (Colégio Marista, Colégio Pódion, Universidade de Brasília e Instituto de Educação Superior de Brasília) por me estimularem a ser curioso.

À minha orientadora, Dra. Stefanie Carlan da Silveira, por me formar intelectualmente com obras e ideias que carregarei para sempre.

Agradeço aos meus amigos e a minha namorada por proporcionarem momentos de descontração, viagens e conversas. Se ninguém ouvisse meus momentos “palestrinha”, estudar seria menos empolgante.

RESUMO

O jornalismo automatizado é o campo de foco desta pesquisa, que busca revisitar os principais conceitos ao redor desta prática que utiliza a geração de linguagem natural para a produção de notícias. A fim de debater as características deste campo de produção jornalística, o conceito de jornalismo automatizado foi analisado dentro do contexto do jornalismo digital, como um desdobramento do jornalismo de dados. Essas variantes de elaboração de notícias unem técnicas de apuração a linguagens de programação, processamento de bases de dados, recursos estatísticos e valores tradicionais da profissão. Para isso, foi realizada uma revisão bibliográfica extensiva e interdisciplinar com o objetivo de contextualizar o jornalismo dentro de um cenário macro de mudanças tecnológicas. Para aproximar a teoria da prática, a pesquisa se propõe a analisar as notícias e os relatos de profissionais envolvidos no projeto de cobertura das eleições municipais de 2020, conduzido pelo portal G1. Desta forma, espera-se mapear quais são as ferramentas, dinâmicas de trabalho e interações necessárias para viabilizar a produção de textos automatizados dentro de um veículo de comunicação.

Palavras-chave: jornalismo automatizado; textos automatizados; datificação; mediação algorítmica; jornalismo de dados.

ABSTRACT

This research aims to study the field of automated journalism through a revision of the main concepts around this practice that employs natural language generation for newsmaking. In order to discuss the characteristics of this type of journalism, the concept of automated journalism was analyzed within the context of digital journalism, as a variation of data journalism. These modern developments of news production unite techniques of reporting with programming languages, database processing, statistical research tools and traditional values of the profession. To this end, an extensive literature review was conducted with the aim of contextualizing journalism within a macro scenario of technological changes. In order to approach both theory and praxis, this research analyzes the coverage of 2020's Brazilian municipal elections made by the news portal G1. Such investigation examines both the news pieces made by natural language generation, as well as the accounts of human professionals enrolled in the project. Therefore, this dissertation aims to create a mapping of tools, work dynamics and interactions necessary to enable text automation in media organizations.

Keywords: automated journalism; automated text generation; datification; algorithmic mediation; data journalism.

SUMÁRIO

LISTA DE FIGURAS & TABELAS

1. INTRODUÇÃO

Em momentos diversos da história, os avanços tecnológicos ameaçaram trabalhadores com a sombra do desemprego estrutural. Desde a primeira Revolução Industrial (HOBSBAWN, 1981), pensadores professavam a chegada da irrelevância do trabalho humano frente às inovações de máquinas. Ao longo do século XIX, operários ingleses chamados de Luditas já expressavam sua indignação contra novos recursos tecnológicos destruindo máquinas de tear hidráulicas. Traços de uma filosofia anti-tecnologia parecem ressurgir sempre que uma onda de mecanização impacta a sociedade de forma ampla, rápida e disruptiva, levando alguns autores a falarem de um neo-ludismo no início do século XXI (BIGGE, 2006; GRAHAM, 2012; JONES, 2013).

Segundo Umberto Eco (2011), vários estudiosos moldam um imaginário pessimista e "apocalíptico" em momentos de profundas mudanças técnicas. Enquanto no outro extremo os chamados "integrados", entusiastas das inovações, anunciam novos equipamentos como o caminho certo para o progresso. Os episódios em que avanços tecnológicos modificam a relação de pessoas com o trabalho não são somente recorrentes, mas sim um dos elementos constitutivos da modernidade[1]. Embora a relação do homem com a máquina seja uma constante, o debate sobre essa dinâmica parece se pautar por posições polarizadas que inflam ânimos de movimentos trabalhistas, sociais e políticos.

No século XXI, a discussão ressurge com a perspectiva de *softwares* de Inteligência Artificial conectados a máquinas automatizarem processos, antes conduzidos tanto por trabalhadores braçais, quanto por empregados escolarizados. Segundo o Relatório sobre o Futuro do Trabalho (2020), 43% dos negócios têm a intenção de reduzir sua força de trabalho em razão da integração de tecnologias. O estudo conduzido pelo Fórum Econômico Mundial estima que 85 milhões de empregos serão afetados pela automação nos próximos cinco anos. Enquanto alguns estudiosos analisam os efeitos da automação sobre as condições de trabalho em si, pautados em grande medida pela economia política, uma perspectiva neomaterialista busca compreender qual é o poder de agência da tecnologia sobre humanos ao mediar nosso

[1] Nem todos os autores aceitam a divisão de tempo conhecida como pré-modernidade, modernidade e pós-modernidade. Um dos principais teóricos que rejeita essas distinções, Bruno Latour, defende em sua obra homônima que *Jamais Fomos Modernos*. Porém, essa pesquisa escolhe cunhar esses termos de divisão temporal pela facilidade que promovem ao leitor e pela ampla utilização por outros teóricos. A separação entre modernidade e pós-modernidade pode ser compreendida a partir da obra de Jean-François Lyotard, *A condição pós-moderna.*

consumo e disseminação de saber (LATOUR, 2005; LEMOS, 2019; GRUSIN, 2015; HJARVARD, 2014).

O jornalismo também se insere dentro deste cenário de inovação. Os efeitos da informatização, do advento da internet à digitalização de rotinas produtivas, criaram um cenário de "mudança de horizonte" (KAWAMOTO, 2003, p. 1). 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 inteligência artificial, o jornalismo automatizado, campo de foco desta pesquisa.

A fim de compreender as características deste campo de produção jornalística, o conceito de jornalismo automatizado será analisado 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 de caso empírica, foi escolhido o exemplo da cobertura feita pelo portal G1 nas eleições municipais de 2020[2]. Naquele ano, o projeto realizou duas coberturas com o auxílio de Inteligência Artificial. No primeiro turno, foram publicados os resultados dos pleitos com informações sobre os vitoriosos. Em votações que foram para o segundo turno, foram publicados dois textos, um para cada concorrente. Já em um segundo momento, o projeto dedicou-se a reportar no dia 1º de janeiro sobre a posse dos prefeitos e vereadores, mas dessa vez, com mais informações no texto produzido automaticamente.

A iniciativa fez a cobertura do evento político em mais de 5.568 municípios brasileiros em menos de 24 horas, nas duas ocasiões. Ela foi descrita pelo próprio veículo como inédita e viabilizada "graças a um modelo de automação que se utiliza de Inteligência Artificial criado em conjunto com a área de Tecnologia da Globo". Sendo assim, a partir da revisão bibliográfica e da análise empírica, este trabalho busca responder o seguinte problema de pesquisa: 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 5 mil notas de jornalismo político em um único dia?

[2] Disponível em: https://g1.globo.com/politica/eleicoes/2020/noticia/2020/12/30/g1-publica-textos-sobre-posse-de-prefe. Acesso em 06 de Março de 2021.

O que o texto das matérias de cobertura eleitoral publicadas de forma automatizada diz sobre essa forma de produção jornalística?

Iniciativas que, como essa, envolvem recursos de análise de dados e Inteligência Artificial, têm se tornado cada vez mais recorrentes ao longo da última década (CODDINGTON, 2014). Muitos desses projetos consistem no uso de *softwares* e bancos de dados, que na prática já fazem parte da rotina de jornalistas pelo menos desde de 1990 (LINDEN, 2017). Para entender o que mudou, é preciso avaliar as características desse processo, tal qual os aspectos da notícia como produto final e mapear as técnicas que saíram de empresas de tecnologia para dentro de redações.

O processamento de grandes volumes de dados para "delimitar a forma de uma história" (GRAY et al, 2014, p. 4) é uma definição de jornalismo de dados que apresenta esta área enquanto instrumento para a apuração jornalística. De forma semelhante, o jornalismo automatizado é descrito por Carlson (2014, p.226) como processos algorítmicos que convertem dados em narrativa jornalística, porém, esta forma de produção precisaria de uma "ação limitada da intervenção humana para além das escolhas iniciais de programação". Outra característica própria da área seria a transferência de autoridade da figura humana para um algoritmo, que passa a ocupar um território antes exclusivo para o homem (HARARI, 2018).

Tal processo de transferência de autoridade para algoritmos, alinhado a um crescente emprego de *softwares* de análise de dados, compõem um fenômeno econômico amplo, que engloba diversas áreas de atuação profissional. A ideia central deste processo é quantificar o comportamento de usuários em ambiente digital como forma de agregar valor a um produto, com fins de predição e monitoramento social (MAYER-SCHOENBERGER & CUKIER, 2013). Tal fenômeno é denominado por Van Dijck (p. 39, 2017) como 'datificação' e carrega consigo características ideológicas que afetam tanto o mercado, como formas de produção de conhecimento (ZUBOFF, 2019). O jornalismo se vê afetado pela datificação nas duas pontas, enquanto organização que busca entender o comportamento do público por meio da análise de dados, mas também para gerar um de seus produtos finais: a notícia (LATZER, 2016).

A partir de uma revisão prévia de conceitos como jornalismo automatizado (DORR, 2015; GRAEFE, 2016), repórter *robot* (CARLSON, 2014), notícias escritas por máquinas (VAN DALEN, 2012), jornalismo computacional (DIAKOPOULOS, 2019), notícias automatizadas (CARREIRA, 2017) e jornalismo algorítmico (DORR, 2015), é possível notar que a característica central do campo é a 'autonomia' para a criação de conteúdo jornalístico.

Para Carlson (2016, p. 226) foram os "avanços na Inteligência Artificial que tiraram escritores não-humanos da teoria para a prática, tendo o jornalismo como principal indústria

afetada". A característica primordial da automação para a produção jornalística se dá no âmbito de uma atividade-chave para a profissão: a escrita. Entender o que é jornalismo automatizado passa por dimensionar o mesmo grau de não intervenção humana descrita por Groover (1980) em processos industriais, mas levando em conta que o jornalismo é mais do que o ato de escrever e publicar. Como aponta Diakopoulos (2019, p. 97), o que pode ser chamado de "produção de conteúdo automatizado" deve ser na prática "orquestrado" por repórteres de carne e osso.

Dar ênfase na característica da "não assistência humana" dentro do jornalismo automatizado é relevante para compreender o que muda com a introdução da Inteligência Artificial na produção de conteúdo noticioso, porém, desconsiderar a dimensão humana do fazer jornalístico nesta área representa pintar um retrato incompleto por duas razões. Primeiro, é possível para o algoritmo "produzir um texto preliminar que será complementado por autores humanos" (WOLKER-POWELL, 2018, p. 86). Segundo, porque estas tecnologias "não são capazes de produzir textos sem a interferência humana" (DORR, 2015, p. 9), uma vez que a etapa de desenvolvimento do código e fase de ajustes exigem a participação de profissionais da área e são fundamentais para o resultado final. "[...] o futuro pós-humano do jornalismo parece sugerir um estado híbrido no qual computadores e autores de carne e osso se misturarão de formas ainda estão por se desenvolver" (CARLSON, 2016, p. 227, tradução nossa)[3].

Entender a possibilidade de colaboração entre atores humanos e não-humanos em veículos de comunicação é parte essencial desta pesquisa. O objetivo é investigar as dinâmicas produtivas que ocorreram dentro do projeto de cobertura automatizada das eleições municipais de 2020, feito pelo portal G1, a partir da interação entre atores humanos e recursos tecnológicos. Para alcançar esse propósito, foram determinados sete objetivos específicos que visam esmiuçar o jornalismo automatizado em sua dimensão prática e teórica.

- Analisar as características de reportagens produzidas em uma cobertura jornalística automatizada;
- Investigar como ocorreu o preparo preparo do algoritmo e quais são os principais conceitos empregados neste processo;
- Identificar qual foi a contribuição dos jornalistas para os preparativos;

[3] "The argument offered in this chapter is not that humans will be displaced from journalism. Instead, the posthuman future of journalism alluded to in the title suggests a hybrid state in which computer and human-authored stories intermingle in ways yet to be developed".

- Distinguir o trabalho da equipe de tecnologia da equipe de jornalismo, sem deixar de observar a colaboração entre ambas;
- Elencar quais outras ferramentas digitais além dos algoritmos foram utilizadas e quais foram a sua função;
- Especificar o que ocorreu em cada etapa, da conceituação do projeto a publicação das reportagens;

Fica patente nos objetivos listados acima que esta pesquisa encara a automação de notícias como um fenômenos sociotécnicos e, portanto, dotado de causas diversas e procedimentos complexos. A fim de observar a cooperação entre esses atores diversos,, se faz necessária tanto a descrição do funcionamento de ferramentas digitais, quanto das atividades executadas por jornalistas, programadores, revisores e etc. Para alcançar uma compreensão detalhada de um trabalho multidisciplinar, foi escolhida a Teoria Ator-Rede (TAR) como marco teórico para este estudo, por permitir a associação do trabalho humano e não-humano dentro de uma rede tecida por interações. Como coloca Latour (2005, p. 72, tradução nossa[4]), "além de determinar e servir como pano de fundo para a ação humana, coisas podem autorizar, custear, encorajar, permitir, sugerir, influenciar, bloquear, tornar possível, proibir e assim por diante".

O pesquisador André Lemos (2020) também defende em sua perspectiva neomaterialista da comunicação que as mediações só podem ser compreendidas a partir de sua materialidade. Ou seja, é impossível descrever qualquer processo comunicacional sem considerar os dispositivos que agem como mediadores da intersubjetividade dos humanos. Não é a intenção subjetiva dos homens a essência do processo, tão pouco os agenciamentos das coisas. Fundamentalmente, a comunicação existe somente nessa dimensão híbrida da ação de humanos e não-humanos, agindo de maneira associada para criar sentido.

Essa produção conjunta de sentido ocorre no jornalismo automatizado com o uso de *softwares* de Geração de Linguagem Natural (DÖRR, 2015), tecnologias que geram textos em linguagem natural a partir de um processo algorítmico que converte bancos de dados estruturados em enunciados, numa velocidade de milhares de páginas por segundo. Simplificadamente, este tipo de *software* é capaz de retirar informações de uma base de dados e preencher relatórios com espaços criados em um *template* sob medida para exposição desses

[4] In addition to 'determining' and 'serving' as a 'backdrop for human action', things might authorize, allow, afford, encourage, permit, suggest, influence, block, render possible, forbid, and so on (LATOUR, 2005, p. 72).

dados[5]. Em sua forma mais complexa, o *software* pode fazer decisões sobre o uso de sinônimos, dando preferência a um tipo de vocábulo em detrimento de outro, se conformando dentro de um determinado estilo textual.

De acordo com a Automated Insights (2018), principal empresa prestadora de serviços de Geração de Linguagem Natural, esse tipo de *software* consiste em transformar dados em narrativas escritas. O primeiro pré-requisito é ter acesso a grandes volumes de dados e estruturá-los. O segundo é definir o formato da narrativa por meio de *templates*, também chamados de fluxos de trabalho baseados em regras, ou seguir por uma abordagem baseada em intenção, onde a máquina está 'livre' para escrever dentro de um conjunto de parâmetros. O último passo é realizar o *template*, ou a intenção, com os dados estruturados em um texto corrido. Segundo Graefe (2016), esses dados estruturados são frequentemente planilhas, mas podem também estar organizados em outros formatos como arquivos ".HTML", "JSON" ou "CSV". Essa observação se mostra verídica ao observar o *Wordsmith*, principal produto da Automated Insights, que recebe apenas arquivos do tipo "CSV".

O uso de banco de dados para investigação já é um insumo básico para o jornalismo de dados, assim como para iniciativas de transparência de governos e empresas desde a década de 1990 (CODDINGTON, 2015). O que parece mudar no jornalismo automatizado é mais o processo algorítmico de redação da notícia e menos a abordagem de apuração documental com protagonismo para os números no lide. Teóricos da computação, tal qual McDonald (2010) e Reiter (2012), dividem esse processo em três etapas: (1) planejamento de documento, (2) microplanejamento e (3) realização do documento. No (1) planejamento de documento, todos os dados são enumerados para se definir 'o que eles' comunicam e 'em que ordem' devem ser priorizados, segundo critérios de valor notícia. Já na etapa (2) microplanejamento, se decidem as nuances de linguagem, como especificar expressões de referência, ou sinônimos para se referir a pessoas e organizações, evitando repetições. Por último, a (3) realização do documento gera o texto seguindo parâmetros gramaticais e sintáticos, sendo considerada a etapa na qual o jornalista menos tem influência sobre.

O recurso da GLN foi empregado no projeto do G1 em conjunto à captação de dados públicos disponibilizados pelo Tribunal Superior Eleitoral (TSE), segundo anúncio feito pelo portal[6]. Observar a extensão em que esta forma de Inteligência Artificial (IA) foi usada, em

---

[5] Ver *"A Comprehensive Guide to Natural Language Generation"*. Disponível em: https://medium.com/sciforce/a-comprehensive-guide-to-natural-language-generation-dd63a4b6e548. Acesso em 06 de Março de 2021.

[6] Disponível em: https://g1.globo.com/politica/eleicoes/2020/noticia/2020/11/12/em-iniciativa-inedita-g1-publica-textos-com-resu

combinação com outros recursos de programação e ferramentas digitais, faz parte dos objetivos desta pesquisa.

Uma tecnologia similar foi usada pelo jornal britânico The Guardian em 2020 e o resultado foi a publicação do artigo "Um Robô escreveu este artigo: você está com medo, humano?"[7] (tradução nossa). O programa de *open source* GLN batizado de GPT-3 foi desafiado a escrever todo um artigo respondendo à questão: "Robôs são pacíficos?" O artigo foi publicado no jornal sem nenhuma edição e instigou outros veículos de imprensa a fazerem experimentos similares. O *New York Times*[8], por exemplo, cedeu espaço à IA em sua coluna *Modern Love,* para narrar um pouco de sua vida amorosa. A coluna é um espaço tradicionalmente dedicado a crônicas sobre relacionamentos amorosos, e a IA se mostrou capaz de ao menos começar a escrever sobre o assunto.

Todos os experimentos servem para demonstrar a viabilidade de iniciativas jornalísticas que buscam alcançar um alto grau de autonomia na produção de notícias, embora o refinamento da Inteligência Artificial e os questionamentos que ela deverá responder sejam cuidadosamente pensados por um entrevistador humano e programadores. Mesmo com a característica da automatização sendo central para este campo de estudo, a relevância do fator humano deve ser observada e compreendida (DALBEN, 2020).

No caso do projeto do G1, também se sabe pelo anúncio do portal que aconteceu interferência humana em ao menos dois momentos. Primeiro, os preparativos foram feitos por uma equipe de programadores e jornalistas para a estruturação da base de dados do TSE, que alimentaria informação para a redação do texto pelos *softwares*. Segundo, antes dos textos serem publicados no site, após serem redigidos pela IA, um grupo de revisores trabalhou para corrigir possíveis erros, embora não se saiba quantos revisores foram empregados para tal tarefa e se a totalidade dos textos passaram por essa revisão. Também foi anunciado pelo veículo que a fase dos preparativos envolveu "profissionais de diversas áreas" e que o projeto foi "trabalhado por meses".

Fora as escassas informações disponibilizadas pelo anúncio, também se sabe o tipo de informação que foi priorizado na redação e, portanto, na estruturação da base de dados. Primeiramente, na ocasião da cobertura do primeiro turno, foi contemplado: o nome do candidato eleito, o partido, o número de votos recebidos, bem como a idade, o estado civil, o

---

ltado-da-eleicao-em-cada-uma-das-5568-cidades-do-brasil-com-auxilio-de-inteligencia-artificial.ghtml. Acesso em 7 de Março de 2021.

[7] Disponível em: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3. Acesso em 06 de Março de 2021.

[8] Disponível em: https://www.nytimes.com/2020/11/24/science/artificial-intelligence-gpt3-writing-love.html. Acesso em 06 de Março de 2021.

grau de instrução, a profissão e o patrimônio declarado. Depois, na ocasião da posse dos vereadores e prefeitos, os textos incluíram todas as informações anteriores, somadas de: ocupação e patrimônio do vice-prefeito, qual partido deteve a maior bancada, a quantidade de vereadores por partido, a relação de vereadores eleitos e também informações sobre o município, como população, Produto Interno Bruto (PIB) e o Índice de Desenvolvimento Humano (IDH).

Sobre a colaboração entre humanos e inteligências artificiais, existe ainda uma segunda indagação sobre os impactos nas relações de trabalho. Para além da questão técnica, o emprego dessas tecnologias pode gerar opiniões e angústias diversas em profissionais de imprensa. Muitos desses trabalhadores podem enxergar a substituição de algumas tarefas executadas na redação por máquinas como um caminho para perder o seu trabalho (DALBEN, 2020). Principalmente se a opinião geral dos profissionais for de que as tarefas essenciais para o trabalho de reportagem e edição podem ser executadas por máquinas com um padrão de qualidade equivalente ao de um jornalista experiente. A própria palavra robô frequentemente carrega uma significação distópica, que está associada a uma adversidade com humanos (DALBEN, 2020). Na tela dos cinemas, *blockbusters* como o Blade Runner (1982), o Exterminador do Futuro (1984) e Matrix (1999) retratam repetidamente a Inteligência Artificial acoplada a criaturas de ferro e silício que intimidam a soberania do *homo sapiens* na Terra. Assim como os luditas do século XIX (JONES, 2013), trabalhadores modernos passam a nutrir um sentimento de ameaça diante da possibilidade de máquinas minarem seu *status* e seu sustento. Uma ponte de mais de três séculos parece transportar uma ansiedade típica de cenários de inovação tecnológica.

Porém, não cabe a apenas esta pesquisa discutir os potenciais efeitos da inteligência artificiais na empregabilidade dos diversos profissionais que compõem um veículo de comunicação (repórteres, editores, produtores, fotógrafos e etc). No entanto, é possível listar a partir de entrevistas com os envolvidos no projeto quais são as atividades das rotinas produtivas jornalísticas que são mais impactadas pela automação. Por ‘atividade’ deve-se entender não a profissão como um todo, mas as pequenas tarefas que são recorrentes no jornalismo, como anotar perguntas antes de uma entrevista, ou gravá-la durante. Embora não seja o objetivo desta pesquisa decifrar os ânimos e ansiedades de profissionais de imprensa em relação à manutenção futura de seus empregos, observar pragmaticamente (fazer-como) os impactos da automação nesta atividade, pode oferecer um vislumbre sobre como estes postos de trabalho tendem a mudar.

A importância do estudo do jornalismo automatizado na prática reside justamente em sua característica de ser uma novidade. Segundo Graefe (2016), a ideia de automatizar a produção de textos já existe há meio século, mas sua execução com *softwares* de NLG em redações começou em 2015. Para o autor, esta forma de produção jornalística pode ser vista como estando em uma fase experimental, ou em uma fase inicial de expansão de mercado. Seja qual for a situação, a pesquisa voltada para o emprego de algoritmos em veículos de comunicação se coloca como uma forma de compreender um possível fator de mudança para as dinâmicas produtivas tradicionais. Vale destacar que no mesmo ano de 2015 o Fórum Mundial de Editores[9] listou a área como uma das principais tendências para as redações.

Para além do fator novidade, é importante considerar o potencial de crescimento de iniciativas similares à do G1. O uso de *softwares* nas redações age no sentido de cobrir áreas que, do contrário, não seriam cobertas por falta de funcionários disponíveis (GRAEFE, 2016). A *Associated Press*[10], por exemplo, iniciou um projeto em 2014 de automação para a cobertura de campeonatos esportivos universitários que, inicialmente, não recebiam atenção em detrimento de campeonatos profissionais. Em um cenário de enxugamento das redações (CHRISTOFOLETTI, 2019), deve ser considerada a possibilidade que a automação possa funcionar como um fator de compensação nos quadros das redações e portanto, a análise de sua aplicação ganha uma utilidade tanto acadêmica quanto comercial

[9] Ver "9 top trends in newsrooms around the world". Disponível em: https://www.journalism.co.uk/news/9-top-trends-in-global-newsrooms/s2/a565371/. Acesso em 9 de Março de 2021.
[10] Ver "NCAA to grow college sports coverage with automated game stories". Disponível em: https://www.ap.org/press-releases/2015/ap-ncaa-to-grow-college-sports-coverage-with-automated-game-stories. Acesso em 9 de Março de 2021.

2. MARCO TEÓRICO E PROCEDIMENTOS METODOLÓGICOS

A primeira instância metodológica do trabalho é a inserção da pesquisa em um quadro teórico de referência sobre tecnologia, jornalismo e automação. Para isso, foi feita uma pesquisa bibliográfica extensiva (STUMPF, 2006) a fim de identificar quais autores e pesquisadores podem auxiliar na construção do cenário e da fundamentação da dissertação. A revisão de literatura tanto de autores nacionais, quanto internacionais, é essencial para entender se a cobertura analisada aqui se insere dentro de parâmetros produtivos similares a outras iniciativas previamente estudadas. Ademais, a exploração da literatura vigente mune o presente estudo dos conceitos necessários para analisar os processos descritos pelas outras metodologias.

A fim de reunir um material suficientemente representativo da produção acadêmica sobre o tema, realizou-se uma revisão sistematizada, conduzida em quatro idiomas: inglês, português, espanhol e francês. As palavras-chave[11] correspondentes ao assunto da automação no jornalismo foram inseridas no Google Acadêmico. Tais palavras foram selecionadas, antes de serem traduzidas, a partir do que o autor identificou como termos recorrentes em textos sobre o jornalismo automatizado. A busca no Google Acadêmico foi feita mediante a aplicação dos filtros de "Ordenar por Relevância" e "Ordenar por Data". Esta peneira nas publicações auxiliou a selecionar para a revisão de literatura os textos de maior visibilidade e, ao mesmo tempo, fazer uma leitura cronológica das publicações dentro do campo.

Para analisar objetivamente os detalhes de um projeto de jornalismo automatizado, esta pesquisa foi guiada pela Teoria Ator Rede, mais especificamente, a abordagem metodológica neomaterialista proposta por André Lemos (2020). Dentro desta perspectiva epistemológica, entende-se que a produção coletiva de sentido e recursos tecnológicos continuamente se relacionam segundo a lógica da 'permanente mutação' (CATTANI-HOLZMANN, 2006). Ambos se alteram reciprocamente em um turbilhão que afeta indivíduos, empresas, governos e instituições.

É a partir desses distúrbios provocados no dia-a-dia pela mecanização, tecnologias da informação e dispositivos digitais que o discurso contemporâneo se molda, reage e tenciona significados de forma dinâmica. Para Umberto Eco (2011), há duas maneiras predominantes de como o discurso sobre inovações tecnológicas se organiza. De um lado, um grupo de

[11] Os termos pesquisados foram "jornalismo automatizado", "robô jornalista", "repórter robô", "jornalismo algorítmico" e "notícias automatizadas". As traduções para outras línguas requisitaram algumas mudanças nos termos como "information automatisée" em francês e "periodismo ciborg" em espanhol.

pessimistas, por vezes tradicionalistas, por vezes antissistema, assume um posicionamento adverso a mudanças profundas. Enquanto no outro extremo, os entusiastas anunciam novos equipamentos como uma forma de emancipação do homem, controle do meio e único caminho possível para o progresso. Os posicionamentos polarizados jogam um véu sobre as possibilidades de compreender as implicações reais de uma nova tecnologia na sociedade.

Para Lemos e Bitencourt (2021), essa polarização persiste nos dias atuais em campos intelectuais e pode ser percebida em pesquisas acadêmicas. Se por um lado, inovações tecnológicas são abordadas em termos de salvacionismo e da utopia tecnológica, uma corrente menor encara algumas inovações apenas como produtoras de impactos negativos nos campos da segurança, vigilância e da política. Ambas as posições falham em reconhecer a materialidade existente nas mediações, tal uma caneta que precisa ser pressionada sobre um papel, antes mesmo do conteúdo de um texto poder ser interpretado por outrem.

De maneira complementar, Morozov (2018) aponta a existência de um discurso hegemônico nascido de grandes empresas de tecnologia que é ecoado por meios de comunicação e formadores de opinião mundo afora. Tal discurso defende o **solucionismo tecnológico**, um deslumbramento quanto à capacidade das empresas de *Big Tech* promoverem soluções para problemas de natureza trabalhista, política ou econômica. Esse discurso utópico também falha em descrever objetivamente as mudanças provocadas pelas TICs.

Em uma breve análise da história do jornalismo, percebe-se que o desenvolvimento da área sempre esteve ligado à evolução da tecnologia. Começando por relatos de viagens do período das navegações, passando pela Prensa de Gutemberg, até chegar ao *feed* de notícias da internet, a relação entre técnica e fazer jornalístico se apresenta como íntima (BRIGGS-BURKE, 2009). A relevância das tecnologias para o jornalismo é tal que os próprios campos, ou gêneros (jornalismo impresso, o jornalismo de radiodifusão e o jornalismo digital) estão ligados a uma tecnologia específica (PRIMO-ZAGO, 2015).

> A tecnologia sempre exerceu um papel importante na apuração e na produção de notícias. Seja ao escrever anotações em uma folha, ou gravar eventos em uma fita de vídeo, ou telefonar para um entrevistado, jornalistas estão acostumados a usar uma variedade de ferramentas técnicas para adquirir dados brutos que eles usam para reportar suas histórias (PAVLIK, 2001, p. 49, tradução nossa)[12].

O jornalismo tem sofrido com a ‘disrupção tecnológica’, processo no qual as mudanças provocadas pelas tecnologias digitais aumentam em ritmo e escala. Santos (2016),

[12] “Technology has always played an important part in the newsgathering and production process. Whether scribbling notes on a page, recording an event on videotape, or taping a telephone interview, journalists are accustomed to using a variety of technical tools to acquire the raw data they use to tell their stories”.

corrobora a ideia de que as tecnologias ocupam um lugar de mediação entre as fontes de poder que se articulam no meio social. Tais fontes de poder em sua esfera comunicacional acabam sendo influenciadas, condicionadas e habilitadas por sistemas técnicos, como por exemplo o desenvovimento de software. Logo, o processo de disrupção tecnológica que afeta o jornalismo pode ser visto como uma mudança tecnolopolítica que altera várias esferas da vida.

Sobre a relação entre disrupção tecnológica e trabalho, Örnebring (2010) argumenta que historicamente as organizações midiáticas sempre buscaram inovações como forma de ganharem competitividade. A *Reuters*, por exemplo, usava barcos a vapor e linhas de telégrafos próprias para dar furos de notícias antes de seus competidores. Nesta perspectiva, a relação entre tecnologia e jornalismo obedece ao **discurso da velocidade**. O imperativo 'busque mais velocidade' vê no avanço técnico uma forma de agregar valor ao conteúdo, mas também de tornar a empresa jornalística cada vez mais competitiva. Por essa lógica, o estreitamento entre tecnologia e fazer jornalístico expressa tanto uma tendência histórica quanto uma necessidade mercadológica.

Para estudiosos da história da mídia, como Örnebring (2010), a própria emergência da atividade profissional jornalística, assim como a demanda da informação neste formato, podem ser creditadas ao surgimento de uma sociedade industrial. A organização do trabalho ao redor de uma linha de montagem, dentro das indústrias, se reproduz em alguma medida dentro das redações também. Nesta visão, os aspectos materiais de uma sociedade que passa a produzir sua riqueza em fábricas e a se organizar em aglomerados humanos urbanos é que tornam necessário e viável o fazer jornalístico. Como Medina (1998) descreve, a veiculação de informação dependia de avanços tecnológicos para acontecer, ao mesmo tempo que ajudava a criar uma espacialidade e temporalidade comum à sociedade industrial.

> Com essas duas variáveis, tempo e espaço, a informação jornalística se alicerça na sociedade urbana e industrial. Vencida uma das principais limitações humanas, tempo/espaço, ninguém tem dúvida ao atribuir a vitória aos recursos tecnológicos que veiculam a informação. E, logo se percebe também que os próprios avanços tecnológicos fazem parte das necessidades da industrialização, o que reforça a informação, no caso, jornalística, como decorrência normal do sistema econômico que está na base (MEDINA, 1988, p.16).

A partir dessa perspectiva, observa-se que o movimento de disrupção tecnológica é abrangente, inerente a sociedades modernas, não afetando apenas o jornalismo, mas o mercado de trabalho como um todo, todavia, com consequências particulares para o exercício da profissão periodística. Por exemplo, segundo o pesquisador Pavlik (2000), os avanços tecnológicos têm impactado a prática do jornalismo em ao menos quatro áreas: (1) como

jornalistas trabalham; (2) o conteúdo da notícia; (3) a estrutura organizacional dos veículos; (4) a relação entre veículos, jornalistas e seus públicos. Esta pesquisa se foca nas duas primeiras áreas de impacto a fim de entender como notícias produzidas por algoritmos, enquanto pessoas produzem algoritmos, se afetam mutuamente.

A relação entre o objeto técnico e o jornalista profissional sempre foi de agência direta. Tal relação se repete, com suas particularidades, entre a produção de notícias e Tecnologias da Informação e da Comunicação (TICs), no advento do que Lewis e Westlund (2015) denominam de 'jornalismo fundido à tecnologia'. Ou seja, a notícia passa a ser produzida e distribuída de acordo com a institucionalização da valorização tecnológica. Na medida que essa mudança drástica ocorre, uma nova perspectiva para o estudo do jornalismo se manifesta ao rejeitar a ideia de que jornalismo é somente aquilo produzido por jornalistas.

> A tecnologia é tipicamente vista como um instrumento que auxilia os processos jornalísticos. Os artefactos digitais, no entanto, raramente são considerados como participantes ativos. Tautologicamente, o jornalismo é definido como uma prática dos jornalistas. Mas o jornalismo não seria o mesmo sem o papel desempenhado pelos artefactos tecnológicos (PRIMO-ZAGO, 2015, p. 38).

Para André Lemos (2020), existe uma perspectiva dominante dentro dos estudos de Comunicação, e até mesmo dentro das Ciências Sociais, que enxerga os instrumentos como meros artefatos submissos às vontades do homem. Essa visão antropocêntrica coloca o homem como único ator a ser analisado, dentro de um fenômeno social mais complexo que engloba objetos, plataformas, animais e espaços geográficos. Por outro lado, o tecido da sociedade e a comunicação que esta produz podem ser compreendidos como uma mistura de atores e materialidades com origens diversas.

Atentar para a materialidade no fenômeno comunicação é entender o objeto como ser atuante, capaz de gerar agenciamentos no comportamento de outros integrantes do processo comunicativo. Observar os objetos como seres que agem, sejam eles o papel de uma revista, as antenas de uma televisão, a tela de um computador, ou o ar que vibra durante uma conversa face a face, é estudar o que *eles fazem* dentro do processo de comunicação. Pois no final das contas, não existe comunicação destituída de materialidade (FELINTO, 2001).

Em busca de superar essa limitação antropocêntrica, tem se notado um crescente emprego da Teoria Ator Rede em estudos sobre tecnologia e jornalismo, como uma forma de repensar os papéis e relações desempenhadas entre jornalistas e máquinas (THURMAN et al, 2019). A perspectiva epistemológica por trás da Teoria Ator Rede, inicialmente formulada pelo sociólogo Gabriel Tarde (1843 - 1904), e posteriormente organizada por diversos pesquisadores, entre eles o antropólogo Bruno Latour, enxerga em qualquer atividade social,

seja ela o trabalho, ou não, uma interação contínua e mutuamente modificadora entre atores humanos e não humanos.

A Teoria da *Assemblage* (DELANDA, 2016) parte de uma noção similar de verdade relacional para promover uma forma de encarar fenômenos complexos, tais quais os comunicacionais. Segundo esta visão ontológica, tudo existe em uma multiplicidade descrita em termos heterogêneos, que estabelecem ligações e relações entre eles, atravessando períodos, gêneros e regimes de naturezas distintas. Logo, a *assemblage* tem como sua única unidade a ideia de co-funcionamento, tal qual uma simbiose. O trabalho do intelectual, ou do pesquisador, não é se ocupar de filiações, mas de alianças e composições. Ou seja, seres de origens distintas como homens, animais, paisagens, objetos e valores se unem na composição de qualquer processo.

No caso do trabalho, se por um lado o instrumento é alterado graças a uma necessidade humana, por outro, o homem que o empunha também é modificado pelas características do instrumento (LATOUR, 2012). Ambos os atores, tanto humanos quanto não-humanos, criam no processo de comunicação associações pragmáticas que afetam o seu estado final.

> Se a ação se limita ao que os humanos fazem de maneira "intencional" ou "significativa", não se concebe como um martelo, um cesto, uma fechadura, um gato, um tapete, uma caneca, um horário ou uma etiqueta possam agir. Talvez existam no domínio das relações sociais. Em contrapartida, se insistirmos na decisão de partir das controvérsias sobre os atores e atos, qualquer coisa que modifique uma situação fazendo diferença é um ator (LATOUR, 2012, p. 108).

Os atores sempre agem e, portanto, são *actantes*. Essa característica ativa das partes que compõem qualquer fenômeno cria relações que se afetam mutuamente, chamadas por Latour de agenciamentos. Quando se fala em mediação algorítmica, como é o caso do jornalismo veiculado em plataformas, ou aquele que se utiliza delas para apuração, se questiona qual o agenciamento que essa entidade virtual tem sobre os usuários e vice e versa. Esses algoritmos digitais, assim como qualquer outro *software*, plataforma digital ou página *web*, também são dotados de sua materialidade. São suas linhas de código, seus formatos de arquivo, suas unidades de medida, seus *inputs* e as informações que estes algoritmos misturam, rearranjam e separam (ANDERSON, 2012).

A comunicação digital e a multiplicidade de objetos técnicos que a compõem (*laptops*, *smartphones*, câmeras, roteadores, interfaces, bancos de dados e outros) convida o olhar do pesquisador à perspectiva da neo materialidade (LEMOS, 2020), ou seja, a análise do fenômeno comunicacional como uma associação de diferentes *actantes*, cada qual dotado de

uma materialidade e de relações de agenciamento. Identificar actantes não-humanos é, para Lemos, realizar um *Inventário*, enquanto para Latour é o trabalho de compor uma *Lista*.

O jornalismo, como parte integrante deste modo de comunicação, não é diferente. Seu processo de produção de notícias engendra diferentes dispositivos e profissionais que se afetam mutuamente, na medida que integram um contexto social mais amplo, gerado por incontáveis associações. Para Anderson e Mayer (2015), o estudo que toma por foco esses "objetos do jornalismo" não desconsidera o contexto social no qual a tecnologia é empregada, mas sim, oferece uma nova porta para o compreender o social.

Estudar a relação próxima entre o fazer jornalístico e a tecnologia que o envolve, tampouco é enquadrar a área dentro de um determinismo tecnológico (ELLUL, 1954)[13]. Para esta área da sociologia, os valores de uma sociedade, suas estruturas e seus processos comunicacionais seriam unicamente impulsionados pela tecnologia existente. Por esta perspectiva, o jornalismo e qualquer outra atividade profissional seriam apenas consequências dos artefatos técnicos de uma época. Seria o extremo oposto da abordagem antropocêntrica que Lemos, Latour, Felinto e outros denunciam. Considerar tanto a natureza dos objetos técnicos, quanto o agenciamento entre atores humanos e não-humanos é fazer uma média entre uma sociologia antropocêntrica, que ignora a materialidade dentro da comunicação, e o determinismo tecnológico que a super-estima.

Os diferentes objetos técnicos dentro de uma fábrica (correias, prensas, esteiras e alavancas), ou de uma redação (computadores, impressoras, televisores e mesas), ou de uma simples busca feita no Google (navegadores, roteadores, cabos de fibra ótica e servidores), operam em grupo. Cada dispositivo se relaciona a outros dispositivos formando o conjunto técnico. É impossível imaginar uma atividade econômica moderna que seja destituída de seu conjunto técnico. Na realidade, a forma como diferentes atividades econômicas, ou grupos de pessoas, compõem e operam o seu conjunto técnico diz respeito a uma cultura técnica. Esta cultura está ligada a uma temporalidade, ou seja, a um momento histórico em que a tecnicidade do objeto e de seu conjunto tomou certa configuração.

> Essa relação do indivíduo humano com o indivíduo técnico é a mais delicada de se formar. Pressupõem uma cultura técnica, que introduz atitudes diferentes daquelas do trabalho e da ação (correspondendo trabalho a compressão dos elementos e ação a compressão dos conjuntos). Trabalho e ação têm em comum a predominância da finalidade sobre a causalidade. Em ambos os casos, o esforço está relacionado a um resultado a ser obtido (SIMONDON, 1969, p. 187).

[13] Jacques Ellul foi um sociólogo e filósofo que publicou diversos trabalhos sobre o impacto da tecnologia moderna sobre a liberdade humana. Ele é comumente visto como o maior representante dos chamados ‘deterministas tecnológicos’, mas tão pouco é o único pensador a se enquadrar nessa alcunha.

Em última instância, o objeto técnico precisa para existir de uma coerência com o espaço geográfico no qual está sendo empregado. A transmissão de uma reportagem pela televisão tem que considerar o tempo para que o *link* entre a equipe *in loco* e o estúdio seja feito. As variáveis para conexão deste *link* podem ser afetadas por questões como distância entre estúdio e equipe, assim como as condições meteorológicas. Os objetos técnicos existem, portanto, em um *meio associado,* que é ao mesmo tempo técnico e geográfico. Em uma breve síntese, é possível compreender que um conjunto técnico (todo) é composto de objetos técnicos (parte), operando em um meio associado, com influência de uma cultura técnica, ao mesmo tempo influenciando e sendo influenciada por uma tecnicidade que progride de um estágio mais abstrato para um mais concreto.

Tanto na *listagem* de Bruno Latour, na *epistemologia neomaterialista,* de André Lemos, ou no *modo de existência dos objetos* de Simondon, os atores não-humanos são deslocados para o foco das reflexões, sem deixarem de partilhar a mesma arena onde os humanos se encontram. Não existimos neste mundo sozinhos, pois "a vida técnica não consiste em dirigir máquinas, mas existir do mesmo nível que elas" (SIMONDON, 1989, p. 117). Estudar processos de mediação algorítmica que se desenvolveram dentro da cibercultura é olhar para as diferentes facetas desses objetos, em interação com os atores humanos.

Em sua gênese, algoritmos digitais se desdobraram em *softwares* de tradução, recomendação, classificação[14] e em recursos de Geração de Linguagem Natural que são empregados dentro do jornalismo para a redação automatizada de textos. Ou seja, a tecnicidade dos algoritmos digitais foi tensionada por uma cultura técnica até desembocar na aplicação que aqui compreendemos como jornalismo automatizado. Para analisar a 'mudança de horizonte' (KAWAMOTO, 2003) provocada pelas mediações algorítmicas na produção jornalística, esses conceitos precisam ser levados em conta. Como a tecnicidade dos algoritmos digitais é empregada pela cultura técnica de um veículo jornalístico? Como os recursos de Geração de Linguagem Natural passaram de um modo abstrato, meramente esquemático, para a concretude de reportagens produzidas e publicadas por esses algoritmos?

De forma similar à Teoria Ator Rede, a filosofia da técnica de Simondon também indica uma modificação relacional e mútua entre os seres humanos e não-humanos. Enquanto essa modificação é válida para a filosofia e para sociologia, também pode ser estendida para os diferentes campos da Comunicação. O teórico Gumbretch (2010), por exemplo, afirma que a materialidade na comunicação é potencializada na pós-modernidade em função do

[14] Para compreender a ampla aplicação de algoritmos digitais na contemporaneidade, ver: O'NEIL, Cathy. Algoritmos de destruição em massa. Editora Rua do Sabão, 2021.

sentimento de um mundo "não mais fundado na figura central do sujeito" (p. 391). O jornalista que ao sentar em frente a um computador, colocar um gravador sobre uma mesa, ou telefonar para uma fonte, o faz acoplado a um objeto técnico.

Há um exemplo nativo do século XXI que ilustra bem o acoplamento entre dispositivo, profissional e conteúdo: o jornalismo para dispositivos móveis. Com a popularização dos *smartphones* a partir de 2008[15], a diagramação das notícias passa a sofrer crescentes adaptações para o formato vertical da tela desses dispositivos. Em poucos anos, esses aparelhos portáteis passaram a ser o principal canal de acesso a leitura de notícias. Ao mesmo tempo, novas dinâmicas de apuração e edição do produto jornalístico surgem influenciadas pelos *apps* (aplicações) contidos nos telefones. A propagação de um pequeno aparelho com dimensões entre 11,5 cm de altura por 6,21 cm de largura alterou em menos de uma década a produção, a apresentação e produção do jornalismo (SILVEIRA, 2017).

As mudanças drásticas que tais acoplamentos promovem no mundo exigem reestruturações de quadros teóricos para explicar novas controvérsias. As tecnologias da informação, ou de inscrição, não são apenas instrumentos para a produção de sentido, mas antes o horizonte no qual o sentido pode ser construído (FELINTO, 2006). Ao colocar a materialidade como questão central na comunicação, no lugar do 'espírito', do 'significado', ou de 'forças ocultas', se busca uma noção de verdade relacional que conecta o social e o material.

Seguir a materialidade é uma maneira de rastrear o fenômeno comunicacional, ou jornalístico, a partir das diferentes associações que o formam. Um exemplo contido na obra "Reagregando o Social" (2012) pode ilustrar como seguir os objetos técnicos é uma forma cientificamente eficaz para compreender as relações que se formam dentro da comunicação. Considere Alice, uma cidadã francesa votando nas eleições gerais de seu país. Ela pega um jornal sob sua mesa de café e confere notícias sobre o pleito para decidir em quem irá votar. Momentos depois, ela se desloca para um local de votação. Pega uma cédula e se dirige para uma das cabines, entregando-a para um dos mesários que a coloca em uma urna eleitoral. Horas mais tarde, liga a sua televisão e confere os resultados das eleições. Um infográfico projetado na tela de sua televisão contabiliza quantos votos cada coligação recebeu e informa Alice sobre o resultado do pleito.

[15] O lançamento do Iphone 3GS e seu sucesso de vendas é considerado um marco na popularização dos *smartphones* por ter sido o primeiro modelo a agregar a tecnologia touchscreen, com câmera traseira, conexão 3g e um processador de 600 MHz, o mais rápido disponível no ano de seu lançamento.

Para entender a relação da garota Alice com a 'França como um todo' é preciso que uma materialidade nos permita rastrear a interação entre o local (eleitor) e o global (país). Essa interação é justamente a comunicação, enquanto o rastreador são os objetos técnicos que medeiam essa relação: jornal, cabine de votação, cédula, infográfico e televisão. Neste caso, seguir essas diferentes materialidades é deslocar as conexões que são estabelecidas entre diferentes locais, até vislumbrar algo que pode ser chamado de 'contexto'.

De forma semelhante à contabilidade de uma eleição, a produção de jornalismo sobre eleições também conta com dispositivos que nos permitem rastrear o processo. A produção de notícias automatizada em uma cobertura política não é diferente. Há uma interação material entre votos em urnas eletrônicas, banco de dados do Tribunal Superior Eleitoral, modelos de textos jornalísticos, algoritmos de Geração de Linguagem Natural e portais online de veículos. A materialidade do jornalismo automatizado é justamente o que permite estudar como atores humanos e tecnológicos compõem o fenômeno comunicacional. Para isso, é necessário identificar quais são os dispositivos e profissionais em jogo, quais são suas interações e o que eles produzem sinteticamente.

Portanto, dentro da perspectiva epistemológica neomaterialista da comunicação associal (LEMOS, 2020), o objetivo é fazer a descrição dessa materialidade. Inicialmente se faz necessário identificar o *Modo* no qual se insere a questão do jornalismo automatizado. Ao mesmo tempo, os objetos técnicos empregados no projeto precisam ser *inventariados*. Nesta etapa, o objetivo é compreender os campos de tensões que envolvem o projeto do G1. Os modos de existência dos diferentes atores precisam ser descritos para definir qual a controvérsia existente dentro desta iniciativa de jornalismo automatizado. Esses questionamentos foram organizados em seis unidades de análise que servirão de norte para a aplicação da metodologia nos dados coletados durante a pesquisa. São eles:

1. **Como foram decididas quais informações entraram no texto?**
2. **Quais são os tipos de profissionais envolvidos?**
3. **Como as tarefas foram divididas entre os diferentes profissionais?**
4. **Qual é a funcionalidade de cada ferramenta digital empregada por cada tipo de profissional?**
5. **Como o trabalho com a inteligência artificial transfere princípios e autoridade entre algoritmos e jornalistas?**

A fim de responder aos questionamentos das unidades de análise, dois métodos são empregados dentro do guarda-chuva metodológico neomaterialista, para responder diferentes aspectos do problema de pesquisa. Primeiramente, a análise de similitude (MARCHAND-RATINAUD, 2012) será utilizada para medir o grau de similaridade lexical e semântica dos textos produzidos no projeto do G1. Originalmente, a metodologia proposta por Marchand e Rautinaud empregam o *software* IRAMUTEQ, porém, a análise do conteúdo com módulos em Python dispõem 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 fazem necessárias: **a coleta** e **a identificação** de padrões textuais.

Para realizar a coleta de uma quantidade grande de textos - 5.568 textos foram publicados, mas nem todos serão obtidos - será usado o método de raspagem automatizado com *Python*. Mais especificamente, o módulo *Beaultiful Soup* vai ser empregado, tendo em vista sua finalidade para extrair dados de páginas HTML[16]. Em sequência, para identificar padrões de similaridade entre as estruturas textuais será empregado, também em Python, o módulo de expressões regulares (*regex*[17]). Essa função permite medir quantitativamente a recorrência de expressões, ordens de palavras e extensão do texto. Ambos os scripts empregados para análise de similitude podem ser encontrados no "**ANEXO III**". Os textos coletados, por sua vez, serão salvos em um repositório online.

É importante atentar para o fato de que todos os textos obedecem ao formato textual do lide jornalístico. Por esse motivo, a análise de similitude será complementada pela análise da cobertura jornalística (SILVA-MAIA, 2011), a fim de reconhecer as marcas de apuração fixadas nas notícias. A intenção é não só observar as semelhanças e diferenças lexicais entre os textos, mas traçar a origem das informações que ali se encontram. Com uma análise quantitativa da recorrência de palavras, de natureza lexical, é possível observar a capacidade da Geração de Linguagem Natural[18] de produzir textos com diferentes informações - mesmo que as reportagens sigam um número limitado de modelos (*template*) para redação automatizada. De forma complementar, a análise da cobertura jornalística nos permitirá traçar as origens e das informações presentes nos textos.

---

[16] Ver "Beautiful Soup Documentation". Dispoinível em: https://tedboy.github.io/bs4_doc/. Acesso feito em 22 de maio de 2022.
[17]Ver "Regular expression operations". Disponível em: https://docs.python.org/3/library/re.html. Acesso feito em 22 de maio de 2022.
[18] No capítulo "**3.5. A GERAÇÃO DE LINGUAGEM NATURAL"** o conceito é discutido e fundamentado em detalhes.

O procedimento se inicia já na coleta dos textos feitos de páginas *online* para arquivos *offline*, a partir da utilização do módulo *Beautiful Soup* em *Python.* As informações que foram escolhidas para a raspagem são, respectivamente, o título, o subtítulo (*soutien*) e o corpo do texto (lide, sublide e complemento). Para que as informações fiquem organizadas para análise, as três informações mencionadas acima serão salvas em listas catalogadas dentro de um dicionário que leva o nome da cidade como chave[19]. Em seguida, a recorrência de padrões nos textos e marcas de apuração serão investigadas segundo as duas metodologias de análise textual que foram mencionadas.

Pela natureza do projeto do G1, acredita-se que as milhares de notícias que compõem o *corpus* são jornalísticas e, portanto, seguem uma estrutura de pirâmide invertida. Portanto, elas seguem características formais da escrita jornalística. Por esse motivo, a análise de similitude precisa ser complementada por uma investigação que observe as marcas de apuração do jornalismo estampadas no texto, como por exemplo a organização do lide e dos parágrafos complementares. Para isso, a análise da cobertura jornalística (SILVA e MAIA, 2011. p 32) será utilizada com o objetivo de "investigar como um determinado veículo estrutura a cobertura de acontecimentos específicos, verificando as marcas das técnicas e estratégias de apuração e composição da matéria jornalística". Este levantamento buscará identificar aspectos dos três níveis do texto jornalístico:

1º nível – Marcas da apuração

2º nível – Marcas da composição do produto

3º nível – Aspectos do contexto de produção

A primeira categoria de análise recai exclusivamente sobre a matéria jornalística – tomada de forma isolada –, explorando indícios do método de apuração e da estratégia de cobertura. Já o segundo nível oferece uma visão um pouco mais aberta do objeto, agora focando na ordem em que os elementos surgem no produto. O terceiro nível atua como uma grande angular – não capta detalhes, mas oferece um plano geral do objeto, captando aspectos da dimensão organizacional e do contexto sócio-histórico cultural em que se insere a produção jornalística.

[19] Listas, dicionários, chaves e valores são parte do jargão de estruturação e análise de dados em *Python*. Elas serão explicadas em detalhe no terceiro capítulo.

Devido ao aspecto qualitativo da metodologia Silva e Maia (2011), não será possível submeter milhares de notícias à análise de cobertura jornalística[20]. Para isso serão escolhidas randomicamente notícias para a investigação. No 1º nível, as **marcas da apuração** serão mostradas a partir de uma lista dos dados contidos em textos com a indicação de sua origem. É uma hipótese desta pesquisa que todos os dados provieram de bases online, resta identificar quais são essas bases e quais partes delas estão presentes no produto final. Já no segundo nível, as **marcas da composição do produto** serão observadas segundo a ordem em que os dados entraram nos textos e em qual formato. Por último, os **aspectos do contexto de produção** serão apresentados de maneira transversal ao longo de todo o capítulo **3**, a partir de correlações que podem ser feitas entre as marcas dos textos e episódios conjunturais das eleições de 2020.

Primeiramente, a pesquisa busca investigar os preparativos da tecnologia empregada na cobertura do G1. A pesquisa trabalha com a hipótese de que meses de trabalho prévio de uma equipe multidisciplinar foram necessários para adequar a tecnologia às necessidades do veículo. Para estudar essa fase do processo, serão conduzidas entrevistas em profundidade com dois jornalistas e dois programadores que estiveram envolvidos no projeto do início ao fim. A fim de aproximar os relatos e compará-los, as entrevistas seguirão 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). Para dar espaço aos relatos dos jornalistas e entrevistados, sem criar complicações de deslocamento e agenda, as entrevistas serão feitas na modalidade on-line em encontros síncronos. A entrevista em profundidade (DUARTE, 2005) é aplicada com o objetivo de descrever a experiência dos profissionais que trabalharam no projeto. Os dois métodos, a análise de similitude e as entrevistas em profundidade, buscam dissecar a rede sociotécnica, listar os diferentes atores humanos e não-humanos, assim como mapear os agenciamentos que se formam entre eles em suas diferentes formas de mediação.

O método da entrevista em profundidade será empregada com três propósitos. Primeiro, como parte da metodologia neomaterialista da comunicação associal. Segundo, como forma de diferenciar as contribuições feitas por jornalistas e programadores. Terceiro, a fim de fazer o inventário das ferramentas digitais que foram usadas pelos diferentes profissionais ao longo de todo o projeto, com breves descrições de suas aplicações. O emprego da metodologia da entrevista busca sanar as três intenções citadas acima, tendo em

[20] O caráter quantitativo-qualitativo desse método dificulta a automação para milhares de textos.

vista sua função de“explicar a produção da notícia em um veículo de comunicação e identificar motivações para uso de determinado serviço” (DUARTE, 2005. p 63). O caráter exploratório deste método será essencial para complementar uma “descrição de processos complexos” no qual os funcionários do G1 estiveram envolvidos. As perguntas destinadas aos entrevistados podem ser conferidas no “**ANEXO I**”.

3. JORNALISMO E AUTOMAÇÃO

Da última década do século XX para as primeiras décadas do século XXI, a sociedade vive uma crescente onda de digitalização, caracterizada pela popularização das TICs, aumento global do acesso à *word wide web* e crescimento da relevância do *software* em atividades de produção tanto culturais, quanto econômicas. O *software* “toma o comando” na sociedade contemporânea e se transforma na era da informação em um elemento comum entre diversas atividades profissionais (MANOVICH, 2014, p. 1). Com o jornalismo não tem sido diferente.

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. As novas ferramentas dão à prática jornalística suas próprias características e implicam em uma remodelação quase integral do fazer jornalístico. Tal fenômeno pode ser compreendido, segundo Salaverría (2015), nas seguintes categorias: (1) contexto histórico e de mercado, (2) processo de inovação, (3) mudanças na rotina produtiva, (4) novos desafios profissionais para os jornalistas e (5) popularização do conteúdo gerado por usuários.

Ao considerar essas cinco categorias é factível compreender que não houve apenas mudanças no conteúdo em si, mas na dinâmica de apurar e redigir notícias. Embora seja possível analisar por uma perspectiva de *newsmaking* como a rotina produtiva nas redações foi alterada, também é possível fazer um recorte relacional sobre como a tecnologia age como mediadora do trabalho jornalístico. É certo que as tecnologias digitais mudaram rotinas dentro das organizações jornalísticas, mas como elas impactam a relação do jornalista com o seu trabalho e na sua interação com o mundo? (SALAVERRÍA, 2015). A atuação do indivíduo-jornalista no meio digital foi afetada por novas mediações, novas competências e formas de agregar valor à notícia. Como Bardoel-Deuze (2001) apontam, as competências centrais dos profissionais foram alteradas.

> Interatividade, personalização de conteúdo, hipertextualidade e multimidialidade estão redefinindo o jornalismo a partir do uso da Internet, mas seus componentes têm implicações para o jornalismo como um todo. Tal impacto pode ser contextualizado quando se considera o jornalismo online como o catalisador para a mudança na profissão como um todo; enquanto que o jornalista costumava depender de uma organização de mídia para oferecer emprego em tempo integral e, portanto, segurança no emprego, o novo jornalismo é um trabalho com múltiplas habilidades, formatos e padrões de emprego que funcionalmente diferenciam a profissão de uma maneira mais holística (BARDOEL-DEUZE, 2001, p. 7).

A transição do jornalista para o meio digital instigou diversas mudanças no conjunto de habilidades deste profissional em função dos novos formatos para a notícia e da característica da comunicação em rede. Ao mesmo tempo, esse cenário de transição criou novas possibilidades de explorar formatos narrativos para o jornalismo na medida que ferramentas digitais eram dominadas por profissionais de imprensa. 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 (BARBOZA, 2003; QUADROS, 2006; MIELNICZUK, 2004; MACHADO, 2005).

Entre as diversas tendências propiciadas pela ida do jornalismo para o ambiente digital, Santos (2016) lista alguma dessas: o jornalismo em rede (*networked journalism*); o conteúdo gerado por usuários (*crowdsourcing and user-generated content*); a mineração de dados, análise de dados, visualização de dados e mapeamento (*data mining, data analysis, data visualization and mapping*); o jornalismo visual (*visual journalism*); o jornalismo de ponto de vista (*point of view journalism*) e o jornalismo automatizado (*automated journalism*). As ramificações que a produção jornalística sofre na sociedade em rede são consequências da introdução de novas ferramentas e competências que os profissionais devem adquirir.

Essa necessidade por novas habilidades no trabalho jornalístico, alinhada a destreza e familiaridade com instrumentos digitais exige dos profissionais uma *inteligência híbrida* (SANTOS, 2018). Um dos argumentos centrais dessa pesquisa é que processos cada vez mais técnicos (digitais) requerem *profissionais híbridos.* O jornalista bem-sucedido em fazer sentido do ambiente virtual é aquele que expande o seu arsenal de métodos, jargões e conceitos. No processo de conquistar o mundo computacional, o profissional deve ele mesmo ser modificado por ele.

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 de informações, passa a encarar na contemporaneidade 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 a mais importante. O novo desafio vai ser enfrentado, segundo as características do jornalismo digital, com o uso de *softwares*, mas dessa vez as competências exigidas para operar esses *softwares* e dar sentido a um volume avassalador de informação, vão provocar um novo desdobramento no jornalismo.

> Olhar para a literatura é perceber um breve indicativo acerca das práticas do jornalismo na contemporaneidade, em que há, ao menos, três conjunturas: aquela assistida por computador – jornalismo digital –, a baseada em dados que são consultados, recuperados e extraídos – jornalismo de dados e de automação (IOSCOTE, 2021, p. 179).

O jornalismo de dados se manifesta nesse 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; TRÄSEL, 2014). O jornalista enquanto investigador, guiado por princípios tradicionais de vigília das instituições (*watchdog*), soma ao seu arsenal valores *hackers* da cibercultura e o emprego de *softwares* para diminuir o ruído de bancos de dados, a fim de dar sentido às informações. A principal barreira para o exercício do jornalismo de dados pela vasta massa de profissionais da imprensa são os conhecimentos das ciências de dados, como linguagens de programação e o tratamento de bancos de dados, que não integram o rol tradicional de habilidades do jornalista.

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. 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. Já para apresentar essas informações, o jornalismo de dados vai usar a multimidialidade característica do meio em recursos como infográficos, GIFs, texto, áudio e vídeos.

O jornalismo de dados enxerga na *web* uma fonte de bancos de dados estruturados, ou não, contendo histórias inéditas com informações que recebem um respaldo quantitativo. Como Träsel (2014) descreve, esse profissional nativo do ciberespaço pode se deparar com arquivos públicos, ou com dados não estruturados que, se organizados, permitem a ação de 'entrevistar planilhas'. Desta forma, o jornalismo digital se desdobra em um novo fazer que mescla recursos próprios da computação e da ciência da informação, mas com o potencial de reportar e interpretar a própria web (CODDINGTON, 2015).

Para alguns autores, a análise feita a partir da navegação em plataformas digitais pode esclarecer 'dinâmicas de sentimentos' (DIAKAPOULOS-SHAMMA, 2010). Por essa perspectiva, a investigação de comportamentos online para a elaboração de reportagens busca explicar processos sociais complexos. Essa tendência de pesquisa e valorização da quantificação do comportamento on-line para fins de monitoramento se insere dentro de um fenômeno mais amplo denominado de *datificação* (MAYER-SCHOENBERGER-CUKIER, 2013).

Segundo Van Dicjk (p. 39, 2017), a popularização da datificação denuncia uma ideologia de "crença generalizada na quantificação (...) por meio de tecnologias de mídia on-line". O fenômeno da 'datificação' carrega consigo características ideológicas que afetam tanto o mercado, como formas de produção de conhecimento (ZUBOFF, 2019). O jornalismo, junto de outros campos como o marketing, a política e até mesmo a própria academia, estariam caminhando rumo a um estreitamento discursivo-ideológico com as chamadas *Big Techs* - grandes empresas de tecnologia. Essa aproximação pode explicar como o jornalismo de dados agrega valor à notícia, ao passo que permite compreender o advento de novas formas de produção de notícias.

A centralidade do *software* para o exercício da atividade jornalística mencionada por Manovich (2014) continua seguindo o *discurso da velocidade* apontado por Örnebring (2010) na busca de tornar a produção de notícias mais rápida e competitiva. O caminho encontrado por algumas empresas midiáticas de grande porte parece ser 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 (DORR, 2015), uma recém-surgida maneira de produzir notícias desponta com a característica da 'autonomia' para a criação de conteúdo jornalístico.

3.1. AUTOMAÇÃO & INTELIGÊNCIA ARTIFICIAL

De acordo com a Teoria da Computação (HOPCROFT et al, 2002), a automação deriva do emprego de um autômato, que por sua vez é um sistema composto de: (1) estado, (2) entrada e (3) memória. Um interruptor pressionado, por exemplo, sai de um estado de desligado para o de ligado por meio de uma entrada, o resultado desse processo fica memorizado. Procedimentos mais complexos podem ser processados por um autômato, como análises textuais, ou cálculos numéricos, por meio de relações lógicas. O emprego de autômatos alinhado ao estudo das 'gramáticas formais' feito por Noam Chomsky na década de 1950 permitiu o advento da comunicação com máquinas por meio do *software*.

Segundo o engenheiro mecânico Mikel P. Groover (1980, p. 61, tradução nossa[21]) a automação é definida como "a tecnologia pela qual um processo, ou procedimento, é realizado sem a assistência humana, sendo implementado por um programa de instruções combinado com um sistema de controles". Ou seja, a automação enquanto processo em si carrega consigo três características fundamentais: (1) uma fonte de força para performar alguma ação, (2) programa de instruções *e* (3) sistema de controle de *feedbacks*.

Quando se tratam de tecnologias digitais, a fonte da força de energia é em sua totalidade elétrica. Logo, interessa compreender a dinâmica entre (2) e (3). As ações performadas por um processo automatizado são provindas de um programa de instruções, que define quais partes de um produto serão combinadas dentro de um ciclo de trabalho. A ação de combinar cada parte é dividida entre as chamadas 'etapas de processamento', que juntas compõem os ciclos de trabalho. O sistema de controles (3) por sua vez executa o programa de instruções que dá início ao processo, ou procedimento automatizado, ao passo que recebe *feedbacks* do *output*. Em suma (3) funciona como uma via de mão dupla que executa as orientações estabelecidas pelo programa de instruções, simultaneamente colhendo 'comentários' sobre o procedimento instaurado.

Os processos de automação alcançam um novo grau de sofisticação com o surgimento da linguagem de programação (1955), memória RAM (1984) e microcomputação (1971). Em consonância com os desenvolvimentos para a manufatura, os avanços tecnológicos também transbordam para além dos muros das fábricas e para dentro de lojas, escritórios, universidades, redações e, por fim, lares. A introdução dos computadores pessoais pela Apple Computer (1978) com o lançamento do Apple I propôs que esses sofisticados aparelhos

[21] Automation is the technology by which a process or procedure is accomplished without human assistance. It is implemented using a program of instructions combined with a control system that executes the instruction.

adentrassem o ambiente doméstico, não se restringindo mais a instalações militares, universidades e empresas. Por meio do computador pessoal, cresce o acesso aos *softwares*, que por sua vez ampliam as vivências em contato direto com a lógica da automação.

Enquanto um *software* é programado para executar tarefas em uma troca entre instruções e *feedbacks*, uma inteligência artificial é programada para aprender a executar uma tarefa (IBM, 2020). Em artigo seminal *"Computing Machinery and Intelligence"*, Alex Turing, conhecido como pai da ciência da computação, propõe o "Teste de Turing" para responder à pergunta: Computadores podem pensar? A concepção do teste avalia se um computador poderia se passar por uma pessoa em uma interação às cegas. Mais tarde, Stuart Russell e Peter Norvig (2004) definem Inteligência Artificial como um campo amplo que engloba lógica, matemática do contínuo, probabilidade, percepção e eletrônica. Ambos os autores resumem IA como a atuação de um agente inteligente artificialmente criado.

> Definimos a IA como o estudo de agentes que recebem percepções do ambiente e executam ações. Cada agente implementa uma função que mapeia sequências de percepções em ações, e abordaremos diferentes maneiras de representar essas funções, tais como sistemas de produção, agentes reativos, planejadores condicionais em tempo real, redes neurais e sistemas de teoria de decisão (RUSSEL e NORVING, 1994, p. 7).

Os autores montaram um esquema para reunir as diferentes funções da IA imaginadas desde Turing até os seus contemporâneos. As quatro categorias encontradas são: (1) Pensando como um humano, (2) Pensando Racionalmente, (3) Agindo como um humano e (4) Agindo racionalmente. Segundo esta esquematização, o Teste de Turing, por exemplo, se enquadraria na primeira categoria. Se um computador é capaz de convencer um humano que pensa como tal, sem que o sujeito saiba que se trata de uma máquina, então o computador 'pensa'. Um *software* de Inteligência Artificial que busca executar atividades comuns a um repórter, ou editor, por exemplo, terá que 'pensar' e 'agir', variando entre a racionalidade e o mimetismo de atitudes humanas.

Aprofundar essa discussão significa inevitavelmente enfrentar questões filosóficas e até existenciais sobre a natureza do pensar e agir no mundo. O que é inteligência? O que é consciência? O que é originalidade? Essas são algumas das perguntas que costumam aparecer logo de início quando comparações são traçadas entre seres humanos e computadores. Não cabe a esta pesquisa adentrar nessa discussão. Obras inteiras já foram escritas tentando responder cada um desses questionamentos, certamente sem conseguirem dar um desfecho definitivo. Mas para fins de elucidação, cabe aqui fazer uma pequena diferenciação.

Em *Homo Deus* o filósofo Harari (2016) define inteligência como **a capacidade de resolver problemas** - sejam quais forem esses problemas, ou quais forem os 'resolvedores'. Inteligência e consciência são, portanto, qualidades que podem estar entrelaçadas, ou podem ser independentes. O desenvolvimento de computadores e da inteligência artificial provocam um "desacoplamento" da inteligência em relação à consciência (HARARI, 2016, p.309).

Permanece a indagação do que é uma inteligência artificial. Para responder essa pergunta, vale retomar o episódio histórico no qual o conceito surgiu, quando o 'desacoplamento' deu os seus primeiros passos. Em 1956, um grupo de matemáticos e outros pesquisadores da Faculdade de *Dartmouth* organizaram uma conferência para discutir a possibilidade de máquinas 'aprenderem' (RUSSELL-NORVIG, 2004). Hoje, a célebre *Conferência de Dartmouth* é tida como o primeiro evento em que o termo "inteligência artificial" foi cunhado. Os participantes eram em sua maioria matemáticos e estatísticos, mas também estavam presentes colegas de outras disciplinas no esforço de criar uma agenda ampla de investigação ao redor deste conceito.

Alguns dos intelectuais mais proeminentes ali presentes como John McCarthy, Marvin Minsky e Herbert Simon, eram matemáticos ativos no desenvolvimento da teoria dos jogos. Desta forma, a inteligência artificial surge com a ideia de que 'raciocinar' e 'jogar' são atitudes similares, ambas influenciadas pela matemática e pela estatística. A ambição dos pesquisadores tomou forma ao longo das décadas subsequentes na forma de programas que desafiavam a destreza humana em jogos como xadrez, cartas, *Go*, dama e outros, mas na segunda metade do século XIX o desenvolvimento de IAs estava limitado pela capacidade de processamento dos computadores. O *hardware* não havia amadurecido o bastante ainda para concretizar a visão dos conferencistas de *Dartmouth*.

Segundo Stuart Russell e Peter Norvig (2004), os avanços significativos em inteligência artificial começam a acontecer no final da década de 1990 e início dos anos 2000. O aumento na capacidade de processamento dos computadores, a disseminação de desenvolvedores de *software* e a formação de projetos interdisciplinares de pesquisa fazem o campo decolar. As aplicações de IA deixam de focar em jogos e passam a extrapolar para toda e qualquer área. Com o progresso dessas tecnologias surge uma possibilidade. A automação que já era parte da economia industrial pode se servir de máquinas cada vez mais competentes em 'raciocinar'. Se no princípio as máquinas das linhas de produção realizavam um trabalho essencialmente físico (GROOVER, 1980), agora elas podem realizar um esforço cada vez mais cognitivo (RUSSELL-NORVIG, 2004).

A história da inteligência artificial continua sendo escrita, e a visão dos pesquisadores de *Dartmouth* permanece como uma inspiração inicial, que pode ser aplicada em jogos, no mercado financeiro, em governos e, por que não, em redações. Todo esforço comunicativo pode ser influenciado pela inteligência artificial, dado o seu caráter interdisciplinar e multisetorial. Para os mais entusiastas, as indagações de Alex Turing em 1950 funcionaram como uma profecia, pois agora máquinas pensam e se comunicam.

3.2. O CONCEITO DE JORNALISMO AUTOMATIZADO

Para a automação da redação de notícias funcionar, é preciso mobilizar processamento computacional, recursos de análise de dados, algoritmos digitais 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. Muitas dessas empreitadas consistem no uso de aparelhos e aplicações digitais que na prática já faziam parte da rotina de jornalistas desde 1990. As mudanças na forma de apuração e publicação do jornalismo envolvem tecnologias que saíram de empresas de *Big Techs* para dentro de redações. A partir dessa contextualização, resta compreender o que é conceitualmente a automação de produção de conteúdo nas redações.

O jornalismo automatizado é definido por Carlson (2016, p. 417, tradução nossa[22]) 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. Por “intervenção humana limitada” não fica claro se jornalistas teriam a capacidade de revisar os textos, por exemplo, ou complementá-los durante o processo de publicação. Uma segunda possibilidade que fica fora da definição é o uso de aprendizagem de máquina por parte do algoritmo para aperfeiçoar a escrita. A definição de Carlson põem enfoque na "conversão de dados”, subentendida como o ato de redação de “narrativas” em formatos jornalísticos diversos (nota, reportagem, coluna e etc).

Por outro lado, a descrição 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’.

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

> 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 (tradução nossa, p. 14).[23]

O ponto positivo da definição de Graefe é ceder ao processo de automação no jornalismo quatro etapas que vão desde a reunião dos dados, até a publicação do texto, porém, o conceito também não abre espaço para a colaboração entre atores humanos e não-humanos durante o processo de redação e publicação. Assim como Carlson, Graefe não descreve como a intervenção humana pode ocorrer na fase de programação, ou se ciclos de aprendizagem de máquina fazem parte do processo.

De forma similar, o conceito de *robô jornalista* também coloca destaque na autonomia de algoritmos para a escrita de textos jornalísticos. "Um processo em que escritores robôs independentemente examinam informação e produzem conteúdo noticioso de acordo com algoritmos programados por humanos" (KIM-KIM, 2018, p.342, tradução nossa[24]). *Robô jornalista* ou *repórter* é um conceito amplamente cunhado por teóricos e pesquisadores (CLERWALL, 2014; LATAR, 2014, 2015; LEVY, 2012; RUTKIN, 2014), mas criticado como uma conceituação banal que apela para a imagem de autômatos antropomórficos escrevendo textos jornalísticos (LINDEN, 2017). Outro problema da definição é igualar 'robô' e 'algoritmo' sem justificar a aproximação ou diferenciá-los tecnicamente. Novamente, o papel do humano se limita às escolhas iniciais de programação e o conceito não abre espaço para interação de profissionais da imprensa e algoritmos, ou um ciclo de *feedback* entre os dois ao longo do processo.

Entender o que é a automação no jornalismo passa por dimensionar o mesmo grau de não intervenção humana descrita por Groover (1980), válido para qualquer atividade laboral, mas levando em conta que o jornalismo é mais do que o ato de escrever e publicar. Segundo Diakopoulos (2019. p. 97) o que pode ser chamado de "produção de conteúdo automatizado" deve ser na prática "orquestrado" com repórteres. Para Carlson (2016, p. 226), foram os "avanços na inteligência artificial que tiraram escritores não-humanos da teoria para a prática, tendo o jornalismo como principal indústria afetada", porém, alguns autores se esquecem de dimensionar o fator humano dentro do processo, por um efeito retórico que chama atenção

[23] 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.

[24] A process in which robot writers independently examine information and produce news content according to algorithms programmed by humans.

para uma realidade quase distópica. Na via contrária, estudos de caso de iniciativas reais com quadros teóricos mais próximos da Teoria Ator-Rede, a Teoria da *Assemblage* e a Comunicação Homem-Máquina sugerem uma amálgama dos papéis entre ambos os agentes.

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 informação estatística e um conjunto de *stock phrases*, sem a interferência de humanos jornalistas" (p. 648, tradução nossa)[25]. É interessante notar como o autor tira a atenção do processo (jornalismo automatizado) e desloca para o produto: a notícia. Se o fator humano é totalmente excluído do fazer jornalístico, ou ao menos do interesse, então faz sentido olhar somente para o resultado final. A definição de Van Dalen foi a mais antiga encontrada nesta revisão de literatura e, talvez pela escassez de estudos da época, carece de pontuações sobre o que é 'automaticamente gerar notícias'.

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 não contrapõe a ideia de jornalismo automatizado proposta por Carlson (2016), mas prefere dar foco em 'notícia' por concluir que somente esse tipo de conteúdo jornalístico pode ser automatizado, excluindo gêneros opinativos e reportagens. Isso ocorre, pois, a reportagem se apoia na subjetividade do repórter, exigindo uma postura analítica e crítica da realidade que é inviável à automação. "Já 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). O professor Nilson Lage (2021) também aborda a técnica jornalística em sua dimensão industrial, que se liga diretamente à apresentação das informações no lide.

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[26], com um maior grau de complexidade, seriam

[25] Algorithms can now automatically generate news stories on the basis of statistical information and a set of stock phrases, without interference from human journalists.
[26] O termo original cunhado por Caswell e Dorr (2018) é "event-driven narratives".

necessários bancos de dados igualmente mais complexos, com tipos de informações que hoje não são usuais. Ou seja, se for do interesse de grupos de mídia estender a automação da produção de notícias para a produção de reportagens, a tecnologia corrente já permite. Para tal, o jornalismo teria que se defrontar com as chamadas Metodologias de Codificação, recursos que convertem diferentes tipos de dados entre si[27].

Já Linden (2016) chama atenção em sua conceituação de jornalismo automatizado para as características do processo, em detrimento das características do produto. O autor elenca dois estágios de *input* e *output* que são necessários para a redação automatizada ocorrer, fazendo uma aproximação entre a Teoria da Computação e as Teorias da Comunicação. Um terceiro ponto é destacado por Linden como fundamental para a área: a qualidade final dos textos. Um lide produzido por um algoritmo deve ser indistinguível de um lide produzido por um humano.

> Processo automatizado guiado por algoritmo que utiliza conjuntos estruturados de dados sobre esportes, mercado imobiliário ou mercado de ações como *input* para para criar notícias como *output*. Plataformas avançadas de Geração de Linguagem Natural transformam os dados em textos indistinguíveis dos que um humano poderia escrever que é indistinguível do que um jornalista humano poderia produzir (LINDER, 2016, p. 125)[28].

Em concordância com Linden, Caswell e Nicholas Dorr (2018) destacam que a qualidade dos textos produzidos pelos *softwares* de Geração de Linguagem Natural é impressionante. O potencial disruptivo desta tecnologia reside na capacidade de combinar tanto a qualidade dos textos, com uma quantidade sobre-humana de publicações, mas ao mesmo tempo que a quantidade de publicações é uma vantagem, a incapacidade dos editores humanos de lidar com esse volume pode ser um contratempo.

> Exemplos iniciais de uso da Geração de Linguagem Natural para automatizar o jornalismo são em sua maioria restritas a textos relativamente curtos de assuntos limitados, mas com resultados impressionantes em termos de qualidade e quantidade. O texto produzido é geralmente indistinguível de textos escritos por humanos e o número de documentos textuais gerados excede substancialmente a capacidade de editores processá-los (CASWELL-DORR, 2018, p. 2, tradução nossa)[29].

---

[27] Ver "Overview of Encoding Methodologies". Disponível em: https://www.datacamp.com/tutorial/encoding-methodologies. Acesso feito em 28 de maio de 2022.

[28] "(...)algorithm-driven automated processes using structured sets of data from sports, real estate and stock markets as input to create news items as output. Platforms of advanced natural generation language transform data into text indistinguishable from what a human person would write (p.125)."

[29] "Early examples of the use of NLG technology to automate journalism are mostly confined to relatively short texts in limited domains, but are nonetheless impressive in terms of both quality and quantity. The text produced is generally indistinguishable from text written by human writers and the number of text documents generated substantially exceeds what is possible from manual editorial processes."

Os pesquisadores Ufarte-Ruiz (2020) detalham que a qualidade dos textos produzidos pelos algoritmos se assenta em valores como a neutralidade, veracidade, boa sintaxe, coerência e concisão. Entretanto, os autores sugerem que essa qualidade implica em um *trade-off* que abre mão de fatores como originalidade, variedade, estilo e ritmicidade. Embora os textos careçam desses valores, o espaço para melhora ainda existe e já foi verificado ao longo dos anos de aplicação dentro do jornalismo.

Apesar da qualidade dos textos feitos de forma automatizada, é possível identificar algumas diferenças entre as notícias escritas por seus parentes de carne e osso. Em uma análise de conteúdo comparativa feita entre notícias produzidas por máquinas e humanos, os pesquisadores C. Tandoc *et al* (2022) apontaram as diferenças que existem entre ambas. Primeiro, as notícias escritas por máquinas não trazem aspas de entrevistados humanos, se apoiando 99.4% das vezes em fontes documentais. Os assuntos das notícias escritas por máquinas são sempre específicos (97.9%), enquanto humanos conseguem redigir sobre assuntos gerais (52.7%). Quanto ao tamanho, os textos feitos por humanos são significativamente mais longos. Por último, as notícias feitas por humanos trazem mais interpretação na forma de opinião e análise.

No entanto, existem pontos em comum entre os produtos das duas origens. Quanto ao formato, tanto máquinas, quanto humanos escrevem mais de 95% das vezes no formato de pirâmide invertida. Ambos os grupos de textos trazem valores-notícia como a novidade e assuntos recentes. Por fim, os textos são semelhantes no tanto de informações que trazem sobre o contexto no qual a notícia se insere. Uma confirmação da ideia de uma descrição de cenário, ou informações de fundo (*background information),* pode ser organizada em um *template* e redigida por uma máquina de forma eficaz.

Mesmo com a qualidade dos textos sendo um fator determinante para a emergência do jornalismo automatizado enquanto prática legítima, uma abordagem materialista exige que a definição seja pautada pelo processo em si. Diferentemente de Linden, Nicholas Dorr (2015) inicia suas pesquisas sobre a área cunhando o conceito de *jornalismo algorítmico* e o atrelando especificamente à produção de notícias automatizadas com o emprego de *softwares* de Geração de Linguagem Natural. A definição de Dorr detalha 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 (tradução nossa, p. 702)[30].

As três etapas mencionadas por Dorr são as mesmas presentes na definição de Latzer et al (2016) sobre o processo algorítmico, mencionada no capítulo "Algoritmos e Jornalismo". Também é notável como a conceituação de Dorr 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*.

Diferente dos conceitos anteriores sobre jornalismo automatizado, que incorrem na afirmação de 'intervenção limitada' ou na 'não intervenção', Dorr se distancia do determinismo tecnológico e se aproxima da TAR na tentativa de envolver atores humanos e não-humanos no processo. Por outro lado, o termo *jornalismo algorítmico* não é tão recorrente na literatura como *jornalismo automatizado* e *robô repórter* justamente por não fazer uma conexão direta entre jornalismo e automação.

Em uma abordagem similarmente mais moderada, Torrijos (2021) propõe o termo *jornalismo semi-automatizado*. A razão de incluir o prefixo 'semi' é justamente destacar a mútua relação entre atores humanos e não humanos no processo. O autor também aponta que a prática jornalística é composta por rotinas produtivas e que toda rotina implica em repetições. Essas repetições são precisamente o objeto da automatização. Para que algo seja automatizável, é necessário que a tarefa seja específica e repetitiva. Existe nessa relação um *input*, o comportamento produtivo do jornalista, para a identificação de padrões, 'tarefas repetitivas', como meio para se chegar a notícia (*output*).

---

[30] "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 (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."

> Automação de notícias implica, por um lado, em identificar e selecionar uma série de rotinas e tarefas repetitivas que ocorrem diariamente em redações, para depois programá-las em algoritmos e *bots*. Por outro lado, se deve criar e aproveitar conjuntos de dados estruturados por meio da programação para facilitar a produção instantânea de textos para uma cobertura noticiosa específica (TORRIJOS, 2021, p. 125, tradução nossa)[31].

Há um segundo ponto interessante na definição do *jornalismo semi-automatizado* quanto à relação entre forma e conteúdo. O assunto de uma cobertura automatizada deve sempre ser específico, igualmente a natureza da tarefa repetitiva. A redação de um lide é uma tarefa específica e repetitiva, tal qual as informações que estão presentes no texto no caso de uma cobertura esportiva, econômica ou política. Como é visto no recorte unicamente eleitoral do G1, ou em outros projetos de automação em editorias de esportes, cidades e política (ver mais exemplos no subcapítulo **3.5**).

A razão da cobertura ter a sua especificidade é explicada por Latar (2018) como a incapacidade da inteligência artificial de 'pensar' além de *framework* previamente estabelecido. Nessa perspectiva, a criatividade de jornalistas humanos seria a capacidade cognitiva de cruzar diferentes tipos de experiências e domínios do conhecimento. Para Latar (2018), esse cruzamento leva a *insights* inesperados, enquanto por outro lado previsibilidade é tudo que programadores esperam dos algoritmos que criam. A limitação dos algoritmos seria, portanto, o outro lado da moeda de sua vantagem.

Torrijos (2021) sustenta uma ideia similar ao apontar que o assunto da cobertura é limitado à fonte de dados estruturados que servem de *input,* alimentando o algoritmo para gerar histórias. Ou seja, 5 mil textos podem ser produzidos em uma cobertura eleitoral, ou 11 mil sobre uma liga de basquete, tudo redigido e publicado em alta velocidade, mas não há multiplicidade de assuntos dentro desse grande volume. O *trade-off* ocorre entre quantidade *versus* variedade.

A principal restrição do conceito de Torrijo (2021) é não mencionar o que acontece entre a entrada e a saída dos dados, o *throughput*. São os procedimentos computacionais específicos que ocorrem na etapa intermediária que diferenciam um algoritmo de outro (LATZER et al, 2019). Da mesma forma que o *throughput* diferencia as formas de mediação algorítmicas listadas na **Tabela I**, ele também distingue aplicações de produção de conteúdo entre elas.

---

[31] news automation implies, on the one hand, identifying and selecting a series of routines and repetitive tasks that are undertaken on a daily basis in newsrooms and wich can be coded by bots and algorithms; and, on the other, creating and leveraging structured data sets that, through programing, facilitate the instant production of texts for a specific news coverage

Tabela I - Tipologia de Algoritmos de Seleção por Funcionalidade de Aplicação

Tabela I - Tipologia de Algoritmos de Seleção por Funcionalidade de Aplicação	
Tipo	**Exemplo**
Aplicação de Busca	Google, Bing, Yahoo;
Aplicação de Agregação	Google News;
Aplicação de Vigilância	PhotoDNA para detectar conteúdo criminoso, como pornografia infantil;
Aplicação de Previsão	Difusão de doenças (Google Flu Trends);
Aplicação de Filtragem	Norton (filtro de spam), Net Nanny (classificação indicativa);
Aplicação de Recomendação	Spotify (música), Netflix (filmes);
Aplicação de Pontuação	Ebay's buyer e Seller reviews (sistemas de reputação);
Aplicação de Produção de Conteúdo	Quakebot (jornalismo algorítmico);
Aplicação de Alocação	Google Adsense (publicidade computacional)

Fonte: Latzer et al, 2016

Como vemos ao longo da pesquisa, é possível especificar em diferentes categorias os impactos e tipos de mediação que os algoritmos produzem. Embora alguns autores se dediquem a criar definições sob medida para essas aplicações, Diakopoulos (2019) vai no sentido contrário ao defender uma conceituação mais ampla, que una o estudo de caso das iniciativas de automação no jornalismo às noções da Teoria da Computação. O autor propõe o termo *jornalismo computacional,* por entender que para tratar da mediação algorítmica dentro do jornalismo se faz necessário um entendimento alargado.

> [...] jornalismo computacional como produção de informação e conhecimento com, por e sobre algoritmos que assumem valores jornalísticos. Enquanto outros, incluindo eu mesmo, preferiram no passado outras definições, gostaria de enfatizar que o jornalismo computacional envolve explorar as formas em que algoritmos são tanto projetados e incorporados em notícias informativas e suas práticas de produção (DIAKOPOULOS, 2019, p. 27, tradução nossa[32])

É notável como o autor se coloca em uma posição de repensar os conceitos que já utilizou no passado, como *jornalismo automatizado* e *robô repórter*, para defender uma

[32] (...) computational journalism as information and knowledge production with, by, and about algorithms that embraces journalistic values. While others, including myself, have proffered other characterizations in the past,26 here I wish to emphasize that computational journalism involves exploring the relationship between the underlying values of journalism and the ways in which algorithms are both designed and incorporated into news information production practices.

definição que englobe o emprego de algoritmos em três frentes: (1) projetados para as práticas de produção de notícias informativas; (2) incorporados nas práticas de produção de notícias informativas; (3) alvo, ou "pauta", da produção de notícias informativas. Diakopoulos emprega um conceito mais abrangente com duas razões. Primeiro, há uma necessidade de disseminar o estudo e promover um diálogo entre pesquisas sobre mediações algorítmicas no jornalismo. Segundo o autor, o *jornalismo computacional* remete à interdisciplinaridade da área, ou seja, leva pesquisadores a buscarem mais referências e a pensarem fora da Teoria da Comunicação para problematizar seus objetos de estudo.

No entanto, a definição de Diakopoulos é demasiada ampla para pesquisas que buscam somente estudar o emprego de algoritmos de Geração de Linguagem Natural na produção de conteúdo jornalístico. A dimensão da mediação algorítmica como '(8) aplicação de produção de conteúdo', descrita por Latzer et al (2016) e detalhada por Dörr (2015), se mostra específica o bastante para tratar de iniciativas de automação no jornalismo que dizem respeito unicamente a redação de conteúdo noticioso, porém, como mencionado acima, o conceito de jornalismo algorítmico falha em estabelecer uma relação direta entre automação e jornalismo.

Pelos motivos expostos até aqui, a presente pesquisa propõe uma definição que faça uma média da literatura corrente, considerando as potencialidades e limitações de cada autor. 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).

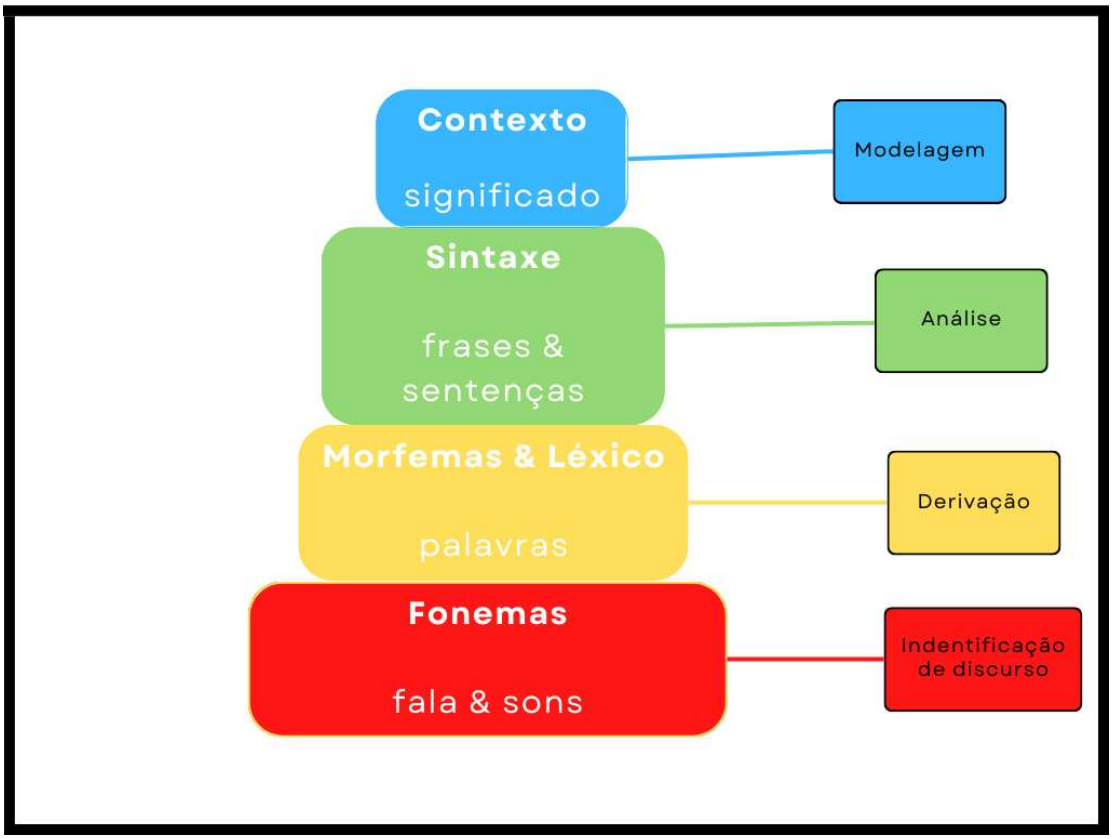
Todos os aspectos citados na definição desta pesquisa, foram abordados ao longo da revisão de literatura, com uma exceção: a Geração de Linguagem Natural. Esse campo, chamado por McDonald (2010) como "tímido", é uma das áreas de menor visibilidade dentro da computação. Para compreender a emergência dessa área de estudos, seus aspectos teóricos, suas limitações práticas, assim como suas aplicações comerciais, é necessário recorrer a definições da Linguística Computacional.

3.3. LINGUAGEM E COMPUTAÇÃO

O campo da Computação Linguística possui várias aplicações possíveis. Em linhas gerais, a disciplina consiste na engenharia de computar a língua escrita, ou falada. Cada nicho carrega suas próprias características em razão do tipo de padrão que busca reconhecer, ou reproduzir. Os principais campos são a Compreensão de Linguagem Natural, o Processamento de Linguagem Natural e a Geração de Linguagem Natural (CLARK et al, 2012). Os três subcampos se comportam como os três lados que compõem um mesmo triângulo. Enquanto a Compreensão de Linguagem Natural encontra formas de interpretar a fala humana, o Processamento identifica padrões no discurso que a Geração reproduz.

Os três campos da computação linguística não devem ser vistos como aplicações essencialmente distintas, tendo em vista que muitas vezes são empregados em etapas diferentes de um mesmo processo. O poder destas aplicações reside no potencial de fazer computadores se comunicarem com humanos em sua própria linguagem, assim como aumentar a escala de tarefas texto-direcionadas. Tudo isso é possível graças a uma visão sistemática do funcionamento da linguagem humana, alinhada à computação.

Figura I - Os blocos que constituem a linguagem e as etapas de computação



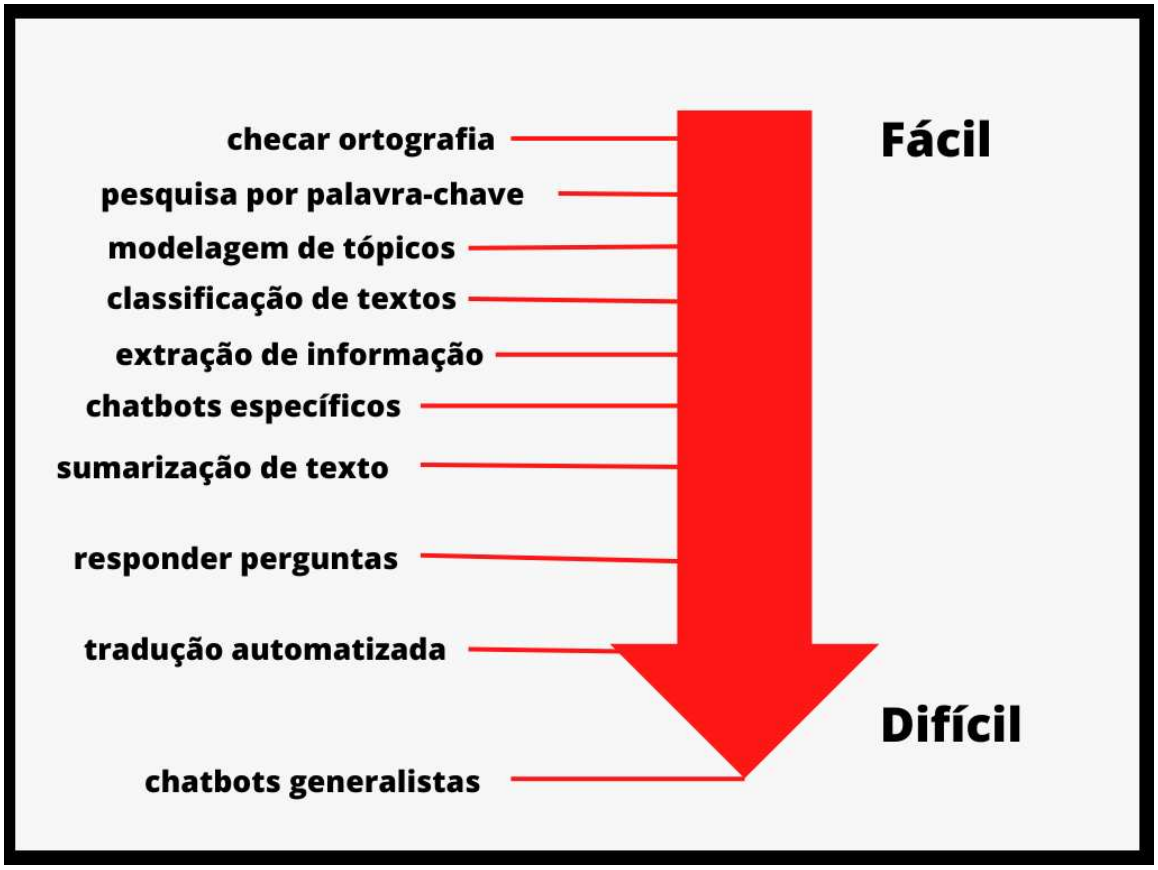
Fonte: Vajjala et al (2020)

Exemplos de programas de Compreensão de Linguagem Natural se popularizaram com assistentes virtuais como Siri (*Apple*), Organa (*Microsoft*) e Alexa (*Amazon*). Essas tecnologias são capazes de identificar padrões no discurso por meio do reconhecimento de voz, depois inferir significado e prover uma resposta. As etapas que antecedem a resposta são exatamente o processo de Compreensão.

Já o Processamento de Linguagem Natural é ainda mais difundido, sendo encontrado em motores de busca, como o Google, filtros de e-mail, sistema de *autocomplete* para textos, *chatbots*, sistemas de classificação de texto e análise de sentimento em redes sociais[33]. É a área da computação linguística mais propagada e que mais se destina poder computacional para aperfeiçoar os modelos e *softwares*. As diversas utilidades de Processamento podem ser compreendidas a partir da sua complexidade.

[33] Ver *How is NLP Used to Conduct Sentiment Analysis*. Disponível em: https://datasaur.ai/blog-posts/how-nlp-conduct-sentiment-analysis. Acesso feito em 07 de fevereiro de 2023.

Figura II - Nível de complexidade dos sistemas de Linguagem Natural



Fonte: Vajjala *et al* (2020)

Por exemplo, a popular aplicação de Inteligência Artificial Generativa, o *Chat GPT-3*, cumpre todas as categorias mencionadas na figura acima. A ferramenta é mais conhecida por funcionar como um *chatbot* generalista, mas que também é capaz de executar checagens de ortografia, traduções, sumarização de texto e modelagem de textos.[34] A tecnologia do *Chat-GPT* é um bom exemplo de como os diferentes campos da computação linguística se relacionam de forma intrincada, pois quando o usuário dá entrada no *prompt* com o seu texto, a ferramenta opera a Compreensão de Linguagem Natural. O treinamento da IA, por sua vez, é feito mediante o Processamento de Linguagem Natural, segundo anúncio da empresa. Entretanto, no momento que o usuário recebe uma resposta no *chat*, a aplicação está atuando como um Gerador de Linguagem Natural. A partir desse exemplo, pende o questionamento de como a etapa de Processamento e Geração funcionam.

Segundo Vajjala *et al* (2020) o Processamento Linguagem Natural pode ser posto em prática de três formas, por (1) Regras & Heurística, (2) Aprendizado de Máquina e (3)

[34] Ler "**ChatGPT: Optimizing Language Models for Dialogue"**. Disponível em: https://openai.com/blog/chatgpt/. Acesso feito em 9 de fevereiro de 2023.

Aprendizado Profundo[35]. As aplicações feitas por (1) Regras & Heurística consistem em programar normas explícitas do que deve ser identificado em um texto. Por exemplo, constatar que uma palavra é masculina, de acordo com uma biblioteca de palavras classificadas por gênero, se for antecedida pelo artigo "o" antes do nome. A precisão desses sistemas depende das decisões que são tomadas com base nas regras, se não houver regra para certo caso, o sistema falha. Este tipo de aplicação é tido como a forma de mais baixa complexidade para criar algoritmos linguísticos.

Uma característica de aplicações que funcionam por Regras & Heurística é que elas remetem aos primórdios do campo da Inteligência Artificial (VAJJALA et al, 2020), assim como da própria computação. As primeiras tentativas de desenvolver sistemas de Processamento de Linguagem Natural exigiam que programadores incorporassem regras gramaticais nos programas. Tais sistemas precisavam ser alimentados por dicionários, *thesaurus* e enciclopédias. Outra forma de aperfeiçoar as regras desse sistema foi a criação de bancos como o Wordnet[36], base de dados contendo a relação semântica entre as palavras. Essas relações podem ser resumidas em sinônimos, hipônimos e merônimos. Os sinônimos são os termos com significado similar. Os hipônimos carregam o tipo de relação, como por exemplo futebol e basquete que têm como denominador comum serem esportes, enquanto os merônimos capturam a parte da relação em conexão com o todo, como por exemplo 'pétala' sendo merônimo de 'flor'. Todas essas informações servem de parâmetro em um sistema de processamento por Regras & Heurística.

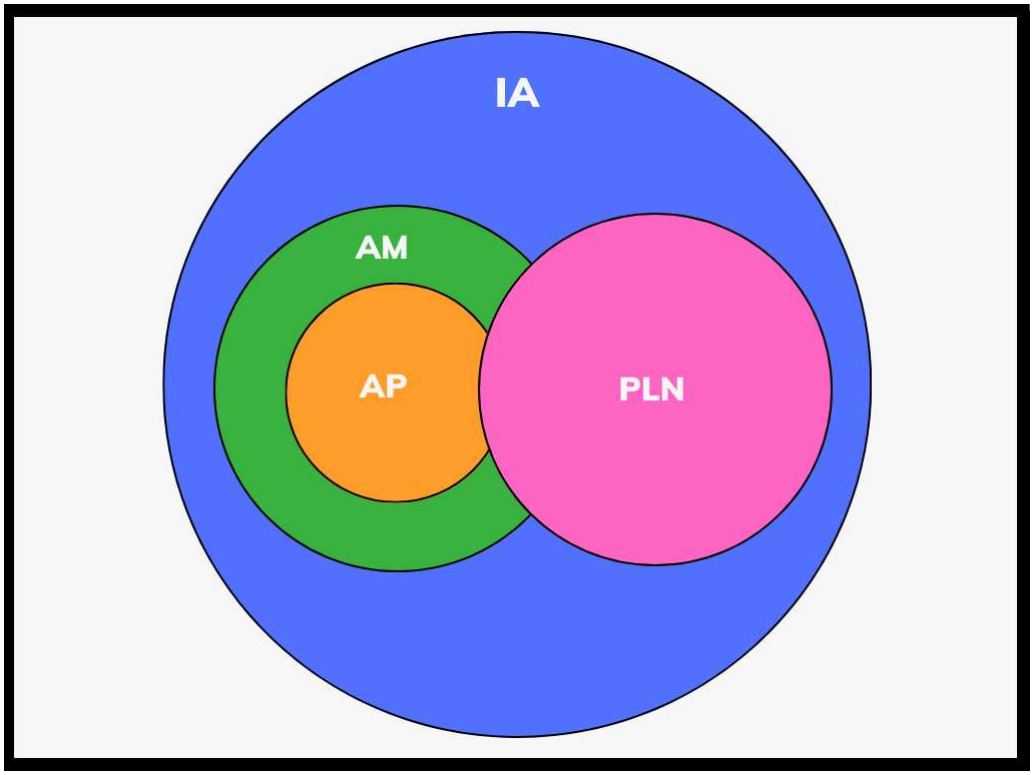
Já o (2) Aprendizado de Máquina (AM) e o (3) Aprendizado Profundo (AP) são os dois campos mais complexos da Inteligência Artificial. No (2) Aprendizado de Máquina, os algoritmos aprendem a executar tarefas automaticamente com base em amplos conjuntos de exemplos, chamados de 'dados de treinamento'. Isso normalmente é feito com um programador criando uma representação numérica, chamada de *feature,* que é aplicada pelo algoritmo aos dados de treinamento para identificar padrões nestes exemplos.

Enquanto o (2) Aprendizado Profundo segue uma arquitetura de redes neurais que tenta emular o comportamento de um cérebro humano. Segundo Bishop (2014), uma rede neural é um modelo de aprendizagem composto por nós de processamento interconectados. Tais nós são treinados por grandes volumes de dados a performar tarefas específicas mediante o ajuste de proximidade entre as conexões. As redes neurais podem ser pensadas como uma

[35] *Machine Learning* e *Deep Learning* são os termos em inglês mais usados para se referir a estas duas aplicações da Inteligência Artifical.
[36] Ver **"WordNet"**. Disponível em: https://wordnet.princeton.edu/. Acesso feito em 09 de fevereiro de 2023.

memória computacional organizada em camadas. Cada fração de um padrão identificado pelo algoritmo é salvo em uma camada, que pode ser depois recombinada para identificar diferentes objetos. Diferentemente do Aprendizado de Máquina, as *features* são determinadas pelo próprio algoritmo a partir de um vasto conjunto de dados de treinamento, que depois são aplicadas na classificação de dados.

Figura III - Como a Inteligência Artificial e se relaciona ao PLN



Fonte: Vajjala *et al* (2020)

Para ilustrar melhor a aplicação destas formas de Inteligência Artificial, vale retomar o exemplo do *Chat-GPT*. O modelo foi treinado utilizando o Aprendizado de Máquina[37] e Aprendizado Profundo, a partir de um banco de textos extraídos da internet. Ao todo, estipula-se que a aplicação foi alimentada com 570 GB de dados, totalizando 300 bilhões de palavras. A partir desses vários exemplos, o modelo é capaz adivinhar por portabilidade qual é a próxima palavra que irá aparecer em uma frase.

Para o *Chat-GPT* chegar ao ponto de sofisticação de gerar respostas precisas, o treinamento foi dividido em duas etapas. Primeiro, um time de pessoas o alimentava com perguntas como, por exemplo: "Qual é a cor de um pato?". Se o modelo respondesse que "O

[37] Ver **"ChatGPT: Everything you need to know about OpenAI's GPT-3 tool"**. Disponível em: https://www.sciencefocus.com/future-technology/gpt-3/. Acesso feito em 11 de fevereiro de 2023.

pato é roxo", o time retornava a resposta correta para que o programa se aperfeiçoasse. Em um segundo momento, o modelo era levado a fornecer uma série de respostas para a mesma pergunta, com o time de humanos ranqueando as melhores e as piores. Assim, a Inteligência Artificial poderia também aprender por comparação. Tal processo de aprendizado por comparação se configura como a estrutura de redes neurais, definida por Bishop (2014)[38].

Embora aplicações como o *Chat-GPT* tenham se difundido e alcançado relativa popularidade na atualidade, tecnologias como esta estão ainda na sua infância. Mesmo dentro da Computação Linguística as aplicações da Geração de Linguagem Natural são descritas por McDonald (2010) como a área que mais demorou para se desenvolver, tanto academicamente, quanto no âmbito comercial. Boa parte dos avanços no campo foram feitos ao longo da primeira década do século XXI. Em síntese, a aplicação consiste no "processo de renderizar pensamento em linguagem" (p. 121, tradução nossa)[39], tendo sido estudada e elaborada simultaneamente por filósofos, neurologistas, linguistas e engenheiros da computação.

A definição ampla do autor faz jus, por um lado, à capacidade dos programas em transformar 'intenção' em enunciados, sem determinar ao certo o que pode ser o pensamento e a linguagem. Uma concepção inicial errônea é que o Gerador de Linguagem Natural é um Processador de Linguagem Natural 'ao contrário', ou seja, com as etapas invertidas do processo que mecanismos de busca executam. Na realidade, McDonald afirma que os processadores enfrentam problemas analíticos, enquanto os geradores incorrem em desafios de **construção e planejamento**. O que é sabido no Processamento são as palavras, enquanto a intenção do falante é desconhecida. A partir do escaneamento das palavras, uma metodologia baseada em múltiplas hipóteses trabalha para chegar em uma aproximação da intenção do falante. Questões de ambiguidade e subespecificação são os principais entraves a serem superados.

A Geração de Linguagem Natural trabalha com um fluxo informacional[40] contrário, indo 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

[38] A metodologia de treinamento do Chat-GPT é chamada de *Reinforcement Learning from Human Feedback (RLHF)* e mistura diferentes técnicas de Aprendizagem de Máquina e Profundo, acompanhada pelo supervisionamento humano. Para saber mais, ver: "Illustrating Reinforcement Learning from Human Feedback (RLHF)". Disponível em: https://huggingface.co/blog/rlhf. Acesso feito em 11 de fevereiro de 2023.
[39] Natural language generation (NLG) is the process by which thought is rendered into language.
[40] O fato do fluxo informacional ser contrário implica em etapas distintas, mas não inversamente proporcionais ao Processamento de Linguagem Natural.

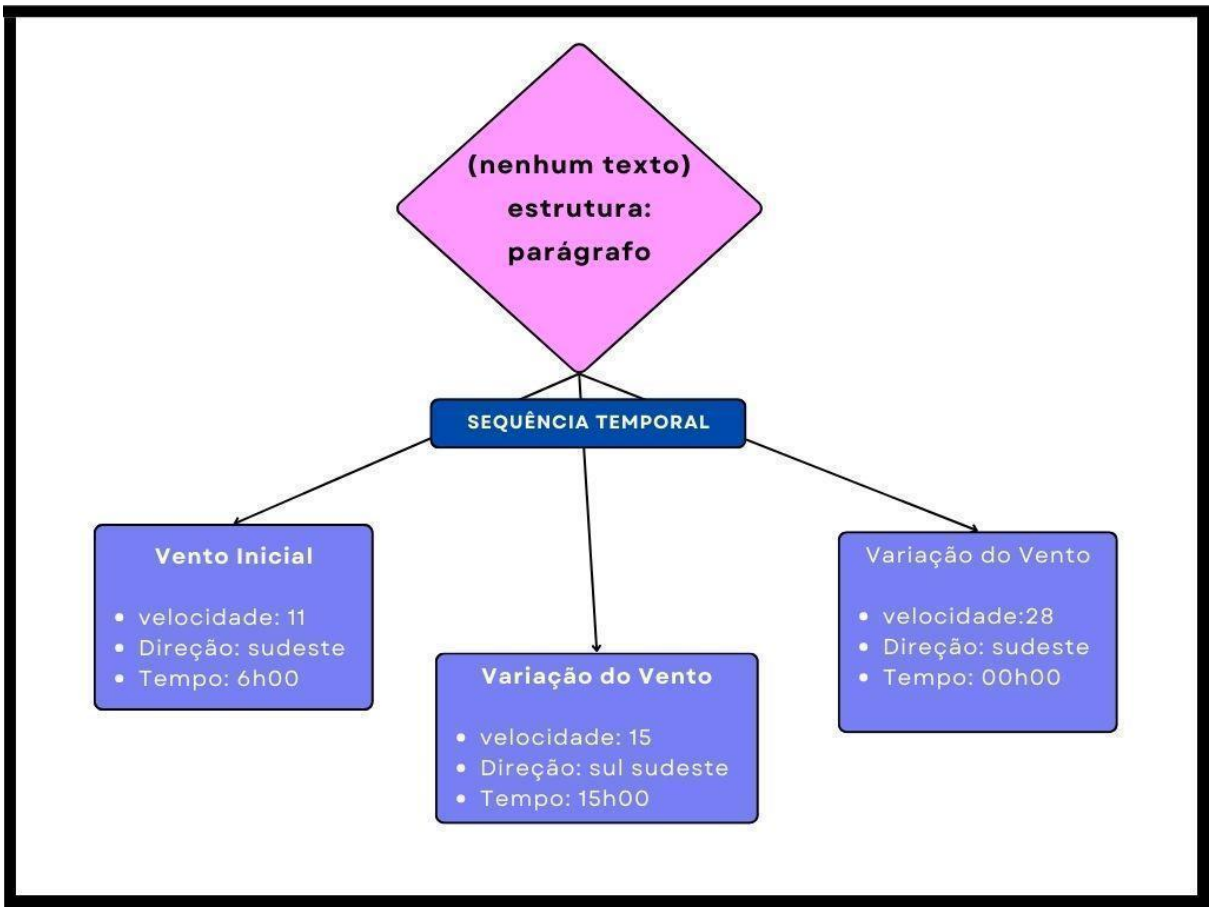
retóricos, todos pré-adequados a uma gramática[41]. Por meio da formatação das palavras no texto redigido, o programa estabelece a prosódia[42] do discurso. Segundo McDonald, isso tudo ocorre em três estágios.

> Geralmente, o processo de geração é dividido em três partes que com frequente se subdividem em três programas: (1) identificar os objetivos do enunciado, (2) planejamento de como os objetivos serão alcançados a partir dos recursos comunicacionais e (3) realizar o planejamento na forma de texto (MCDONALD, 2010, p. 121, tradução nossa)[43]

O pesquisador Ehud Reiter (2012) também vai ao encontro dessa separação do processo em três etapas, mas com uma definição mais simples para os estágios: (1) Planejamento do Documento, (2) Microplanejamento e (3) Realização. O primeiro estágio consiste em decidir qual informação comunicar (determinação de conteúdo), assim como a organização da informação (estruturação do documento). De uma perspectiva do *software*, Reiter explica que o *input* do Planejamento do Documento é o *input* de todo o sistema de Geração de Linguagem Natural. Já o *output* é geralmente constituído de uma *árvore de mensagens*, sendo as mensagens pedaços de informações que podem ser linguisticamente expressos em cláusulas, ou frases.

[41] A Computação Linguística como um todo, parte do estudo de gramáticas formais.
[42] Prosódia é o ramo da linguística que estuda a entonação, sendo mais característico da linguagem oral. Dentro da linguística computacional, a ideia de prosódia está ligada à intenção comunicativa tanto em linguagem oral, quanto escrita.
[43] Modulo a number of caveats discussed later, the process of generation is usually divided into three parts, often implemented as three separate programs: (1) identifying the goals of the utterance, (2) planning how the goals may be achieved by evaluating the situation and available communicative resources, and (3) realizing the plans as a text.

Figura IV - Árvore de Mensagens



Fonte: Reiter (2012). Exemplo do *software* SumTime.

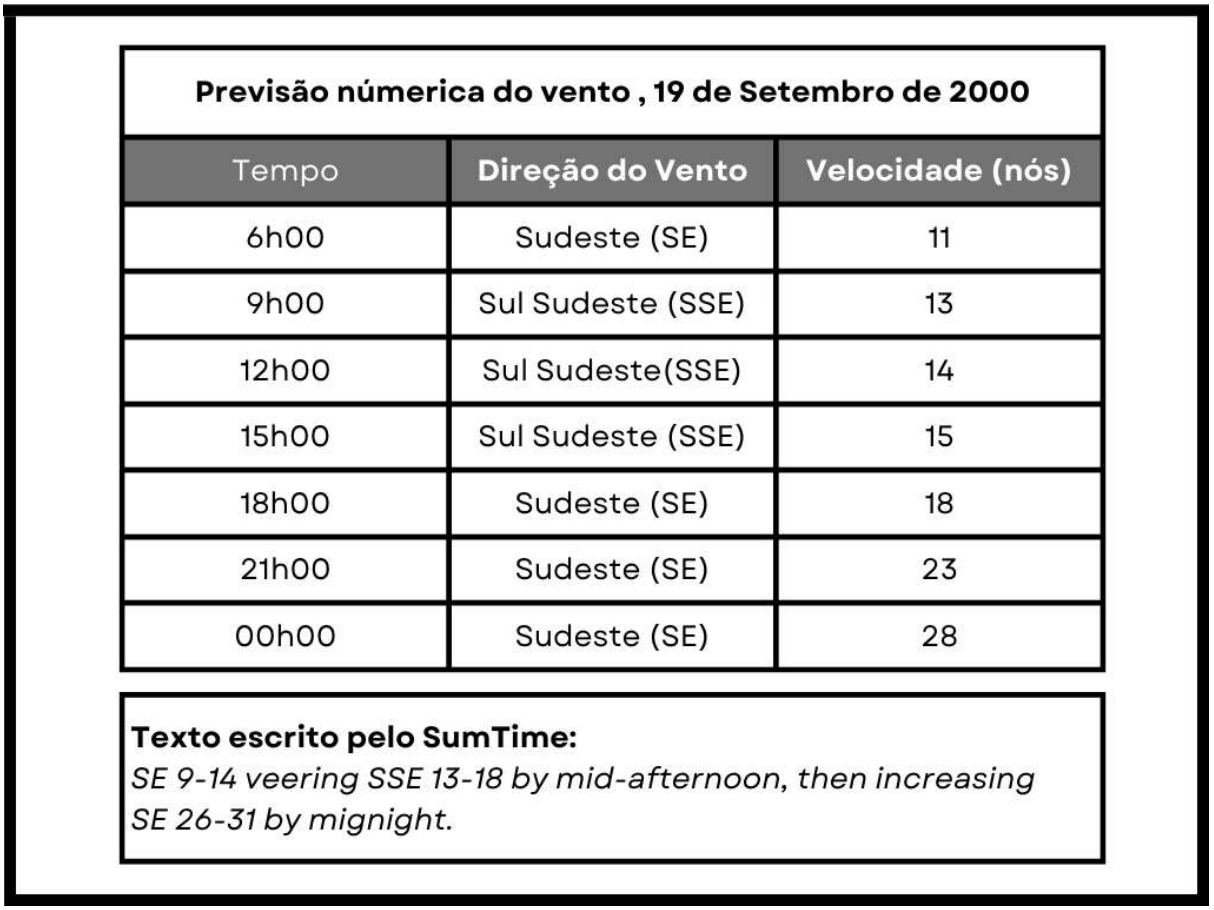
Como o trabalho de Reiter explica, a raiz constitui uma interseção na qual não existe nenhuma mensagem, apenas a especificação de que a estrutura é um parágrafo. Essa interseção tem três ramificações que se desencadeiam em uma sequência temporal. Cada ramificação, no exemplo, comunica uma mensagem singular sobre o fluxo do vento. A primeira ramificação é sobre o estado inicial do vento, enquanto as duas ramificações que seguem comunicam estados de mudança - todas contém parâmetros de velocidade e direção de tempo.

Uma abordagem possível na fase do planejamento é tentar imitar o que autores humanos de fato fazem (REITER et al, 2003), a partir de uma investigação detalhada sobre o seu trabalho. Neste processo, é possível analisar *corpus* de textos já escritos para encontrar padrões, ou mesmo conduzir entrevistas e sessões colaborativas com os profissionais. Um desafio para essa colaboração é identificar padrões sem explicitamente imitar o raciocínio dos autores, uma vez que máquinas não comportam aspectos subjetivos do processo.

A fase do (2) Microplanejamento consiste em decidir como as informações são expressas no texto que será gerado. Para isso, quatro tipos de escolhas devem ser feitos: a) **Escolhas Lexicais**, decidem quais palavras devem ser usadas para expressar conceitos centrais e os dados; b) **Escolhas de Referência**, decidem quais expressões de referência identificam objetos individuais no discurso; c) **Escolhas Sintáticas**, decidem a estrutura das sentenças; d) **Agregação**, decidem quantas mensagens devem ser expressadas em cada sentença. O *input* do (2) Microplanejamento é o (2) Planejamento do Documento, enquanto o *output* é a **especificação do texto**. Essencialmente, esse *output* é uma árvore em que as ramificações determinam a estrutura textual.

Por último, a (3) Realização gera um texto concreto, baseado nas informações selecionadas na etapa do planejamento e moduladas pelas escolhas linguísticas do microplanejamento. No estágio da realização, o *input* é condicionado a um formalismo gramatical específico, que varia de língua para língua. Os 'realizadores' também suportam as funções de *sobre-geração* e *seleção*, ou seja, um mecanismo cria várias possibilidades de textos, enquanto outro seleciona uma delas. Essas duas funções explicam a possibilidade de personalização do conteúdo jornalístico, mencionado no capítulo "**3.2. AUTOMAÇÃO & INTELIGÊNCIA ARTIFICIAL**", sobre a redação de notícias distintas para um mesmo resultado de uma partida de futebol. Os resultados do realizador no exemplo do *SumTime* podem ser vistos na figura abaixo.

Figura V - Frase escrita pelo *software* SumTime



Previsão númerica do vento , 19 de Setembro de 2000		
Tempo	Direção do Vento	Velocidade (nós)
6h00	Sudeste (SE)	11
9h00	Sul Sudeste (SSE)	13
12h00	Sul Sudeste(SSE)	14
15h00	Sul Sudeste (SSE)	15
18h00	Sudeste (SE)	18
21h00	Sudeste (SE)	23
00h00	Sudeste (SE)	28

Texto escrito pelo SumTime:
SE 9-14 veering SSE 13-18 by mid-afternoon, then increasing SE 26-31 by mignight.

Fonte: Reiter (2012)

Como resultado, o *software SumTime* produziu em um processo algorítmico (passo-a-passo) uma frase curta com três estados de variação do vento. Em termos de estilo e extensão, McDonald não coloca uma limitação específica para os gêneros textuais que a Geração de Linguagem Natural pode produzir. Embora os casos sejam em sua maioria de textos curtos e descritivos, com grande presença de dados quantitativos, tais quais relatórios meteorológicos.

> Hoje, o que um gerador pode produzir varia de uma palavra só a frases com respostas, ou etiquetas em um diagrama, ou múltiplas sentenças e observações dentro de um diálogo, até várias páginas de explicações. Tudo depende da capacidade do programa e de seus objetivos, ou seja, a máquina 'locutora' e as demandas particulares de cada contexto (MCDONALD, 2010, p. 121, tradução nossa)[44].

[44] Today, what a generator produces can range from a single word or phrase given in answer to a question or as label on a diagram, through multi-sentence remarks and questions within a dialog, and on to multipage explanations and beyond depending on the capacity and goals of the program it is working for—the machine 'speaker' with something to say—and the demands and particulars of the context.

A publicação "*A* Survey *of Natural Language Generation”*, feita pelos pesquisadores Dong *et al* (2022), busca atualizar a definição de Geração de Linguagem Natural defendida por McDonald (2010) e Reiter (2012). Paras o *survey*, o campo pode ser definido para além de suas etapas, como um processo de produção de texto para cumprir objetivos comunicativos específicos. Os textos que são gerados podem variar desde uma única frase dada como resposta a uma pergunta, até observações de várias frases e perguntas dentro de um diálogo, ou mesmo explicações de página inteira.

É interessante reparar na definição de Dong et al (2022) que não há qualquer limitação quanto ao formato ou extensão do texto. Há possibilidades de geração tanto para diálogos (*chatbots*) quanto para relatórios, ou notícias, porém, os pesquisadores defendem que existem somente três caminhos para se chegar à Geração:

1. **Dados-para-texto**
2. **Texto-para-texto**
3. **Imagem-para-texto**

Os diferentes caminhos variam de acordo com o seu ponto de partida. As aplicações de dados-para-textos partem de informações estruturadas para seja qual for o tipo de redação desejada. Já as aplicações de texto-para-texto se dividem em três subcategorias: **(a) abreviação de texto, (b) expansão de texto e (c) síntese textual**. Nas duas primeiras, o objetivo é destilar informação de grandes conjuntos de texto e expandir a redação a partir de pequenas frases, respectivamente. Já a função de (c) síntese textual tem por objetivo transferir um estilo textual de um exemplo para o outro, sem retirar as informações primordiais.

As aplicações de imagem-para-texto são baseadas em tecnologias similares às aplicações de dados-para-textos, afinal de contas um *pixel* também é um dado, porém elas se diferenciam em sua funcionalidade. O objetivo é explicar, ou resumir, o conceito visual de uma imagem, ou vídeo. Isso pode ser feito por três vias: legenda de imagem, legenda de vídeo, ou narração visual. As legendas de imagem são resumos feitos a partir de um arquivo visual, enquanto a legenda de vídeo faz o mesmo processo para uma série de imagens. Entretanto, a narração visual identifica não só a correlação entre objetos de uma única imagem, mas também fornece a relação lógica entre imagens sequenciais consecutivas.

Para além dos três caminhos para a Geração de Linguagem Natural, existem também os métodos computacionais para executá-la. Ao todo, os chamados métodos gerais da Geração de Linguagem Natural se organizam em oito categorias:

Tabela II - Métodos Gerais de Geração de Linguagem Natural

Métodos Gerais de Geração de Linguagem Natural
Redes Neurais Recorrentes
Transformer
Mecanismo de Atenção
Mecanismo de Apontar e Copiar
Expressões-chave e *templates*
Rede Adversária Generativa
Rede de Memória
Rede Neural de Grafos
Modelos pré-treinados

Fonte: Dong et al (2022)

Todas as formas da Geração de Linguagem Natural envolvem o planejamento, determinação e realização do conteúdo. Estas três etapas são seguidas com o intuito de dar forma as intenções do falante, resultando em palavras e marcadores sintácticos ordenados de acordo com tais intenções. Juntamente com a sua aplicação, a situação e o discurso, fornecem a base para fazer escolhas entre as formulações e construções que a lingua proporciona. Quando o assunto é Geração de Linguagem Natural não se deve perder de vista que o esforço é a construção deliberada de um texto. Esse objetivo só é alcançado mediante **decisões** que são feitas pelo algoritmo entre um vocábulo, em detrimento de outro, um verbo, em detrimento de outro, um sinônimo, em detrimento de outro, e assim por diante.

A forma do texto produzido pelo gerador, como enfatizam tanto McDonald quanto Reiter, depende dos objetivos que condicionam a configuração do programa. Como visto acima, esses objetivos são em grande medida determinados na fase de (1) Planejamento do Documento. Para o jornalismo automatizado, é esse ponto de encontro entre os dados estruturados que alimentam o algoritmo e as prioridades editoriais. A importância de uma informação no texto se reflete em decisões de discurso, como optar por uma abordagem descritiva versus uma explicativa, ou mesmo resolver se um dado vem antes, ou depois numa frase, ou parágrafo.

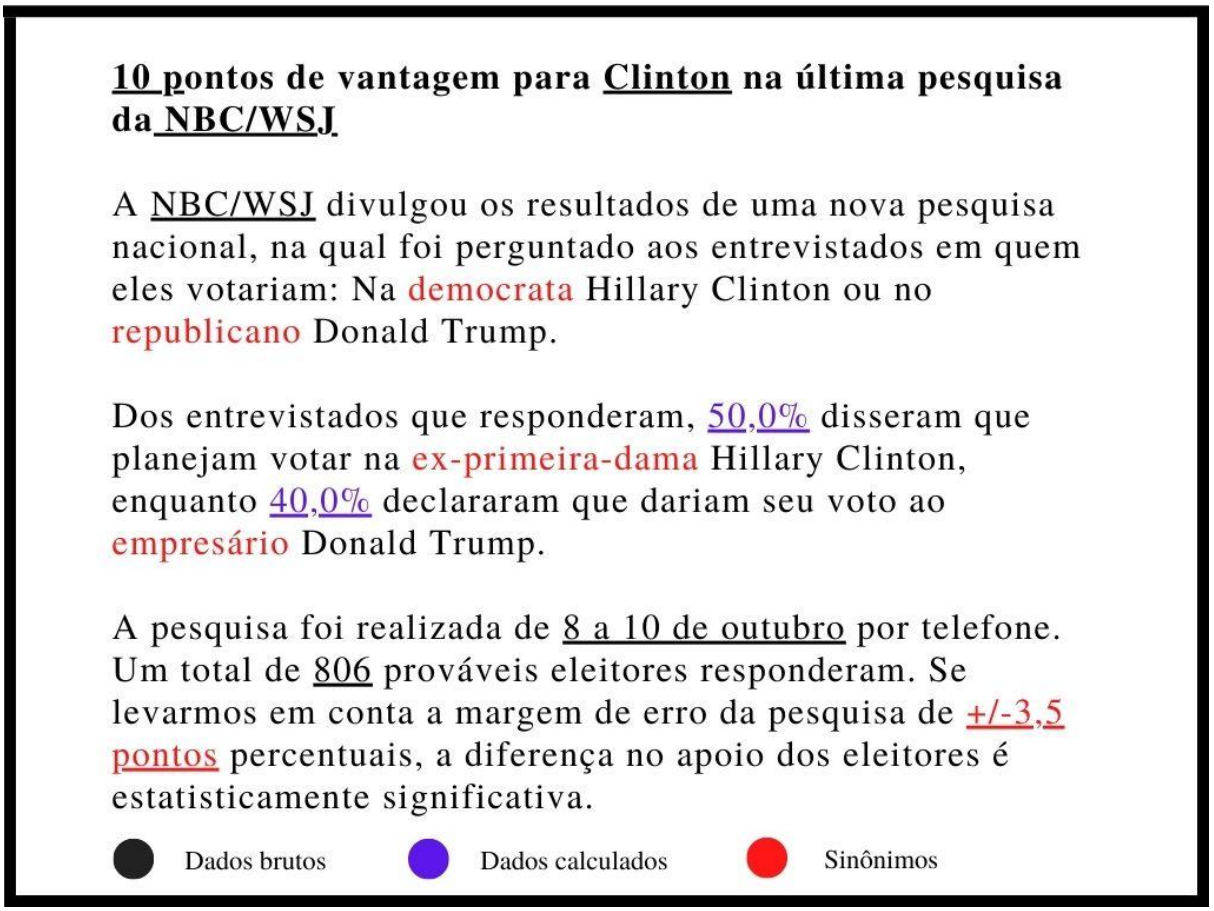
Como explica Diakopoulos (2019, p. 90, tradução nossa)[45], se o (1) Planejamento do Documento é decidir o que comunicar, o processo é impactado por "o que o leitor está interessado em ler e o que o escritor está tentando conquistar (por exemplo, explicar ou persuadir), a partir de limites tais quais os dados e o espaço disponíveis para o texto". Em outras palavras, o planejamento enumera quais dados serão expressos e quais fatos terão prioridade, o que no jornalismo se configura como critérios de noticiabilidade. É neste momento que aspectos éticos, ou até mesmo ideológicos, do jornalismo agem como filtros no interior dos algoritmos de Geração de Linguagem Natural. Por exemplo, em uma notícia política, o veículo jornalístico irá decidir quais dados sobre os candidatos e o pleito são mais interessantes para o leitor, assim como quais são dispensáveis.

Já no próximo estágio, o (2) Microplanejamento, impacta a diversidade e a complexidade da linguagem no *output*, a partir das escolhas lexicais, sintáticas referenciadoras e agregadoras mencionadas acima. Ou seja, para o jornalismo é esta etapa que influencia na dimensão estética do produto final: a notícia. Segundo Diakopoulos (2019), é no Microplanejamento que o algoritmo incorpora parâmetros de estilo, se conforma a um gênero textual específico e determina o 'tom' do que será escrito. Um outro tipo de escolha relevante para o jornalismo automatizado é determinar como integrar sinônimos. Voltando ao exemplo da cobertura política, nesse momento faz a escolha entre repetir o nome de um candidato, se referir pelo nome do partido (pemedebista, petista e etc) ou por nome do cargo (atual prefeito, ou ex-deputado e etc). A seleção cuidadosa de como integrar os sinônimos é um fator chave para o grau de monotonia, ou dinamismo, do texto.

Por último, a fase da (3) Realização é a conformação de todo o processo a especificações linguísticas de um idioma. É neste momento que se define regras de concordância verbal, declinação dos adjetivos e o plural ou singular dos substantivos. A Realização é apontada tanto por Diakopoulos, quanto por McDonald, como o aspecto mais robusto e bem estudado da Geração de Linguagem Natural. Em razão desta solidez, existe pouca influência jornalística e decisões editoriais na forma como o algoritmo opera neste momento. É como se o potencial do algoritmo fosse máximo nesta instância, enquanto o do jornalismo é mínimo.

[45] Deciding what to communicate is impacted by what the reader is interested in, what the writer is trying to accomplish (for example, explain or persuade), and by constraints such as the available space and data.

Figura VI - Exemplo de notícia escrita por Geração de Linguagem Natural



Fonte: Diakopoulos (2019)

Para além das etapas, existem os diferentes formatos e as situações em que a Geração de Linguagem Natural pode ser aplicada. Sobre os formatos, Diakopoulos (2019) destaca que a produção de conteúdo automatizada pode ser aplicada a imagens e vídeos, embora seja mais comumente encontrada em textos. Isso se explica pois no caso de conteúdos gráficos, a árvore de decisão do algoritmo precisa incorporar etapas de edição de cores e formas dos elementos visuais. Por exemplo, já existem exemplos de geradores que colocam faixas de identificação de locais e personagens em vídeos.

A constatação de que a produção de outros formatos (ilustrações, vídeos, gráficos e etc) pode ser automatizada, leva a dedução de que os diversos exemplos de textos automatizados mencionados nesta pesquisa se mostram como aplicações seminais desta tecnologia. Como aponta Diakopoulos (2019), existe uma boa chance de que na medida que programadores e veículos de comunicação se familiarizem com o processo, a automação extrapole para além da redação de textos.

No momento, com as aplicações da Geração de Linguagem Natural sendo em sua maioria texto-direcionadas, elas se concentram em esforços de síntese, ou ampliação de conteúdos que foram previamente escritos. Sobre as situações em que a tecnologia pode ser usada, McDonald (2010) e Reiter (2012) elencam seis tipos:

1. **Escrita de texto automatizada:** redigir frases e parágrafos descritivos, como o demonstrado pelo *SumTime*. Essa aplicação surgiu para a produção de relatórios meteorológicos, mas já é aplicada em múltiplas áreas, como o jornalismo.
2. **Sumarização:** tornar um conjunto de publicações em um registro de eventos numericamente ordenados. Essa prática já foi observada em registros médicos, dados financeiros, dados esportivos, meteorológicos e em redes sociais.
3. **Gerar rascunhos iniciais:** descreve, ou resume, o conteúdo de um conjunto de documentos. O recurso já foi empregado em manuais de instrução, documentos jurídicos, diagnósticos clínicos e informes administrativos.
4. **Gerar explicações de raciocínio em Inteligências Artificiais:** com a operação de sistemas de aprendizado de máquina sendo muitas vezes um mistério, até mesmo para os programadores, esse recurso busca reconhecer decisões tomadas pela máquina para criar avisos.
5. **Gerar textos para persuadir usuários:** têm por objetivo motivar usuários, ou reduzir suas ansiedades, com um caráter puramente interativo.
6. **Dar suporte para usuários com deficiências:** por exemplo, permitir deficientes visuais examinarem gráficos, ou ajudar pessoas com mudez a criar histórias. Esse tipo de aplicação é ainda experimental e carece de exemplos reais.

A partir dos diversos casos que serão mencionados no subcapítulo **3.5** é possível inferir que o jornalismo se beneficia tanto da **1. Escrita automatizada de textos** quanto para **3. Gerar rascunhos iniciais.** Diakopoulos acrescenta que existe um grande potencial, ainda pouco explorado, na aplicação da **2. Sumarização** para o jornalismo. Eventos complexos com diversos episódios poderiam ser elencados em um sumário, que induz o leitor a se informar por diferentes pontos de partida. Como, por exemplo, uma notícia que traz em sua conclusão

uma seleção de suítes[46]. A sumarização pode atuar em uma atividade que já é difundida dentro do jornalismo digital: a curadoria de conteúdo.

Como podemos observar, a Geração de Linguagem Natural e o jornalismo apresentam vários pontos de contato. No que diz respeito às etapas do processo do *software*, existem momentos em que existe uma troca entre os valores da profissão jornalística (éticos, estéticos e discursivos) com o funcionamento do recurso algorítmico. Essa dinâmica atualmente se resume na imensa maioria dos casos a formatos textuais, mas não se limitando exclusivamente a eles, com a possibilidade no horizonte de geração para formatos em fotos e vídeos. Essa tecnologia encontra-se na sua primeira década de vida dentro do ambiente editorial e já dá sinais de que pode ser aplicada em diversas situações dentro do jornalismo.

Como foi exposto ao longo deste capítulo, o processo de planejamento e criação dos *softwares* automatizados de conteúdo são passíveis de serem ajustados aos princípios editoriais de uma organização. Os tomadores de decisão dentro de uma redação, sejam eles repórteres, ou editores, são convidados a partilhar seus valores profissionais com cientistas de dados e engenheiros de *software*. O esforço é de transmitir o *know how do ofício jornalístico* para um algoritmo. O público também cumpre seu papel, tendo em vista que a audiência é o alvo desenhado a partir do que o jornalista julga ser relevante para ela.

Os valores jornalísticos de diferentes ordens são incorporados no sistema, desde aspectos estéticos como a construção de uma frase, até preferências de valor notícia, como qual base de dados usar, como calcular os dados, como as informações são apresentadas, em qual ordem. Este exercício de decidir o que entra e o que sai acaba por esbarrar em questões seculares para o jornalismo, como "novidade", "importância", "relevância" e "imprevisibilidade". Os cientistas da computação fazem o trabalho de codificar os protocolos do fazer jornalístico, quebrando-as em partes mínimas para servirem de *input* a uma máquina. faria, para o algoritmo.

A viabilidade da automação para o trabalho editorial pressupõe que ele próprio é dotado de tarefas mecânicas. A repetição e sua forma são as chaves para determinar como o sistema deverá operar, logo, o papel de um jornalista na criação de aplicações de Geração de Linguagem Natural é precisamente o de expor quais são essas ações repetitivas e sistemáticas.

[46] Suíte para o jornalismo remete a reportagens que remetem a fatos anteriormente publicados. Ver Manual de Redação da Folha de S.Paulo (1996) . Disponível em: https://www1.folha.uol.com.br/folha/circulo/manual_producao_s.htm. Acesso feito em 27 de setembro de 2022.

3.4. ALGORITMOS & JORNALISMO

Os algoritmos hoje influenciam praticamente todas as etapas de produção jornalística, desde o filtro de uma caixa de e-mail que decide qual será ou não lido por um repórter, passando por escolhas editoriais em função do comportamento da audiência nas redes, até o estágio final de consumo da notícia. Enquanto eles são vistos por diversos autores como 'caixas-pretas', os algoritmos são ao mesmo tempo objetos técnicos e construções sociais que carregam múltiplos agenciamentos.

Nenhum algoritmo existe em isolamento, mas sim em uma concatenação de atores, *actantes*, instituições e valores. Como defende Annany (2016, p. 7, tradução nossa[47]), os algoritmos que atuam na *web* são em si redes que podem ser chamadas de *Network Information Algorithm* (NIA), ou Algoritmo de Informação em Rede. Essa tecnologia pode ser descrita como uma *assemblage* de "código computacional institucionalmente situado, práticas humanas e normativas lógicas que criam, sustentam e significam relações entre pessoas e dados, por meio de um processo minimamente observável de ação semi-autônoma".

O estudo das mediações algorítmicas parte de um reconhecimento do poder da tecnologia, que compõem uma esfera de discussões entre acadêmicos e jornalistas sobre a capacidade desses "seres" em (re)estruturar realidades sociais e fenômenos comunicacionais. O estado da literatura permite observar que o papel dos algoritmos é estudado em ao menos quatro dimensões: (1) algoritmos como agentes (MACHILL-BEILER, 2008), (2) algoritmos como instituições (NAPOLI, 2013), (3) algoritmos como ideologia (MOZOROV, 2020; VAN DIJK, 2017; HARARI, 2018) e (4) algoritmos como *gatekeepers* (JURGENS et al. 2011; WALLACE, 2018).

Embora os algoritmos sejam imbuídos de uma lógica própria, eles podem carregar valores éticos e estéticos do jornalismo. A disseminação da mediação algorítmica no jornalismo acontece em função de um contexto mais amplo de digitalização da vida como um todo. Conforme as informações passam a ser crescentemente produzidas e consumidas em ambientes virtuais, algoritmos passam a mediar essas trocas, assumindo uma considerável centralidade no mundo contemporâneo (MUSIANI, 2013). A influência é tamanha que para Napoli (2014) o jornalismo vive uma virada algorítmica.

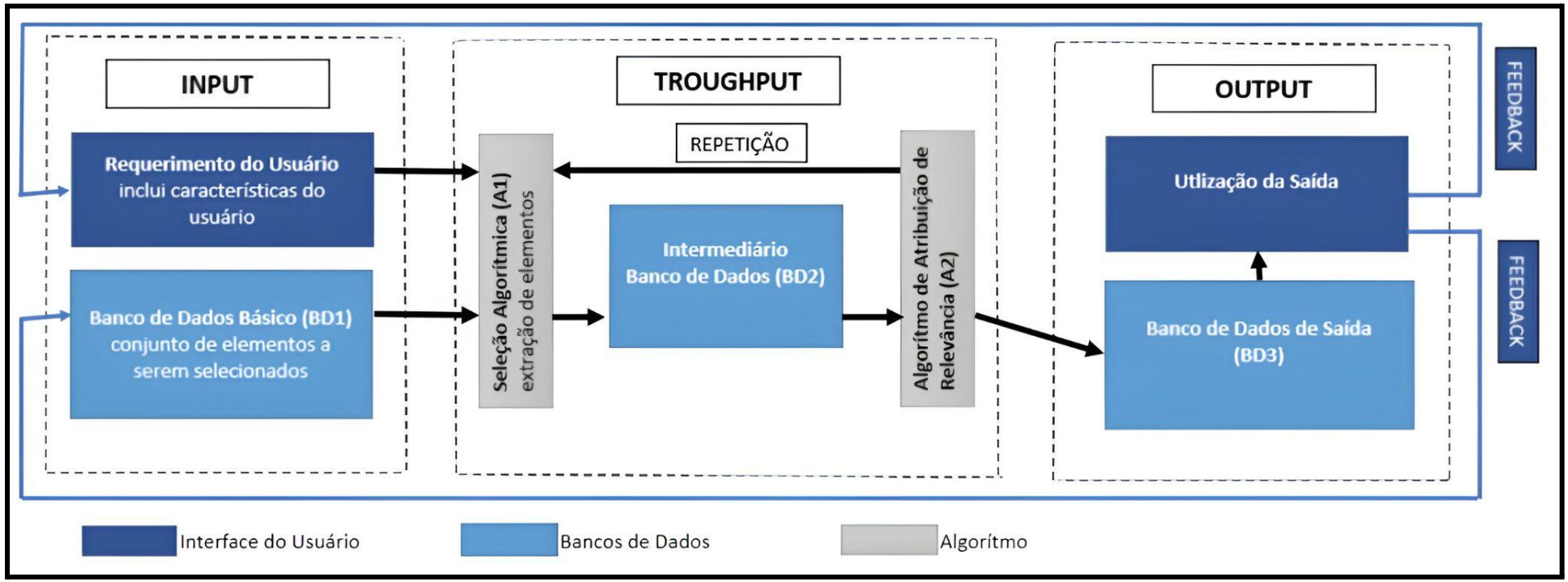
[47] With these relationships in mind, I define an NIA as an assemblage (DeLanda 2006; Latour 2005) of institutionally situated computational code, human practices, and normative logics that creates, sustains, and signifies relationships among people and data through minimally observable, semiautonomous action. Although code, practices, and norms may be observed individually in other contexts, their full ''meaning and force ... can only be understood in terms of relations with other modular units'' (Chadwick 2013, 63).

Não há somente uma definição amplamente aceita do que é um algoritmo. Porém, há duas escolas dominantes da literatura: uma social e outra técnica (ZAMITH, 2019). Definições técnicas tipicamente focam em aspectos materiais do código, da capacidade computacional ou da arquitetura de dados. Já as definições sociais dão destaque aos processos que alimentam, e retroalimentam, os objetos técnicos associados aos algoritmos, que por sua vez causam impactos na esfera pública, organizacional e institucional. Poucas pesquisas buscam concatenar essas duas abordagens, embora haja uma tendência crescente dentro da TAR e da Comunicação Homem-Máquina em mesclar esses dois tipos de explicações (LEWIS et al, 2019).

Com as conceituações variando entre abordagens materiais e contextuais, Latzer et al (2016) propõem uma definição de algoritmo como "uma série finita de regras precisamente descritas ou um processo para resolver um problema" (p. 4, tradução nossa)[48]. Quando o algoritmo é encarado como processo, ele pode ser compreendido como uma atividade em três etapas: uma fase inicial (*input*) é submetida a procedimentos computacionais específicos (*throughput*), produzindo um resultado (output).

[48] finite series of precisely described rules or processes to solve a problem.

Figura VII - Modelo Algoritmo de Seleção de na Internet



Fonte: Baseado em Latzer et al (2014)

Segundo os mesmos autores, os algoritmos podem ser compreendidos por sua utilização, ao afirmarem que "embora as formas de operação sejam diferentes nos detalhes, todas essas aplicações são caracterizadas por uma funcionalidade em comum: elas automaticamente selecionam informação e atribuem relevância" (p. 1, tradução nossa)[49]. Tendo as aplicações algorítmicas um ponto em comum, é possível diferenciá-las por sua operacionalidade em nove categorias, conforme vistas previamente na **Tabela I** (ver capítulo 3)**.**

Na oitava categoria proposta pelos autores, é possível observar que o jornalismo é nominalmente citado como área que exemplifica a produção de conteúdo por algoritmos. Chamado na pesquisa de jornalismo algorítmico, o exemplo do *Quakebot* consiste em uma aplicação criada pelo *Los Angeles Times* que cumpre duas funções. Primeiro, ele avalia notificações do Serviço Geológico dos Estados Unidos e envia alertas para os membros da redação do jornal sobre abalos sísmicos na forma de uma notícia gerada automaticamente. O artigo só é gerado se a notificação obedece a uma série de critérios, como localização e intensidade do tremor. Se algum editor do *Times* decidir que a notícia tem valor, ela pode ser publicada na forma como foi feita, ou complementada por um humano[50]. Mas seria a produção de conteúdo noticioso a única aplicabilidade dos algoritmos para o jornalismo?

Para Zamith (2019), esses mediadores atuam no jornalismo em ao menos quatro níveis: (1) colhendo informação, (2) automatizando a produção, (3) analisando a audiência e (4) personalizando o conteúdo. João Canavilhas (2022) aponta um quinto nível que seria a (5) busca por novas tendências de notícias. Um sexto nível ainda é proposto por Thurman et al (2019) que são as (6) interfaces de comunicação com os usuários-leitores, comumente chamadas de *chatbots*.

[49] Although their modes of operation differ in detail, all of these applications are characterized by a common basic functionality: They automatically select information elements and assign relevance to them.
[50] Disponível em: https://www.latimes.com/people/quakebot. Acesso feito em 7 de setembro de 2022.

Figura VIII - Mediação Algorítmica no jornalismo

Fonte**:** Canavilhas (2022), Zamith (2019) e Thurman et al (2019)

No primeiro nível, é possível mencionar as Aplicações de Busca on-line, já mencionadas na **Tabela I,** como mediações algorítmicas que auxiliam os jornalistas a reunir informações sobre determinado assunto. Embora seja importante destacar que a primeira etapa não se resume a isso. Por exemplo, os disparos automáticos de alertas sobre a publicação de uma página podem se enquadrar nessa etapa. Alguns veículos chegam a criar seus próprios robôs e *web crawlers* para percorrer bases de dados, sites governamentais, a fim de acusar possíveis 'furos', ou informações úteis. Um exemplo disso é o *Rui Barbot* criado pelo veículo Jota[51], jornal especializado em cobertura jurídica, para notificar a tramitação de um processo do Supremo Tribunal Federal (DALBEN, 2018).

O segundo nível diz respeito ao foco desta dissertação, o processo de automatização da produção de notícias e sua eventual publicação. A discussão teórica sobre o que é a

[51] Ver "JOTA lança robô Rui para monitorar tempo que STF leva para julgar processos". Disponível em: https://www.jota.info/dados/rui/prazer-rui-barbot-24042018. Acesso feito em 1 de junho de 2023.

automatização no jornalismo, o estado da pesquisa acadêmica sobre o assunto, assim como exemplos práticos, serão abordados no próximo subcapítulo.

Já a análise de audiências (3) é talvez um dos aspectos da mediação algorítmica com maior impacto comercial para o jornalismo. Racionalizar e quantificar o comportamento da audiência é visto pelos veículos como uma oportunidade de aumentar lucros, reduzir custos e estreitar relações com o leitor, compreendendo seus gostos e preferências (CHRISTIN, 2020). Os sistemas de análise de audiência empregados pelas empresas jornalísticas podem ser descritos como "medição, coleção, análise e reporte de dados digitais concernentes a como o conteúdo jornalístico é consumido e interagido com" (ZAMITH, 2018, p. 421, tradução nossa)[52]. Alguns exemplos desses sistemas são o *Chartbeat*, *Google Analytics*, *Parse.ly* e *Adobe Analytics*.

O estágio de (4) personalização do conteúdo é tanto uma tendência identificada nas redações desde o final da década de 1990, como uma das características centrais do jornalismo digital (MIELNICZUK, 2002). Aqui, as escolhas editoriais implicam em segmentar a produção de notícias para atingir tipos específicos de consumidores de notícias. Ocorre dessa forma uma quebra de paradigma no jornalismo tradicional, onde o foco em propagar um senso de importância compartilhada entre o público (THURMAN, 2011) se transforma em um esforço para satisfazer o senso de relevância individual (ANDERSON, 2011). Os algoritmos permitem que isso seja feito em larga escala, com a distribuição de notícias para os seus tipos ideais de espectadores. Muitos estudos discutem o impacto negativo desta mediação algorítmica em razão da formação das bolhas digitais[53].

Em seguida, a busca por (5) tendências de notícias é, em síntese, uma forma de usar algoritmos como termômetro da sociedade em rede. Stainer (2014) identifica a *Wikipedia Live Monitor*, ferramenta para assistir a edições no site em tempo real, como recurso usado por jornalistas para identificar possíveis *breaking news*. De forma similar, o *Google Trends* também pode ser usado para identificar um aumento abrupto na busca por alguma palavra-chave. Esse aumento quantitativo e repentino no interesse de usuários por um assunto, pode servir de pontapé inicial para a pauta de um repórter, por exemplo, ou para toda uma cobertura.

---

[52] enable the measurement, collection, analysis, and reporting of digital data pertaining to how content is consumed and interacted with.

[53] O termo "bolhas digitais" provém do inglês *filter bubble*. A expressão se popularizou a partir da publicação em 2011 do livro"*The filter bubble: What the Internet is hiding from you*", escrito pelo ativista Eli Pariser. Inicialmente o termo se referia aos diferentes resultados de busca no *Google*, personalizados de acordo com o histórico de cada usuário. Mais tarde o termo passou a ser amplamente cunhado para denominar a personalização de conteúdo em qualquer plataforma digital, cada vez mais associado a um sentido de manipulação, polarização política e empobrecimento do debate público nas redes sociais.

Por último, os *chatbots* (6), interfaces de comunicação com os usuários, são frequentemente encontrados em contextos externos ao jornalismo, tal qual em serviços de atendimento ao cliente, ou publicidade. Porém, segundo Thurman et al (2019), esse tipo de mediação algorítmica já foi colocado em prática no caso de duas organizações jornalísticas: a *Australian Broadcasting Corporation* (ABC) e a *British Broadcasting Corporation* (BBC). A implementação de *chatbots* ocorre em razão do distanciamento do público de meios de comunicação tradicionais (televisão, rádio e mídias impressas) em direção a redes sociais como *Facebook* e *WhatsApp*. Os chatbots representam, portanto, um esvaziamento de formas de comunicação pública e o crescimento de meios privados, onde impera a regra dos meios digitais de personalização e conteúdo direcionado.

Seja o uso de algoritmos para interagir com o público (6), ou o emprego de mecanismos de busca na etapa da apuração (1), as seis formas de mediações algorítmicas identificadas na literatura sugerem que o jornalismo é perpassado por essas tecnologias de ponta a ponta. A pesquisa de Santos (2018b) é enfática em estimular a busca de explicações mais conectadas com a ciência da computação para discorrer sobre transformações atravessadas pelas tecnologias digitais. Outro levantamento, feito por Ioscote (2021) a partir das produções acadêmicas feitas no Brasil sobre o estudo de algoritmos, Inteligência Artificial e automação, conclui que dentro da pesquisa sobre jornalismo há uma crescente preocupação com o emprego dessas tecnologias entre 2010 e 2020. Porém, boa parte das publicações são sobre o impacto do algoritmo dentro da atividade jornalística (perspectiva social), com pouca investigação sobre aplicações especializadas (perspectiva técnica). Investigar as características da Geração de Linguagem Natural, por exemplo, seria estudar a parte mais nebulosa das aplicações, pela via da perspectiva técnica.

A partir da pesquisa de Ioscote (2021) é possível concluir que a bibliografia nacional destina pouca atenção para a (2) produção automatizada de jornalismo, apontada por Zamith (2019), em seus aspectos formais. Grosso modo, poucos pesquisadores buscam entender o que é e como funciona um algoritmo, tal qual a Geração de Linguagem Natural, se detendo mais em identificar seus efeitos no jornalismo. Embora o reconhecimento do impacto das mediações algorítmicas na atividade profissional permita, em alguma medida, iniciar uma discussão, para adentrá-la é necessário descrever alguns casos concretos no qual notícias foram escritas por máquinas. A tarefa de dar um panorama histórico da automação no jornalismo e reunir alguns exemplos é objetivo do próximo capítulo.

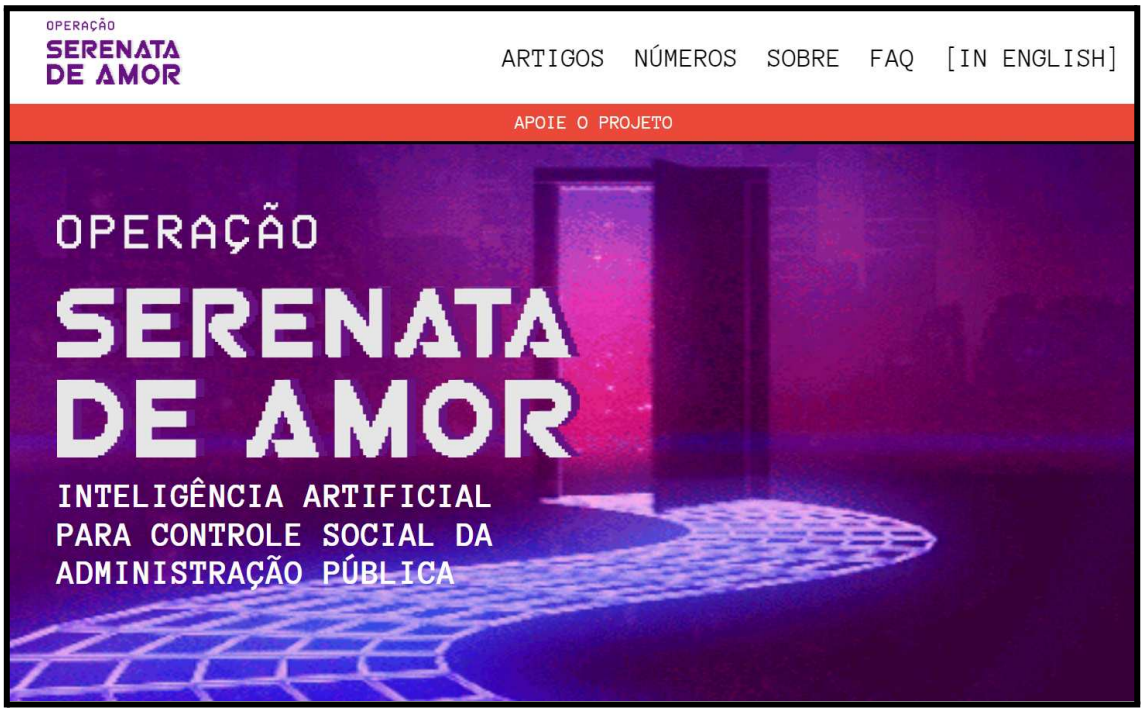
3.5. AUTOMAÇÕES NAS REDAÇÕES

Em 2020, o Portal G1 realizou a cobertura dos resultados das eleições municipais de 5.568 municípios brasileiros em menos de 24 horas, com o auxílio de *softwares* de inteligência artificial 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 automatizada das notícias.

Três anos antes da iniciativa, um outro projeto batizado de 'Operação Serenata de Amor' utilizou recursos de Inteligência Artificial para fiscalizar os gastos dos Deputados Federais que são reembolsados pela Cota para Exercício da Atividade Parlamentar (CEAP). Os valores gastos pelos parlamentares em alimentação, transporte, hospedagem, cursos e assinaturas são publicados na plataforma de Dados Abertos da Câmara dos Deputados, para depois serem submetidos a um *software* de aprendizado de máquina chamado *Rem osie*. O robô identifica compras suspeitas e as publica no *Twitter* para posterior investigação de jornalistas, ativistas ou qualquer entusiasta do monitoramento de contas públicas. A iniciativa se intitula como "primeiro robô jornalista do Brasil" e já serviu de fonte para publicações em 90 veículos jornalísticos[54].

[54] Disponível em: https://serenata.ai/. Acesso feito em 7 de setembro de 2022.

Figura IX - Plataforma da Operação Serenata de Amor

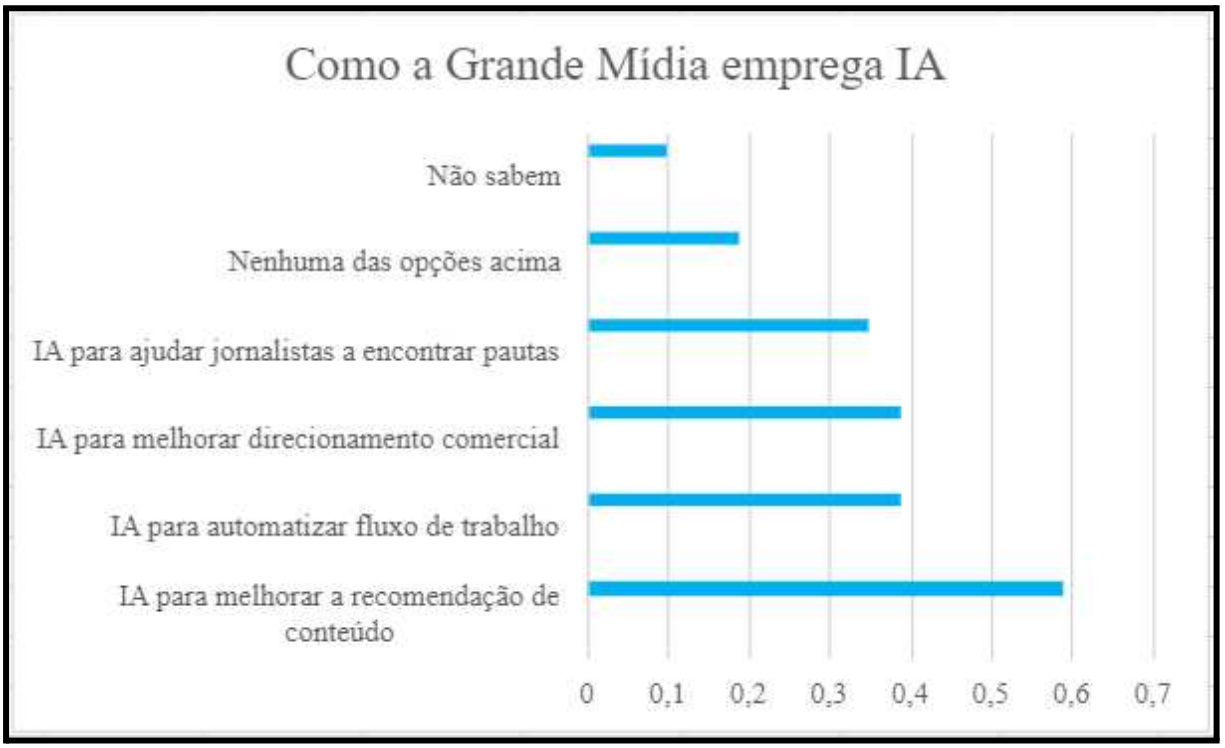


Fonte: https://serenata.ai/

Embora a ideia de que tecnologias são capazes de escrever histórias, com algum grau de autonomia, tenha ares de distopia e ficção científica, alguns casos práticos confirmam a sua viabilidade. O relatório do Instituto Reuters de 2018 "Mídia, Tendências e Expectativas Tecnológicas" revela que três quartos dos entrevistados, funcionários de meios de comunicação, declararam utilizar inteligência artificial em algum processo dentro da organização.[55] Do total, 39% declararam empregar a inteligência artificial especificamente para automatizar processos dentro das rotinas produtivas dos jornalistas. O universo de entrevistados envolveu 194 participantes que ocupam cargos de chefia em veículos tradicionais, ou nativos digitais. A grande proporção de veículos que confirmaram a automação no jornalismo como uma realidade já posta, mostra que a tendência está se consolidando ao menos na chamada *Grande Mídia*, majoritariamente de países do Norte Global, assim como em portais de jornalismo online.

[55] Ver "Journalism, Media, and Technology Trends and Predictions 2018". Disponível em: https://www.digitalnewsreport.org/publications/2018/journalism-media-technology-trends-predictions-2018/.

Figura X - Como a Grande Mídia emprega Inteligência Artificial



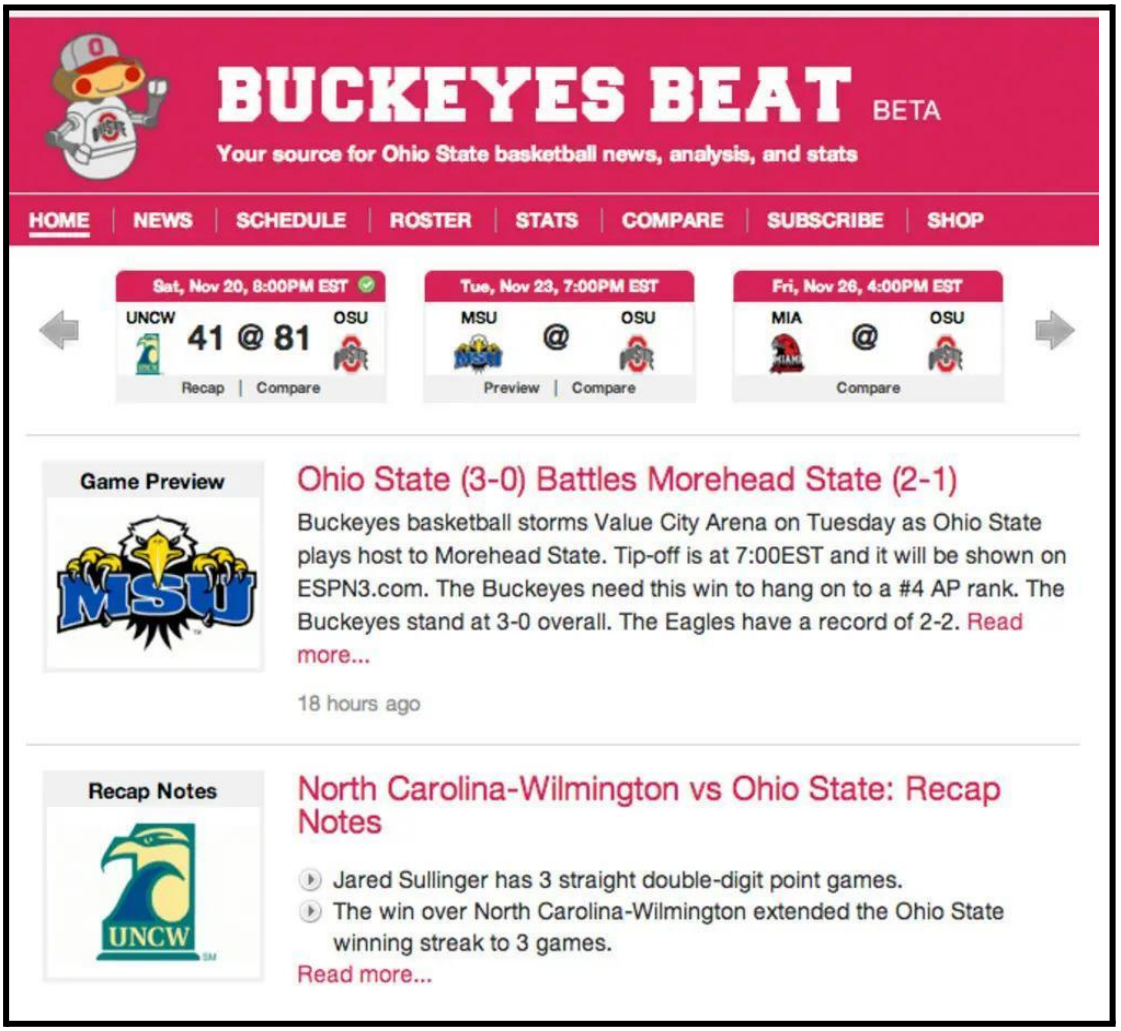
Fonte: Journalism, Media, and Technology Trends and Predictions 2018 - Instituto Reuters.

A inovação dos textos redigidos por máquinas não começou no século XXI. Segundo Linden (2017), o recurso da automação para contar histórias tem mais de 40 anos de idade. O primeiro caso remonta da década de 1960 com a geração de resumos a partir de relatórios meteorológicos para a previsão do tempo. Mais tarde, o campo migrou para as áreas dos esportes, finanças e medicina na década de 1990, mas somente no meio corporativo.

Já Van Dalen (2012) propôs como marco inicial o anúncio da Reuters em 2006 de que usaria algoritmos para compilar reportagens financeiras em seu site. Um segundo momento relevante ocorreu em 2010 com o lançamento da *Statsheet*, uma empresa de tecnologia digital gerenciadora de 345 sites, responsável por fornecer uma cobertura automatizada da divisão de basquete universitário estadunidense (NCAA). Mais de 15 mil notícias sobre resultados de partidas foram publicadas pela iniciativa.

Os conteúdos gerados pela *Statsheet* variavam entre prévias dos jogos, resultados das partidas e outros tipos de artigos, todos gerados por algoritmos. A tecnologia desenvolvida pela empresa era capaz de combinar - ou como de se diz na computação, concatenar - os dados dos jogos com um estoque de sentenças. Uma vez que as informações de uma partida eram digitalizadas, um *script* as puxava de um banco de dados e inseria, de forma automática, em um modelo de texto.

Figura XI - *Statsheet*, um dos primeiros casos de notícias escrita por máquina.



Fonte: *When the Software Is the Sportswriter*. escrito por Randall Stross, publicado no *New York Times* em novembro de 2010.

De maneira similar a Van Dalen, Silver (2014) defende que o jornalismo de dados - enquanto precursor do jornalismo automatizado - surge com a cobertura de ligas de *baseball* que têm a tradição de quantificar as jogadas dos times e jogadores nas chamadas *stats (*abreviação de estatísticas).

Não é muito claro em que momento a automação de textos migrou de agências meteorológicas, ou do setor financeiro, ou de ligas esportivas, para dentro de redações, porém, é notável como áreas com um alto grau de quantificação em seus discursos capitanearam esse processo, antes dele ser colocado em prática por veículos jornalísticos. As aplicações recentes de projetos de jornalismo automatizado em grandes empresas jornalísticas demonstram essa propensão. Embora seja importante pontuar que somente poucas editorias

até o momento comportam a automação das notícias, sendo elas: esportes, finanças, clima, crime e resultados eleitorais (CASWELL-DORR, 2018).

Neste sentido, é possível alegar que essa automação está diretamente relacionada à quantificação da narrativa jornalística. A presença crescente de dados no lide jornalístico, defendida por Meyer (2002) como o jornalismo de precisão, ou a virada quantitativa de Coddingtion (2015), indicam uma busca por alicerçar a objetividade jornalística em números. Santos (2016) corrobora essa ideia ao afirmar que os conteúdos baseados em informações numéricas podem ser extraídos mais facilmente de campos já quantificados, como torneios esportivos, ou as apostas do mercado financeiro. Essa quantificação se encaixa com a automação, na medida em que ela adentra no lide sempre nas mesmas posições dentro do texto, em um lugar de destaque.

Anderson (2018) vai além ao afirmar que o esforço em colocar dados dentro de notícias vem de um intuito do jornalismo em emular as ciências naturais, de certa maneira, pegando a credibilidade deste segundo campo emprestada. Seja qual for a explicação, a emergência de um discurso jornalístico quantitativo, por um lado, e da automação da produção de notícias, por outro, se mostram como fenômenos intrinsecamente conectados.

Se a motivação que impulsiona o jornalismo de dados é a objetividade e a credibilidade, resta compreender o que motiva a automação. Para o autor Aljazairi (2016), a fórmula capitalista do corte de gastos em função do aumento de lucro pode ser o motivo. Nesta visão, o motor por trás da automação é meramente econômico. Embora seja difícil explicar somente em termos econômicos a trajetória de convergência entre diferentes disciplinas (computação, ciências sociais, matemática e etc) e áreas (esportes, meteorologia, finanças, política eleitoral) para que o jornalismo automatizado fosse viabilizado. Segundo Aljazairi (2016), é possível que haja em paralelo à redução de gastos, uma busca por agregar mais valor ao trabalho do repórter. Poupar o tempo deste profissional com a produção de notícias, vistas como meros insumos, para destinar mais esforços a textos mais analíticos e de maior profundidade, tais quais as reportagens. Örnebring (2010) vai parcialmente ao encontro dessa avaliação afirmando que a tecnologia, no geral, 'alivia' o jornalista de trabalhos mecânicos, mas para Aljazairi, permanece a dúvida se a automação no jornalismo é uma ameaça ou uma oportunidade.

Já para Lazter et al (2016), os algoritmos na indústria midiática (jornais, produtoras de filmes, canais de televisão, gravadoras e etc) 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.

Thurman, Dorr e Kunert (2017) apontam que a razão econômica primária por trás da automação no jornalismo seja uma adequação dos modelos de negócios a novas pressões econômicas, como a queda de receita provindas de anunciantes, porém, existem ao menos quatro motivos que foram elencados em entrevistas com profissionais de organizações midiáticas tradicionais, como Reuters, CNN e BBC, para o investimento na área: (1) reduzir custos, (2) aumentar a velocidade, (3) expandir a cobertura e (4) produzir conteúdo para dispositivos móveis. Ali e Hassoun (2019) elencam ainda mais dois fatores que são a (5) abundância de informações no meio digital, exigindo em excesso a capacidade de jornalistas de dar sentido a esse volume, e (6) a busca por credibilidade em um cenário de baixa confiança em instituições jornalística. Os seis motivos listados pelos autores permitem olhar para os benefícios pouco óbvios da automação.

Em (3) expandir a cobertura, entende-se que a produção automatizada de jornalismo permite a veículos realizar coberturas que, outrora, simplesmente não seriam feitas por falta de capital humano. Um exemplo disso é a cobertura da *Statsheet* de ligas universitárias de basquete, tendo em vista que frequentemente só ligas profissionais recebem um jornalista para uma cobertura *in loco* de todas as partidas. Ao ampliar a cobertura, existe também a possibilidade de personalização, como apontado por um jornalista esportivo da Reuters que foi entrevistado pela pesquisa. “Você poderia automaticamente gerar uma notícia, por exemplo, sobre a partida de futebol *Leicester versus Liverpool*, enviado uma versão para o torcedor de Leicester, uma para o de Liverpool e uma terceira para leitores neutros” (THURMAN et al, 2017, p.11, tradução nossa)[56].

A quarta razão, (4) produzir conteúdo para dispositivos móveis, se dá graças a uma característica estética própria da era dos *smartphones*: textos curtos. A notícia produzida por algoritmos frequentemente reúne dados em um texto conciso, menor do que uma reportagem, ou um artigo. Segundo um editor da *Reuters* entrevistado por Thurman et al (2017), "ninguém que ler 5 mil palavras na tela de um celular” (p. 11, tradução nossa)[57]. O público leitor que acessa notícias por dispositivos móveis tem uma inclinação menor a ler longas

[56] “You could automatically generate a story about, for example, Leicester versus Liverpool [soccer match] and send a different version of the story to Liverpool and Leicester, and have a third one for neutrals”.
[57] “People don’t want to read 5000 words on their phone. They want a really nice picture and two or three paragraphs on what’s happened. This is what we’re thinking currently about our automated future”

reportagens, com vários recursos visuais e narrativos, nas telas de seus telefones, preferindo textos curtos com galerias de fotos e vídeos curtos. A inclinação de jornalistas produzirem textos cada vez menores é uma característica notada desde o advento do webjornalismo em meados da década de 1990 e início dos anos 2000 (MIELNICZUK, 2003). O futuro da automação no jornalismo estaria, portanto, associado a "leituras de dois ou três parágrafos".

Se o texto diminui, por um lado, o volume de informações disponíveis aumenta exponencialmente. A tentativa de lidar com a (5) abundância de informações no meio digital é uma das características centrais do jornalismo de dados (GRAY et al, 2014) que parece ter sido herdada pela vertente do jornalismo automatizado. De forma similar, a (6) a busca por credibilidade em um cenário em que a legitimidade do jornalismo é contestada, também se apresenta como um denominador comum. Enquanto a pós-verdade entra em cena, veículos de comunicação tentam remediar a questão colocando ênfase na objetividade, na cientificidade, logo, em dados.

Figura XII - Motivos para Perseguir a Automação

Fonte: Thurman, Dorr e Kunert (2017); Ali e Hassoun (2019); *Tendências e Expectativas Tecnológicas* (2018).

Apesar dos diversos motivos econômicos, estéticos e profissionais para se adotar a automação na produção de notícias, a pesquisa de Thurman, Dorr e Kunert (2017) conclui que a opinião geral dos jornalistas é de que o estado atual da tecnologia é limitado e, portanto, não produz impactos econômicos consideráveis na dinâmica de empresas jornalísticas. Embora ainda não seja possível estabelecer uma relação de causa entre o recém-surgido jornalismo automatizado e as mudanças em modelo de negócios de empresas jornalísticas, a literatura corrente permite identificar dois valores que impulsionam a área: a busca por velocidade e o ganho de escala (CASWELL-DORR, 2018; LATZER et al 2016; LINDEN, 2017; VAN DALEN, 2012; THRUMAN et al, 2017).

É possível argumentar que perseguir esses dois valores é uma constante dentro da própria modernidade e da história da mídia. O discurso da velocidade que impulsiona veículos a fazerem o possível para publicar um furo primeiro, distribuindo essa notícia para o maior número possível de leitores (ÖRNEBRING, 2010). Entretanto, é interessante notar como essa busca por velocidade e ganho de escala parece produzir uma controvérsia singular. A ubiquidade dos computadores, das conexões em rede e, por consequência, das mediações algorítmicas, resulta em uma capacidade humana de tomada de decisão aumentada, e ao mesmo tempo diminuída, ou substituída, por algoritmos (BROUSSARD, 2018). O processo de passar de princípios éticos, formatos estéticos e poder de decisão de uma atividade até então exclusivamente humana, a escrita, para seres não-humanos denota um processo de transferência de autoridade (HARARI, 2018). Em função da contemporaneidade desse processo, é complexo compreender o que essa transferência representa para o futuro distante do jornalismo.

O aumento do poder de decisão de algoritmos é uma experiência mais antiga para alguns setores da economia, como por exemplo o mercado financeiro[58] ou a indústria automobilística[59]. Ambas apresentam aplicações comerciais da tecnologia que datam da década de 1970 e 1960, respectivamente. Mas no caso de empresas midiáticas, a comunicação feita com o emprego de inteligência artificial e automação de rotinas por meio de algoritmos, é uma tendência recente, com os primeiros casos surgindo em meados dos anos 2000 (VAN DALEN, 2012).

---

[58] Disponível em: https://www.nasdaq.com/articles/nasdaq%3A-50-years-of-market-innovation-2021-02-11. Acesso feito em 10 de setembro de 2022.
[59] Disponível em: https://www.automate.org/blogs/the-history-of-robotics-in-the-automotive-industry#:~:text=Automotive%20Automation%20Booms%20in%20the,the%20Stanford%20Arm%20was%20developed. Acesso feito em 10 de setembro de 2022.

Em 2020, a Bloomberg, veículo especializado em notícias sobre finanças, estabeleceu um serviço separado e totalmente dedicado à redação automatizada de notícias, chamado *Bloomberg Automated Intelligence (BAI)*. A empresa que conta com uma média de 5.000 notícias publicadas por dia passou a se apoiar amplamente na automação para dar conta da produção. Segundo seu editor-chefe, John Micklethwait, um terço de todo o conteúdo produzido pela empresa é automatizado em algum nível. O serviço BAI aproveita os bancos de dados da Bloomberg sobre flutuações nos mercados para alimentar 500 *templates* de notícias, que são então disponibilizados para leitura dos assinantes (Bloomberg, 2021).

O relatório publicado pelo Laboratório *Nieman* da Universidade de *Harvard* aponta uma direção semelhante. Há uma aplicação progressiva da automação de tarefas por robôs e algoritmos em redações. Segundo o relatório, veículos de imprensa buscam na automação uma forma de expandir suas coberturas, engajar mais sua audiência e ter mais agilidade para emplacar furos. Editores entrevistados pela pesquisadora Clarence Lecompte (2015) declararam que tirar o trabalho mecânico de redigir notícias meramente descritivas, libera os jornalistas para se dedicarem mais ao trabalho de reportagem.

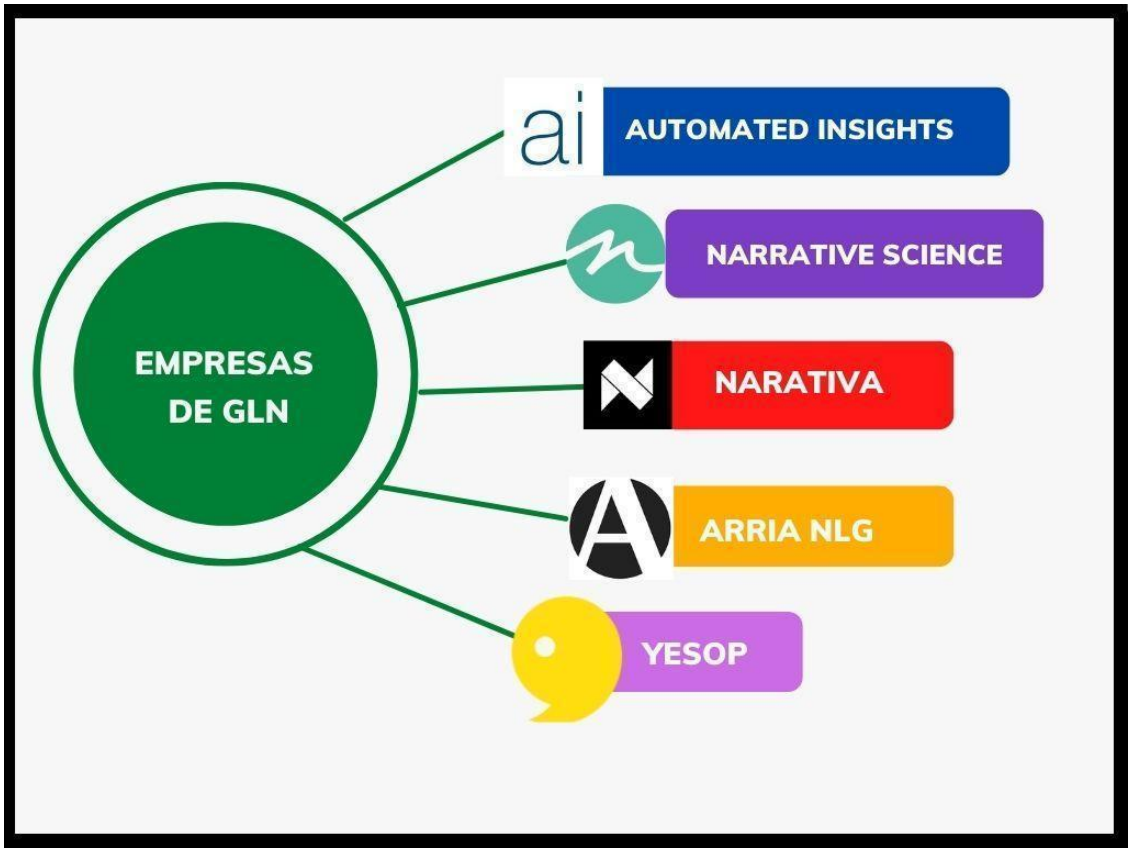
Segundo o levantamento exploratório feito por Carreira (2017), 59 veículos de mídia já executaram projetos de automatização da produção de notícias sobre oito tópicos de cobertura (política, finanças, esporte, previsão do tempo, crime, viagem, trânsito e entretenimento) em 10 países da América do Norte, Europa e Ásia. Organizações tradicionais de jornalismo como a *Associated Press, Reuters, Los Angeles Times, Washignton Post, Deutsche Welle* e Grupo Globo têm experimentado o emprego de um conjunto de ferramentas digitais para automatizar a produção de notícias.

A aproximação entre veículos consolidados e a automação também está presente em uma pesquisa apoiada pelo Instituto Reuters. O pesquisador Alex Fanta (2017) identificou que 11 de 15 agências de notícias operando na Europa empregavam tecnologias de automação em alguma instância. A conclusão da pesquisa aponta para uma maior tendência das agências em investir na automação do que outros tipos de organizações jornalísticas (emissoras, revistas, veículos regionais e outros). Essa inclinação se explica pelo papel das agências, as 'mídias das mídias', em fornecer coberturas amplas. O fato disso ter se viabilizado na Europa, por outro lado, pode ser compreendido a partir de fatores econômicos e de disponibilidade de mão de obra especializada.

As ferramentas usadas pelo jornalismo automatizado, *softwares* que transformam conjuntos de dados em prosa jornalística, nem sempre são construídas dentro do ambiente editorial. A revisão da literatura permite identificar com frequência iniciativas privadas

especializadas nas áreas de Geração de Linguagem Natural e Processamento de Linguagem Natural[60] sendo terceirizadas por organizações de mídia. Empresas como a *Automated Insights*, *Narrative Science*, Narrativa, Arria NGL e Yseop não se declaram como organizações jornalísticas e sim como companhias de tecnologia. O mercado para provedores desse tipo de serviço se mostra constituído de poucos *players* até o momento (DORR, 2015, GRAEFE, 2016; SÁNCHEZ-RUIZ, 2020, VEEL, 2018).

Figura XIII- Empresas provedoras de serviços de Geração de Linguagem Natural



Fonte: Dorr (2015), Graefe (2016); Sánchez-Ruiz (2020), Veel (2018).

A forma como essas empresas se apresentam para o mercado mostra uma retórica singular. Segundo o pesquisador Kristin Veel (2018), a partir das peças publicitárias divulgadas por essas empresas, é possível identificar como os dados são encarados como material bruto para a construção de narrativas. Partindo do clichê "os dados são o novo petróleo", companhias como a *Automated Insights* anunciam que "dados estruturados são o

[60] A definição e fundamentação do que são essas tecnologias será feita no próximo capítulo "2.7. Geração de Linguagem Natural".

combustível para o *Wordsmith*" (tradução nossa)[61]. De um ponto de vista literário, Veel destaca como as companhias de tecnologia sugerem que as narrativas já existem, antes de qualquer tipo de trabalho, em bancos de dados dispersos pelo mundo virtual, precisando apenas ser desveladas pelos algoritmos de Geração de Linguagem Natural.

> A forma como essa terminologia é usada como no exemplo da peça de marketing da Narrative Science "Contar histórias escondidas em seus dados", comunica a impressão para o potencial cliente que a existência das narrativas precede a implementação das técnicas oferecidas, ou seja, que narrativas estão de alguma forma imbuídas nos dados, com a única necessidade de emergirem por meio do processo de Geração de Linguagem Natural e automatizadas com Inteligência Artificial (VEEL, 2018, p. 4, tradução nossa)[62].

Llorente (2016) revela que além de grupos de mídia, o portfólio de clientes dessas empresas de tecnologia se estende para o *e-commerce*, serviços de logística, indústria de energia, laboratórios farmacêuticos e o mercado imobiliário. Embora a produção de textos por máquinas ainda soe como algo saído da histeria do debate público com a Inteligência Artificial, e portanto como um fenômeno recente, a variedade de setores em que ela é aplicada revela o contrário. Llorente (2016) e Veel (2018) observam como o crescimento dos serviços de Geração de Linguagem Natural aponta para um cenário de ubiquidade destas tecnologias.

As empresas de mídia, em seu próprio mérito, se constituem como clientes de primeira hora de companhias como a *Automated Insight*. Fundada em 2010, esta já era citada no mesmo ano por estudiosos do tema como provedora de serviço para veículos midiáticos (VAN DALEN, 2012). Embora ainda não haja uma produção acadêmica extensa sobre a conexão entre serviços de Geração de Linguagem Natural e grupos de mídia, o estudo ainda incipiente dessas iniciativas em organizações jornalísticas mostra que esses *softwares* conseguem operar com um considerável grau de autonomia nas etapas de redação e publicação das reportagens, para além das escolhas iniciais de delimitação de pauta, definição da estrutura textual e programação.

---

[61] "Structured data is the fuel for Wordsmith – use your data to power your narratives by leveraging our API'. Acesso feito em 20 de setembro de 2022. Disponível em: www.automatedinsights.com

[62] "The way in which this terminology is used in marketing material such as this example or Narrative Science's 'Tell the stories hidden in your data' (Narrative Science, 2017) conveys an impression to the lay customer that the narratives exist prior to the implementation of the offered techniques – i.e. that the narratives are somehow embedded in the data and only need to be carved out, which is enabled through NLG and automated with AI."

> Há uma forma de jornalismo que se dedica em recolher informação noticiosa de correntes de inteligência coletiva presente em plataformas e redes sociais. O jornalismo orientado para a tecnologia, em que os atores humanos e os atores tecnológicos se tornam interdependentes, está ganhando proeminência. Um exemplo disso é o jornalismo automatizado, uma vertente que a partir de algoritmos criados por pessoas é capaz de automatizar parte da produção de notícias para produzir milhares de produtos jornalísticos (BRAUN-ZAMITH, 2019, p. 2).

A emergência do jornalismo automatizado implica em novos desafios para a categoria dos profissionais de imprensa, ao mesmo tempo que suscita opiniões e angústias dentro de redações. Os pesquisadores Ali e Hassoun (2019) declaram que já existe uma ala de pessimistas e outra de otimistas dentro da indústria midiática e acadêmica sobre os impactos da automação no jornalismo. A ala dos pessimistas vê a inovação como uma ameaça à manutenção de empregos. O enxugamento de empregos em veículos, assim como a chamada precarização do trabalho, é uma realidade que preocupa o setor. Uma segunda aflição estaria relacionada com uma exigência, considerada excessiva por parte dos empregadores, de jornalistas com um alto nível de literacia digital e domínio de programação para conseguir operar nesse cenário.

Já a ala dos otimistas acredita que o jornalismo automatizado é uma possibilidade de elevar a qualidade do trabalho jornalístico, assim como a sua relevância. Entre as possíveis melhoras, Ali e Hassoun (2019) mencionam que o jornalista fica habilitado a lidar com um fluxo maior de informações, combater *fake news* de forma automática, fazer checagem de fatos com o auxílio de algoritmos, adequar a escrita de textos aos manuais de redação e personalizar o conteúdo para os sub-nichos de sua audiência. Para a ala dos otimistas, a automação do jornalismo não trará um cenário de escassez de empregos, mas um emprego com diferentes características e grau de complexidade.

Entre otimistas e pessimistas há uma clara distância entre o que se acredita que uma inteligência artificial pode de fato executar. É importante observar, portanto, as limitações de recursos algoritmos na produção de notícias, Ali e Hassoun (2018) elencam três: (1) criatividade, (2) vigilância (*watchdog*) e (3) viés algorítmico. A limitação de criatividade em algoritmos ocorre pela ausência de uma subjetividade na leitura de mundo, como de uma intersubjetividade para propor angulações, ganchos e pautas. A tecnologia seria incapaz por si só de fazer uma entrevista, descrever o cenário de um desastre, ou acidente, e de despertar reações emocionais no público leitor.

A limitação da vigilância está relacionada à natureza do jornalismo de ser "uma força essencial para manter a sobrevivência do sistema social" (ALI-HASSOUN, p. 44, 2019,

tradução nossa)[63]. Esse papel público do jornalismo está associado à capacidade de identificar eventos alarmantes, inesperados, que constituem de certa forma uma perturbação na esfera social. Por uma razão similar à (1) limitação de criatividade, a (2) limitação da vigilância ocorre por uma incapacidade de algoritmos fazerem conexões com experiências pregressas, externas ao seu quadro de referências. Tal inabilidade resulta em não reconhecer o inesperado, ou a novidade, para produzir denúncias e, portanto, vigiar as instituições (*watchdog*).

Já a limitação de (3) viés algorítmico ocorre em função da tecnologia digital reproduzir a ideologia de quem o configurou, mesmo que seus criadores o façam inconscientemente. O conceito desse tipo de viés é apresentado por Cathy O'Neil (2021) e Safiya U. Noble (2018) como a transferência de preconceitos do homem para a máquina, que o reproduzem acriticamente. Um exemplo, em 2015 o *app* Google Fotos etiquetou a imagem de duas pessoas negras como sendo de gorilas[64]. A empresa mais tarde se desculpou pelo erro. Enquanto isso, pesquisadores especularam a causa do erro como uma falta de imagens de pessoas negras no conjunto de dados que alimentou o processo de aprendizagem de máquina do algoritmo. Noble (2018) afirma que essa limitação na forma de vieses, na maioria dos casos, de gênero, raça e classe, é inerente ao processo de criar uma inteligência artificial e não pode ser solucionada sem uma interferência externa.

Para além das três restrições na implementação dos algoritmos mencionados acima, o jornalismo automatizado acarreta desafios profissionais de natureza ética. Autores como Clerwall (2014), Montal e Reich (2017), Ali e Hassoun (2019) e Monti (2019) elencam ao menos quatro obstáculos: (1) transparência; (2) checagem de fatos; (3) justeza (*fairness*) e (4) utilização e qualidade dos dados.

O desafio da (1) transparência envolve comunicar ao leitor de onde veio a base dados estruturados usados para a produção automatizada da notícia. Em suma, deixar a referência clara para os leitores é vista como uma necessidade para manter a objetividade e credibilidade jornalística, atrelada tanto à organização quanto ao algoritmo. Para Ali e Hassoun (2019) consiste em "ser aberto sobre a forma como os dados foram coletados e usados, incluindo as informações que foram deliberadamente deixadas de fora" (p. 44, tradução nossa)[65].

---

[63] Journalism is an essential force to maintain the survival of the social system.

[64] Ver "Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition". Disponível em: https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=5e7b3280713d

[65] This term essentially refers to being open about the way data is collected and used, as well as eschewing unnecessary data collection.

Por outro lado, a (2) checagem dos fatos é para Clerwall (2014) garantir a credibilidade e objetividade das fontes utilizadas. Da mesma forma que, no combate às *fake news,* a checagem é essencial, assim como um recurso para que jornalistas e redações não sejam acusados de calúnia e difamação. Isto se mantém verdade para o jornalismo automatizado, com a preocupação extra que uma informação equivocada poderá ser replicada em centenas de milhares de publicações, tendo em vista o aspecto do ganho de escala mencionado anteriormente.

Em (3) justeza (*fairness*), o desafio consiste em evitar publicações que invadam a privacidade dos dados de usuários, a fim de mitigar manipulações sociais e opressões estruturais. A digitalização crescente da vida acarreta em questões de Proteção Geral de Dados Pessoais. O icônico escândalo da *Cambridge Analytica*[66] e suas implicações eleitorais é um exemplo da história recente de como dados pessoais podem ser utilizados de maneira manipuladora e sem o consentimento dos usuários, ou de má fé, para ganhos políticos e econômicos. No momento que veículos jornalísticos passam a usar um algoritmo para tornar públicas informações disponíveis online, eles devem considerar princípios de *netsecurity* tais quais empresas de tecnologia.

Por último, a (4) utilização e qualidade dos dados dizem respeito à precisão das informações presentes nas notícias. De forma muito similar ao desafio da checagem, o quarto desafio ético do jornalismo de dados consiste em uma preocupação prévia com a objetividade das informações. O trabalho de algoritmos é repetitivo e sistemático, para o bem ou para o mal. Um dado sem precisão será reproduzido múltiplas vezes sem julgamento por parte da tecnologia.

De maneira similar às preocupações de qualidade e checagem dos fatos, os pesquisadores Sánchez e Ruiz (2020) apontam ainda mais uma possibilidade ameaçadora para o jornalismo profissional: a automação da produção de *fake news*. Se notícias com precisão e qualidade podem ser produzidas automaticamente em grande volume, o contrário também é válido. Em uma dimensão estética, as notícias falsas se utilizam de um *aspecto de jornalismo* para promover desinformação, manipulação política e dissenso. Casos concretos de *bots* utilizados para disseminar notícias falsas são múltiplos, vis a vis a interferência Russa

[66] Ver: Privacidade Hackeada: Até onde a LGPD pode nos proteger de escândalos como o da Cambridge Analytica?. Disponível em: https://www.netsecurity.com.br/privacidade-hackeada-ate-onde-a-lgpd-pode-nos-proteger-de-escandalos-como-o-da-cambridge-analytica/

nas eleições estadounidense de 2016[67], ou o Inquérito das *Fake News* no Brasil[68]. Entretanto, no lugar da automação para a distribuição, existe a possibilidade da automação para a redação das *fake news*.

Figura XIV - Desafios Éticos da Automação

Fonte: Clerwall (2014), Montal e Reich (2017), Ali e Hassoun (2019), Monti (2019) e Sánchez e Ruiz (2020)

Em resumo, o surgimento do jornalismo automatizado vem acompanhado de novos desafios éticos e limitações algorítmicas provenientes da dinâmica entre os diferentes atores. As mudanças tecnológicas exigem discussões de pesquisadores e da categoria profissional sobre como contorná-las. A imprensa como um todo vive há algumas décadas uma transição histórica em função das tecnologias digitais (PAVLIK, 2001). O jornalismo automatizado se apresenta como mais um capítulo no qual novos obstáculos emergem, exigindo soluções sem precedentes.

[67] Ver: Russian trolls who interfered in 2016 U.S. election also made ad money, report says. Disponível em: https://www.nbcnews.com/politics/national-security/russian-trolls-who-interfered-2016-u-s-election-also-made-n1013811

[68] Ver: Inquérito do STF sobre fake news: entenda as polêmicas da investigação que provoca atrito entre Bolsonaro e a Corte. Disponível em: https://www.bbc.com/portuguese/brasil-52824346

4. OS ATORES NO JORNALISMO AUTOMATIZADO

Ao longo dos capítulos anteriores vimos discussões teóricas que explicam o surgimento dos textos automatizados. Disciplinas e áreas do conhecimento tão distintas quanto a Computação, Linguística, Estatística e a Comunicação tiveram que confluir para tornar viável que sistemas, algoritmos e IAs fossem capazes de redigir textos. Essa mistura rica e complexa acabou por desembocar nas redações jornalísticas, que vivem à sua própria maneira um turbilhão de mudanças causadas pelas tecnologias digitais. Na gênese do jornalismo digital, passando pelo jornalismo de dados, chegamos ao jornalismo automatizado. Resta saber na prática, a partir de um caso concreto, como a automação de notícias é feita.

O primeiro projeto de jornalismo automatizado do Brasil foi a cobertura das Eleições Municipais de 2020[69], feita pelo portal G1. O veículo *online* 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. O portal estabelece relação com as cinco redações próprias do Grupo Globo situadas no Rio de Janeiro, em São Paulo, Brasília, Belo Horizonte e Recife, assim como as afiliadas distribuídas por todos os outros estados da Federação.

A Cobertura das Eleições Municipais de 2020 foi feita pelo portal em âmbito nacional. A fim de trazer em detalhes como essa iniciativa foi implementada, esta pesquisa traz o relato dos profissionais responsáveis pelo projeto. Felipe Grandin, Tiago Reis, Hector Iankovski e Rafael Muniz foram os jornalistas e cientistas da computação que viabilizaram esta cobertura automatizada do início ao fim. Cada um deles foi entrevistado por uma média de 1 hora e seus depoimentos são fundamentais para compreender a práxis do jornalismo automatizado. Em paralelo, o produto final dos seus esforços (as notícias) também serão analisados ao longo deste capítulo. Ao todo, **2.966** mil notícias publicadas foram coletadas e irão ter o seu conteúdo analisado mediante uma combinação de duas metodologias. Uma análise de similitude será conduzida para identificar padrões nos textos, enquanto uma análise de cobertura jornalística é feita para compreender as marcas da apuração. Os detalhes destes procedimentos podem ser vistos no capítulo "**2. MARCO TEÓRICO E PROCEDIMENTOS METODOLÓGICOS**". Todos os depoimentos e notícias que

[69] A partir da extensa revisão de literatura, esta pesquisa conclui que nenhum veículo brasileiro havia tocado um projeto de jornalismo automatizado antes do G1.

compõem o *corpus* desta pesquisa, assim como questionários de entrevistas e outros instrumentos, podem ser conferidos no **ANEXO I, ANEXO II**, **ANEXO III**, e **ANEXO IV**, respectivamente.

A partir da fundamentação teórica, alinhada às entrevistas em profundidade com os profissionais do G1, espera-se identificar quais são as principais etapas desta iniciativa de automação de notícias. Os relatos dos executores do projeto permitiram a identificação de **seis atores fundamentais para a geração automática das notícias**, especificamente no caso escolhido para ser aprofundado neste trabalho. 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[70] 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.

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. Ambos iniciaram as primeiras conversas sobre cobertura de eleições municipais em 2019, mas só definiram o escopo do projeto no início de 2020. De acordo com Tiago Reis, a ideia surgiu de uma referência de um veículo estrangeiro, a BBC, e do contato direto com os profissionais de tecnologia da Globo.

> "A gente tinha desde 2019 um contato em reuniões periódicas com a área de Tecnologia lá da Globo. Foi em uma dessas reuniões em que a gente foi apresentar o nosso projeto de eleição. A gente falou: queremos muito fazer esse tipo de coisa com texto automatizado, vocês acham que é possível? E aí em um primeiro momento eles falaram que precisavam estudar para pensar se é possível, como que eles fariam e tal. Então esses primeiros papos começaram ali numa dessas reuniões periódicas que a gente tinha com esses setores de Tecnologia e no começo do ano a gente foi desenvolvendo uma tentativa de protótipo do que seria esse texto" (REIS, 2022)

A referência da BBC que serviu de inspiração para Tiago Reis foi a cobertura das eleições gerais do Reino Unido em 2019[71]. A eleição ocorreu no dia 12 de dezembro e as contagens dos votos foram feitas ao longo da noite, até a madrugada do dia 13. Ao todo, 690

[70] 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.

[71] Ver "*General Election 2019: How computers wrote BBC election result stories". Disponível em: https://www.bbc.com/news/technology-50779761*

notícias foram publicadas. Segundo o anúncio do veículo, todos os artigos foram checados por um editor humano antes de serem publicados. De acordo com o responsável do projeto, Robert McKenzie, "o projeto foi sobre fazer um jornalismo que não poderia ser feito somente com humanos" (tradução nossa)[72]. Essas duas características se repetiram no projeto idealizado por Tiago Reis, no G1.

Dentro da equipe editorial do 'núcleo de dados, *fact-checking* e projetos especiais' Tiago Reis trabalhou lado a lado com o então repórter Felipe Grandin. Os dois participavam das reuniões quinzenais com a equipe de ciência de dados e trabalhavam na "definição do *template*", ou em "fazer um texto padrão". Anos mais tarde, em 2022, Tiago Reis já havia deixado a Globo, enquanto Grandin coordenava um segundo projeto de geração automática de textos, dessa vez para as eleições presidenciais e legislativas do mesmo ano.[73] O relato de Grandin sobre o projeto piloto de 2020 e sua sequência em 2022 permite identificar a progressão do jornalismo automatizado dentro da Globo, mesmo que em caráter seminal.

Em 2020, a equipe editorial era composta, portanto, de duas pessoas, um repórter e um editor que tiveram uma tarefa dupla. Primeiro, definir o *template,* um conceito que iremos abordar em detalhes nos próximos subcapítulos. Concomitantemente, Tiago Reis e Felipe Grandin também definiram 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. Nas palavras de Felipe Grandin, essas bases eram duas.

> "Tudo isso são informações que tinham sobre os candidatos na base do TSE, que é basicamente o Divulgacand. É uma base do TSE que é onde você vê as informações dos candidatos. Tem as outras bases que são de resultados, que daí é o principal de onde a gente puxou quem foi eleito e quem não foi eleito, quantos votos teve e como ficou o percentual" (GRANDIN, 2022)

Essas duas bases de dados foram escolhidas por sua robustez técnica, confiabilidade, noticiabilidade e formatação homogênea[74]. As duas bases escolhidas são administradas pela mesma instituição: o Tribunal Superior Eleitoral (TSE). A primeira é chamada oficialmente

[72] "This is about doing journalism that we cannot do with human beings at the moment," said Robert McKenzie, editor of BBC News Labs.

[73] Ver "Textos e vídeos automatizados, jogos eleitorais, propostas, pesquisas, números e Fato ou Fake: as páginas especiais do g1 nas eleições". Disponível em:https://g1.globo.com/politica/eleicoes/2022/noticia/2022/10/03/textos-e-videos-automatizados-jogos-eleitorais-propostas-numeros-e-fato-ou-fake-as-paginas-especiais-do-g1-nas-eleicoes.ghtml

[74] 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".

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 de dados é o “Resultados”, uma coleção de arquivos que podem ser acessados por recorte de ano, região, pleito e tipos de cargos. As duas bases estão conformadas dentro do “Portal de Dados Abertos do TSE”, que disponibiliza 143 conjuntos de dados ao todo sobre os pleitos realizados no Brasil desde 1933.

As informações identificadas por Tiago Reis e Felipe Grandin como relevantes (ver capítulo **4.2)** , dotadas de “critérios de noticiabilidade”, entraram no *template*. As duas bases precisaram ser “concatenadas”[75], 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. Tanto Hector Iankovski, quanto Rafael Muniz, estavam lotados em diferentes departamentos dentro da estrutura organizacional da empresa, porém, colaboraram de maneira próxima na parte técnica de geração de textos. Hector Iankowski fez o papel de “gerente de projeto”, em suas palavras. O editor Tiago Reis confirma a atuação de Iankowski como “a pessoa que ficou responsável por pensar nessa parte de Processamento de Linguagem Natural. Ele era meu ponto focal na Tecnologia”.

O trabalho da equipe de tecnologia feito pela dupla Hector Iankowski e Rafael Muniz pode ser resumido em três partes. Primeiro, colher os dados indicados pela equipe editorial e tratá-los. Segundo, desenvolver o (4) algoritmo[76] de geração dos textos em linguagem de programação, para realização do *template*. Terceiro, distribuir o *output* desse algoritmo para o (5) o sistema de gerenciamento de conteúdo. Essas três etapas englobam o processo de Geração de Linguagem Natural no jornalismo, segundo Diakopoulos (2019). No caso do G1, esse procedimento resultou na pré-publicação das notícias, com mais uma etapa pendente para a distribuição efetiva dos textos para o público.

Por meio do relato de Hector Iankovski enquanto programador do algoritmo, assim como de Rafael Muniz na arquitetura dos dados, é possível descrever com mais detalhes a

[75] Concatenação é um termo usado em computação para designar a operação de unir o conteúdo de duas cadeias de textos

[76] O algoritmo é mencionado na fala dos entrevistados ora como sistema, ora como gerador, ora como script, ora como código. Todos esses conceitos possuem pontos de contato, mas podem significar por vezes significar coisas distintas. As notas dos seguintes capítulos podem elucidar algumas dessas dúvidas.

parte computacional do jornalismo automatizado. O jargão usado pela dupla de desenvolvedores foi a chave para especificar as partes do processo. Para acompanhar a linguagem dos entrevistados, esta pesquisa recorreu a dois anos de estudos de Teoria da Computação e Análise de Dados, feitos em paralelo à dissertação. O trabalho de tradução dos conceitos é uma das metas estabelecidas nos objetivos específicos expostos no capítulo "1.. Explicar a terminologia técnica, como o que é "*pipeline*', '*data lake*', 'ingestão de dados' , 'disparo de eventos' e outros, é a forma que esta investigação encontrou para fazer o *Inventário* do algoritmo e esclarecer o seu *modo de existência* (LEMOS, 2020; SIMONDON, 1980).

A partir das entrevistas realizadas, percebeu-se que a atuação da (3) 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 é 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.

De acordo com Hector Iankovski, o produto final do algoritmo era um arquivo postado dentro do sistema de gerenciamento de conteúdo, chamado dentro da Globo de "CMA"[77]. Esse arquivo estava acompanhado de *tags,* um tipo de identificador presente em todo conteúdo *web*, que pode ter tanto uma função estilística, como marcar o espaço entre um título e um texto, quanto organizacional, como localizar aquele conteúdo dentro de uma estrutura do site. No caso dos arquivos das mais de 5 mil notícias, as *tags* serviram para organizar os arquivos dentro do (5) sistema de gerenciamento de conteúdo. "O nosso *output*, a nossa publicação, era um arquivo que a gente gerava para o sistema de publicação" (IANKOVISKI, 2023). O editor Tiago Reis corrobora a dinâmica entre o algoritmo de Geração de Linguagem Natural e o *software* de conteúdo. "O G1 tem um publicador que se chama CMA, onde você coloca o texto lá, aperta o botão e publica. A gente precisava de uma interface ainda de como esse texto ia entrar no publicador" (REIS, 2022).

A partir dessa integração entre o (4) algoritmo e (5) o sistema de gerenciamento de conteúdo retornava o produto jornalístico de volta às mãos humanas. As *tags* funcionam

[77] Os tipos de *software* para publicação *web*, chamados de *Content Management System (CMS)*, também são referidos por alguns fornecedores como *Content Management Application (CMA)*. A rede Globo usa a segunda terminologia.

como o sinalizador de qual afiliada, ou filial da Rede Globo, ficaria com a sua remessa de notícias.

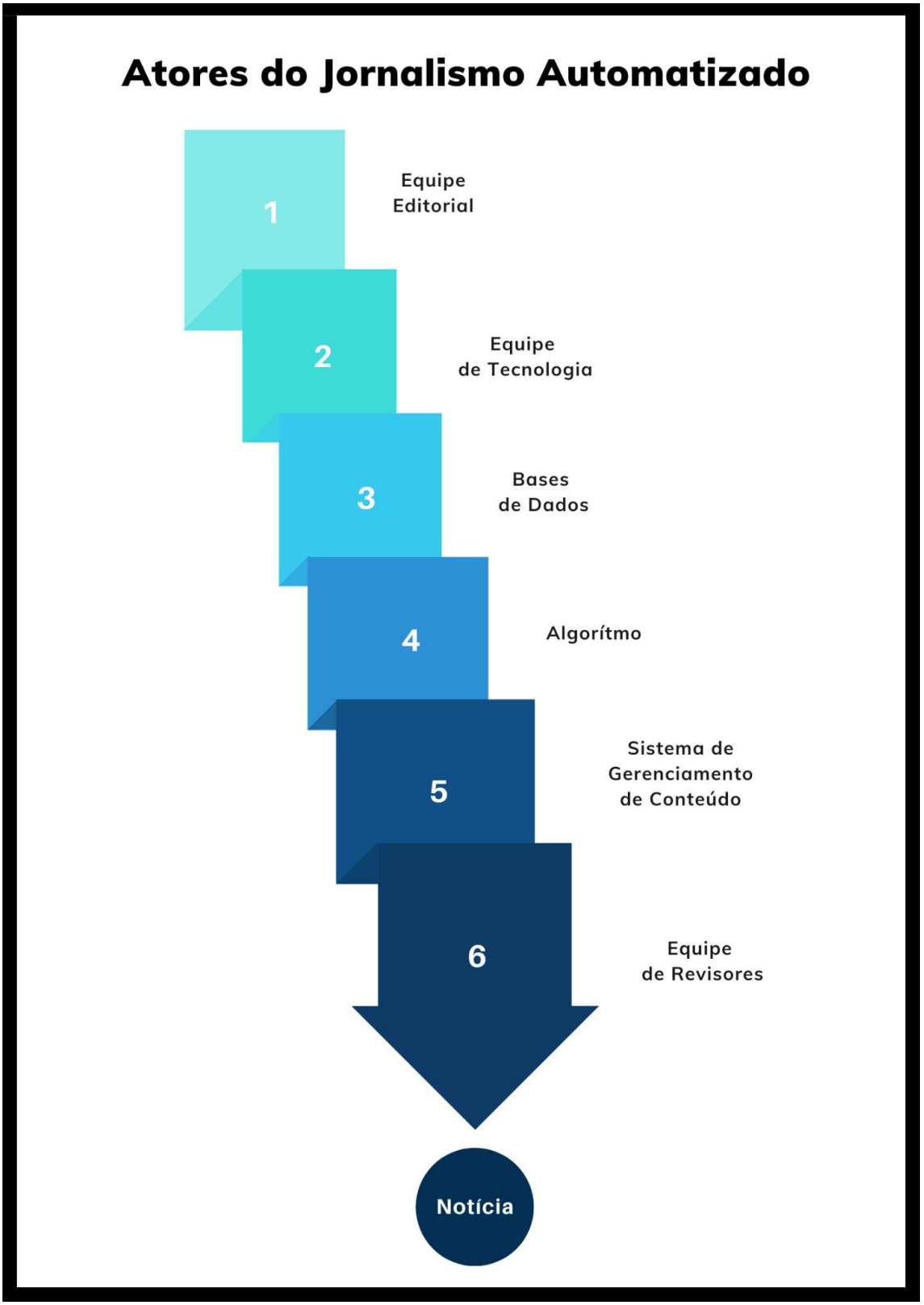
> "A gente acordou com eles, qual era essa formatação, se eu não me engano eram umas 8 ou 9 *tags*, com informações que eram pertinentes para eles fazerem essa vinculação dentro do CMA. A primeira informação, por exemplo, era um código da cidade. A segunda era o nome da cidade. A terceira era sobre o estado, porque ali dentro do CMA eles faziam uma distribuição" (IANKOVISKI, 2023)

A partir dessas 8 ou 9 tags, o (5) sistema de gerenciamento de conteúdo era capaz de disponibilizar a notícia gerada (título, linha fina e texto) para o jornalista da afiliada correspondente. Eles recebiam um conjunto de notícias para as cidades que faziam parte da sua área de cobertura. Com a distribuição feita dentro do sistema, as 5 mil notícias pré-publicadas surgiram nas telas dos membros da (6) equipe de revisores. Jornalistas de todos os estados do Brasil, cumpriram plantão no dia das eleições para checar e publicar manualmente os textos. Segundo Tiago Reis, a equipe "teve que fazer uma operação de guerra envolvendo todas as afiliadas do G1. A gente envolveu mais de 200 jornalistas pelo Brasil, porque temos afiliadas em todos os estados e alguns estados mais uma afilhada". Esses profissionais ficaram incumbidos de conferir e publicar dezenas de notícias, em alguns casos centenas de notícias, mas com a responsabilidade de fazer inclusões de informação. Em alguns casos, os textos foram complementados, ou mesmo descartados, em detrimento de uma notícia mais detalhada. No subcapítulo posterior "**4.4. TEXTOS ESCRITOS POR MÁQUINAS NAS ELEIÇÕES DE 2020**" a análise do conteúdo das notícias torna latente os casos que se afastam do padrão gerado pelo algoritmo.

Os relatos dos quatro entrevistados, acrescidos da investigação dos textos, habilitaram a pesquisa a identificar esses seis atores. São eles: (1) equipe editorial, (2) equipe de tecnologia, (3) as bases de dados, (4) o algoritmo, (5) o sistema de gerenciamento de conteúdo e a (6) equipe de revisores. Os depoimentos vão de encontro às etapas de automação de texto apresentadas na fundamentação teórica, como Reiter (2012), Mc Donald (2010), Diakopoulos e Nicholas Dorr (2015)[78]. A visão geral do projeto apresentada aqui é o ponto de partida para aprofundar na contribuição de cada um desses actantes. Segundo essa descrição do caso do G1, foi possível elencar aspectos práticos do jornalismo automatizado.

[78] Para esses autores, a etapa de revisão e complementação dos textos não é fundamental, mas é cabível.

Figura XV - Atores no Jornalismo Automatizado presentes no caso do G1



Fonte: Esta pesquisa

As três equipes e os três atores não-humanos (bases, algoritmo e sistema) elencados acima tiveram que trabalhar de forma coordenada para entregar uma notícia para cada cidade do Brasil, no dia das eleições. Embora não seja possível determinar uma só origem para o projeto, a automação só foi possível graças à vontade de jornalistas que estavam inteirados sobre desenvolvimentos tecnológicos que poderiam impulsionar o seu trabalho. No subcapítulo seguinte temos a perspectiva desses profissionais sobre a automação de notícias.

4.1 O TRABALHO EDITORIAL NO JORNALISMO AUTOMATIZADO DO G1

De acordo com o editor responsável pelo núcleo de dados, *fact-checking* e projetos especiais do G1 em 2020, Tiago Reis, a idealização do projeto surgiu de uma referência tirada de um veículo estrangeiro. Já de acordo com o repórter de dados, Felipe Grandin, o ponto de partida foi a familiaridade que a equipe tinha com projetos especiais de cobertura eleitoral. Duas iniciativas internas tocada pela equipe de Tiago Reis e Felipe Grandin, chamadas de "Jogo Eleitoral"[79] e "Fato ou Fake"[80], foram mencionadas como entregas que já eram feitas e que, portanto, serviram de ponto de partida. Ambas as iniciativas já empregavam técnicas do jornalismo de dados com uma cobertura eleitoral ampla. O Jogo Eleitoral, por exemplo, é uma ferramenta de *newsgame* que postula dez perguntas para o usuário e retorna uma sugestão de qual candidato mais se adequa ao perfil. Parte dos dados oficiais dos candidatos provèm da mesma fonte: a base do TSE.

O cientista de dados, Hector Iankovski, conta outra história. Segundo ele, tudo começou com uma conferência de mídia global, na qual cientistas da computação e jornalistas da Globo se encontraram e partilharam suas expertises. A partir dessa interação, seus predecessores na equipe de tecnologia criaram um projeto piloto que não chegou a ser publicado. "Eu soube que existia uma outra área na Globo que também fazia alguns projetos de inovação e experimentação. Eles fizeram um projeto piloto de geração automatizada de textos. Por coincidência, duas áreas distintas da Globo foram até um congresso em Amsterdam [...] Eles participaram dessa conferência e a área que ficava envolvida mais com o jornalismo achou o projeto interessante" (IANKOVISKI, 2023). Para o cientista, áreas distintas da mesma corporação precisaram se encontrar neste congresso para a ideia surgir, indicando a interdisciplinaridade da iniciativa desde a origem.

[79] Ver "Jogo Eleitoral". Disponível em: https://especiaisg1.globo/politica/eleicoes/2022/jogo-eleitoral/
[80] Ver "Fato ou Fake". Disponível em: https://g1.globo.com/fato-ou-fake/eleicoes/

Não há um relato único de como o projeto se iniciou, mas a fala dos entrevistados explicita duas origens. Primeiro, a equipe editorial estava habituada a trocas com a equipe de tecnologia, assim como ambas contavam com um vocabulário em comum. “Esses primeiros papos começaram ali numa dessas reuniões periódicas que a gente tinha com o setor de Tecnologia e no começo do ano e a gente foi desenvolvendo uma tentativa de protótipo do que seria esse texto”(REIS, 2022). A partir dessas conversas, o conceito "NLP" surgiu e a equipe editorial instigou a equipe de tecnologia a estudar a viabilidade da automação de textos. A questão do jargão técnico partilhado pelas equipes também surgiu no relato dos jornalistas Tiago Reis e Felipe Grandin.

> “Isso ajudou a conversar com eles e a entender tudo muito mais fácil. As conversas eram muito mais fluidas e a gente percebia isso nas próprias reuniões quinzenais, porque o pessoal que participava dos outros setores da Globo, não era da área de dados e não tinham essa expertise [...] então com certeza esse contato nosso com a linguagem de programação ajudava muito nessas conversas” (REIS, 2022)

O repórter Felipe Grandin também comenta que essa trajetória no jornalismo de dados facilitou o diálogo. Ambos os profissionais empregam o que o pesquisador Márcio Carneiro dos Santos (2018b) enquadra como uma *inteligência híbrida*, ou seja, provinda da confluência de duas áreas do conhecimento. O hibridismo destes profissionais é o que permite dinâmicas de trabalho aparentemente distintas, ciência da computação e jornalismo político, se encontrem para gerar uma síntese. O segundo ponto que viabilizou o projeto, por parte dos jornalistas, é que já havia uma familiaridade da equipe editorial com os dados eleitorais, assim como o *modus operand*i do TSE na publicação dessas informações.

> “A primeira ação na verdade foi eu mandar para eles quais eram as informações que a gente ia precisar e de que fontes de dados. Então assim, a gente precisou olhar todas as informações possíveis que estão lá no TSE do candidato, a gente queria também acessar as informações do IBGE. Eu fui listando ali algumas fontes que a gente poderia usar” (REIS, 2020)

Se a fase de idealização surgiu com os conceitos Cobertura Eleitoral e Processamento de Linguagem Natural convergindo, o próximo passo foi definir as fontes de informação. A equipe editorial decidiu quais fontes eram confiáveis, quais dados eram necessários, o que ficaria no texto e o que sairia[81]. Essa série de escolhas, acrescida da ordem em que as

[81] Sabemos pelo relato de Tiago Reis que as primeiras versões do texto traziam dados demográficos sobre cada cidade, advindos do IBGE, além de vários dados sobre cada candidato. A versão inicial teria ficado muito longa e, como é comum no jornalismo digital, textos menores são preferíveis. Porém, está pesquisa não sabe ao certo quais informações ficaram de fora, por não ter tido acesso aos rascunhos iniciais.

informações adentraram a notícia, se conformam dentro do processo de **(1) Planejamento do Documento**, como a primeira etapa de qualquer processo de Geração de Linguagem Natural segundo McDonald (2010) e Reiter (2012). Neste ponto é que se define a 'intenção do falante' quanto ao seu 'algo a dizer'. Para o jornalista Tiago Reis, o planejamento do documento era pensar em versões que dessem conta de reportar um resultado eleitoral, a partir das fontes de dados que ele estava habituado. A filtragem das possibilidades foi o exercício principal nesta etapa, segundo o editor.

> "A gente queria também acessar as informações do IBGE. Eu fui listando ali algumas fontes que a gente poderia usar. Depois, você vai ver que no fim a gente basicamente usou só as informações do TSE. A gente resolveu simplificar porque ia ser muito complicado ter várias fontes ali ao mesmo tempo" (REIS, 2022)

É explícito como somente bases do Governo foram consideradas noticiáveis pela equipe. "No começo a gente queria mostrar em quais bases públicas a gente poderia conseguir as informações" (REIS, 2022) Apesar de Institutos de Pesquisa e consultorias também gerarem dados sobre eleições, nenhuma dessas fontes foi cogitada. Enquanto os jornalistas decidiam quais informações eram credíveis e noticiáveis, se impunha uma necessidade de adaptar o conteúdo às limitações da inovação. Quando Tiago Reis fala sobre "simplificar" é em razão deles não estarem habituados àquela tecnologia específica, pois o processo poderia ficar "complicado demais".

Há um grau de desconhecimento que guiou a ação da equipe com uma dose de cautela. Essa ideia se repete no relato de Tiago Reis em vários momentos. "A gente fez ali o que era possível fazer na primeira vez e já era muita coisa, já era um esforço de inovação" (REIS, 2022). Por "inovação" não há somente a interpretação de pioneirismo, ou do sentido positivo de "novidade", mas também no sentido negativo, de como os problemas que surgem e se amontoam de maneira imprevisível. Pegar dados de mais de uma base era algo que os jornalistas estavam habituados em seu dia a dia, por exemplo em projetos do "Jogo Eleitoral" ou "Fato ou Fake", porém, quando a Geração de Linguagem Natural se apresentou pela primeira vez frente à equipe editorial, eles parecem ter escolhido a via da precaução.

Ao retornar ao exercício de "simplificar" e "escolher" o que seria publicado a partir das bases de dados, Tiago Reis e Felipe Grandin explicam que, no projeto do G1, atuaram como filtros nessa etapa dentro do processo de automação das notícias. Este aspecto tem conexão com um ponto que surge na literatura especializada sobre jornalismo de dados: a mudança de paradigma na profissão. Se tradicionalmente o jornalista agia como um caçador

de informação, hoje o profissional opera como um filtro de informações já publicadas. A mudança de paradigma diz respeito a um contexto de escassez de dados, que muda drasticamente para um cenário de abundância. Meyer (1991, p.1) coloca nominalmente no fluxo de trabalho deste novo jornalista o dever de “filtrar as informações e transmiti-las", a fim de lidar com o “crescimento explosivo de informações disponíveis". Este aspecto que marca o surgimento do jornalismo de dados se mantém na linha de base do jornalismo automatizado. Outro ponto pertinente é que no exercício de filtrar as informações, os jornalistas estavam pré-definindo a apuração que seria executada pelo algoritmo. Ou seja, estavam agindo no primeiro nível de uma cobertura jornalística, de acordo com Silva e Maia (2011), estabelecendo as **marcas de apuração** da notícia.

Depois de fazerem este trabalho de seleção, Tiago Reis e Felipe Grandin partiram para o próximo passo. Os jornalistas já haviam esmiuçado as bases de dados públicas que achavam pertinentes e noticiáveis, veio a vez criar um modelo de texto, chamado de *template.* Vale explicar em detalhes o que é o *template*, pois é neste ponto que os relatos de Hector Iankovski e Tiago Reis se encontram, reunindo tanto a tecnicidade da geração de textos, quanto o trabalho editorial. Das diversas formas de Geração de Linguagem Natural, o uso de expressões-chaves é um dos caminhos apresentados por Dong *et al* (2022). Com este método é possível realizar texto-para-texto, ou dado-para-texto como método para **expansão de texto[82].** 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. Tudo isso é definido em sequência ao (1) Planejamento do Documento, na fase de (2) Microplanejamento. O *template* expressa, portanto, as escolhas linguísticas que o algoritmo GLN terá de executar para produzir o resultado desejado.

A definição deste ‘modelo de texto’ abarca tanto o que foi filtrado das bases de dados, o estilo editorial do veículo, como as possibilidades (somadas aos entraves) que a cobertura pode encontrar. Para Tiago Reis e Felipe Grandin, o *template* era o território de contribuição para a equipe de tecnologia. Quando questionado sobre essa etapa do trabalho, Tiago Reis explica que é neste ponto que o trabalho editorial passa a ser um ponto de partida para os cientistas da computação.

[82] Para mais informações sobre as formas de Geração de Linguagem Natural, ver o capítulo “**3.3. LINGUAGEM E COMPUTAÇÃO”**

> "A gente foi conversando a partir desse modelo que eu mandei, desse *template*. Então eu mandei esse rascunho, no e-mail mesmo, com essas várias lacunas do que eu achava que se deveria preencher. No começo, como eu falei, era um texto bem maior com várias informações [...] Tinha algumas informações a mais que a gente acabou não usando. E aí a gente foi burilando isso junto com o pessoal da tecnologia para ver o que fazia sentido e o que era mais fácil fazer. A partir dali a houve mais uma preocupação de como funcionaria esse Processamento de Linguagem Natural no texto, ali para mudar feminino e masculino, para mudar algumas construções do texto" (REIS, 2022)

O repórter Felipe Grandin assegura que o trabalho da equipe foi de "construção do texto", com o intuito de que o *template* gerasse um resultado "idêntico" ao que um humano produziria.

> "Então assim, o esforço todo na elaboração do *template*, foi para que não houvesse essa diferença entre essa matéria e uma matéria que a gente fosse escrever normalmente sobre o resultado. É lógico que tem coisas que a gente talvez fosse escrever na própria matéria e que não tinham na base do TSE. Por exemplo, alguns detalhes do que aconteceu no dia, ou uma aspa de alguém. Isso não teria como incluir, mas o texto ali factual é uma matéria normal do G1" (GRANDIN, 2022)

Quando Felipe Grandin fala sobre extinguir qualquer "diferença" entre o texto automatizado e o escrito por uma repórter, ele suscita uma questão de estilo. É premente que a notícia em sua dimensão estética esbarra em escolhas literárias. O estilo que um repórter pode empregar em seu texto afeta o senso de objetividade da reportagem, ou corrobora. Essas escolhas são em um primeiro momento guiadas pelo lide, mas em um segundo momento são retomadas pelos manuais de redação, que variam de veículo para veículo. Tiago Reis enfatiza que essas pequenas decisões foram parte do processo. "Tudo passou por mim, se tinha zero depois da vírgula, se era caixa alta ou baixa, se iria usar o cifrão, se o patrimônio viria com vírgula mais x's zeros depois" (REIS, 2022). Dentro da interdisciplinaridade da automação de notícias, o estilo do texto era uma incumbência da equipe editorial. Falar sobre estilo e escolhas tão pequenas quanto usar um ponto ou uma vírgula como separador, parece entrar em pormenores e detalhes dispensáveis, mas é esse o tipo de transferência de autoridade (HARARI, 2018) que se estabelece entre atores humanos e não-humanos.

A questão do estilo e da formatação presente na fala de Tiago Reis também revela o segundo nível de análise da cobertura jornalística. Segundo Silva e Maia (2011) essas decisões expressam as **marcas da composição do produto.** Neste caso, as marcas de composição do produto são transferidas de um humano, para o algoritmo, chegando por fim na notícia. Tudo isso intermediado pelo *template*. Felipe Grandin destaca em seu relato como a formatação dizia respeito a como o dado entraria no *template*.

> "Então por exemplo, vai fazer o texto, tem que estar lá indicado no texto qual dado vai entrar. Vai ser o candidato a prefeito, vai ser o mais votado. Tinha que ter uma indicação ali de qual dado vai entrar. Qual a idade, qual é a UF (Unidade da Federação), qual é o município, tinha que ter ao longo do texto, entre colchetes, qual o dado que vai entrar e quais são as possibilidades"(GRANDIN, 2022)

É possível inferir que essas "possibilidades" de como o dado entraria, assim como as questões de pontuação e tamanhos de letras elencadas por Tiago Reis, se enquadram dentro do processo de **(2) Microplanejamento** da Geração de Linguagem Natural (MCDONALD, 2010; REITER, 2012). O desafio da formatação se mostra no panorama feito pelo editor do projeto como uma pauta constante nos encontros com a equipe de tecnologia. "Tinha algumas coisas que para eles não faziam sentido, de formatação de texto mesmo. A gente teve essa conversa para poder chegar num texto no padrão do G1 e no padrão jornalístico, que fizesse sentido", relembra Tiago Reis. As conversas sobre a padronização do texto são recorrentes nessas reuniões e isso, se repete, segundo Tiago Reis, nos testes que foram conduzidos com o algoritmo de geração. "Antes da eleição, claro, eles fizeram umas simulações e mandaram para a gente dar uma olhada. Eu dividi com a minha equipe que era o Grandin, a Clara e a Gabi. Os três eram mais de jornalismo de dados mesmo. Então os três ali me ajudaram a revisar o texto para ver se tinha alguma coisa muito estranha e para dar sugestões. Então teve esse trabalho também".

O quesito salientado por Gradin é que a formatação também é um ponto de contato entre *template* e base de dados. Essa sobreposição é apresentada como a "limpeza". O repórter coloca a padronização dos valores na base como uma questão fundamental, que por sua vez permite que símbolos e palavras sejam ajustados para a forma com que a Globo costuma redigir as notícias.

> "Nesse ponto a qualidade da base é fundamental. Se você tiver uma base suja, você teria que fazer outro processo. É que assim, no caso da base do TSE você já faz um processo de formatação dos dados antes. A estruturação dos dados também acontece antes [...] tem um processo ali de extrair os dados que a gente quer e formatar do jeito que a gente quer, para depois poder jogar para o algoritmo" (GRANDIN, 2022)

Felipe Grandin reitera que a limpeza, tanto enquanto processo de formatação, como etapa de pré análise dos dados, é de suma importância.

> "Você já ter o dado limpinho ali, não precisa daquele processo de ter um processo de limpeza enorme entende? Deixar os nomes iguais, ver se não tem um 'NaN'[83] ali diferente, ou campo vazio, que não era para ter. Então com certeza ajuda muito você ter uma base limpa. Você minimiza muito o erro (...) O mais difícil de trabalhar com bases de dados é a limpeza." (GRANDIN, 2022)

Em algumas partes do seu relato, Gradin destaca que a qualidade da base no quesito "limpeza" é um dos fatores que viabiliza a automação de notícias. Outras fontes de informação, como a base DataSUS, já eram de familiaridade da equipe, mas não apresentavam a qualidade da limpeza. Os dados aparecem ora com um divisor, ora com outro, ora com uma pontuação, ora com outra. Essa inconstância é um entrave tanto para a estruturação dos dados, quanto para a automação das notícias.

> "Ah, sim. Se não fosse por uma base boa, a gente nem ia fazer. Pensa que podem ter vários problemas [...] O TSE tem um sistema bem robusto, então eles vão mandando os dados e os dados vão chegando certinho, não cai a conexão e não chega pela metade. Esse tipo de coisa acontece sem erro. Então é bem confiável. Isso é uma coisa, a outra coisa é a base em si. A base do TSE é bem limpinha. Não tem coisas fora de formatação, entendeu? Não tem erros ali dentro que você não identifica na hora. Então ela é uma base bem padronizada, bem formatada. Ela é bem limpa e você pode trabalhar direto com ela" (GRANDIN, 2022)

Neste ponto, vale relembrar que Tiago Reis foi o editor do projeto em 2020, no primeiro momento em que o G1 publicou uma cobertura automatizada, porém, na eleição seguinte, Tiago Reis havia deixado o veículo e Grandin foi responsável pela continuidade do projeto. Em 2022, o pleito ocorreu para o Legislativo e o Executivo, com uma publicação por cidade, indicando a proporção dos votos da população para os candidatos. A pertinência de mostrar a experiência de Grandin na função que foi exercida, em um primeiro momento, por Tiago Reis, se dá em contextualizar as suas falas sobre o jornalismo automatizado.

Tanto Tiago Reis, quanto Felipe Grandin, partilharam de atuações profissionais muito similares ao longo dessas duas iniciativas de geração automática de notícias. O "modelo de texto", "*template*", "rascunho" e "definição do modelo" são expressões que se repetem na narrativa de ambos. Ao retornar a questão do *template*, fica latente uma das preocupações centrais da equipe editorial: antecipar os cenários possíveis dentro daquela cobertura. A fase de **(1) Planejamento de Documento** (MCDONALD, 2010; REITER, 2012) seria melhor descrita aqui como um planejamento de *templates*, pois a equipe teve que prever um número de cenários possíveis. Como fica evidente na explicação de Tiago Reis, uma eleição pode ter

[83] NaN é uma sigla em inglês para *Not a Number* (não é um número). É um termo corrente da ciência de dados que se refere a dados faltantes. Quando se tem uma célula vazia, é de praxe indicar que ali não tem nenhum valor com esta sigla.

nº número de desfechos e eles precisavam de um *template* para cada situação. “Foram quatro rascunhos. Um para a vitória em primeiro turno. Empate com decisão só no segundo turno. Candidatura impugnada e uma para quando estivesse aguardando decisão do TSE”(REIS, 2010). Felipe Grandin corrobora essa ideia ao mostrar alguns dos casos e excessões que surgem em uma cobertura eleitoral

> “Tinham questões sobre se estivesse impugnada a candidatura. Como é que avisa? Tem outras variáveis que a gente tem que prever o que vai acontecer. Uma candidatura pode não estar valendo por vários motivos. A gente juntou todas elas [...] Essas informações tem na base do TSE, aí tem uma frasezinha para cada uma. Daí o texto era publicado com a devida ressalva” (GRANDIN, 2023)

O aspecto antecipatório, ou preditivo, é uma marca da automação industrial, segundo Groover (1980), que parece se repetir na definição do *template* na geração de textos. Dada uma quantidade finita de insumos e um número **x** de produtos que se almeja construir, deve se determinar **x** caminhos dentro da linha de produção, livres de perturbações. Dessa forma, o conhecimento que a equipe editorial possuía da base do TSE foi instrumental para compreender que havia quatro cenários possíveis, previstos nas linhas e colunas da fonte, portanto, esses quatro cenários exigiam quatro *templates*.

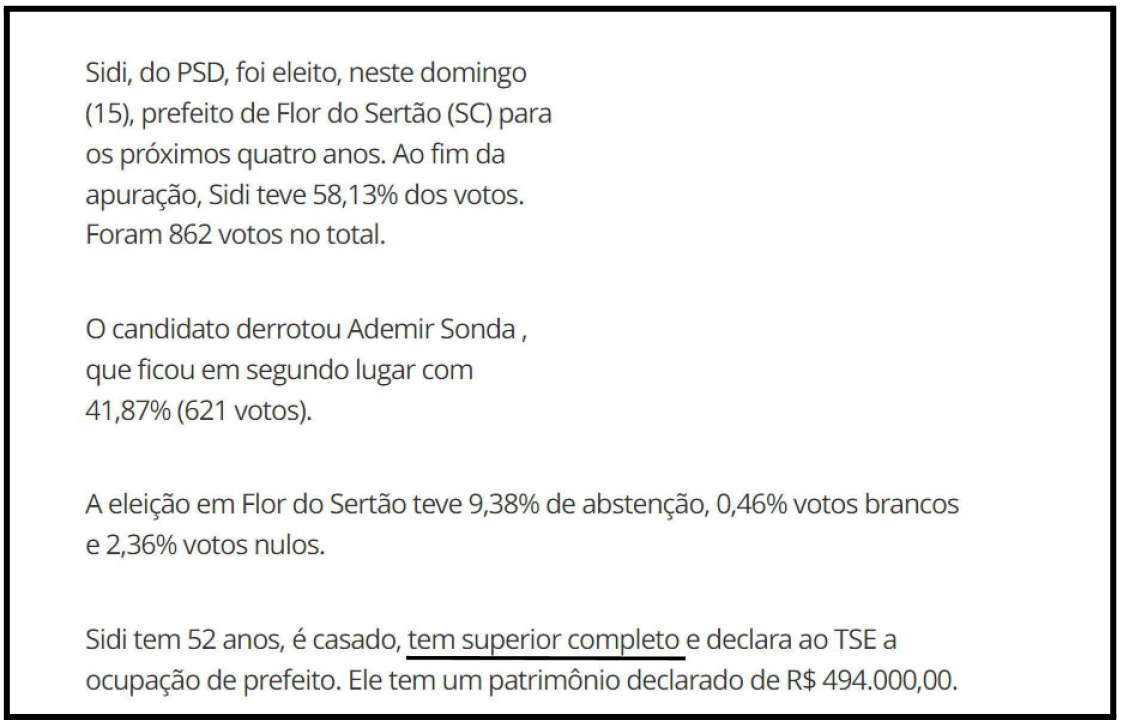
O trabalho de prever as possibilidades não difere de uma delimitação de escopo da cobertura. Um conceito que atravessa o *template* e a notícia gerada, com o algoritmo no meio do caminho, é a árvore de decisão**,** ou árvore de mensagens (REITER, 2012). “Então, você tinha que prever, por exemplo, se ganhou no primeiro ou no segundo turno, se é homem ou é mulher. Tudo isso é feito por uma árvore de decisão. Se ganhou no primeiro turno, vai ser esse texto aqui, se não, vai ser o outro. Aí vai ter essa proposição aqui, ou aquela proposição ali” (GRANDIN, 2022). No esclarecimento de Felipe Grandin, os quatro *templates* mencionados por Tiago Reis funcionam como galhos. As decisões que são tomadas no nível das frases e expressões podem ser compreendidas como ramos desses galhos (ver **Figura IV** para exemplo).

As escolhas feitas no nível das sentenças são exemplificadas pelo editor na descrição dos candidatos. Uma das informações presentes no repositório 'Divulgacand' era a descrição do grau de escolaridade dos concorrentes.

> “A gente percebeu que lá na base não tinha só se era fundamental e superior completo. Tinha também se só lê e escreve. Dai tinha que mudar a palavra que acompanha, por que não dá para ficar ‘tem lê e escreve’. Então tinha várias informações que a gente tinha que ter noção de tudo que poderia vir da base do TSE” (REIS, 2022)

Essa explicação retoma o princípio da antecipação, mas apresenta um caso concreto de **(2) Microplanejamento** na Geração de Linguagem Natural (REITER, 2012). Por uma questão de concordância, o nome do candidato não poderia vir acompanhado do verbo "ter" se referir ao "lê e escreve", embora esse verbo fosse adequado para todas as construções, como pode ser visto no exemplo abaixo.

Figura XVI - Notícia com resultado da eleição no município Flor do Sertão (SC)



Sidi, do PSD, foi eleito, neste domingo (15), prefeito de Flor do Sertão (SC) para os próximos quatro anos. Ao fim da apuração, Sidi teve 58,13% dos votos. Foram 862 votos no total.

O candidato derrotou Ademir Sonda , que ficou em segundo lugar com 41,87% (621 votos).

A eleição em Flor do Sertão teve 9,38% de abstenção, 0,46% votos brancos e 2,36% votos nulos.

Sidi tem 52 anos, é casado, tem superior completo e declara ao TSE a ocupação de prefeito. Ele tem um patrimônio declarado de R$ 494.000,00.

Fonte: Esta pesquisa

Para contornar esse desafio da concordância, o *template* deveria prever algum tipo de variabilidade que tratasse dessa condição. Como por exemplo substituir o "ter" pela palavra "sabe", toda vez que o algoritmo encontrasse "lê"[84], porém, não foi isso o que aconteceu. No exemplo abaixo, fica visível que o erro de concordância que Tiago Reis menciona foi publicado. Esta pesquisa identificou que esse equívoco se repete **15** vezes ao longo de toda a

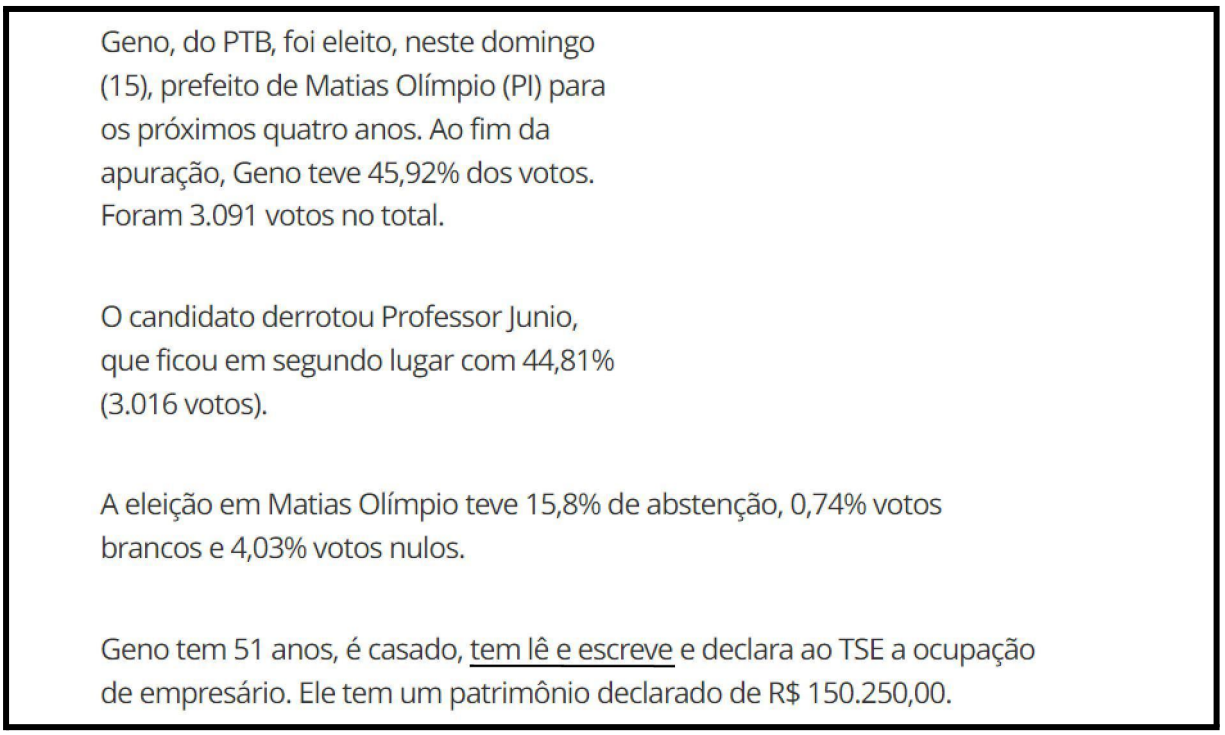
[84]A pesquisa identificou que esse tipo de reconhecimento de palavras, ou partículas, é geralmente feita com um método chamado Regular Expressions, ou pela abreviação regex. Em Python, por exemplo, isso poderia ser feito com as duas linhas de código abaixo:

```
pattern = r"lê"

matches = re.findall(pattern, input_string)
```

cobertura do primeiro turno (menos de 0,3% dos casos) . Em algumas cidades, como na eleição para Pavussu (PI), o verbo “tem lê e escreve” de fato foi substituído por “é alfabetizado”, sem causar problemas de concordância. Não é possível determinar, para além de meras especulações, por que esse erro foi gerado e publicado. Infelizmente, esta pesquisa só identificou o erro após as entrevistas com os integrantes do projeto. Durante os depoimentos, os entrevistados se mostraram ignorantes de que algum equívoco havia sido gerado e publicado de maneira recorrente.

Figura XVII - Notícia com resultado da eleição no município Flor do Sertão (SC)



Geno, do PTB, foi eleito, neste domingo (15), prefeito de Matias Olímpio (PI) para os próximos quatro anos. Ao fim da apuração, Geno teve 45,92% dos votos. Foram 3.091 votos no total.

O candidato derrotou Professor Junio, que ficou em segundo lugar com 44,81% (3.016 votos).

A eleição em Matias Olímpio teve 15,8% de abstenção, 0,74% votos brancos e 4,03% votos nulos.

Geno tem 51 anos, é casado, tem lê e escreve e declara ao TSE a ocupação de empresário. Ele tem um patrimônio declarado de R$ 150.250,00.

Fonte: Esta pesquisa

Havia uma outra escolha que deveria ser estabelecida no nível da frase, segundo Tiago Reis, que foi bastante desafiadora. A expressão referente a profissão de cada candidato poderia exigir preposições e verbos diversos.

> "A informação da profissão foi bem complicada. A gente teve que ver todas as profissões possíveis para ver se tinha alguma ali que fugia do padrão e ficava estranho. Tanto o grau de instrução como todas as outras informações tiveram que avaliar as possibilidades dentro do Processamento de Linguagem Natural" (REIS, 2022)

As profissões dentro do conjunto de notícias são de fato diversas. 'Agrônomo', ‘servidor público’, ‘professor’ e ‘empresário’ são alguns dos exemplos encontrados. Elas aparecem sempre precedidas por “declara ao TSE a ocupação de”, cuja concordância funciona tanto para candidatos homens quanto para mulheres.

Essas decisões que refletiam o que estava proposto no *template*, levando em conta as adaptações necessárias para as informações da base do TSE, ocuparam fortemente Tiago Reis e Felipe Grandin. A tônica do relato da equipe editorial deixa evidente a preocupação constante em evitar esses erros de digitação, formatação e concordância. Por isso, um outro papel que eles tiveram foi acompanhar os testes que estavam sendo rodados em paralelo pela equipe de tecnologia. Conforme explica Felipe Grandin, as mudanças no *template* se davam em função dos testes e vice-versa.

> “A gente foi fazendo e refazendo o *template*, fazendo testes com o texto, por exemplo: a gente faz um *template* básico primeiro, aí manda para o departamento de tecnologia o *template*, eles rodam aquele com o *script* do algoritmo, pegam a base da última eleição (2018) e rodam com os dados antigos. Quando foi chegando mais perto, o TSE soltou uma base *fake* também. Eles criaram essa base de candidatos *fake* para fazer testes" (IANKOVSKI, 2023).

O cientista de dados Hector Iankovski corrobora a recorrência dos testes ao longo do projeto. Ele comenta que em alguns casos, essas verificações eram feitas com informações inventadas pelos próprios programadores. “Tinha esse esforço de criar informações randômicas, a gente não estava preocupado ainda com aquele teste em tempo real. Então a gente criou informações que colocavam os nomes das pessoas da equipe, com a porcentagem da eleição para testar mesmo" (IANKOVSKI, 2023).

A preocupação premente em realizar verificações e acompanhar os testes, por parte da equipe editorial, espelha dois princípios éticos do jornalismo: objetividade e credibilidade. A inquietação, segundo as confirmações de Tiago Reis, era ainda mais aguda devido a uma característica própria da automação: todo produto é **escalado**, tanto os aspectos desejados, quanto os malquistos. Se uma característica deficiênte está programada para acontecer, ela será reproduzida até que se altere os parâmetros. Groover (1980) em sua descrição dos processos industriais, coloca a testagem do ‘programa de instruções’ como etapa fundamental para evitar a replicação de erros. Por esse motivo também, no caso de produtos físicos, se realizam testes de conformidade e metrologia antes dos objetos chegarem nas prateleiras. A notícia automatizada se trata de um bem-imaterial, enquanto produto industrial, mas mesmo assim passa por uma ‘programa de instrução’ e é passível de testagem.

Uma analogia válida para entender a preocupação da equipe editorial com os testes das notícias é a prática do *recall*[85]. Na indústria automobilística, por exemplo, quando um automóvel apresenta um defeito que foi escalado para todos os veículos, a companhia compra de volta esse produto. Muitas vezes o defeito se limita a uma peça, mas o risco de aquela falha gerar acidentes, com danos reais aos bem-estar físico dos condutores, justificam o *recall*. Isso sem falar nas indenizações milionárias que esse tipo de problema pode resultar. Quando Tiago Reis fala no receio de deixar passar algum erro, é justamente para não afetar a credibilidade do projeto, ou do veículo, assim como a dificuldade que seria fazer os ajustes manualmente.

Para além dos testes pré-eleição, a próxima e última responsabilidade da equipe editorial era monitorar a geração de textos no dia do pleito. Este trabalho envolveu acompanhar indiretamente a equipe de revisores, entre as notícias que o algoritmo redigia e os arquivos que iam para o sistema de gerenciamento de conteúdo. "No dia da apuração a gente fez uma operação de guerra na redação. Nós tínhamos muito receio de publicar os textos diretamente no ar. Depois que a gente estruturou o projeto e fez o Processamento de Linguagem Natural, nós vimos que estava funcionando" (REIS, 2022).

Esta "operação de guerra" à qual o editor se refere consistia em mobilizar jornalistas de todo o país para atuarem como revisores. Cada equipe das afiliadas do G1 de todos os Estados receberia sua remessa de textos automatizados, dentro do sistema de gerenciamento de conteúdo, para conferir se não havia nenhum erro e, em seguida, publicá-los manualmente. "A gente envolveu mais de 200 jornalistas pelo Brasil, porque temos afiliadas em todos os Estados e em alguns Estados mais de uma afiliada. Por exemplo, São Paulo, a gente tem em Sorocaba, Ribeirão Preto, Campinas e Itapetininga. No Pará, a gente tem em Santarém e a gente tem em Belém. Em Minas, a gente tem o triângulo Mineiro. São umas 55 afiliadas se eu não me engano.", especifica Tiago Reis. Os detalhes de como o algoritmo distribuiu os textos automatizados, por meio do publicador, para as afiliadas do G1 será explicado nos próximos capítulos. O trabalho da equipe de revisores também será abordado nos itens subsequentes deste trabalho, porém, vale aqui ressaltar que o papel de monitorar a geração de textos por parte do algoritmo, tal qual a distribuição dos textos entre os revisores, ficou sob a responsabilidade da equipe editorial.

[85]Ver "*Takata Airbag Recall: Everything You Need to Know"*. Disponível em: https://www.consumerreports.org/cars/car-recalls-defects/takata-airbag-recall-everything-you-need-to-know-a1060713669/

O trabalho de monitoramento da automação foi, segundo Tiago Reis, um esforço de coordenação. De acordo com o seu relato, os textos que correspondiam a um Estado deveriam ser distribuídos entre o número de afiliadas daquele Estado, levando em conta o que seria humanamente possível de ser feito com o pessoal disponível naquela redação.

> "Em algumas áreas, como eram muitas cidades, na Bahia por exemplo tem muitas cidades, eles não iam conseguir dar conta dos textos que eles tinham lá. A gente pegou uma afiliada que fica em São Paulo para subir alguns textos nesses Estados, onde tinham muitos e muitos textos. Então tinha um pessoal lá que estava responsável pelos da Bahia, alguns pelo Rio Grande do Sul, que também tinha muitos textos. Tinha alguns de Minas [...] Enfim, então a gente pensou e fizemos uma conta do que cada uma dessas afiliadas ia ter que publicar" (REIS, 2022)

Há pelo menos dois pontos de interesse nesta tarefa de monitoramento que a equipe do núcleo de dados desempenhou. Primeiro, Tiago Reis utilizou uma "planilha" e fez "uma conta" para distribuir as notícias entre as afiliadas. Algo que é muito típico de jornalistas de dados, como o pesquisador Marcelo Trasel (2014) coloca, é *entrevistar planilhas* e ter a objetividade matemática como guia. Faz parte do método desses profissionais o emprego desta ferramenta nas rotinas de apuração. Aqui, a planilha é usada para uma função distinta, mais próxima de uma coordenação de revisores, ou mesmo a distribuição do trabalho do algoritmo. Em segundo lugar, esse trabalho de monitoramento, segundo Felipe Grandin, pode ser em alguma medida dispensável. "Os textos saíram redondos assim como a gente estava esperando, tanto que esse ano [2022] a gente não revisou antes, a gente publicou direto" (GRANDIN, 2022). O ano que ele se refere é 2022, quando o projeto teve a sua segunda edição, dessa vez para as eleições legislativas e presidenciais. Ou seja, toda a "operação de guerra" tocada por Tiago Reis, com os seus 200 jornalistas à disposição, foi simplesmente considerada desnecessária após o sucesso da inovação.

Felipe Grandin não chegou a entrar em detalhes sobre qual foi a sua atuação no dia da geração dos textos, em 2022. Então sabe-se que houve algum monitoramento, não se pode inferir que os textos foram gerados e publicados sem nenhuma intervenção humana, porém, é possível concluir que essa publicação mobilizou bem menos pessoas. Essa observação sugere que o "receio" mencionado por Tiago Reis em "publicar direto", diminuiu consideravelmente, de um ponto de vista organizacional, depois que um projeto piloto foi realizado a contento. Essa comparação sugere um avanço em termos de autonomia do trabalho do algoritmo dentro do veículo.

Conforme Felipe Grandin explica, a mudança na forma de monitorar o trabalho algorítmico acompanhou também uma modificação no formato do conteúdo jornalístico. Na

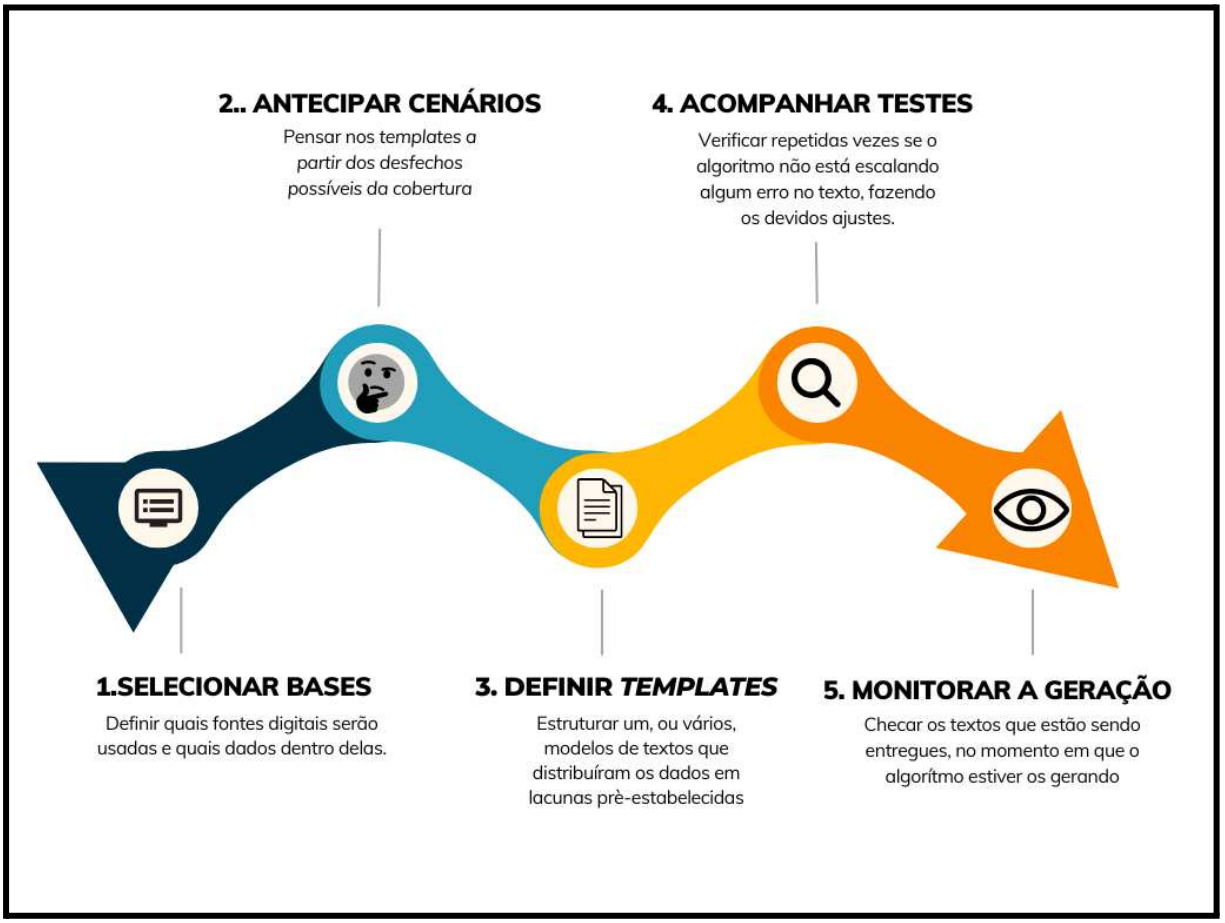
primeira edição do projeto, em 2020, a equipe se sentiu confortável somente em publicar textos, contanto que esses textos passassem pela revisão de outros jornalistas. Esse cenário mudou drasticamente em 2022, quando o veículo já acumulava alguma experiência com o jornalismo automatizado e ousou fazer duas mudanças. Primeiro, a equipe de Felipe Grandin tomou a liberdade de publicar os conteúdos sem passar por revisores. Segundo, eles decidiram explorar um novo formato, junto dos textos, automatizando a produção de vídeos.[86]

> "Esse ano [2022] a gente fez a mesma coisa com os vídeos automatizados. Foi a primeira vez que a gente publicou esse ano [2022]. Então a gente revisou por lote. Geramos 10 vídeos, confere, gerou mais 10, gerou mais 30, confere. Depois que chegamos em um número X, a gente falou 'bom, agora tá tudo certo '. Se está tudo certo e não teve nada errado, pode rodar e publicar. A diferença para mim pessoalmente foi que agora eu toquei a parte editorial. Em 2020, tinha outra pessoa tocando e eu estava ali dando apoio. O principal avanço foram os vídeos. Porque além do texto, a gente fez vídeos automatizados" (GRANDIN, 2022).

É notável o fato de que a automação foi extrapolada de um formato (texto) para dois (texto e vídeo) ao longo de dois anos. Para o editor do projeto na cobertura de 2020, uma ambição da equipe era inserir fotos nas notícias, mas a tecnologia de então não permitia. "A gente gostaria de ter tido vários elementos a mais, né? Mas que a gente não conseguiu nesse primeiro momento. Por exemplo, colocar foto foi impossível [...] Então a gente optou por ter só um texto direto ali mesmo, nesse primeiro momento" (REIS, 2022). Se as fotos eram "impossíveis" em 2020, o audiovisual já era factível em 2022.

[86] O *corpus* deste trabalho engloba notícias e relatos sobre o projeto do G1 de 2020, conforme explicitado no capítulo "Procedimentos Metodológicos". Embora não seja a pretensão desta pesquisa fazer um estudo comparado sobre as iniciativas de automação de notícias, julgou-se pertinente incluir esta parte do relato sobre o jornalista que participou de ambas as iniciativas.

Figura XVIII - O trabalho editorial no jornalismo automatizado na cobertura das eleições de 2020



Fonte: Esta pesquisa

Ao mesmo tempo em que se observa um entusiasmo dos jornalistas em explorar as possibilidades de automação, nota-se também o medo prudente de encarar algo novo pela primeira vez. Várias partes do relato da equipe editorial comunicam objetivamente como foi lidar com um projeto de jornalismo automatizado, havendo consonância com teoria sobre a Geração de Linguagem Natural, principalmente no que diz respeito ao planejamento, microplanejamento e realização das intenções dos falantes.

Aproveitando os relatos salpicados de impressões subjetivas, a pesquisa questionou Tiago Reis e Felipe Grandin sobre qual foi o sentimento final de trabalhar lado a lado com uma máquina. Também foi perguntado se eles sentiram a manutenção de seus empregos ameaçada, ou mesmo sua autoridade enquanto jornalistas. Para o editor, essa ideia era motivo de piada. “Primeiro que a gente brincava com isso na redação. A gente falava que os robôs iam roubar nossos empregos. Nós fazíamos muita brincadeira com isso” (REIS, 2022). Tiago Reis foi além ao expor o que acha da automação.

> “Respondendo a sua pergunta, sobre esse projeto específico, eu acho que não, porque se fosse assim a gente não teria feito esses textos. Se não tivesse sido automatizado, só haveria textos para as principais cidades, entende? Agora, para as 5.568 cidades, ninguém teria feito. Eu não acho que seria um problema de vamos substituir as pessoas, porque a gente simplesmente não conseguiria fazer com as pessoas que a gente tinha. Então não é nem essa a questão [...] Eu vejo mais nesse sentido, como uma tarefa complementar ao trabalho do jornalista, mais do que uma substituição do papel do jornalista. Nesse caso, acho que é um exemplo bem claro de não substituição e sim de complementaridade” (REIS, 2022).

O argumento do editor tangencia o fato de que muitos dos textos automatizados foram depois reforçados por repórteres humanos dentro do sistema de gerenciamento de conteúdo, ou mesmo foram descartados em função de notícias totalmente escritas por humanos. A questão das variações nas notícias será discutida no subcapítulo “**4.4. TEXTOS ESCRITOS POR MÁQUINAS NAS ELEIÇÕES DE 2020”**. Feita essa consideração, é latente como a opinião de Tiago Reis tende a uma visão mais integrada à tecnologia, do que avessa a ela (ECO, 2011). O editor expõe a óbvia limitação humana perante um volume grande de informações, assim como a própria limitação de pessoal da empresa para fazer frente ao desafio. Quando Tiago Reis fala em "complementaridade" há no contexto do seu argumento, uma referência à inclusão de públicos frequentemente ignorados, pertencentes às pequenas cidades que participam do jogo eleitoral, mas são frequentemente deixadas de lado. Da mesma forma, o editor defende que haja espaço para o trabalho algorítmico e humano dentro da mesma cobertura, com os atores agindo ombro a ombro, especialmente para dar conta de informar populações historicamente ignoradas pelos grupos de mídia.

O relato de Grandin não se distancia da visão de Tiago Reis, com a diferença que, para o repórter, o algoritmo é não só um colaborador, mas um “potencializador” do esforço editorial. Segundo ele, a ferramenta é bem-vinda para fazer o jornalismo dar conta de coberturas amplas.

> “Colaboração com algoritmo. É um tipo de aplicação que não nos substitui. O trabalho que a gente fez usando o algoritmo nesse caso foi para fazer o que não dá para fazer no braço, entendeu? A gente não tá pegando o trabalho que alguém faz e a pessoa deixou de fazer esse trabalho. A gente está pegando uma coisa que não existe, que não dá para fazer, que não tem gente o suficiente, não tem tempo suficiente, não tem recurso e a gente está usando um algoritmo para fazer isso. Então o algoritmo, nesse caso, é o potencializador do nosso trabalho.” (GRANDIN, 2022)

Ficam evidentes na fala de Felipe Grandin fatores como o volume e a temporalidade do próprio jornalismo serem incentivos para o emprego de algoritmos, ao passo que a automação também se apresenta como solução frente às limitações econômicas e trabalhistas

dos veículos. Os dois primeiros fatores obedecem ao *discurso da velocidade*, exposto por Örnebring (2010), como o imperativo que guia as inovações em empresas midiáticas: mais volume em menos tempo, é igual a maior competitividade. Já as duas outras limitações, expostas por Grandin, também são de ordem econômica e parecem se relacionar diretamente com a crise financeira que a Grande Mídia vive no século (CHRISTOFOLETTI, 2019). De uma perspectiva unicamente econômica, todas as quatro causas, ou incentivos, apontados pelo repórter para implementar o jornalismo automatizado, são típicas do carácter industrial da produção jornalística. "Os próprios avanços tecnológicos fazem parte das necessidades da industrialização, o que reforça a informação, no caso, jornalística, como decorrência normal do sistema econômico que está na base" (MEDINA, 1988, p. 16).

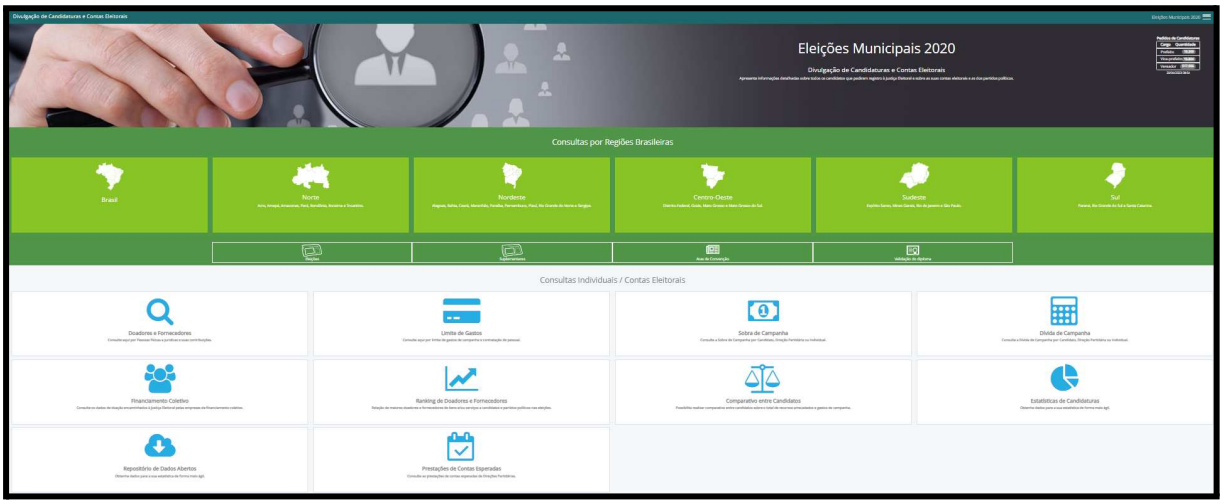
É inegável que haja uma motivação econômica para um veículo como o G1 explorar projetos de jornalismo automatizado e lançá-los em coberturas de grande exposição, como é o caso de uma eleição para o Executivo. Entretanto, as falas dos jornalistas também deixam claras outras motivações, tais como a inclusão, a experimentação, o pioneirismo e a própria interdisciplinaridade[87], que viabilizou o projeto de ponta a ponta. Surpreende o fato de a automação de uma atividade tão própria da profissão, como a escrita, não ter causado nenhum tipo de apreensão nesses dois jornalistas. Ao mesmo tempo que a fala de ambos não deixou de transparecer ao longo das entrevistas, mesmo que implicitamente, um tom de orgulho.

4.2 AS BASES DE DADOS E O JORNALISMO AUTOMATIZADO NO G1

Duas bases de dados foram utilizadas no projeto de jornalismo automatizado das Eleições Municipais de 2020 no G1, ambas pertencentes ao TSE, sendo elas a base **Divulgacand** e a **Resultados**. As duas fontes de informação cumprem o papel de *input* dos dados, descrito por Nicholas Dörr (2015) e Latzer (2014) como a primeira etapa de qualquer processo algorítmico. Elas foram escolhidas e seus valores foram selecionados, a partir do que a equipe editorial julgava ter noticiabilidade. Há também o aspecto da robustez da base, apontado pela equipe de tecnologia.

[87] O argumento é de que a interdisciplinaridade não é apenas o método, ou seja, um meio de se construir um novo produto jornalístico. As trocas entre diferentes profissionais também enseja que uma síntese de seus conhecimentos seja um fim em si.

Figura XIX - Página inicial da Base Divulgacand do TSE



Fonte: Tribunal Superior Eleitoral

A base **Divulgacand** é um dos repositórios online que compõem os esforços de transparência da Justiça Eleitoral eleitoral brasileira[88]. As consultas de informações podem ser feitas por (a) regiões, (b) eleições, (c) eleições suplementares, (d) atas de convenção e (e) validação do diploma. Existe ainda um segundo menu de acesso à base que permite realizar consultas individuais, ou conferir contas eleitorais, segundo recortes de: Pessoas Físicas e Jurídicas e suas contribuições; Limite de gastos de campanha e contratação de pessoal; Sobra de Campanha por Candidato; Dívida de Campanha por Candidato; Direção Partidária ou Individual; Dados de doação encaminhados à Justiça Eleitoral pelas empresas de financiamento coletivo; Relação de maiores doadores e fornecedores de bens ou serviços a candidatos e partidos políticos nas eleições; Comparativo entre candidatos sobre o total de recursos arrecadados e gastos de campanha e Prestações de contas esperadas de Direções Partidárias. Há, portanto, 12 formas de requerer dados públicos pelo portal Divulgacand, sendo que algumas dessas 12 formas acabam chegando à mesma informação, mas por um recorte diferente.

Ao selecionar na base as Eleições 2020, pela consulta de "Estatística de Candidatura" como filtro, é possível ver os números de pedidos de registos totais, candidaturas aptas e inaptas. Das candidaturas aptas para prefeito e vice-prefeito há **39.193** registros. O número condiz aproximadamente com o relato do cientista de dados Hector Iankovski Iankowski, sobre a quantidade de candidatos que o projeto selecionou para a geração dos textos. “Na época que a gente fez a eleição para prefeito, se tratava de 40 a 50 mil candidatos”,

[88] Ver *Divulgação de Candidaturas e Contas Eleitorais*. Disponível em:https://divulgacandcontas.tse.jus.br/divulga/#/ . Acesso feito no dia 20 de Abril de 2023.

(IANKOVISKI, 2023). Pelo volume de dados que 39 mil requisições representam, o projeto se enquadra como um método de ***Big Data***.

De acordo com a equipe de tecnologia, todos esses registros eram baixados constantemente para uma base própria da equipe, onde eles poderiam dar o tratamento adequado antes do algoritmo de geração de textos acessá-las. Conforme relembra o parceiro de equipe de Hector Iankovski, o engenheiro de dados Rafael Muniz, no mínimo uma vez por semana essas bases eram rebaixadas para evitar desatualizações. "A gente configurou o código para puxar esses arquivos uma vez por semana. Um pouco antes das eleições, nós ainda fizemos uma conferida manualmente para ver se estava tudo certinho". O arquivo baixado para os servidores locais era salvo no formato **".csv"** e o método de requisição era **assíncrono**, ou seja, um código automaticamente puxava uma nova base com a frequência semanal definida por Rafael Muniz.

O depoimento da equipe de tecnologia é instrumental para entender como essas bases foram acessadas e os dados brutos foram convertidos em notícias. Entretanto, um outro método de pesquisa é também eficiente para visualizar o resultado do processo descrito pelos profissionais: cruzar as informações dos textos com o repositório do Divulgacand. Para fazer isso, a melhor forma encontrada pela pesquisa foi partir de um exemplo concreto. A cidade de Acrelândia (AC) elegeu em 2020 o prefeito Olavinho do MDB. Como em todos os **39.193** registros eleitorais daquele ano, existe um perfil específico para Olavinho no Divulgacand[89]. Neste perfil existem seis subconjuntos de dados, todos relativos ao candidato e não à eleição em si[90], sendo eles: Consultas, Dados do Candidato, Documentos, Prestações de Contas, Receitas e Despesas. Os três primeiros subconjuntos trazem informações referentes à pessoa física Olavo Francelino de Rezende e sua persona política, Olavinho. Já os três últimos conjuntos de dados disponibilizam informações diversas sobre a campanha em si: despesas descritas, doações recebidas, uso do Fundo Eleitoral Partidário, notas fiscais recolhidas e outros. Ao observarem-se somente os dois primeiros subconjuntos, **Consultas** e **Dados do Candidato**, é possível identificar todas as informações sobre o candidato contidas na notícia da eleição em Acrelândia (AC).

[89] Ver Divulgacand 0 Olavinho. Disponível em: https://divulgacandcontas.tse.jus.br/divulga/#/candidato/2020/2030402020/01120/10001072604. Acesso feito no dia 24 de

[90] O perfil do Divulgacand não trás informações sore o pleito, como números de votos e abstenções, mas apresenta uma tarja sob a foto do candidato informando se ele foi eleito.

“Olavinho tem **51 anos**, é casado, tem **ensino médio completo** e declara ao TSE a ocupação de **empresário**. Ele tem um patrimônio declarado de **R$ 225.573,23”**[91].

Figura XX - Perfil no Divulgacand do prefeito eleito de Acrelândia (AC)

Fonte: Divulgacand (TSE)

Na figura acima vemos que todas as informações de estado civil, grau de instrução e ocupação estão contidas no subconjunto de dados do candidato. Enquanto as informações de patrimônio estão contidas em **Consultas**, na opção de “Listas de Bens Declarados”. O processo que o engenheiro Rafael Muniz e Hector Iankovski descrevem é exatamente essa obtenção de dados, replicada **39.193** vezes, para todos os candidatos. Essas 39 mil requisições eram rodadas semanalmente para não deixar passar nenhuma informação desatualizada. Depois da obtenção, há ainda um segundo processo de ingestão dessas informações em uma *Data Warehouse*, um repositório interno e privado da equipe de tecnologia, porém, esse processo será explicado mais adiante no capítulo “**3.5. O Algoritmo de Geração de Notícias no G1**”.

Ao retornar às bases de dados que foram escolhidas pela equipe editorial, resta descrever a segunda fonte. O repositório digital **Resultados** do TSE também apresenta resultados de eleições desde 1933, retornando arquivos em formato “**.csv**”, ou “**.txt**” a depender do tipo de conjunto de dados que se pretende acessar. A questão do tipo de arquivo é relevante, pois há diferentes *softwares* no mercado para processar e visualizar esses dados.

[91] Ver “Olavinho, do MDB, é eleito prefeito de Acrelândia”. Disponível em: https://g1.globo.com/ac/acre/noticia/2020/11/16/olavinho-do-mdb-e-eleito-prefeito-de-acrelandia.ghtml. Acesso feito em 20 de abril de 2023.

O *Microsoft Excel* e o *Google Sheets*, por exemplo, abrem arquivos do tipo **".xls"**, mas já o *Power BI* não abre. Os dois formatos de arquivos retornados pela base do TSE são formatos mais maleáveis e executáveis por múltiplos *softwares*. Essa justificativa vem no topo da página, ao acessar qualquer um dos conjuntos. "Arquivos de dados com um grande número de linhas (particularmente aqueles com extensão .csv e .txt) podem não ser visualizados em sua totalidade a depender do *software* utilizado"[92].

O formato dos arquivos reflete uma preocupação intrínseca a processos de *Big Data*, processar e visualizar grandes volumes de informações. O que fica evidente é que cada um dos conjuntos acessíveis nesse repositório é, em si, um caso de *Big Data*. O portal disponibiliza para cada pleito, os dados de quatro formas:

1. **Correspondências esperadas e efetivadas - 1º e 2º turno**
2. **Boletim de Urna**
3. **Arquivos transmitidos para totalização**
4. **Resultados**

Não vale aqui entrar em detalhe sobre o que cada conjunto apresenta, até mesmo por que a complexidade de cada um é avassaladora. O Boletim de Urna, por exemplo, apresenta 42 variáveis (colunas quando se visualiza em tabela) com uma linha para cada votação feita em um dado Estado. Os subconjuntos dentro do Boletim de Urna estão divididos por três grandezas: Estado, Turno (1º ou 2º) e Exterior. Há, portanto, um subconjunto para os 26 estados da Federação, mais o Distrito Federal, com um subconjunto para os eleitores que votaram em países estrangeiros. Tudo isso multiplicado por dois, se a eleição teve os dois turnos. A complexidade dos tipos de valores e variáveis de cada conjunto é tão grande, que cada requisição vem acompanhada de uma espécie de manual de instrução. Um arquivo "leiame.pdf"[93] traz uma explanação geral sobre aquele subconjunto com descrições de cada uma das variáveis ali contidas. A breve descrição de como funcionam apenas um dos quatro subconjuntos dá a dimensão do volume de dados com que a equipe de tecnologia do G1 teve que lidar. Conforme Hector Iankovski aponta em seu relato, a equipe executou uma "coleta bastante robusta".

---

[92] Ver Portal de Dados Abertos do TSE. Disponível em: https://dadosabertos.tse.jus.br/dataset/resultados-2020. Acesso feito em 21 de Abril de 2023.

[93] "Leia me" é uma tradução direta de arquivos chamados em inglês de "readme". Eles representam uma forma de comunicação e instrução característica da cibercultura e, principalmente, do movimento open source - entusiasta da computação que acreditam na transparência não só de informação, mas das próprias ferramentas digitais presentes na web.

Dos quatro subconjuntos de dados presentes na base do TSE, somente um é de interesse desta pesquisa. Em (4) **Resultados,** estão presentes os valores restantes que complementam as informações provindas do **Divulgacand**, para compor a geração automática de notícias. Tais valores são referentes, é claro, aos resultados do pleito específico daquele ano, ou turno.

Ao acessar o subconjunto de dados relativo ao estado do Acre (AC), mais precisamente do primeiro turno, é possível fazer o *download* do arquivo "votacao_secao_2020_AC.csv" acompanhado das instruções "leiame.pdf". A descrição do documento de apoio explica, por exemplo, que a base em questão conta com 22 variáveis, logo, 22 colunas. Cada linha corresponde à "apuração por município e zona/seção - Votação nominal/partido por município e zona - Votação por seção eleitoral" (TSE, 2020). Entretanto, com apenas três dessas variáveis é possível descobrir o vencedor do pleito. São elas: "NM_MUNICIPIO", "NM_VOTAVEL e "QT_VOTOS". Colocando um filtro para a cidade de Acrelândia (AC) e incorporando uma fórmula que retorne o nome do candidato (NM_VOTAVEL) com a maior quantidade de votos (QT_VOTOS), chega-se ao resultado: **Olavinho (MDB)**, **2.638** votos.

Também pela variável do nome do candidato, pode-se contabilizar outras informações como número de votos nulos, votos em branco e voto anulado. De acordo com a descrição dessa variável no documento "leiame", os valores podem ser de ordem distintas.

> Pode assumir os valores: Nome do candidato (quando voto nominal ou voto anulado); Nome do partido (quando voto em legenda); Voto em branco (quando voto em branco); Voto anulado e apurado em separado (quando voto anulado e apurado em separado) (TSE, 2020).

A partir da soma do número de votos do candidato mais votado, a porcentagem de votos nominais que ele recebeu em relação aos outros candidatos, mais a porcentagem de nulos, brancos e abstenções, chega-se a um resumo da eleição em Acrelândia (AC). O resultado traz precisamente todos os valores contidos no lide na notícia do G1, como exposto na imagem abaixo.

Figura XXI - Notícia do resultado da eleição para prefeito em Acrelândia (AC)

Olavinho, do MDB, é eleito prefeito de Acrelândia

Ele teve 38,52% dos votos dados a todos os candidatos e derrotou Caetano, que ficou em segundo lugar com 35,11%.

Por G1 AC — Rio Branco
16/11/2020 00h41 · Atualizado há 2 anos

Olavinho, do MDB, foi eleito, neste domingo (15), prefeito de Acrelândia (AC) para os próximos quatro anos. Ao fim da apuração, Olavinho teve 38,52% dos votos. Foram 2.638 votos no total.

O candidato derrotou Caetano, que ficou em segundo lugar com 35,11% (2.405 votos).

A eleição em Acrelândia teve 22,64% de abstenção, 0,76% votos brancos e 2,68% votos nulos.

Fonte: Olavinho, do MDB, é eleito prefeito de Acrelândia. G1 (2020)

Com este processo, ficam pendentes apenas duas operações. A primeira é repetir a contagem para o segundo colocado. No caso de Acrelândia (AC) este é o ex-prefeito Ederaldo Caetano (PP). A segunda operação, e a mais importante, é concatenar esses valores com os dados requisitados da tabela **Divulgacand**, sendo eles o nome do candidato e seu partido. O termo concatenação é um jargão corrente na ciência de dados, que significa em linhas gerais a sobreposição de conjuntos. No caso da cobertura das eleições municipais, esses conjuntos são, justamente, as informações das duas bases de dados aqui descritas.

A partir da listagem de todas essas informações do **Divulgacand** e dos **Resultados**, que compõem as notícias do primeiro turno, é possível determinar as **marcas de apuração** desta cobertura jornalística (SILVA-MAIA, 2011). Se levarmos em conta a ordem em que os dados estão dispostos, também é factível observar as **marcas de composição** destas notícias. Conforme foi visto ao longo deste subcapítulo, a base dados é um ator viabilizador da automação de notícias. A depuração e análise dessas bases, a partir das marcas deixadas nas notícias, é a chave para compreender o produto final da cobertura automatizada.

Tabela III- Marcas de apuração e composição das notícias automatizadas do G1

Tabela III - Marcas de apuração e composição das notícias automatizadas do G1		
Sessão do Texto	**Tipo de Dados**	**Base de Dados**
Título	- Nome (1º colocado) - Partido (1º colocado) - Cargo disputado - Cidade	Divulgacand
Soutien	- Percentual de votos (1º colocado) - Nome (2º colocado) - Percentual de votos (2º colocado)	Resultados Divulgacand
Lide	- Nome (1º colocado) - Partido (1º colocado) - Cargo disputado - Percentual de votos (1º colocado) - Total de votos (1º colocado)	Divulgacand Resultados
Sublide	- Nome (2º colocado) - Percentual de votos (2º colocado) - Total de votos (1º colocado)	Divulgacand Resultados
3º parágrafo	- Percentual de abstenções - Percentual de votos brancos - Percentual de votos nulos.	Resultados
4º parágrafo	- Nome (1º colocado) - Idade (1º colocado) - Escolaridade (1º colocado) - Ocupação (1º colocado) - Patrimônio (R$) (1º colocado)	Divulgacand
5º parágrafo	- Nome do vice (1º colocado) - Partido do vice (1º colocado) - Idade do vice (1º colocado)	Divulgacand
6º parágrafo	- Nome da coligação - Partidos da coligação	Divulgacand
7º parágrafo	- Nome + Partido + Percentual (1º colocado) - Nome + Partido + Percentual (2º colocado) - Nome + Partido + Percentual (3º colocado) - Nome + Partido + Percentual (3º colocado)*	Divulgacand Resultados
* O número de colocados está limitado a oito nomes, mas na maioria dos casos traz os quatro primeiros.		

Fonte: Esta Pesquisa

A tabela acima deixa evidente como diferentes bases de dados foram "concatenadas" para nutrir o algoritmo de geração de linguagem natural. Vale fazer a ressalva de que os dados citados compõem de fato as **marcas de apuração** de toda a cobertura, mas a ordem em que eles aparecem nos textos, portanto as **marcas de composição** variam de acordo com o *template* realizado. No caso, temos a realização do template de vitória em 1ª turno, que representa mais de 90% de toda a cobertura, uma amostragem significativa (ver subcapítulo "**4.4. TEXTOS ESCRITOS POR MÁQUINAS NAS ELEIÇÕES DE 2020**"). Exemplos que mostram as marcas de composição dos outros templates realizados são dados na última parte desta pesquisa.

Embora seja demonstrável que a base de **Resultados** é capaz de retornar os valores contidos na notícia, não foi essa interface utilizada pela equipe de tecnologia para receber os dados no dia da eleição. Isso é explicado por Hector Iankovski e Rafael Muniz em suas falas. O TSE disponibiliza no dia da eleição um sistema dedicado para transmitir as informações apuradas nas sessões eleitorais. Esse sistema é compartilhado com os veículos de mídia que se inscrevem para acessá-lo, a fim de receber os dados em tempo real. É evidente que o mesmo dado transmitido por esse sistema, também é disponibilizado na base Resultados assim que a contagem termina. Essa dinâmica de inscrição dos veículos está presente em um comunicado publicado pela Secretaria de Comunicação e Multimídia (SCM) do órgão no dia 12 de agosto de 2020[94].

> O Tribunal Superior Eleitoral (TSE) começará a orientar, a partir da próxima segunda-feira (17), as instituições e os veículos de comunicação interessados em divulgar os resultados das Eleições Municipais de 2020 sobre os procedimentos que devem seguir para acessar os dados gerados pelo Tribunal (...) Essas orientações possibilitarão que emissoras de televisão e de rádio, portais de internet e a imprensa, em geral, entre outras mídias, possam informar à população, em tempo real, a partir do encerramento da votação, os votos recebidos por cada candidato a prefeito, a vice-prefeito e a vereador no dia da eleição (...) As informações ficarão disponíveis em nuvem. Por questão de segurança, será limitada a quantidade de acessos por segundo de cada interessado ao data center. O número máximo de requisições permitidas será de 300 por segundo (SCM-TSE, 2020).

[94] Ver "Começa fase de orientação de instituições interessadas em divulgar os resultados das Eleições de 2020". Disponível em: https://www.tse.jus.br/comunicacao/noticias/2020/Agosto/comeca-fase-de-orientacao-de-instituicoes-interessadas-em-divulgar-os-resultados-das-eleicoes-de-2020. Acesso feito no dia 22 de abril de 2023.

O comunicado do tribunal foi feito a partir da publicação Resolução n º 23.611/2019, artigos 207 e 213[95]. A importância da norma é criar um conjunto de instruções para que as empresas de mídia possam fazer a cobertura no dia do pleito, evitando problemas técnicos. Ademais, a resolução cria também uma espécie de calendário de atendimento aos veículos que estão "interessados em compartilhar as informações no dia da eleição". O interessante é que neste calendário, os jornais, emissoras, portais e rádios podem participar de simulações de como a apuração será transmitida no dia da eleição.

> No simulado, que terá a duração de alguns dias, os representantes dos veículos de comunicação testam o funcionamento de seus próprios softwares a partir de dados brutos oferecidos pelo TSE. Os arquivos de dados brutos devem ser compreendidos e trabalhados em softwares, pelas instituições, para que possam ser divulgados a seus usuários da maneira que acharem melhor. Os simulados são importantes para analisar vários fatores imprescindíveis para a divulgação dos resultados. Entre eles, está a aplicação das regras, o desempenho e o comportamento da divulgação na geração dos arquivos necessários (SCM-TSE, 2020).

Outro motivo pelo qual o anúncio do órgão é pertinente diz respeito à questão da arquitetura de transmissão de dados, um conceito relevante à *tecnicidade* da automação de notícias (SIMONDON, 1988). Em suma, a forma como o dado é transmitido implica em como ele pode ser processado, para depois entrar no texto. Essa questão é esmiuçada no próximo capítulo, pelos relatos de Hector Iankovski e Rafael Muniz. Entretanto, é importante salientar que o próprio TSE com seu atendimento prestado aos veículos é um ator da automação. Graças a essa transparência técnica, as notícias vão ao ar.

4.3 A EQUIPE DE TECNOLOGIA DO G1 E O ALGORITMO

A dupla Hector Iankovski e Rafael Muniz compunham em 2020 o chamado *squad* de ciência de dados do G1. Hector Iankovski era o cientista de dados, enquanto Rafael Muniz era o engenheiro. O cientista de dados é o responsável por criar o modelo de análise de dados. Sua função é essencialmente analítica. Já o engenheiro de dados é o profissional que se dedica ao desenvolvimento, construção, e manutenção das arquiteturas de dados. O sistema de processamento em si.

Essa dinâmica entre as duas profissões fica evidente na dinâmica da equipe. Conforme Rafael Muniz coloca, ele era responsável por criar o *pipeline* do sistema. O termo

[95] Ver "Resolução n º 23.611/2019, artigos 207 e 213". Disponível em: https://www.tse.jus.br/legislacao/compilada/res/2020/resolucao-no-23-627-de-13-de-agosto-de-2020. Acesso feito no dia 22 de abril de 2023.

pipeline é amplamente usado no setor de tecnologia. Ele pode ser traduzido literalmente como "tubulação", mas na realidade se refere a uma sequência de preparo de um dado conjunto de dados, seguidos do processamento e análise. O cientista de dados Hector Iankovski, por sua vez, desenvolvia o algoritmo de Geração de Linguagem Natural. A dinâmica da dupla também foi marcada pela cronologia de suas carreiras. Hector Iankovski entrou na organização no início de 2020 e começou a acompanhar as conversas com a equipe editorial logo de início. Já Rafael Muniz iniciou o seu contrato poucos meses depois, enquanto Hector Iankovski já tocava o projeto. Isso leva, por exemplo, o editor Tiago Reis a chamá-lo de o "ponto focal" da equipe de tecnologia.

A implementação técnica de todo o processo, segundo Hector Iankovski, começou com estudar o conceito de Geração de Linguagem Natural e desenhar um sistema que atendesse ao escopo da cobertura pensado pela equipe editorial. Isso envolvia distribuir tarefas e mediar demandas entre equipes. "Eu trabalhei nesse projeto meio que como gerente. Eu cuidava de toda a equipe de coleta de dados da Globo, eu que falava com a equipe de distribuição do CMA[96] e o que fazia toda essa parte de implementação do código em Python. Para depois fazer o tratamento da informação, a geração do texto e a criação do arquivo posterior" (IANKOVISKI, 2023)

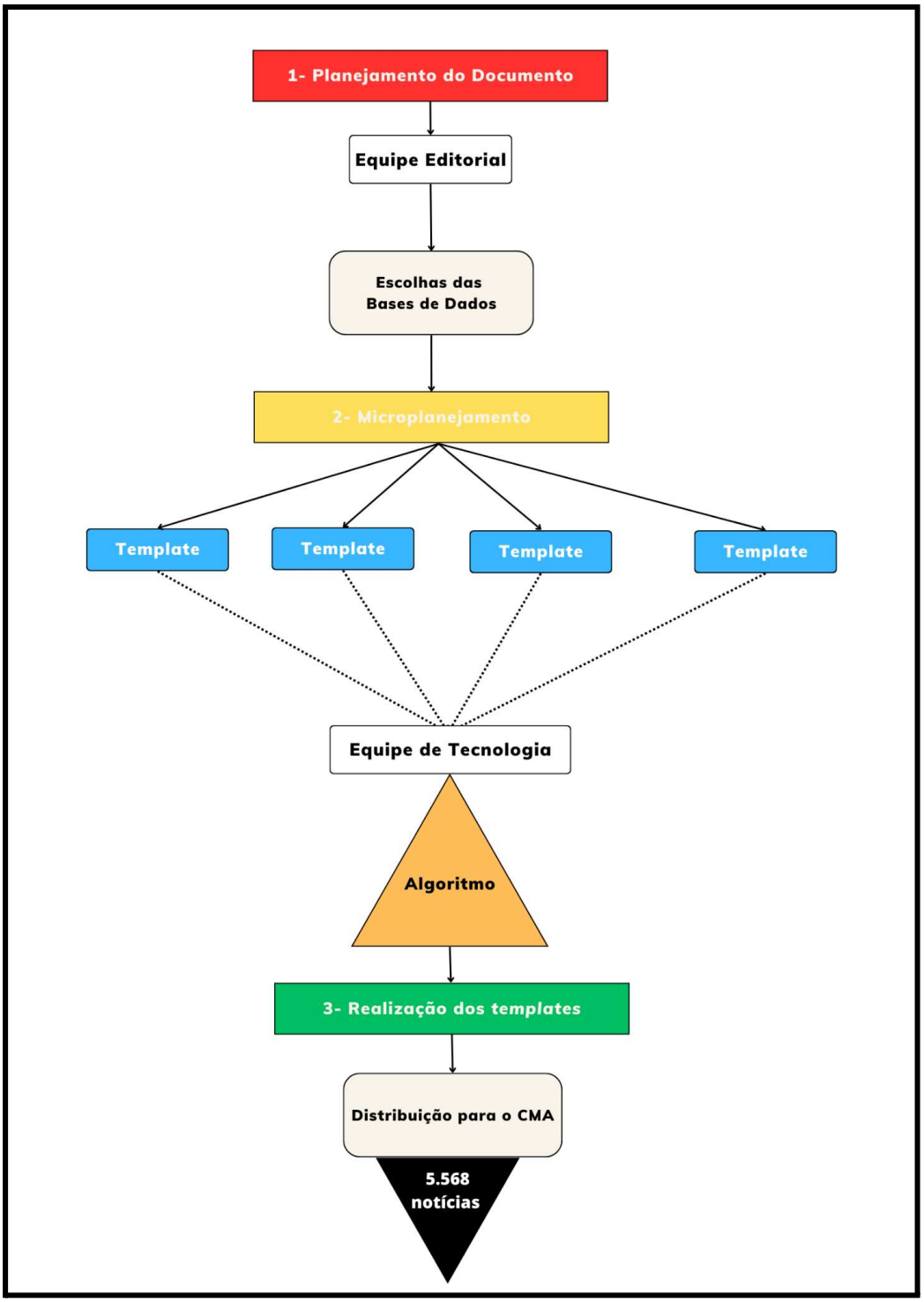
A explicação de Hector Iankovski denuncia a posição intermediária que a equipe de tecnologia assumiu. Ao mesmo tempo que os jornalistas Tiago Reis e Felipe Grandin determinavam o escopo da cobertura, Hector Iankovski e Rafael Muniz cuidavam da parte técnica com o apoio de mais dois departamentos. Os dados em tempo real do TSE vinham de um time que se dedicava exclusivamente a cuidar da interface com o sistema do tribunal. Tal departamento funcionava em paralelo à equipe de tecnologia, mas cumpria um papel indispensável. "O nosso *input* era esse dado que vinha do TSE através do sistema de coleta da Globo. A gente fazia o meio campo ali, que era o tratamento da informação e a geração do texto" (IANKOVISKI, 2023). Essa posição intermediária que o cientista de dados assumiu, exigiu que ele acumulasse um conhecimento já familiar para a equipe editorial: a dinâmica da cobertura eleitoral. Esse contato indica que Hector Iankovski desenvolveu uma *inteligência híbrida* (SANTOS, 2018b) de maneira simétrica àquela adquirida pela equipe editorial (ver **"4.1 O TRABALHO EDITORIAL NO JORNALISMO AUTOMATIZADO DO G1"**).

Simultaneamente, o segundo departamento que possuía contato direto com a equipe eram os responsáveis pelo sistema de gerenciamento de conteúdo. Rafael Muniz e Hector

[96] A sigla remete ao sistema de gerenciamento de conteúdo da Globo.

Iankovski se referem a eles como a "equipe do CMA", os profissionais responsáveis por distribuírem as notícias automatizadas dentro do publicador. O interessante de reparar é que os dois times que a equipe de tecnologia intermediou cumpriam, respectivamente, o papel de *input* e *output* do sistema de geração de textos.

Figura XXII - As etapas de Geração de Linguagem Natural no projeto do G1



Fonte: Esta pesquisa, baseado em Diakapoulos (2019), Reiter (2012) e McDonald (2010)

Ao atuar como um gerente da iniciativa, Hector Iankovski organizou como a automação de textos precisaria ser desenvolvida, mas o trabalho do cientista não começou do zero, conforme ele mesmo explica. “Eu entrei na Globo no começo de 2020. Soube que existia uma outra área na Globo que também fazia alguns projetos de inovação e experimentação. Eles fizeram um projeto piloto de geração automatizada de textos (...) Essa outra área desenvolveu o projeto de uma maneira meio informal e experimental, mas sem uma aplicação prática”. O que o cientista aponta como o pontapé inicial da tecnologia é, portanto, um protótipo que já havia sido preparado em outro departamento de tecnologia do Grupo Globo, mas que nunca tinha sido usado para nada.

O fato de Hector Iankovski ter partido de um protótipo, até então sem aplicação definida, mostra a dinâmica que Simondon (1988) chama de a *gênese dos objetos técnicos*. Esse caminho é sempre percorrido de um estágio de abstração - o projeto piloto neste caso - para um estágio de concretude. O que é característico dessa gênese é que cientistas como Hector Iankovski vão adicionando camadas a uma tecnologia pré-existente, até ela ganhar uma aplicação clara, ou concreta. Desta forma os objetos técnicos vão desenvolvendo a sua individualidade.

O ponto de partida do desenvolvimento técnico do projeto foi, portanto, um algoritmo feito de forma “experimental”. Conforme expõe Hector Iankovski, tal experimento se soma à demanda criada pela equipe editorial, como um ponto de partida para a elaboração de uma tecnologia de Geração de Linguagem Natural que atendesse à uma produção robusta.

> “A ideação do projeto que eu conheci foi assim. O pessoal do G1 falou assim: ó, queremos produzir um texto automático de alguma forma utilizando essa ideia aqui. Como que vai ser feito e de que forma vai ser feito, ficou a nosso cargo. Como pensar na solução e como desenvolver ela, fazer a coleta de dados necessária para isso. O pessoal do jornalismo que participou era o Tiago Reis, o Felipe Grandin e existia uma terceira pessoa, que era o chefe do Tiago Reis, o Franzine. Eles eram os responsáveis no G1 pela parte da publicação e pela curadoria” (IANKOVISKI, 2023)

A "solução" mencionada por Hector Iankovski pode ser dividida em cinco etapas: (1) criação de uma base de dados interna, (2) programação do algoritmo, (3) condução de testes, (4) integração com o sistema de gerenciamento de conteúdo e (5) monitoramento do algoritmo em tempo real. Todas essas etapas já foram de alguma forma abordadas nos capítulos anteriores, mas vale retomá-las pela perspectiva dos membros da equipe de tecnologia.

Tabela IV - O trabalho da equipe de tecnologia na cobertura das eleições de 2020

Tabela IV - O trabalho da equipe de tecnologia na cobertura das eleições de 2020	
Etapa (nº)	**Objetivo**
(1) criação de uma base de dados interna	Inserir todos os dados públicos para um *data lake* próprio;
(2) programação do algoritmo	Escrever o *script* em *Python* de quais *templates* seriam realizados, a partir dos dados, para a Geração de Linguagem Natural;
(3) condução de testes	Acompanhar se os textos estão sendo gerados na velocidade desejada, sem erros, com os parâmetros definidos na etapa anterior;
(4) integração com o sistema de gerenciamento de conteúdo	Garantir que os arquivos contendo os textos estão sendo gerados com os marcadores (*tags)* necessários para distribuição no sistema de gerenciamento de conteúdo
(5) monitoramento em tempo real	Acompanhar em tempo real se os milhares de textos estão sendo gerados pelo algoritmo sem entraves, ou gargalos.

Fonte: Esta pesquisa

Todas essas etapas foram executadas, é claro, para que os textos fossem gerados no dia da eleição sem gargalos. Para isso, uma sequência de métodos digitais e programas deveriam funcionar de maneira coordenada, pré-testada e verificável no dia do pleito. Essa sequência pode ser compreendida como o *pipeline*, termo recorrente no relato da equipe.

De início a (1) criação da base de dados interna ficou sob responsabilidade do engenheiro, Rafael Muniz. O processo consistiu em baixar e subir as bases de dados do TSE para um *Data Warehouse*[97] próprio do G1. Quais bases e quais dados dentro delas eram de interesse do projeto, eram determinadas pelo *template* e, portanto, pela equipe editorial. O sistema escolhido, conforme aponta Rafael Muniz, foi o *Big Querry*, pertencente ao *Google Cloud Platform.* Segundo a empresa, esse é um serviço de "armazenamento de dados corporativo totalmente gerenciado que ajuda a gerenciar e analisar dados com recursos

[97] Outro sinônimo de Data Warehouse que aparece nas entrevistas é o Data Lake. Ambos se referem ao mesmo processo, ou ambiente virtual.

integrados"[98]. Hector Iankovski aponta que o G1 e a Google tinham um acordo comercial, o que facilitava o uso das ferramentas pela equipe. "Na época o Google Cloud Plataform foi o que a gente usou. Existem outras como a AWS e o Azura da *Microsoft*. Mas na Globo a gente tinha uma parceria com a *Google* e a gente fez toda essa coleta e armazenamento das informações nessa base de dados do *Big Querry"*(IANKOVISKI, 2023).

Depois da "ingestão robusta" de dados para gerar base interna da equipe, o cientista de dados teve que elaborar o (2) algoritmo de Geração de Linguagem Natural. Segundo os teóricos Reiter (2012) e Mcdonald (2010), o algoritmo é em essência a **realização do documento**, feita a partir de seu **microplanejamento.** Hector Iankovski explica em detalhe que esse código acessava as informações do *Big Querry* para gerar o texto, a partir do que tinha sido definido no *template*.

> "O *Big Querry* é o banco de dados no caso [...] é um sistema de banco de dados e a linguagem de consulta do banco de dados é o SQL. Dentro do código em Python você pode executar *queries* de SQL, para depois plugar elas dentro ali do *template* que a gente estava utilizando. Então dentro do código Python, a gente tinha a execução de *queries* SQL para consultar informação do banco de dados. Feito a execução do código o nosso *output* inicial, era a gravação da informação novamente no *Big Querry*, para ficar armazenado ali, caso a gente precisasse re-gerar esse texto, ou consultar se a informação estava correta, ou fazer algum outro tipo de consulta. O outro passo depois foi: a gente tinha um outro código em Python posterior a esse salvamento no *Big Querry*, que era fazer a leitura do banco de dados e gerar o arquivo 'json' para ser publicado e enviado para a equipe do CMA" (IANKOVISKI, 2023)

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. Quando Hector Iankovski menciona "a leitura do banco de dados e gerar o arquivo *json*" ele está descrevendo tecnicamente o próprio algoritmo. Vale relembrar que Latzer (2016) aponta que a especificidade de cada algoritmo é definida pelo seu *thoughtput*, ou seja, o que acontece entre um *input* e um *output*. A explicação do cientista é justamente sobre o que ocorre no meio, entre o banco de dados (*input*) e o arquivo json (*output*). Os processos da mediação algorítmica são divididos em duas partes por Hector Iankovski. "Colocamos os sinônimos que a gente precisava para publicação do texto da eleição, criando os *templates* e definido com o pessoal do jornalismo, inserindo as informações com o *fill the blanks* para o preenchimento das lacunas, com as informações coletadas, anteriormente e também com o sistema do TSE em tempo real"(IANKOVSKI, 2023).

[98] Ver *O que é o Big Querry*. Disponível em: https://cloud.google.com/bigquery/docs/introduction?hl=pt-br. Acesso feito em 28 de Abril de 2023.

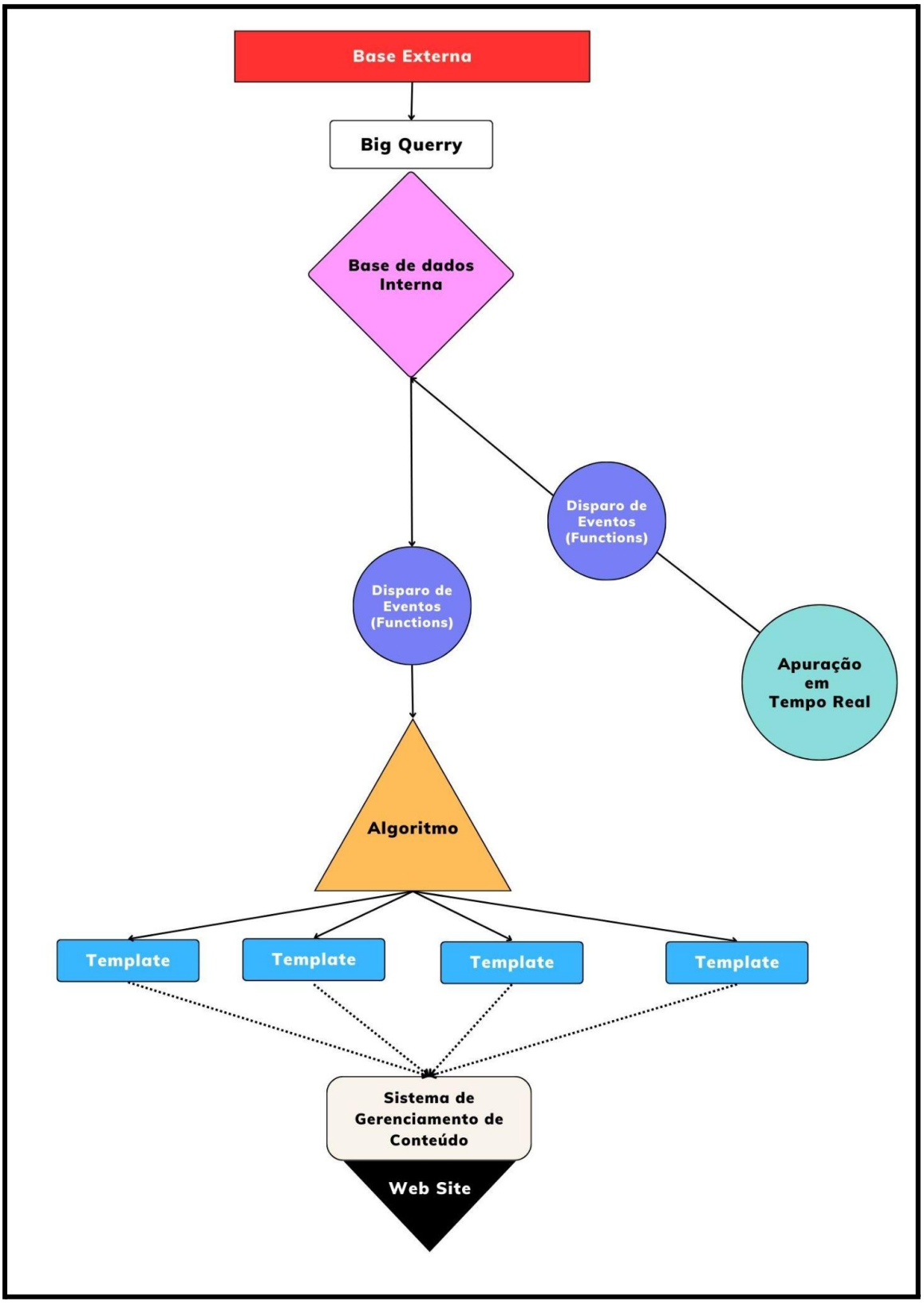
Há somente duas informações nesse relato que dizem respeito ao algoritmo de geração *per si*. Primeiro, os "sinônimos" que se referem às flexões de gênero (ex: prefeito e prefeito) e as maneiras de se referir aos candidatos quanto a sua profissão e escolaridade. Segundo, o *fill in the blanks* enquanto o processo de preencher as lacunas do texto com as informações coletadas pelas consultas (*queries)* feitas ao banco de dados. Esses dois processos detalham a **realização do documento** até chegar em seu *output.*

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 do *pipeline* seja o gatilho para um próximo acontecimento, compondo assim a automação. Rafael Muniz corrobora o relato em que ambos trabalharam em um sistema de disparos, responsável por fazer a integração entre o banco de dados e o Processamento de Linguagem Natural.

> "Dentro do *Google Cloud* existe um serviço chamado *Functions*, que é um sistema de mensageria que recebe e dispara eventos. Dentro dele você pode colocar um código *Python*, ou então pode colocar uma rotina para alguma coisa ser executada ao mesmo tempo. Então dentro do *Functions* a gente disparava assim: 'eu quero que o *Big Querry* faça isso, colete a informação e grave no banco de dados. Aí o próximo passo é o tratamento da informação, para tirar caracteres especiais e deixar no formato que a gente tinha pré-definido. O próximo passo era o disparo da mensagem para execução do código, para a geração do texto de fato" (IANKOVISKI, 2023).

Fica evidente na fala do cientista de dados que toda a geração do texto depende de um conjunto intrincado de tecnologias distintas. 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. Acompanhar tal processo é entender a **materialidade** da cobertura automatizada das eleições municipais de 2020. Essa dinâmica está ilustrada na imagem abaixo, que por sua vez também também lista os objetos técnicos componentes do *Inventário* deste fenômeno comunicacional (LEMOS, 2020).

Figura XXIII - *Pipeline* de Geração de Notícias Automatizadas



Fonte: Esta pesquisa

No dia da eleição, como apontam Rafael Muniz e Hector Iankovski, não haveria tempo hábil de fazer mudanças expressivas no algoritmo ou no *pipeline*. Para que tudo estivesse pronto e verificado, a equipe precisava ter (3) conduzido 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. Então os testes eram constantes sim. Porque no meio ali dos testes a gente pensava em melhorar isso aqui e aquilo ali [...] 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.

Toda essa preocupação de “performance”, evitar “gargalos” e “*delay*” expressam novamente o caráter industrial do projeto. A performance enquanto uma razão entre tempo e produto são um dos motivos apontados por Diakopoulos (2019) para se perseguir a automação de notícias. Os "gargalos” de uma linha de produção surgem aqui no ambiente virtual como dificuldades de processamento. Já o “*delay*” proferido como um temor, algo a ser evitado, é justamente aquilo que se opõem a velocidade enquanto valor (ÖRNEBRING, 2010).

A questão dos testes surge na história de Hector Iankovski e Rafael Muniz como algo bastante atrelado à velocidade de processamento do sistema. Esta velocidade é indissociada do volume. Em outra explicação, o cientista de dados revela que os experimentos tinham justamente o objetivo de revelar se o algoritmo era capaz de gerar os textos no volume esperado. “A gente fazia uma cópia truncada, que gerava vários arquivos iguais e ficava colocando 100 arquivos por vez, ou 200 arquivos por vez, 500 arquivos por vez, para ver se o nosso sistema estava escalando”.

Além destas “cópias truncadas” que Hector Iankovski e Rafael Muniz geravam por conta própria, a equipe também participou dos simulados com o TSE. “A gente fez todo o teste com uma quantidade massiva de arquivos. Se eu não me engano uma ou duas semanas antes, a gente teve um último teste com o TSE. Que era o teste final e oficial. Depois daquilo não teria mais nenhum teste com o TSE, ele seria o formato final que todo mundo receberia” (IANKOVISKI, 2023). Mais uma vez, percebe-se que a transparência do órgão foi um fator que viabilizou a automação de textos. Se não fosse por esse calendário de simulados do Tribunal, a equipe não teria certeza de que o algoritmo estava de fato dando conta do *input*.

Finalizadas as testagens, restou à equipe fazer a (4) integração com o sistema de gerenciamento de conteúdo. 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 distribuição do arquivo dentro do sistema de gerenciamento de conteúdo.

> "Cada cidade tinha um arquivo json gerado. Eles viam uma série de *tags*, que podiam ser utilizadas. A gente acordou com eles, qual era essa formatação, se eu não me engano eram umas 8 ou 9 *tags*, com informações que eram pertinentes para eles fazerem essa vinculação dentro do CMA. A primeira informação, por exemplo, era um código da cidade. A segunda era o nome da cidade. A terceira era sobre o Estado (...) Assim esse texto era distribuído para todos os jornalistas que deveriam ter acesso àquela publicação" (IANKOVSKI, 2023)

O arquivo final "json" gerado, tal qual as *tags*, foi verificado por esta pesquisa no processo de *web scraping*. Para conseguir o maior número possível de notícias publicadas pela cobertura da eleição, foi necessário acessar o site do G1 a partir de cada "estado > afiliada > cidade" para conseguir recuperar a notícia. Ao executar este método automatizado de raspagem dos textos, o arquivo final que esta pesquisa conseguiu é um documento "json" muito similar ao que Hector Iankovski descreve. Como pode se observar abaixo, a partir do documento "Piracicaba_São_Paulo_Notícias.json" foram obtidos os elementos textuais das notícias de todas as cidades[99] daquele "estado > afiliada".

> "Águas de São Pedro": ["João Victor Barboza, do Cidadania, é eleito prefeito de Águas de São Pedro | Piracicaba e Região | G1", "Ele teve 43,73% dos votos dados a todos os candidatos e derrotou Maria Ely, que ficou em segundo lugar com 41,51%.", "João Victor Barboza, do Cidadania, foi eleito, neste domingo (15), prefeito de Águas de São Pedro (SP) para os próximos quatro anos. Ao fim da apuração, João Victor Barboza teve 43,73% dos votos. Foram 1.203 votos no total. O candidato derrotou Maria Ely, que ficou em segundo lugar com 41,51% (1.142 votos). A eleição em Águas de São Pedro teve 23,16% de abstenção, 1,94% votos brancos e 4,46% votos nulos. João Victor Barboza tem 30 anos, é solteiro, tem superior completo e declara ao TSE a ocupação de administrador. Ele tem um patrimônio declarado de R$ 91.408,83. O vice é Dr. Edison Xavier, do PSDB, que tem 60 anos. Os dois fazem parte da coligação Novos Tempos Para Águas de São Pedro, formada pelos partidos PL, CIDADANIA, PSDB, PODE e SOLIDARIEDADE. Veja o resultado após o fim da apuração: João Victor Barboza - CIDADANIA - 43,73%Maria Ely - DEM - 41,51%Moraes do Grande Hotel - PSL - 14,76% * Esta reportagem foi produzida de modo automático com o apoio de um sistema de inteligência artificial e foi revisada por um jornalista do G1 antes de ser publicada. Se houver novas informações relevantes, a reportagem pode ser atualizada. Saiba mais sobre o sistema de inteligência artificial usado pelo G1 em g1.com.br/eleicoes"] (ESTA PESQUISA, 2023)

[99] Segundo a afiliada da Globo, há 18 municípios que são cobertos pelo veículo na região.

O exemplo de Águas de São Pedro mostra uma estrutura de dados chamada de **dicionário**[100]. O próprio nome da cidade é uma **chave** do dicionário, com os seus **valores** postos entre colchetes e separados por vírgula. Tais valores no caso são justamente os elementos textuais da notícia. Observa-se que os parágrafos estão embutidos por meio de "*tabs*", os grandes espaços que precedem a primeira frase. Dentro desta estrutura de **dicionário**, há uma **chave** para cada cidade da afiliada de Piracicaba do estado de São Paulo. Ou seja, o processo automatizado de raspagem de textos retornou, um arquivo muito similar ao que Hector Iankovski descreve como sendo o *output*. No **ANEXO IV**, desta pesquisa encontram-se os links para os arquivos das **2.966** notícias raspadas neste mesmo formato aqui descrito.

Os arquivos "json" foram, portanto, o produto final da equipe de tecnologia. Os 5.558 documentos 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 (5) monitorar a geração dos arquivos no dia do pleito, conforme as urnas foram apuradas e os dados transmitidos do TSE para a Globo. Os integrantes da equipe de tecnologia descrevem essa parte do trabalho com um tom de casualidade, como se fosse a parte menos demandante do projeto.

> "No dia a gente começou a acompanhar a partir das quatro ou cinco horas da tarde, quando começou a chegar aos primeiros arquivos. A gente detectou um pequeno erro num dos formatos de *template*, que a gente tinha feito e testado, mas passou batido por todo mundo que estava acompanhando o desenvolvimento. A gente fez uma alteração rápida ali no código e corrigiu isso, mas no dia mesmo foi mais uma questão de monitoramento" (IANKOVISKI, 2023)

Embora os diversos testes tivessem sido conduzidos, ainda havia no dia do pleito espaço para ajustes no *template*. Conforme **esta pesquisa** demonstrou, um erro de concordância ainda passou despercebido e foi publicado. Segundo Tiago Reis, naquele ano o sistema de transmissão de dados do TSE passou algumas horas interrompido. Isso criou o que a equipe de tecnologia chamou de "o gargalo", uma impossibilidade temporária de gerar os textos na velocidade esperada. "Eu não sei se a gente falou isso, mas a gente teve alguns problemas com um próprio sistema do TSE que interrompeu a transmissão. Em algum momento no meio da eleição eles interromperam o envio dos dados, a gente ficou umas quatro ou cinco horas sem conseguir fazer nada " (REIS, 2022). O relato de Hector Iankovski é similar.

[100] Ver *Data Structures: Dictionaries*. Disponível em: https://docs.python.org/3/tutorial/datastructures.html. Acesso feito em 30 de abril de 2023.

> "A gente teve um grande problema no dia que não foi culpa nossa, foi uma culpa do TSE. Não sei se você lembra, se você pegar as notícias ali de 2020, você vai ver que teve um problema porque o TSE começou a utilizar um outro sistema para contagem e apuração dos votos. Isso gerou um gargalo e a distribuição dos arquivos por um formato ficou impactado [...] Isso atrasou a publicação dos textos. A nossa ideia era publicar os textos até as nove ou dez horas da noite, mas a gente só terminou de publicar os textos na segunda-feira pela manhã" (IANKOVISKI, 2023).

O gargalo relatado tanto pela equipe editorial, quanto pela equipe de tecnologia, é abordado por um comunicado do Tribunal Superior Eleitoral. Na terça-feira (17), uma nota técnica foi publicada pelo órgão admitindo que houve uma lentidão na totalização dos votos, o que Hector Iankovski e Tiago Reis descrevem como a causa do problema de geração dos textos.

> "O Tribunal Superior Eleitoral (TSE) divulgou, nesta terça-feira (17), Nota Técnica a respeito da lentidão da totalização dos votos no primeiro turno das eleições, ocorrido no último domingo (15). O atraso de aproximadamente duas horas e trinta minutos na divulgação dos resultados das Eleições Municipais foi ocasionado por recurso de inteligência artificial existente em um otimizador do banco de dados Oracle, que garante a velocidade no processamento das informações. Apesar disso, os cidadãos brasileiros tiveram acesso ao resultado das urnas em todo o país no mesmo dia da realização da votação, antes da meia-noite" (TSE, 2020).[101]

A lentidão pode ter sido causada por uma incompatibilidade de formatos de arquivo, como afirma Hector Iankovski, ou por um sobrecarregamento do sistema do TSE, como expõe a nota técnica. Não cabe a esta pesquisa precisar qual foi a causa desse gargalo, porém, é de interesse para a compreensão do jornalismo automatizado perceber que essas iniciativas estão sujeitas a serem impactadas por sistemas de terceiros, principalmente no momento da inovação, ou seja, quando o território técnico ainda é parcialmente inexplorado. Essa vulnerabilidade ao imprevisto pode ser explicada pelo que Cattani e Holzmann (2006) chamam de a lógica da *permanente mutação*. Como os conjuntos técnicos estão sempre mudando, ao passo que eles operam de forma interconectada, uma mudança em um ponto causa um "gargalo" no outro. Se uma gama de atores deve operar de maneira coordenada para que a automação funcione, um imprevisto de transmissão de dados é o bastante para inviabilizá-la.

Embora Hector Iankovski e Rafael Muniz sugiram que o monitoramento em tempo real do algoritmo foi a parte menos demandante do trabalho, o episódio do "gargalo" mostra

[101] Ver "*TSE divulga nota técnica sobre o atraso da totalização dos votos no primeiro turno"*. Disponível em: https://www.tse.jus.br/comunicacao/noticias/2020/Novembro/tse-divulga-nota-tecnica-sobre-o-atraso-da-totalizacao-dos-votos-no-primeiro-turno. Acesso feito em 30 de abril de 2023

que ela não foi menos importante. O fato de um erro no *template* ter sido notado durante o (5) monitoramento, mesmo depois da (3) condução de testes, também aponta nesse sentido. As cinco etapas do trabalho da equipe de tecnologia foram indispensáveis para a execução do projeto, cada uma com a sua relevância. Ao levar em conta os processos que Hector Iankovski e Rafael Muniz desenvolveram, é possível chegar a uma percepção mais concreta das dinâmicas do jornalismo automatizado.

4.4 OS TEXTOS ESCRITOS POR MÁQUINAS NAS ELEIÇÕES DE 2020

A análise de similitude das **2.966** notícias que compõem o *corpus* desta pesquisa 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%[102]. Ou seja, para as notícias que cobriam as prefeituras que tiveram o pleito definido no primeiro turno, a correspondência entre os valores foi total. Porém, 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 do documento**, a partir dos *template*s definidos por Tiago Reis e Felipe Grandin.

Essa equivalência entre os dados assevera os relatos da equipe editorial e da equipe de tecnologia sobre a origem das informações. Também reforça a importância da própria base de dados como um ator no jornalismo automatizado. Na mesma medida em que elas compõem o texto da notícia, as bases viabilizam a automação da redação dos textos pelo algoritmo de Geração de Linguagem Natural. A redação dos textos pelo processo algorítmico se confirma como única diferença entre o jornalismo de dados e o jornalismo automatizado (DIAKOPOULOS, 2019).

A análise das notícias também confirma o relato de Tiago Reis sobre os tipos de *templates* executados na automação. Há de fato quatro *templates* que conformam quase toda a cobertura. O primeiro *template*, também o mais frequente, é para vitória em primeiro turno. Segundo, há o *template* de disputa eleitoral prorrogada para o segundo turno. Terceiro, há o *template* para candidatura sub judice, aguardando decisão judicial. Quarto, o *template* de candidatura impugnada. Há ainda uma exceção, um *template* para o caso de empate técnico.

[102] 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.

Assim como há também as notícias que foram complementadas por jornalistas de carne e osso.

Tabela V - Relação de *templates* no projeto de jornalismo automatizado do G1

Tabela V - 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%

Fonte: 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[103]. Ou seja, o candidato que obtiver a maior proporção de votos válidos ganha o pleito, não importando se a diferença entre ele e o segundo candidato é de unidades, ou dezenas de votos, como já aconteceu no passado[104]. 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. No texto da lei está previsto a eleição em sistema majoritário, no caso dos municípios, apenas para cidades com mais de 200 mil habitantes.

Por uma questão puramente demográfica, um total de **94%** (IBGE, 2021) dos municípios brasileiros ficam abaixo da marca dos 200 mil habitantes. Portanto, elas concorrem em um sistema de contagem de votos que é desenhado para ter resultado logo no primeiro turno. A lista das cidades que passam pelo critério constitucional é publicada

[103] Ver "Como funciona o sistema proporcional?". Disponível em: https://www.tse.jus.br/institucional/escola-judiciaria-eleitoral/publicacoes/revistas-da-eje/artigos/revista-eletronica-eje-n.-5-ano-3/como-funciona-o-sistema-proporcional. Acesso feito em 31 de maio de 2023.
[104] Ver "Desempate por idade, vantagem de 1 voto: veja as disputas mais acirradas nas eleições de 2020". Disponível em: https://g1.globo.com/politica/eleicoes/2020/eleicao-em-numeros/noticia/2020/11/16/desempate-por-idade-vantagem-de-1-voto-veja-as-disputas-mais-acirradas-nas-eleicoes-de-2020.ghtml. Acesso feito em 31 de maio de 2020.

previamente pelo TSE. No ano de 2020 ela compunha 95 cidades, entre capitais, ex-capitais, cidades portuárias, industriais e turísticas[105]. O fato de as notícias seguirem esse padrão majoritário de *template* (Vitória em 1º Turno) indica uma **marca do contexto da cobertura** (SILVA-MAIA, 2011). O nível mais profundo a ser analisado em um esforço jornalístico, segundo a metodologia de análise de Silva e Maia (2011), são aqueles que expressam características conjunturais. A realização de um *template* em **96,4%** indica uma característica que diz respeito ao contexto do país, seu território, suas leis e seu sistema eleitoral.

A partir dessa constatação, é possível concluir que a preponderância do *template* **Vitória em 1º Turno** está diretamente associada a um fator que é tanto legal, quanto geográfico. No início desta pesquisa, foi exposto que para a *epistemologia neomaterialista* de André Lemos (2020), assim como para a filosofia da técnica de Gilbert Simondon (1988), as formações geográficas exercem poder de agência sobre os atores. Os agentes técnicos por sua vez só existem em um **Meio Associado,** que nesse caso é o próprio território brasileiro e seus agrupamentos urbanos. A relação entre a frequência de um tipo de notícia e a disposição dos municípios brasileiros é expressada na forma do *template* que o algoritmo mais executou. Tudo isso atravessado pelas regras do jogo eleitoral.

Segundo essas próprias regras, uma minoria de cidades estava apta a seguir para o próximo turno das eleições. Novamente uma **marca do contexto da cobertura** surge quando observamos as notícias que não foram concluídas no 1º turno (SILVA-MAIA, 2011). É o que vemos na segunda linha da tabela de **Disputa em 2º Turno**, onde somente 32 notícias realizam o respectivo *template.* No ano de 2020 haviam 95 cidades habilitadas a terem 2º turno, entre as quais 57 seguiram de fato para o próximo pleito[106]. Parte dessas 57 cidades ficaram de fora do *corpus*, por razões técnicas. O método de raspagem online de textos (ver **ANEXO IV**) foi capaz de reunir **53,5%** do total de notícias publicadas pelo projeto. De toda forma, as 32 notícias analisadas apresentaram o *template* que foi planejado para a situação. Abaixo pode ser visto a semelhança entre o lide do *template* realizado na eleição de Rio Branco (AC) e Piracicaba (SP).

> "Tião Bocalom, do PP, e Socorro Neri, do PSB, vão decidir em 2º turno, no próximo dia 29, quem será o próximo prefeito de Rio Branco (AC). Segundo dados do TSE (Tribunal Superior Eleitoral), Tião Bocalom teve 87.987 votos (49,58% dos votos),

[105] Ver "Eleições 2020: 95 municípios com mais de 200 mil eleitores poderão ter 2º turno em novembro". Disponível em: https://g1.globo.com/politica/eleicoes/2020/eleicao-em-numeros/noticia/2020/11/16/desempate-por-idade-vantagem-de-1-voto-veja-as-disputas-mais-acirradas-nas-eleicoes-de-2020.ghtml. Acesso feito em 31 de maio de 2020.

[106] Ver"Candidatos de 57 cidades disputarão prefeitura no 2º turno". Disponível em:

> ante 40.250 de Socorro Neri – o que representa 22,68% dos votos. A eleição em Rio Branco teve 27,23% de abstenção, 1,80% votos brancos e 3,18% votos nulos" (G1, 2020).
>
> "Barjas Negri, do PSDB, e Luciano Almeida, do DEM, vão decidir em 2º turno, no próximo dia 29, quem será o próximo prefeito de Piracicaba (SP). Com 100% das urnas apuradas, segundo dados do TSE (Tribunal Superior Eleitoral), Barjas Negri teve 56.760 votos (34,33% dos votos), ante 25.786 de Luciano Almeida – o que representa 15,60% dos votos. A eleição em Piracicaba teve 30,51% de abstenção, 8,10% votos brancos e 10,14% votos nulos" (G1, 2020).

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. Abaixo é possível ver o lide realizado pelo *template* **Sub Judice** nos municípios de Brejolândia (AL) e Lamim (MG).

> "Edezio Bastos, do DEM, foi eleito, neste domingo (15), prefeito de Brejolândia (BA) para os próximos quatro anos. Ao fim da apuração, Edezio Bastos teve 53,94% dos votos. Foram 3.983 votos no total[...] A candidatura de Edezio Bastos está indeferida com recurso no TSE. Isso ocorre quando o candidato não está regular e com pedido de registro julgado indeferido; no entanto, há recurso interposto contra essa decisão e aguarda julgamento por instância superior."(G1, 2020)
>
> "Roberto do Juca, do PP, foi eleito no domingo (15) prefeito de Lamim, pelos próximos quatro anos. Ele ficou com 57,66% do votos ao final da apuração. Foram 1.712 votos no total [...] A candidatura de Roberto do Juca está indeferida com recurso no TSE. Isso ocorre quando o candidato não regular e com pedido de registro julgado indeferido; no entanto, há recurso interposto contra essa decisão e aguarda julgamento por instância superior" (G1, 2020)

O quarto 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. 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. Neste caso, o critério de desempate é a idade. A notícia pode ser vista abaixo.

> "Após um empate de votos nas eleições de Kaloré, no norte do Paraná, o candidato mais velho acabou sendo eleito prefeito da cidade, no domingo (15). Edmilson, do PL, será o gestor do município pelos próximos quatro anos. Tanto Edmilson, do PL, quanto Ritinha, do PSD, tiveram 1.186 votos cada, com 37,14% dos votos. No entanto, pelo critério de desempate do Tribunal Superior Eleitoral, o candidato Edmilson foi eleito, por ser mais velho. Ele tem 59 anos, enquanto Ritinha tem 41. A eleição em Kaloré teve 12,66% de abstenção, 0,82% votos brancos e 1,98% votos nulos. Edmilson tem 59 anos, é casado, tem ensino superior completo e declara ao TSE a ocupação de agricultor. Ele tem um patrimônio declarado de R$ 65.151,26. O vice é Spadin, do PDT, que tem 31 anos. Os dois fazem parte da coligação Coligação Vamos Caminhar Juntos, formada pelos partidos DEM, PDT, PSB e PL".(G1, 2020)

No exemplo de Kaloré, observa-se novamente o caráter preditivo do trabalho jornalístico na automação de notícias. Ao definirem os *templates*, os jornalistas precisam antecipar as exceções aos casos majoritários para evitar erros. Se o algoritmo se depara com um caso em que não está previsto uma **decisão**, ele pode travar a geração de outros textos, entrando em loop, ou simplesmente não retornar o texto para aquele caso. Conforme expôs Grandim, isso tudo é feito por uma **árvore de decisão**. Se o candidato obtiver mais votos, seu nome entra no título como ganhador. Mas se o valor foi o mesmo de seu concorrente, por mais raro que é o caso, o algoritmo tem que ser capaz de decidir qual template vai realizar. Vale lembrar que a árvore de decisão corresponde ao ***throughput*** no processo algorítmico (LATZER, 2014). Sendo portanto os meandros da realização do *template* frente às características das informações presentes nas bases de dados (*input*).

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) era homem e 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. Conforme os valores expostos na tabela abaixo, entre os candidatos eleitos no 1º turno, a maioria preponderante foi de prefeitos homens.

Tabela VI - Conjugação de gênero nos templates de jornalismo automatizado do G1

Conjugação de gênero nos templates de jornalismo automatizado do G1		
Gênero	Quantidade	Porcentagem
Masculino	2.489	89,6%
Feminino	339	11,4%

Fonte: Esta Pesquisa

O mais pertinente é que a partir da análise das notícias, observando sua conjugação de gênero, se chegou a um percentual muito próximo da representativa de homens e mulheres na política brasileira. Segundo o TSE[107], no ano de 2020 somente 12% dos municípios elegeram mulheres, apesar de mulheres serem 51,8% da população e 52% do eleitorado. O valor encontrado de 11,4% se aproxima daquele declarado pelo TSE. Não é objetivo desta pesquisa adentrar uma discussão sobre representatividade na política, ou em disputas eleitorais. A intenção de analisar a conjugação é a de expor um processo de Geração de Linguagem Natural, que invariavelmente toca em questões de gênero pela própria estrutura da língua portuguesa. Mas é curioso como algumas dessas questões acabam emergindo quando se analisa um fenômeno sociotécnico de dimensão nacional. A proporção entre candidatos eleitos homens e mulheres, invariavelmente se constitui como uma **marca do contexto da cobertura,** por se tratar de um fenômeno conjuntural (SILVA-MAIA, 2011).

Para além dos *templates* e da conjugação, uma outra característica pode ser observada no conjunto dos textos. Algumas das publicações, principalmente aquelas das cidades mais proeminentes, foram complementadas por humanos. A ideia de 'proeminência' não é de fácil definição, mas nota-se nas cidades que contam com afiliadas da Globo, ou são capitais, a repetição deste padrão. Aqui, a agência humana de jornalistas anônimos se mostra como a causa de textos que seguiram um template, mas foram complementados, ou alterados. Na explicação de Tiago Reis, isso estava previsto para acontecer desde o início do projeto. Algorítmos e humanos escrevendo a quatro mãos.

> **"**Depois que subimos os textos no CMA, eles estavam prontos, mas podiam ser alterados. Via de regra, nas principais cidades, eu digo nas capitais, quase todos devem ter feito isso manualmente. Eu recomendava para eles que os textos automatizados estariam ali, mas é claro que a filial tinha liberdade para preparar mais alguma coisa na véspera" (REIS, 2022).

Na fala de Tiago Reis fica claro mais uma vez o caráter sociotécnico do projeto. Algumas notícias poderiam carregar as palavras do ator não-humano (algorítmico), assim como dos atores humanos. De certa forma, foi isso o que decorreu ao longo de toda a automação. Mas nesses casos, os textos passam ainda por uma edição após a geração do

[107] Ver "Mulheres representam apenas 12% dos prefeitos eleitos no 1º turno das Eleições 2020". Disponível em: https://www.tse.jus.br/comunicacao/noticias/2020/Novembro/mulheres-representam-apenas-12-dos-prefeitos-eleitos-no-1o-turno-das-eleicoes-2020. Acesso feito em 1 de junho de 2023.

algoritmo e postagem no sistema de gestão de conteúdo (CMA). Abaixo é possível ver um exemplo de textos que foi complementado por humanos, em Maceió (AC).

> Alfredo Gaspar de Mendonça, do MDB, e JHC, do PSB, vão decidir em 2º turno, no próximo dia 29, quem será o próximo prefeito de Maceió. Com 100% das urnas apuradas, segundo dados do Tribunal Superior Eleitoral (TSE), Gaspar teve 110.234 votos (28,87%), e JHC teve 109.053 votos (28,56%). A eleição em Maceió teve 148.318 abstenções (25,04%), 21.001 votos brancos (4,73%) e 41.261 votos nulos (9,29%). Alfredo Gaspar de Mendonça Neto tem 50 anos, é formado em direito e pós-graduado em direito público. Foi promotor de Justiça por mais de 20 anos, procurador-geral de Justiça e secretário estadual da Segurança Pública no primeiro mandato do governador Renan Filho (MDB). Ele tem um patrimônio declarado de R$ 341.626,09. O vice é Tácio Melo (PODE), de 40 anos."As minhas primeiras palavras são de gratidão. Gratidão a Deus e gratidão ao povo de Maceió. Eu não sou da política, é a minha primeira eleição. E isso mostra que a minha história de vida está sendo o meu cartão de visita pra o futuro. Eu vou dialogar muito porque eu estou pronto pra, com muita força, ao lado do povo de Maceió, resolver os problemas da minha cidade", disse Gaspar. João Henrique Holanda Caldas, o JHC, tem 33 anos, é advogado, especialista em direito digital e compliance, mestrando em gestão pública e empreendedor. Foi eleito em 2014 como o deputado federal mais votado de Alagoas. Nas eleições gerais de 2018, foi candidato à reeleição, sendo o deputado federal mais bem votado do Brasil, proporcionalmente. Tem um patrimônio declarado de R$ 1.973.391,98. O vice é Ronaldo Lessa (PDT), de 71 anos. "Com certeza a gente tem as melhores propostas para Maceió, nós somos a mudança, e a gente vai conseguir, através do povo, do interesse público e não do interesse particular e de grupos políticos, revelar no segundo turno a vitória que virá das urnas, falada pela população, que é o sentimento real de mudança, com alguém que esteja preparado para poder guiar os destinos da nossa cidade", afirmou. O plano de governo de Alfredo Gaspar tem como base cinco pilares para o desenvolvimento de Maceió, que vão desde implementação de políticas públicas e ações com foco na qualidade de vida, passando pela cidade sustentável, atuando para reduzir as desigualdades socioterritoriais; políticas que devem impulsionar a economia local através da geração de renda [...] Veja o resultado após o fim da apuração: Alfredo Gaspar de Mendonça (MDB): 28,87% JHC (PSB): 28,56% Davi Davino Filho (Progressistas): 25,5% Josan Leite (Patriota): 6,27% Valéria Correia (PSOL): 3,55% Cícero Almeida (DC): 2,59% Ricardo Barbosa (PT): 2,33% Lenilda Luna (UP): 1,28% Cícero Filho (PCdoB): 0,90% Corintho Campelo (PMN) 0,13%”(G1, 2020).

Acima vemos como a notícia trás os mesmos dados adquiridos nas bases de dados **Resultados** e **Divulgacand**, dispostas em ordem muito similar aos textos de outros *templates*. Há o nome do candidato mais votado, assim como do segundo colocado, o partido, o número de votos totais e percentuais, bem como a idade, o grau de instrução, a profissão e o patrimônio declarado. A formatação permanece a mesma. O texto é finalizado com a mesma lista contendo a porcentagem de votos dos outros colocados, padrão que se repetiu em todas as 2.966 notícias. Porém, são notáveis as inclusões que foram feitas por jornalistas. Entre essas inclusões estão as aspas dos dois candidatos, no caso Alfredo Gaspar de Mendonça e João Henrique Caldas (JHC), propostas de plano de governo e informações mais detalhadas sobre a trajetória política dos candidatos.

Os complementos feitos pelos jornalistas/revisores reforçam o ponto defendido ao longo desta pesquisa de que os fenômenos comunicativos ocorrem em uma cooperação entre atores humanos e não-humanos (LATOUR, 2012). Neste projeto de jornalismo automatizado, o agenciamento inicial dos jornalistas Tiago Reis e Felipe Grandin atravessa um conjunto de objetos técnicos até voltar à mão de humanos que podem reescrever uma notícia, originalmente redigida por um algoritmo. O ato complementar às notícias não acontece na maioria das cidades, por haver um entendimento de que isso seria reservado às "cidades mais importantes". Por mais que Tiago Reis não especifique o que essa "importância", há um senso tácito de que a demografia brasileira, mais as regras do sistema majoritário do TSE, trazem essa definição. Mais uma **marca de contexto da cobertura jornalística** se expressa mesmo nos casos excepcionais (SILVA-MAIA, 2011).

A partir da análise do conjunto de 2.966 notícias percebe-se que os *templates* foram preparados, executados e, em alguns casos, complementados manualmente. É patente que nem o portal G1, nem nenhum outro, havia tentado antes de 2020 publicar uma notícia para cada cidade do Brasil, em menos de 48 horas. Essa façanha só foi possibilitada pela combinação de diferentes profissionais, diferentes tecnologias e diferentes instituições. O jornalismo automatizado viabilizou, portanto, um episódio histórico para a cobertura eleitoral do país. É possível argumentar que os textos carecem de elementos visuais, profundidade, ou mesmo de 'estilo', mas é impossível não destacar a escala, abrangência e agilidade do projeto.

Segundo os idealizadores da iniciativa, Tiago Reis e Felipe Grandin, o algoritmo foi o "potencializador" do trabalho jornalístico. Para eles, a cidade de Borá em São Paulo, com seus 800 habitantes, nunca viria um texto sobre sua eleição publicado em um portal de notícias. Porém, esse potencializador não acontece sem reveses. Meses de trabalho prévio foram necessários para um único dia de cobertura. Alguns erros foram replicados pelo algoritmo sistematicamente. Sem falar na dependência que os jornalistas criam em relação a ferramentas e outros profissionais para fazer esse tipo de cobertura.

As mudanças tecnológicas parecem ocorrer com uma dose de tentação e outra de cobrança. Inovações como a geração automática de textos abrem portas para coberturas outrora consideradas impossíveis. Esta pesquisa buscou mostrar de forma instrutiva **como** o jornalismo automatizado pode ser feito, a partir da exposição de um caso concreto. Mas é evidente que a emergência deste campo não acontece sem que alguns perigos venham à tona. No próximo capítulo abordaremos alguns desses perigos, assim como possibilidades de pesquisa, que as notícias automatizadas suscitam.

5. CONSIDERAÇÕES FINAIS

O termo 'velocidade' foi cunhado um total de 20 vezes ao longo desta pesquisa. 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). A fala dos entrevistados não poderia ser mais transparente. Para Felipe Grandin e Tiago Reis, o emprego de algoritmos de redação era a única forma de viabilizar a cobertura de 5.568 municípios. Entretanto, a velocidade também é o principal entrave deste trabalho para definir qual é a dinâmica do jornalismo automatizado. As mudanças tecnológicas acontecem em um ritmo desenfreado. O fenômeno da automação de textos está em pleno devir.

Viu-se no caso do G1 nas Eleições Municipais de 2020 uma forma de automatizar notícias, a partir de uma automação algorítmica por Regras & Heurística (VAJJALA *et al*, 2020), que aplicou uma das nove formas de Geração de Linguagem Natural, o método de *expressões-chave e templates* (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. Para dimensionar a complexidade do problema, há ainda outras oito formas de Geração de Linguagem Natural, outras vias (ex: imagem-para-texto) e outras finalidades, como síntese textual e sumarização. Isso sem falar na geração de outros formatos como vídeos e imagens que tem se popularizado e, sem dúvidas produzem impacto no jornalismo[108].

Em 2022, o G1 refez o projeto, dessa vez para eleições legislativas e presidenciais. Pela fala do então coordenador, Felipe Grandin, sabe-se que a equipe de revisores foi dispensada. O texto gerado pelo algoritmo foi publicado diretamente no portal. Além disso, cada notícia contou com um vídeo automatizado. Duas mudanças significativas para o jornalismo automatizado ocorreram dentro do mesmo veículo no curso de dois anos. O *Chat GPT* 3 também foi lançado nesse ínterim e alcançou 100 milhões de usuários em um mês. É esse tipo de velocidade que 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.

Primeiro, 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

[108] Ao menos a revisão de literatura trouxe um quadro de referências vasto que pode ajudar na classificação de outros casos de jornalismo automatizado.

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). Na cobertura do G1, quatro *templates* foram executados para redigir lides que dessem conta de todos os cenários possíveis daquela cobertura. Vitória em 1º turno, Decisão em 2º turno, Candidatura Sub Judice e Candidatura Impugnada foram os modelos de texto imaginados para conformar todos os desfechos de uma corrida eleitoral. Houve ainda um quinto caso, bastante raro, de Empate Técnico, que foi previsto para lidar com uma situação anômala, mas ainda assim possível.

Também observa-se na realização dos *templates*, 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 Ator Rede, assim como da ideia de *assemblage* (DELANDA, 2016), que se expressa nas notícias a partir dessa redação mista, feita camada a camada por diferentes agentes. A cooperação entre humanos e objetos técnicos aparece no jornalismo algorítmico de Nicholas Dorr, usado para compor o conceito de jornalismo automatizado (ver capítulo 3). 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 que parece à primeira vista uma questão estética, se revela também como uma transferência de autoridade (HARARI, 2018). 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 por que 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 de maneira recorrente na análise do projeto e na literatura especializada. Os membros da equipe editorial repetiram diversas vezes como a cobertura precisa ser **previsível** para 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. A necessidade de antecipar os cenários se expressa no planejamento do *template* (REITER, 2012), a primeira etapa do processo de geração de textos.

Em segundo lugar, 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, evitando problemas de *lag*[109]. Segundo o repórter Felipe Grandin, existem várias bases online no Brasil, mas não foi por acaso que os repositórios do TSE foram usados. A base do Tribunal Superior Eleitoral é 'limpa', 'robusta' e 'transparente'. Para o cientista de dados Hector Iankoski, ela suporta uma requisição de dados substancial, sua arquitetura é conhecida e, por isso, cumpre o seu papel. 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. Esse cenário amplo no qual o jornalismo se insere é chamado por autores como Van Dijk (2017) e Shoshana Zuboff (2021) de **datificação**. Tal acontecimento fica mais claro nas técnicas do jornalismo de dados - também referido como Jornalismo em Base de Dados (BARBOSA, 2007) - alinhadas à Geração de Linguagem Natural para produzir as notícias em grandes quantidades.

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. Isso fica explícito na fala do jornalista Tiago Reis. "Ao meu ver esse projeto não é um exemplo de como robôs podem substituir humanos e sim, de ajudar numa tarefa que ia ser impossível um humano fazer sozinho" (REIS, 2022). A segunda parte desta declaração[110] resume a origem do jornalismo automatizado. A complexidade de certas coberturas torna o desafio impossível para repórteres humanos lidarem por conta própria. O emprego da Geração de Linguagem

[109]O *lag* é o atraso (latência) entre a ação do usuário (input) e a reação do servidor que suporta a base de dados, na função de retornar os valores para o usuário.
[110] Não podemos concluir, a partir dos dados dessa pesquisa, que tipo de consequência a automação de notícias tem para a empregabilidade de jornalistas.

Natural é um instrumento para enfrentar a abundância de informações (MEYER, 2002) e interpretar a vastidão do mundo virtual (CODDINGTON, 2015).

Embora fatores econômicos estejam por trás do advento do jornalismo automatizado, eles se mostraram como laterais. Segundo Aljazairi (2016), a redução de custos é o principal motivo para que a automação de notícias seja buscada por empresas jornalísticas. É óbvio que pelo caráter industrial desse tipo de periodismo, estratégias comerciais podem influenciar um veículo a buscar o jornalismo automatizado como solução para desafios financeiros. Entretanto, isso não apareceu na fala dos entrevistados em nenhum momento. De maneira mais latente, a automação se apresentou como solução para lidar com um problema de *Big Data* (MAYER-SCHÖNBERGER-CUKIER, 2013).

Considerando a relevância das bases de dados, também podemos concluir a **cultura de dados abertos** como um fator-chave para o jornalismo automatizado. Esse princípio da **cibercultura** (GRAY et al, 2014), já presente no *ethos* de jornalistas de dados, se repete várias vezes na fala dos entrevistados. Graças à agenda que o TSE estabelece junto à sociedade civil, graças às instruções nos arquivos "*readme*" e graças aos testes que podem ser conduzidos antes dos pleitos, o G1 foi capaz de gerar textos automatizados. Novamente, as bases de dados demonstram o seu caráter viabilizador. O jornalismo automatizado acontece por conta da transparência de instituições, acessada por profissionais que se sentem no dever **entrevistar planilhas** (TRASEL, 2014) e monitorá-las.

A transparência das instituições também deu espaço para expor as **marca de contexto da cobertura jornalística** (SILVA-MAIA, 2011). O jornalismo automatizado expressa aspectos conjunturais da sociedade em que ele reporta. A questão da formação geográfica brasileira ficou evidente nos tipos de *templates* realizados: cidades menores concorrem em um sistema de votação (majoritário), enquanto cidades grandes concorrem no sistema proporcional. Outra marca de contexto que apareceu nas notícias foi a preponderância masculina na política brasileira. A questão da representatividade se mostrou na conjugação que a Geração de Linguagem Natural executou sobre os *templates*. Segundo o TSE, somente 12% dos municípios elegeram mulheres em 2020, apesar de elas serem 52% do eleitorado. O valor encontrado de 11,4% se aproximou daquele declarado pelo tribunal. O que fica claro é que o algoritmo replica os princípios e vieses que estão presentes nos dados estruturados. Se há mais cidades pequenas com prefeitos homens, os textos estarão de acordo. O produto final é consequência direta do *input*.

Isso nos leva à questão dos erros. Essa pesquisa identificou 15 erros de concordância nas 2.966 notícias analisadas. É premente nas falas dos jornalistas e cientistas da computação

que a preocupação em evitar equívocos acompanhou todo o projeto. Vários testes e revisões foram feitas para que não houvesse erros de ortografia ou concordância. No final das contas, um número pouco expressivo foi de fato para o site, mas mesmo assim, eles foram publicados. Talvez esse seja o maior perigo do jornalismo automatizado: se o ganho de escala é uma das vantagens, escalar "barrigadas", como é dito no jargão informal da profissão, é sua maior desvantagem.

A possibilidade de **multiplicar equívocos** foi uma das cinco questões éticas que esta pesquisa identificou sobre os desafios que o jornalismo automatizado suscita. Autores como Clerwall (2014), Montal e Reich (2017), Ali e Hassoun (2019) e Monti (2019) destacam outros quatro entraves éticos que vem a tona: transparência, checagem de fatos, justeza (fairness), utilização e qualidade dos dados. A transparência envolve fornecer informações claras sobre a origem dos dados usados na produção automatizada das notícias. A checagem de fatos é relevante para garantir a credibilidade das fontes e evitar a propagação de informações falsas. A justeza diz respeito à proteção da privacidade dos usuários e prevenção de manipulações sociais. A utilização e qualidade dos dados dizem respeito à precisão e objetividade das informações apresentadas nas notícias. Por último, a automação da produção de ***fake news*** é uma ameaça adicional ao jornalismo profissional. Todas essas cinco questões podem servir de norte para pesquisadores que busquem investigar os efeitos negativos do jornalismo automatizado.

Uma outra hipótese provável é que novas bases de dados habilitem novas iniciativas de jornalismo automatizado. Não é possível concluir isso a partir do estudo de um único caso, porém, Felipe Grandin declarou em sua entrevista que uma outra fonte de dados robusta era tudo que eles precisavam para repetir o projeto em outras editorias. Vimos ao longo do subcapítulo "**3.5. AUTOMAÇÕES NAS REDAÇÕES**" vários exemplos de outros projetos que foram implementados por já terem acesso a bases de dados consolidada sobre esportes, meteorologia, finanças e política. Isso nos leva a concluir que o jornalismo automatizado **não está limitado a uma temática**, podendo ser replicado para outras editorias. A partir dessa hipótese, outras pesquisas no futuro poderiam verificar a relação entre novas bases de dados e novas automações de notícias.

Em quarto lugar, esta dissertação aponta a *interdisciplinaridade* como aspecto central do jornalismo automatizado. Os atores da equipe editorial e da equipe de tecnologia trabalharam lado a lado para concretizar o projeto. A colaboração entre diferentes profissionais foi fundamental ao longo de todo processo. Todavia, os próprios membros da equipe editorial não são 'jornalistas tradicionais', carregando técnicas e a literacia da

computação e dos meios digitais. Logo, o jornalismo automatizado é descendente de profissionais com uma **inteligência híbrida** (SANTOS, 2018b). O hibridismo destes profissionais é o que permitiu a ideação do projeto, as trocas com os cientistas da computação e, mais importante, que a ciência da computação e jornalismo político se encontrassem para gerar uma síntese. "Com certeza esse contato nosso com a linguagem de programação ajudava muito nessas conversas" (REIS, 2022). O cientista da computação Hector Iankovski corrobora que os conhecimentos de Tiago Reis e Felipe Grandin os destacavam nas reuniões que ocorreram com outras equipes da Globo.

É evidente que inteligência híbrida e interdisciplinaridade podem ser sinônimos. O jornalismo automatizado, assim como diversas outras vertentes do jornalismo digital (KAWAMOTO, 2003), são herdeiros dessas misturas. Por essa razão, objetos de estudo como esse só podem ser compreendidos em sua interdisciplinaridade. Muitos dos pensadores citados na fundamentação teórica são defensores de uma agenda de pesquisa igualmente híbrida (CASWELL-DORR, 2018; DIAKOPOULOS, 2019; MAYER-SCHÖNBERGER-CUKIER, 2013; SANTOS, 2016, LEMOS, 2020; LIMA JUNIOR, 2011). A experiência pessoal do responsável por esta pesquisa foi de conduzir estudos paralelos sobre ciência da computação ao longo de dois anos. Sem estudos intensos sobre análise de dados, linguagens de programação e teoria da computação, não haveria um diálogo produtivo com os entrevistados. A literacia e o vocabulário técnicos são chaves para investigar os efeitos da tecnologia em fenômenos comunicacionais. O mais interessante é que no esforço de compreender um objeto híbrido, o próprio pesquisador teve que se transformar em um híbrido.

Portanto, esta pesquisa faz coro a outros trabalhos que defendem uma agenda interdisciplinar e interdepartamental que busquem compreender os efeitos dos objetos técnicos sobre o social. No caso do jornalismo automatizado, a associação entre diferentes atores é o que concebe a emergência do campo, desde o princípio. Não são poucos os exemplos e as obras que nos levam a olhar para a materialidade dos acontecimentos sociais. Portanto, a ação dos objetos técnicos não pode ficar de fora. Parafraseando Gilbert Simondon (1988, p.117), não existimos neste mundo sozinhos, pois "a vida técnica não consiste em dirigir máquinas, mas existir do mesmo nível que elas".

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, Christopher William. **Apostles of certainty: Data journalism and the politics of doubt.** Oxford University Press, 2018.

ANANNY, Mike. **Toward an ethics of algorithms: Convening, observation, probability, and timeliness**. Science, Technology, & Human Values, v. 41, n. 1, p. 93-117, 2016.

ALI, Waleed; HASSOUN, Mohamed. **Artificial intelligence and automated journalism: Contemporary challenges and new opportunities.** International journal of media, journalism and mass communications, v. 5, n. 1, p. 40-49, 2019.

Artificial Intelligence (AI). **IBM**, 2020. Disponível em <.https://www.ibm.com/cloud/learn/what-is-artificial-intelligence> Acesso em: 16 de fevereiro de 2022.

ALJAZAIRI, Sena. **ROBOT JOURNALISM: THREAT OR AN OPPORTUNITY.** 2016.

BARBOSA, Suzana. **Jornalismo digital em base de dados (JDBD): um paradigma para produtos jornalísticos digitais dinâmicos**. 2007.

Bloomberg. **Using Bloomberg automated news stories to predict market events**, Bloomberg, 2021. Disponível em: https://www.bloomberg.com/professional/blog/using-bloomberg-automatednews-stories-to-predict-market-events. Acesso em: 16 de agosto de 2022.

BISHOP, C. M. (1994). **Neural networks and their applications. Review of Scientific Instruments**.65, article 1803. Disponível em: https://doi.org/10.1063/1.1144830. Acesso em: 09 de fevereiro de 2022.

BRIGGS, Asa; BURKE, Peter. **A social history of the media: From Gutenberg to the Internet**. Polity, 2009.

CASWELL, David; DÖRR, Konstantin. **Automated Journalism 2.0: Event-driven narratives: From simple descriptions to real stories**. Journalism practice, v. 12, n. 4, p. 477-496, 2018.

CARLSON, Matt. The robotic reporter: **Automated journalism and the redefinition of labor, compositional forms, and journalistic authority.** Digital journalism, v. 3, n. 3, p. 416-431, 2015.

CARLSON, Matt. Automated journalism: **A posthuman future for digital news?. In: The Routledge companion to digital journalism studies**. Routledge, 2016

CODDINGTON, Mark. **Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting.** Digital journalism, v. 3, n. 3, p. 331-348, 2015.

COULDRY, N; HEPP, A. **The Mediated Construction of Reality**. Cambridge: Polity Press, 2016.

CHRISTOFOLETTI, Rogério. **A crise do jornalismo tem solução?** ESTAÇÃO DAS LETRAS E CORES EDI, 2019.

CHRISTOFOLETTI, Rogério; LAUX, Ana Paula França. **Confiabilidade, credibilidade e reputação: no jornalismo e na blogosfera**. Intercom-Revista Brasileira de Ciências da Comunicação, v. 31, n. 1, p. 29-50, 2008.

CHRISTIN, Angèle. **Metrics at work: journalism and the contested meaning of algorithms.** Princeton University Press, 2020.

CLARK, Alexander; FOX, Chris; LAPPIN, Shalom (Ed.). **The handbook of computational linguistics and natural language processing**. John Wiley & Sons, 2012.

CLERWALL, Christer. Enter the robot journalist: Users' perceptions of automated content. Journalism practice, v. 8, n. 5, p. 519-531, 2014.

C. Tandoc Jr., E., Wu, S., Tan, J., & Contreras-Yap, S. (2022). O que são notícias (automatizadas)? Uma análise de conteúdo de artigos noticiosos escritos por algoritmos. Media & Jornalismo, 22(41), 103-120. https://doi.org/10.14195/2183-5462_41_6

ECO, Umberto. **Apocalípticos e integrados**. DEBOLSILLO, 2011.

DE BRUYNE, Paul; HERMAN, Jacques; DE SCHOUTHEETE, Marc. **Dinâmica da pesquisa em ciências sociais.** Rio de Janeiro: Francisco Alves, 1991.

DEUZE, Mark. What is multimedia journalism? **Journalism studies**, v. 5, n. 2, p. 139-152, 2004.

DELANDA, Manuel. Assemblage theory. Edinburgh University Press, 2016.

DIAKOPOULOS, Nicholas. **Automating the news**. Harvard University Press, 2019.

DUARTE, Jorge. Entrevista em profundidade. **Métodos e técnicas de pesquisa em comunicação. São Paulo: Atlas**, v. 1, p. 62-83, 2005.

DONG, Chenhe et al. A **survey of natural language generation**. ACM Computing Surveys, v. 55, n. 8, p. 1-38, 2022.

DÖRR, Konstantin Nicholas. **Mapping the field of algorithmic journalism**. Digital journalism, 2015.

FANTA, Alexander. **Putting Europe's robots on the map: Automated journalism in news agencies**. Reuters Institute Fellowship Paper, v. 9, p. 1-23, 2017.

KIM, Daewon; KIM, Seongcheol. **Newspaper journalists' attitudes towards robot journalism. Telematics and Informatics**, v. 35, n. 2, p. 340-357, 2018.

GRAEFE, Andreas. **Guide to automated journalism.** 2016.

GRAY, Jonathan; CHAMBERS, Lucy; BOUNEGRU, Liliana. **The data journalism handbook: how journalists can use data to improve the news.** O'Reilly Media, Inc, 2012.

GROOVER, Mikell P. Automation, Production Systems, and Computer‐integrated Manufacturing 2nd ed. **Assembly Automation**, 2002.

GRUSIN, R. (Ed.), The Nonhuman Turn. Minneapolis: University of Minnesota Press, 2015.

Guzman, Andrea L, and Seth C. Lewis. 2019. "Artificial Intelligence and Communication: A Human–Machine Communication Research Agenda." New Media & Society. doi:10.1177/1461444819858691

_____. **What is human-machine communication, anyway. Human-machine communication: Rethinking communication, technology, and ourselves**, p. 1-28, 2018.

HOBSBAWM, Eric J. The machine breakers. **Past & Present**, n. 1, p. 57-70, 1952.

HOPCROFT, John E.; ULLMAN, Jeffrey D.; MOTWANI, Rajeev. Introdução à teoria de autômatos, linguagens e computação. **Rio de Janeiro: Campus**, 2002.

HJARVARD, Stig. Midiatização: conceituando a mudança social e cultural. Matrizes, v. 8, n. 1, p. 21-44, 2014.

HARARI, Yuval Noah. 21 lições para o século 21. Editora Companhia das Letras, 2018.

HARARI, Yuval Noah. Homo Deus: uma breve história do amanhã. Editora Companhia das Letras, 2016.

HARAWAY, Donna J. **Manifestly haraway**. U of Minnesota Press, 2016.

JONES, Steven E. **Against technology: From the Luddites to neo-Luddism**. Routledge, 2013.

JÜRGENS, Pascal; JUNGHERR, Andreas; SCHOEN, Harald. Small worlds with a difference: New gatekeepers and the filtering of political information on Twitter. In: Proceedings of the 3rd international web science conference. 2011. p. 1-5.

KAWAMOTO, Kevin (Ed.). **Digital journalism: Emerging media and the changing horizons of journalism**. Rowman & Littlefield Publishers, 2003.

KIM, Daewon; KIM, Seongcheol. **Newspaper journalists' attitudes towards robot journalism**. Telematics and Informatics, v. 35, n. 2, p. 340-357, 2018.

LAGE, Nilson. Ideologia e técnica da notícia. Digitaliza Conteúdo, 2021.

LATAR, Noam Lemelshtrich. **The robot journalist in the age of social physics: The end of human journalism?**. In: The new world of transitioned media. Springer, Cham, 2015. p. 65-80.

Lecompte, C. (2015). "**Automation in the Newsroom. How algorithms are helping reporters expand coverage, engage audiences, and respond to breaking news**". En Niemanreports.org: https://niemanreports.org/articles/ automation-in-the-newsroom/

LATOUR, Bruno et al. **Reagregando o Social: uma introdução à teoria Ator-Rede**. EDUSC, 2005.

LATZER, Michael et al. **The economics of algorithmic selection on the Internet. In: Handbook on the Economics of the Internet.** Edward Elgar Publishing, p. 395-425, 2016.

_____; FESTIC, Noemi. **A guideline for understanding and measuring algorithmic governance in everyday life.** Internet Policy Review, v. 8, n. 2, 2019.

LEMOS, André. Epistemologia da comunicação, neomaterialismo e cultura digital. **Galáxia (São Paulo)**, p. 54-66, 2020.

_____; BITENCOURT, Elias Cunha. Antropocentrismo e Comunicação: Análise dos GT da COMPÓS "Epistemologia da comunicação" e "Comunicação e Cibercultura" de 2017 a 2019. Fronteiras-estudos midiáticos, v. 23, n. 1, p. 40-56, 2021.

LEVY, Steven. **The rise of the robot reporter.** Wired, v. 20, n. 5, p. 132-139, 2012.

LEWIS, Seth C.; GUZMAN, Andrea L.; SCHMIDT, Thomas R. Automation, journalism, and human–machine communication: Rethinking roles and relationships of humans and machines in news. Digital Journalism, v. 7, n. 4, p. 409-427, 2019.

_____; WESTLUND, Oscar. Actors, actants, audiences, and activities in cross-media news work: A matrix and a research agenda. Digital journalism, v. 3, n. 1, p. 19-37, 2015.

Llorente, D (2016). **Automatic natural language generation, the new "normal"** Disponível em: https://medium.com/@davidllorente/automatic-natural-language-generation-the-new-normal-cd36ed8976de. Acesso feito em 20 de setembro de 2022.

Linden, C. G. (2017). **Decades of Automation in the Newsroom**. Digital Journalism, 5 (2), 123–140. DOI: 10.1080/21670811.2016.1160791

MACHILL, Marcel; BEILER, Markus; ZENKER, Martin. Search-engine research: a European-American overview and systematization of an interdisciplinary and international research field. Media, Culture & Society, v. 30, n. 5, p. 591-608, 2008.

MARCHAND, Pascal; RATINAUD, Pierre. **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, v. 2012, p. 687-699, 2012.

MANOVICH, Lev. **El software toma el mando**. Editorial UOC, 2014.

MATSUMOTO, Rie et al. **Journalist robot: Robot system making news articles from real world.** In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2007. p. 1234-1241.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big data: A revolution that will transform how we live, work, and think**. Houghton Mifflin Harcourt, 2013.

MCDONALD, David D. **Natural Language Generation: Handbook of Natural Language Processing**, v. 2, p. 121-144, 2010.

MEDINA, Cremilda. **Notícia, um produto à venda: jornalismo na sociedade urbana e industrial.** Summus Editorial, 1988.

MEYER, Philip. 2002. Precision Journalism. 4th ed. Lanham, MD: Rowman and Littlefield.

MIELNICZUK, Luciana. **Jornalismo na Web: uma contribuição para o estudo do formato da notícia na escrita hipertextual**. 2003.

MONTAL, Tal; REICH, Zvi. I, robot. You, journalist. Who is the author? Authorship, bylines and full disclosure in automated journalism. Digital journalism, v. 5, n. 7, p. 829-849, 2017.

MONTI, Matteo. Automated journalism and freedom of information: Ethical and juridical problems related to AI in the press field. Opinio Juris in Comparatione, v. 1, p. 2018, 2019.

MUSIANI, Francesca. **Governance by algorithms**. Internet Policy Review, v. 2, n. 3, p. 1-8, 2013.

NAPOLI, Philip M. The algorithm as institution: Toward a theoretical framework for automated media production and consumption. Fordham University Schools of Business Research Paper, 2013.

NORVIG, Peter; RUSSELL, Stuart. Inteligência Artificial. **Editora Campus**, v. 20, 2004.

NOBLE, Safiya Umoja. **Algorithms of oppression. In: Algorithms of Oppression**. New York University Press, 2018.

O'NEIL, Cathy. **Algoritmos de destruição em massa**. Editora Rua do Sabão, 2021.

ÖRNEBRING, Henrik. **Technology and journalism-as-labour: Historical perspectives**. Journalism, v. 11, n. 1, p. 57-74, 2010.

PARISER, Eli. **The filter bubble: What the Internet is hiding from you**. Penguin UK, 2011.

PAVLIK, John. **Journalism and new media**. Columbia University Press, 2001.

Poynter. 2014. **"L.A. Times reporter talks about his storywriting 'Quakebot'."** Disponível em: http://www.poynter.org/news/mediawire/243744/l`a`times`reporter`talks`about`his`story`writing`quakebot/. Acesso em 10 de março de 2021.

REITER, Ehud; SRIPADA, Somayajulu G.; ROBERTSON, Roma. **Acquiring correct knowledge for natural language generation.** Journal of Artificial Intelligence Research, v. 18, p. 491-516, 2003.

RUTKIN, Aviva. **Rise of robot reporters: When software writes the news**. 2014.

SÁNCHEZ, Juan Luis Manfredi; RUIZ, María José Ufarte. Inteligencia artificial y periodismo. Revista Cidob d'afers internacionals, n. 124, p. 49-72, 2020.

SANTOS, Márcio Carneiro. **Narrativas Automatizadas e a Geração de Textos Jornalísticos: A estrutura de organização do lide traduzida em código**. Brazilian journalism research, v. 12, n. 1, p. 160-185, 2016.

SANTOS, Márcio Carneiro. **Inteligência híbrida e análise de sentimentos: integrando curadoria humana e coleta de dados automatizada para avaliar a comunicação de governo.** Conexão-Comunicação e Cultura, v. 17, n. 33, 2018.
SANTOS, Márcio Carneiro. **Pesquisa Aplicada em Comunicação. O estranhamento da interdisciplinaridade que nos assombra**. Revista Comunicação e Inovação v.19, no 41, 2018.

Silvia, DalBen. **Automated Journalism in Brazil: an Analysis of Three Robots on Twitter.**
DOI: 10.25200/BJR.v16n3.2020.1305.

SILVA, Gislene; MAIA, Flávia Dourado. Análise de cobertura jornalística: um protocolo metodológico. **Rumores**, v. 5, n. 10, p. 18-36, 2011.

SILVER, Nate. **O sinal e o ruído**. Editora Intrínseca, 2013.

SILVEIRA, Stefanie Carlan da. **Conteúdo jornalístico para smartphones: o formato da narrativa sistêmica no jornalismo ubíquo**. 2017. Tese de Doutorado. Universidade de São Paulo.

STEENSEN, Steen; WESTLUND, Oscar. What is digital journalism studies?. Taylor & Francis, 2021.

Steiner, Thomas. 2014. Telling Breaking News Stories from Wikipedia with Social Multimedia: A Case Study of the 2014 Winter Olympics. Available online: https://arxiv.org/abs/1403.4289 (accessed on 21 November 2021).

TRÄSEL, Marcelo Ruschel. Entrevistando planilhas: estudo das crenças e do ethos de um grupo de profissionais de jornalismo guiado por dados no Brasil. 2014.

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

VAN DALEN,, Arjen. "**The algorithms behind the headlines. How machine written news redefines the core skills of human journalists**." **Journalism Practice**. v.6, n. 56, p. 648–658, 2012.

VEEL, K. (2018). Make data sing: The automation of storytelling. Big Data & Society, 5(1), 205395171875668. doi:10.1177/2053951718756686

STROSS, Randall. **When The Software is a Sportswriter.** New York Times. Disponível em: https://www.nytimes.com/2010/11/28/business/28digi.html. Acesso feito em 22 de fevereiro de 2023.

WALLACE, Julian. Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. Digital Journalism, v. 6, n. 3, p. 274-293, 2018.

WOLF, Mauro. **Teorias da comunicação**. 7 ed. Lisboa: Presença, 2002.

WÖLKER, Anja; POWELL, Thomas E. **Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism.** Journalism, p. 146, 2018.

YIN, Robert K. **Estudo de caso: planejamento e métodos**. Porto Alegre: Bookman, 2001.

ZAMITH, Rodrigo; BRAUN, Joshua A. **Technology and journalism**. The international encyclopedia of journalism studies, p. 1-7, 2019.

ANEXO I

Questionário para Equipe de Tecnologia

1. Quando a ideia de automatizar a cobertura de uma eleição surgiu para a equipe de tecnologia da Globo?
2. Havia algum exemplo de outro veículo nacional, ou internacional, como referência quando a ideia surgiu?
3. Quantos meses de trabalho foram necessários para desenvolver o algoritmo?
4. Como a base de dados do TSE foi estudada/esmiuçada para que as informações entrassem no texto?
5. Os dados contidos na base de dados precisavam passar por processos de formatação, como por exemplo colocar ou retirar letras maiúsculas, separar casas decimais por vírgula no lugar de ponto, incluir cifrões. Se sim, como você descreveria a importância da formatação para o processo.
6. A qualidade da base de dados, quanto ao número de células preenchidas versus o número de células em branco, formatação, identificação correta dos dados (features), contribuiu para o resultado final? De certa forma, o que impede hoje que outras bases de dados alimentem algoritmos similares?
7. Como foi definido o formato do texto a partir da sua extensão?
8. Como os aspectos textuais do texto, tal qual título, linha fina e corpo do texto tiveram que ser incorporados ao algoritmo?
9. Houve conversas e predefinições sobre o estilo do texto, quanto ao tamanho das frases, sinônimos que poderiam ser usados, palavras que poderiam iniciar, ou terminar uma frase?
10. Como a equipe lidou com a possibilidade de replicações de erros, sejam ortográficos, ou de concordância verbal e nominal por parte do algoritmo? Tendo em vista que mais de 5 mil textos foram publicados em minutos, não havia um risco de erros dessa natureza escalarem?
11. Houve diálogo constante entre jornalistas, editores e programadores ao longo do projeto?
12. Como ocorreram essas trocas na rotina entre as equipes?
13. Depois que o algoritmo foi programado e testado, houve influência de jornalistas e/ou editores que provocaram mudanças no funcionamento da ferramenta? Em outras

palavras, houve interferência editorial para além das escolhas iniciais de programação?

14. Visto que o algoritmo estabelece um ciclo de feedback entre as informações de entrada e saída, que neste caso busca emular a redação de um jornalista humano, como a tecnologia "aprendeu" com o profissional de carne e osso?
15. Na noite da apuração, a equipe precisou fazer algum tipo de acompanhamento?
16. Quais foram as ferramentas digitais, entre linguagens de programação e softwares que foram indispensáveis para o desenvolvimento do algoritmo?
17. Quando o assunto é Geração de Linguagem Natural, há uma referência frequente ao software GPT-3, que inclusive já foi utilizado por outros sistemas. O sistema desenvolvido pela Globo teve inspiração nesse software? Caso contrário, qual são as diferenças?
18. Há extensão do território nacional e o tempo de apuração das urnas pelo TSE interferiu no tempo de processamento dos dados e na posterior publicação automática dos textos?

Questionário para Equipe de Jornalismo

1) Havia algum exemplo de outro veículo nacional, ou internacional, como referência quando a ideia surgiu?
2) Quantos meses de trabalho prévio foram necessários para os jornalistas desenvolverem o projeto?
3) Como a base de dados do TSE foi estudada/esmiuçada para selecionar quais informações entrariam no texto?
4) Como a ordem em que as informações entram no texto foi definida? Por exemplo, o estado civil do candidato aparece antes de seu grau de escolaridade, que aparece antes do seu patrimônio declarado.
5) Essa ordem das informações no texto representa critérios de noticiabilidade do veículo, ou refletem uma tentativa de dar fluidez ao texto?
6) Como as informações disponíveis na base de dados do TSE influenciaram na delimitação do texto?
7) Foram discutidas questões sobre extensão do texto, como por exemplo número máximo de palavras ou caracteres para cada parágrafo?

8) Como as questões estilísticas do texto jornalístico foram discutidas com a equipe de tecnologia?
9) Como foram delimitadas as palavras que acompanham os dados do TSE, sinônimos, termos para iniciar, ou terminar uma frase?
10) Como foram definidas as questões de formatação no texto, como por exemplo incluir ou retirar letras maiúsculas, separar casas decimais, colocar cifrões em valores?
11) A equipe de jornalistas se reuniu com frequência com a equipe de tecnologia? Como esses encontros tomaram forma ao longo de todo o projeto?
12) A colaboração entre jornalistas, editores, programadores e o algoritmo em si foi constante?
13) Houve uma necessidade dos jornalistas de se familiarizar com novas ferramentas digitais, softwares ou até mesmo conceitos de programação para conseguirem participar do projeto?
14) Quais foram os instrumentos dos jornalistas para participar desse projeto, entre softwares, editores de texto, portais, sites e assim por diante?
15) Houve uma preocupação em realizar com a inteligência artificial? Os jornalistas contribuíram com feedbacks nesses testes?
16) Todos os textos foram revisados por ao menos um jornalista segundo o anúncio do site. Como esse processo se deu de fato? Cada jornalista ficou com uma cota de textos para revisar em um determinado período de tempo?
17) Depois da revisão, houve algum tipo de erro, seja ortográfico, ou de concordância, ou até mesmo informações faltantes, que foi mais frequente?
18) Muito se fala sobre o fantasma do desemprego estrutural nos estudos sobre o impacto da automação no jornalismo. Para alguém que trabalhou diretamente com um algoritmo que redigiu milhares de texto em um período curto de tempo, como fica a percepção desta ameaça?

ANEXO II

Entrevista decupada com Hector Iankovski

[00:19]
Com você eu queria ver sobretudo a parte de processamento de linguagem natural, que é a lacuna técnica que eles não souberam me descrever, até porque parece ter ficado mais com você mesmo essa parte técnica. Correto?

[01:00]
Isso, essa parte mais técnica de Machine Learning e Data Science foi nossa responsabilidade de desenvolver. Nós junto ao pessoal do COE, em integração também com outros times da Globo. Nós fomos responsáveis por fazer a implementação técnica mesmo do projeto, a partir da demanda que eles nos levantaram. Fizemos esse projeto pioneiro no Brasil, pelo menos eu acho que a nossa iniciativa foi uma das primeiras que teve no Brasil de geração automática de textos. Não só o projeto foi bem inovador por se tratar da geração automática, mas também pelo volume,. Porque a gente fez uma batelada de textos em tempo muito curto. Contamos com o apoio do jornalismo para fazer a revisão e a publicação desses textos. Então foi foi um projeto bem bacana de trabalhar e bem interessante.

[01:51]
Quando que surgiu a ideia para você, ou a demanda? Vou contar um pouquinho da história para você. Eu entrei na Globo no começo de 2020. Soube que existia uma outra área na Globo que também fazia alguns projetos de inovação e experimentação. Eles fizeram um projeto piloto de geração automatizada de textos. Por coincidência, duas áreas distintas da Globo foram até um congresso em Amsterdam, que é um congresso de mídia global que tem todo ano. Eles participaram dessa conferência e a área que ficava envolvida mais com o jornalismo achou o projeto interessante. Houve essa sinergia e eles falaram poxa, a gente poderia utilizar isso de uma forma produtiva na Globo.

[02:51]
Essa outra área desenvolveu o projeto de uma maneira meio informal e experimental, mas sem uma aplicação prática. Eles fizeram um teste e viram um potencial nisso. E aí apresentaram nesse congresso que como eu falei era voltado para a mídia em geral.

[03:15]
Aí essas outras pessoas envolvidas no jornalismo da Globo acharam essa iniciativa bacana, porque foi um um projeto feito em casa. Então isso ganhou uma relevância bastante importante dentro da empresa. Daí eles criaram essa ideia e queriam trazer isso para o jornalismo usar de alguma forma.

[03:34]
Entre as várias ideias que foram surgindo, veio a ideia de usar esse projeto piloto para uma iniciativa de eleição. Que é um trabalho bastante intenso, bem grande e que precisa de uma coleta de dados bastante robusta. Na época que a gente fez a eleição para prefeito, se tratavam de 40 a 50 mil candidatos. Lembra que a gente queria fazer a coleta de todos esses dados, de todos esses candidatos e fazer a uma publicação de texto da forma mais rápida possível. Então surgiu essa e era um cenário perfeito para a gente tentar utilizar essa solução feita em casa.

[05:39]
Eles te demandaram isso, você estando ali no como o cientista de dados. Eles te deram algum exemplo assim de um veículo nacional como referência?

[06:06]
Eles tinham essa ideia de fazer a produção automatizada de textos. Não havia nenhuma exemplo. Depois quando eu comecei a minerar um pouco e buscar referências, eu encontrei algumas referências no mercado americano. Se eu não me engano o Washington Post foi um dos veículos que fez uso disso na eleição de 2018 para presidente dos Estados Unidos.

[06:46]
Eu posso buscar depois as referências que eu encontrei na época e te passar, sobre o próprio jornal falando sobre como foi o projeto, como foi o desafio que eles tiveram em fazer essas publicações.

[06:59]
Eu posso te passar depois, mas eu não me recordo exatamente, qual que era o escopo que eles estavam trabalhando, se eu não me engano, eles tinham uma geração de mapas a geração de zonas de votação também. Então eu vou encontrar e te passar, porque foi essa mais ou menos a referência que eu tive.

[07:24]
Beleza, mas você pegou isso por conta própria, né? Isso foi quantos meses antes do dia da eleição?

[07:36]
A eleição foi em novembro, né? Por causa da pandemia a gente começou a pensar no projeto por volta de Maio e Junho. Então por volta de cinco meses antes da eleição.

[07:55]
A gente não trabalhou o tempo todo no projeto porque a gente não trabalhava só com isso. A gente tinha outros projetos em paralelo também, fazendo outras atividades. Então não foi o

tempo todo trabalhando em cima desse projeto. Na realidade a gente tinha outras demandas e outras atribuições. O projeto em si teve uma duração menor.

[08:19]
É que quando se trabalha em um projeto com equipes diversas, com coleta de dados diversas, tem muito vai e volta. Você cria uma demanda. Manda para uma equipe e eles retornam. Então não é aquele trabalho intensivo em cima do projeto. Tudo vai desenrolar um pouco mais devagar assim. Não é algo que você senta e em um mês você programa tudo, ou implementa tudo e coloca para funcionar. Então leva um tempinho e às vezes por causa dessas adversidades e outros trabalhos paralelos.

[08:52]
Entendi, quando você começou ali em maio, você começou fazendo o que exatamente?

[08:58]
A gente começou a pensar primeiramente em estudar a solução que havia sido desenvolvida por esse pessoal que fez o projeto experimental. Estudar a solução e como a gente poderia utilizar ela para o nosso problema. Se você quiser ir um pouco mais a fundo na parte mais técnica, existem vários formatos de geração de texto automatizado. Tem o formato fill the blanks, em que você cria um modelo único e simplesmente preenche ali com informações que você trata, coleta, ou gera na hora. Tem um outro modelo baseado em templates. Foi esse o que a gente utilizou. É um segundo nível um pouco mais avançado.Tem também o modelo generativo, que é o que o pessoal está fazendo hoje com o Chat-GPT.

[09:51]
Em português tem um modelo chamado PTT 5. Então dentro de cada uma das categorias tem várias formas de utilizar. Então a gente começou primeiramente entendendo como que era a solução que eles haviam criado. Como ela se encaixava na resolução do nosso problema e como a gente iria reaproveitar ela. Foi mais ou menos isso, a gente começou a estudar a solução e em seguinte a gente começou a discutir com jornalismo, o que seria interesse deles colocarem na notícia. O que colocar na publicação dos textos, né? A informação sobre o candidato, a profissão dele, a escolaridade, as informações sobre a cidade. A gente começou a fazer essa interação com o jornalismo para entender qual que era o escopo do projeto. O que eles queriam ver publicado.

[10:50]
A gente optou por usar essa ferramenta de templates.

[10:55]
Porque o jornalismo nos colocou que apesar de ser feito metade do trabalho com a ferramenta de criação de texto e consolidação com os dados que a gente coletaria posteriormente, eles não tinham essa segurança de fazer uma publicação automatizada. Então a gente tinha o texto gerado automaticamente para publicar no portal do G1. Mas eles gostariam de fazer uma revisão primeiramente do texto. Aí um texto automaticamente gerado, com muitas variáveis,

com muitas possibilidades de parágrafos, de conteúdos e tudo mais, ficaria muito difícil. A gente publicou 5.530 textos mais ou menos, era um para todas as cidades do país. Então, eles não queriam um modelo muito plástico.

[11:47]
Porque ficaria muito difícil de fazer essa revisão no momento da publicação. Apesar de ter uma equipe grande de jornalistas, eu acho que não existia mais de 20 ou 30 jornalistas fazendo a revisão e a publicação. Ficaria inviável fazer essa revisão de imediato, porque a gente queria fazer a publicação logo em seguida. Então a gente optou por esse modelo de template. Porque é um formato mais quadradinho. Você tem ali um escopo definido. Você tem possibilidades bem definidas e as variáveis que você coloca no texto também são bem previsíveis. Então ficaria mais fácil para o jornalista fazer essa revisão posteriormente, quando eles fossem efetivamente publicar o texto.

[12:31]
Entendi, pelo que eu estudei um pouco desse modelo de template, me parece haver três tipos de informação que entram no texto. Uma é mais fixa, uma que vem de outro lugar, os preenchimentos de sinônimos. Correto? Isso, perfeito, exatamente. Então a gente tinha quatro ou cinco parágrafos que eram definidos fixamente. O primeiro parágrafo era uma descrição de quem era o ganhador, mais o segundo colocado na eleição e informações sobre o resultado da eleição mesmo, com a porcentagem de votos.

[13:22]
Depois eram informações sobre a vida pessoal do candidato, que a gente coletou a partir de informações do TSE. Depois eram informações gerais da cidade, alguma coisa assim. Então esse era o primeiro nível. O formato quadradinho. Já as variações de sinônimo e de palavras eram no sentido de variações de gênero. Porque você tinha prefeito e prefeita, funcionário público e funcionária pública. Variações de gênero e de artigo. Haviam também variação de profissão, no sentido de todas as palavras que tinham a flexão de gênero, por exemplo, candidata e candidato, professor, professora.

[14:26]
Então todas essas variações de gênero a gente tinha previsto. Também haviam algumas variações de sinônimos de palavras. Então a gente criou um leque de palavras que a gente utilizava para fazer variações pequenas no texto. A gente tinha um formato padrão, mas tinham algus sinônimos ali para fazer essa alternação. Eles ficavam como que em bibliotecas? Mais ou menos. Eles ficavam dentro do template, você tinha o léxico de palavras que você criava. E aí você fazia um tratamento nessas informações para colocar um sinônimo ou outro, meio que sorteando eles conforme a lógica que a gente criava dentro do código.

[15:23]

Entendi. Então de fato, você tinham dimensões do texto que eram 100% estáticas, é qual seriam essas mais ou menos?

[15:34]
O começo Se não me engano, por exemplo a cidade, o "nome da cidade tem seu novo prefeito eleito". Então eram coisas assim. A paragrafização do texto era fixa.

[15:57]
Tinha essas variações de sinônimos de palavras e de gêneros, que a gente fazia alternação conforme necessário usando uma lógica definida. Aí o terceiro level, que a gente tinha eram as informações que a gente coletou do TSE anteriormente. Mas no processo seguinte do projeto, que eu não descrevi ainda, houve a coleta

[16:20]
de todo dado que o candidato no momento de fazer a inscrição tinha submetido. Para se candidatar a prefeito, a vereador, ou a qualquer coisa que seja, o cara deve fornecer informações ao TSE. Então a gente coletou todas essas informações, por exemplo, a profissão da pessoa era coletada,. A gente criou um banco de dados e limentou esse banco de dados com essas informações todas que a gente tinha de todos os candidatos possíveis, que poderiam ser eleitos no Brasil.

[16:50]
A gente armazenou tudo isso no banco de dados posteriormente pra utilizar e inserir no preenchimento das lacunas.

[17:06]
Então a idade da pessoa, a profissão, o patrimônio declarado, tudo que a pessoa concede ao TSE. Coisas assim bem pessoais, bem ligadas ao candidato. Então seriam três níveis, né? Os sinônimos que eram bibliotecas que estavam dentro do código?

[17:35]
Você tem as partes estáticas e essa terceira parte você poderia chamar de coleta de dados? É o que veio da base do TSE. Como é que foi o seu contato abase do TSE? Isto, a coleta em si dos dados a gente tinha uma equipe de engenharia de dados que fez essa fez a conexão com a base do TSE e a coleta das informações.

[18:27]
A gente conectava através de uma API. Que é para fazer a solicitação dos dados e a coleta. A gente não fez essa coleta em tempo real. No TSE, existe uma data limite, que é por volta de um ou dois meses anteriores a eleição, em que todos os candidatos devem fornecer essas informações obrigatóriamente ao TSE. Então posterior a essa data limite a gente fez uma uma coleta completa da base de dados.A gente fez a coleta e ingestão dessa base de dados dentro do nosso banco de dados, assim como todo o tratamento dela.

[19:10]
A gente montou a nossa base de dados conforme nossas necessidades, fazendo tratamento de informação e limpando a informação. Por exemplo, colocando a idade no formato de número e não de texto. Às vezes vem em string. Fazer o tratamento de assentos, caracteres especiais e tudo mais. Então esse foi o tratamento básico que a gente fez nela. Como foi assincro, não foi em tempo real quando a gente tava produzindo os textos a gente fez toda essa ingestão e criou uma base de dados dentro do nosso próprio ambiente. Na época o Google Cloud Plataform foi o que a gente usou. Existem outras como a AWS e o Azura da Microsoft. Mas na Globo a gente tinha uma parceria com a Google e a gente fez toda essa coleta e armazenamento das informações nessa base de dados do Big Querry, do Google Plataform.

[20:28]
Você poderia me falar um pouco mais sobre o próprio código e o próprio formato dos arquivos em que você salvou os dados? A gente começou a fazer testes da maneira como a gente deveria implementar a solução, para ter um melhor aproveitamento e para que ela fosse escalável. Porque a publicação dos textos não era linear. A gente começa com a apuração em poucas cidades, por volta das 17h tem um pico grande, por volta das 8h e 9 h tem outro e depois vai reduzindo e ficando só cidades menores. Então a gente fez uma batelada de testes para ver qual seria a melhor solução.

[21:22]
A gente fez isso para que não tivesse um gargalo, ou um acúmulo de eventos e isso acarretasse um delay. Ou seja, uma demora na publicação e na geração dos textos.

[21:35]
Como eu falei a gente utilizou a plataforma da Google, para fazer tanto a ingestão dos dados, quanto toda a arquitetura do sistema para publicação dos textos. Então a gente criou ultilizando a linguagem Python, para escrever os scripts de publicação dos textos. Existem vários serviços e várias ferramentas diferentes dentro da plataforma, que a gente utilizava para para fazer esse tratamento e manipulação de informação.

[22:03]
Os dados ficavam armazenados no Big Querry, que é o banco de dados mais utilizado para armazenar esse tipo de informação.

[22:29]
A gente tinha dentro da Globo, como a eleição um dos momentos mais importantes, que se tem uma demanda muito grande por informação em tempo real, também é criado lá dentro um sistema para coleta e captação dos dados do TSE no momento da apuração. Então deixa eu ver se entendi, já existiam boa parte dos dados capturados uns dois meses antes, que eram

os dados relativos ao Divulgacand. E aí existia a informação do pleito em si que ia ser capturada em tempo real? Isso, existia a apuração das urnas.

[23:30]
Os dados dos candidatos já estavam no Big Querry e ele já estavam formatados tratados?

[23:40]
Isso, exato. Como eu estava explicando, a informação usada na direção na Globo não foi só utilizada por nós. Toda a publicação de gráficos no G1, publicação de gráficos na TV aberta, no Globo News e tudo mais é feito via um sistema dedicado especificamente para isso. Ele transmite essa informação para que seja distribuído na Globo de uma forma única. Para que não seja uma equipe coletando de uma forma e outra equipe coletando de outra. Então é criado um sistema e tem uma equipe que trabalha nisso, fazendo a ponte entre a Globo e o TSE.

[24:34]
Porque isso. Porque o TSE tem uma demanda muito grande pela divulgação e publicação desses dados em tempo real. Então eles têm alguns gargalos que podem ocasionar problemas. Então somente empresas e serviços cadastrados anteriormente podem fazer essa coleta em tempo real sem ter um time out.

[24:59]
Esse sistema que a Globo cria internamente faz toda essa captação e coleta diretamente do TSE. O TSE divulga através de uma API esses dados e a gente faz toda essa coleta e captação de uma forma única, com a distribuição posterior para todas as equipes que querem utilizar esses dados. Então a nossa equipe do jornalismo, a equipe de publicação do Globo News todo mundo faz uso dessa mesma fonte. Esses dados para nós também eram ingestados no Big Querry. Então a gente coletava eles e fazia um tratamento ali em tempo real, armazenava eles no banco e disparava ali um gatilho de vários processos em tempo real.

[25:58]
No momento que a gente recebia a informação do candidato um arquivo recebido do TSE, a gente disparava um evento que fazia o armazenamento da informação no Big Querry. Posteriormente a gente fazia um outro disparo para fazer o tratamento e a limpeza desses dados. Outro gatilho disparava para a geração a partir do código. Outro gatilho para salvar o arquivo e posteriormente fazer o envio para publicação. O gatilho do código era para fazer a criação do texto mesmo.

[26:51]
Mas eu entendo que esse script já estava pronto anteriormente, correto? Isso, a gente dispara um evento e esse evento roda esse código. No caso, o código faz a geração do texto para a publicação, conforme as informações que a gente recebeu e capturou, concatenando também com as informações que a gente já tinha na base do Big Querry.

[27:14]
Como que esse sistema interno aí da Globo de puxar os dados do TSE aparecia para você?

[27:29]
A gente recebia dentro da plataforma do Google em um sistema de armazenamento de arquivos. Para simplificar, é como se você fosse copiar um texto do teu Windows Explorer de uma pasta para outra. Então esse sistema da Globo fazia essa geração do formato que a gente tinha definido, gerava lá distribuía para todos os sistemas. Eles depositavam esse arquivo em um sistema da Google para depósito de arquivos e a partir do depósito de arquivo disparavam um evento para iniciar o nosso processo do nosso lado.

[28:08]
Eles distribuiam um arquivo em formato XML.

[28:22]
Esse arquivo não era o arquivo original do TSE, já era um arquivo é tratado e formatado no padrão que a Globo precisava. Cada equipe solicitava do jeito que ia precisar do arquivo. Eles coletavam a mesma informação para todo mundo, mas distribuíam com os requisitos que foram definidos anteriormente. Então nós recebíamos um arquivo no formato e com uma quantidade de informações. O pessoal de geração de gráficos, ou da geração de imagens para TV, recebia a mesma informação, mas em um outro formato com outro com outro escopo de conteúdo. Eram escopo de variáveis.

[29:21]
É uma equipe composta por uma empresa terceira que é contratada pela Globo e algumas pessoas internas da Globo que fazem o gerenciamento desse projeto. Então a equipe de engenharia de dados, que eu falei anteriormente, era uma equipe que ficava dentro do COE e que trabalhava junto comigo. Vamos falar de quando o arquivo chegava pra gente. Ele era depositado no nosso sistema de armazenamento de arquivos, a partir desse momento a nossa equipe de engenharia de dados que desenvolvia todo o processo de tratamento, coleta de informação e gravação no Big Querry.

[29:59]
Então o seu contato no final das contas com a base de dados do TSE foi indireto. É que na verdade no dia da eleição a gente já estava com tudo pronto. O que o TSE faz dois ou três meses antes da eleição é uma batelada de testes.

[30:21]
Então todas aquelas pessoas que são cadastradas e são vinculadas ao TSE nos meses anteriores são convidados a participar desses testes. O que são feitos nesses testes é mais ou menos uma simulação do que vai acontecer no dia da eleição. Então eles publicam lá alguns dados fictícios sobre candidatos e fazem uma distribuição massiva de arquivos para ver se

está tudo funcionando. A ideia é ver se o sistema que a gente desenvolveu e testou está recebendo o arquivo e ver se o arquivo está chegando no formato correto, se ele não está corrompido e também para testar o sistema que a própria Globo desenvolveu.

[31:11]
A partir desses testes a gente também fazia os testes do nosso lado. Então a gente recebia esse arquivo da equipe que tinha integração com o TSE. A gente recebia esse arquivo e fazia os nossos testes.

[31:24]
A gente fazia simulação para ver se o que a gente estava desenvolvendo no nosso lado também estava compatível com o que a gente receberia no dia do pleito. Então a equipe que fazia essa integração com o TSE, também já tinha desenvolvido toda a parte deles, que consiste em incorporar o dado, tratar ele e distribuir para as equipes. A gente já tinha então um arquivo modelo, do que seria recebido no dia da eleição. Então no dia a gente não fez nada basicamente. No dia a gente só acompanhou o pleito e identificou se tinha algum problema, se os arquivos estavam chegando e se estava tudo de acordo com o que a gente precisava. Mas tudo isso já tinha sido feito com antecedência. Eu acho que com um mês de antecedência a gente já tinha o sistema pronto e funcionando, como ele deveria ser no dia da eleição.

[32:13]
Então essa informação que chegou em tempo real do TSE, no dia a gente acompanhava, para ver se o arquivo estava sendo gerado certo e se os dados estavam sendo ingestando no Big Querry. Nos viamos se eles estavam de acordo com o que a gente estava recebendo e se eles estavam sendo salvos de uma forma correta, se eles estavam íntegros e se eles estavam condizentes como que a gente estava recebendo. Mas isso era só o modo de monitoramento. Todo essa preparação já tinha sido feita dois três meses antes de isso acontecer. Então a gente recebeu os arquivos testes do TSE com dados fictícios gerados por eles, mas era a simulação exata de como aconteceria no dia da eleição. Porque no dia da eleição a gente não tinha tempo para corrigir, ou fazer qualquer coisa que precisasse ser feita. Isso demandaria horas de trabalho.

[33:11]
Todos esses testes, toda essa distribuição de arquivos do TSE, foi feita com bastante antecedência para todas para todo mundo. Para que todo mundo possa se preparar e possa ver como as coisas estão funcionando, se não estão funcionando, se é detectado algum delay, alguma latência na distribuição do arquivo, se o arquivo tá chegando de acordo e tudo mais. Tem uma série de requisitos que são feitos ao TSE com um ano de antecedência. Eles públicam um modelo de como vai ser a distribuição do arquivo. Então você já sabe que o arquivo vai ser no formato xyz e vai conter um determinado escopo de informações.

[34:11]

Tanto a equipe que faz a captação dos dados, quanto nós, já sabemos como está sendo feito o desenvolvimento. Mas eu acho que com quatro ou seis meses de antecedência, isso já tá consolidado. Aí com dois ou três meses de antecedência é feito esse testes que eu te falei em tempo real, de como seria no dia da eleição.

[34:33]
Então você fez esse teste no dia em que o TSE produziu os dados fakes? Mas com o mesmo tipo de arquivo, no mesmo endereço, na mesma plataforma? Antes disso vocês conduziram testes internos? Sim, a gente já tinha essa ideia desse esqueleto do arquivo. Então nós mesmos criamos arquivos testes para fazer o pipeline da execução. Então a gente criou o arquivo XML do jeito que a gente queria, com as informações fictícias, que a gente tinha. Também tinha esse esforço de criar informações randômicas, a gente não estava preocupado ainda com aquele teste em tempo real.

[35:37]
Então a gente criou informações que colocava os nomes das pessoas da equipe, com a porcentagem da eleição para testar mesmo. A gente não fazia uma geração massiva de dados. A gente fez isso com dois ou três meses de antecedência do pleito.

[36:06]
A gente fazia uma copia truncada, que gerava vários arquivos iguais e ficava colocando 100 arquivos por vez, ou 200 arquivos por vez, 500 arquivos por vez, para ver se o nosso sistema estava escalando. Para ver se o sistema estava dando conta do throughput de informações que a gente estava recebendo. Para ver se ia sair tanto no formato desejado, como no volume esperado. Porque o pipeline era composto de várias etapas. A primeira era o depósito do arquivo e o disparo dessa mensagem para o Big Querry.

[36:43]
Dentro do Google existe um serviço chamado functions, que é um sistema de mensageria que recebe e dispara ecentos. Dentro dele você pode colocar um código Python, ou então pode colocar uma rotina para alguma coisa ser executada ao mesmo tempo. Então dentro do functions a gente disparava assim: "eu quero que o Big Querry faça isso, colete a informação e grave no banco de dados. Aí o próximo passo é trato o tratamento da informação, para tirar caracteres especiais e deixar no formato que a gente tinha pré-definido. O próximo passo era o disparo da mensagem para execução do código, para a geração do texto de fato. O outro passo era o a ingestão desse texto gerado para o Big Querry, em um outro escopo de tabelas que a gente tinha para guardar essa informação. Aí disparava uma outra mensagem para fazer a geração do texto no formato que a gente tinha pré-definido com o pessoal que trabalhava junto ao G1.

[37:43]
O próximo evento era depositar esse arquivo no nosso sistema de gerenciamento de arquivos. Aí o outro passo era disparar uma mensagem para enviar esse arquivo para o sistema de

publicação do G1. Então tem todo esse pipeline de teste que a gente tinha que fazer, porque assim você tem a informação entrando, tudo isso tem que ser ingestado da maneira correta. Se você não fizer isso, o arquivo quebra no meio. Agora eu coloquei 50 arquivos e quero ver se as 50 informações vão estar disponíveis lá no banco de dados. Agora eu coloquei 100 arquivos e quero ver se essas informações foram salvas. Quero ver se o texto foi gerado sem nenhum gargalo. Quero ver se não teve, por exemplo, um texto que foi gerado com cinco segundos o outro foi gerado com 10 segundos..

[38:43]
Quando a gente colocava 200 arquivos ao mesmo tempo, a gente queria ver se a fila de processamento era executada de uma forma condizente com o que a gente esperava no dia. Para que não tivesse nenhum gargalo e não tivesse nenhum problema de execução também. Porque no dia o volume seria bastante grande. O caminho não podia ter um único gargalo para a publicação. A gente estava bastante preocupado com essa questão do desempenho de performance do nosso pipeline de execução. Então os testes eram constantes sim. Porque no meio ali dos testes a gente pensava em melhorar isso aqui e aquilo ali. Melhorar a geração do código, a geração do texto. Se a gente fazia algum tratamento no dado, a gente tinha que testar todo o processo de novo. Não é só colocar um arquivo lá e ver se sai no final. Tinha que testar com uma batelada de 100 a 500 arquivos para ver se funcionava.

[39:54]
Em algum momento vocês fizeram um teste com o mesmo volume de 5 mil textos?

[40:02]
Não, porque pelo histórico que a gente tinha e pelas conversas que a gente teve com o TSE, em nenhum momento a gente imaginou que chegariam 5 mil textos de uma vez só. Porque não é o comportamento esperado no dia . O processamento do TSE das urnas eletrônicas tem uma curva bem diferente. No começo são pouca cidades. Depois concentra por volta das 18 ou 19 horas. Ai depois para o final vão ficando apenas algumas cidades. Então a gente não esperava que tivesse 5 mil textos recebidos e 5 arquivos de uma vez só. Era um comportamento que a gente não não imaginava que seria passível de acontecer.

[40:58]
Ao longo da tua fala eu fui anotando algumas ferramentas aqui que você mencionou. Eu quero saber se a gente conseguiria listar todas. Pode sr? Se tiver faltando alguma algum complemento aí depois chamada por e-mail.

[41:16]
Tranquilo, mas você mencionou o Google Cloud para armazenar os dados. É que o Google Cloud é uma é uma plataforma. É como se fosse um um conjunto de serviços que existe para você fazer diversas diversas atividades, desde armazenar dados até executar códigos e receber

eventos, ou receber arquivos. Então seria como se fosse a grosso modo o sistema operacional. Existem diferentes ferramentas dentro dessa plataforma. Existem vários serviços. Como eu falei, há o serviço de banco de dados, de execução de código, de tratamento de eventos e de mensageria.

[41:59]
O Big Querry está dentro do Google Cloud, assim como o Google Cloud Functions, o Google Storage, são todos serviços dentro do Google Cloud Platform.

[42:37]
Todo o nosso pipeline de execução estava dentro do Google Cloud Platform. As nossas comunicações externas, onde a gente tinha contato com outro sistemas, era o sistema de produção de arquivos lá da Globo, que fazia a coleta do dado do TSE.

[43:59]
Pensa que o Google Cloud Plataform é um cercadinho com uma empresa lá dentro funcionando. O arquivo que eles entregavam era como se fosse o correio entregando uma correspondência para você, a gente recebia isso e processava dentro. O nosso output, a nossa publicação, era um arquivo que a gente gerava para o sistema de publicação da Globo. Esse sistema de publicação da Globo é um sistema terceiro, correto?

[44:33]
É um outro sistema. Eu não me lembro direito do nome, eu não vou lembrar agora. Mas tá se você tiver anotado aí seria legal.

[44:48]
É o sistema CMA.

[44:53]
Eu lembro com um outro nome que não era CMA mas talvez seja a sigla que eles utilizavam. Mas era o sistema de publicação que o jornalista usa para criar o texto e para colocar imagens, para colocar gráficos e outras informações. Ele é conectado junto ao portal do G1. É o que faz a publicação do texto.

[45:20]
Outra informação que eu não tenho é como esse sistema funciona, como ele é feito como que é. Eu não tenho a menor ideia também. Como eu falei para você o nosso input era esse dado que vinha do TSE através do sistema de coleta da Globo. A gente fazia o meio campo ali, que era o tratamento da informação e a geração do texto. O nosso output no caso era um outro arquivo que a gente depositava e enviava pra equipe do CMA. Esse texto era ingestado e automaticamente caía dentro do CMA lá do jornalista que acessava. Não sei se aparecia no

inbox dele, ou se esse texto estava lá pronto para ele revisar e fazer a publicação manualmente.

[46:17]
Perfeito, está bem claro agora. O próprio sistema da Globo fazia a conexão com o TSE. O que vocês tinham captado antes vinha do DivulgaCand e já estava salvo no Big Querry. Provavelmente em arquivos json, correto? Não, o Big Querry é um sistema de banco de dados. Então a gente tem as tabelas do banco de dados prontas para consultar a gente usa linguagem SQL. Ou você pode usar a linguagem SQL para fazer a consulta do banco e plugar essas informações no texto que a gente gerava.

[47:00]
Era SQL no Big Querry, mas esses bancos de dados eles ficavam em algum formato específico?

[47:09]
O Big Querry é o banco de dados no caso. Eu não sei como que é a implementação do Big Querry lá dentro do Cloud, mas é um sistema de banco de dados e a linguagem de consulta do banco de dados é o SQL. Você acabou de trazer uma informação nova que é a consulta em SQL. Mas a implementação dos textos pelo script era em Python? Isso, perfeito. Dentro do código em Python você pode executar querries de SQL, para depois plugar elas dentro ali do template que a gente estava ultilizando.

[48:09]
Então dentro do código Python, a gente tinha execução de querries SQL para consultar informação do banco de dados.

[48:22]
E o output é um arquivo que vocês entregavam para a equipe do CMA? Isso, feito a execução do código o nosso output inicial, era a gravação da informação novamente no Big Querry, para pra ficar armazenado ali, caso a gente precisasse regerar esse texto, ou consultar se a informação estava correta, ou fazer algum outro tipo de consulta.

[49:03]
O outro passo depois foi, a gente tinha um outro código em Python posterior a esse salvamento no Big Querry, que era fazer a leitura do banco de dados e gerar o arquivo json para ser publicado e enviado para a equipe do CMA.

[49:24]
Então a equipe do CMA recebia um arquivo de json já com os textos? Já com texto pronto. Isso foi um acordo que a gente fez, a nossa geração do arquivo era para cada cidade. Então

cada cidade tinha um arquivo json gerado. Eles viam com uma série de tags, que podiam ser utilizadas.

[50:06]
A gente acordou com eles, qual era essa formatação, se eu não me engano eram umas 8 ou 9 tags, com informações que eram pertinentes para eles fazerem essa vinculação dentro do CMA. A primeira informação, por exemplo, era um código da cidade. A segunda era o nome da cidade. A terceira era sobre o estado, porque ali dentro do CMA eles faziam uma distribuição. Por que a eleição acontecia em inúmeros estados. Tinhamos 27 estados para fazer a publicação dos textos. Assim esse texto era distribuído para todos os jornalistas que deveriam ter acesso àquela publicação. A gente fazia esse roteamento da informação para cada Estado receber os textos, ou cada Regional receber os textos que eram designados para aquela regional. Existem as afiliadas e as unidades da Globo. Por exemplo Belo Horizonte, Recife, Brasília, São Paulo e Rio de Janeiro são da Globo. As outras são afiliadas.

[51:05]
Essas tags funcionavam, por exemplo, você tem uma afilhado da região de Grande Matão, Estado de São Paulo, essa tag fazia aparecer o texto na pagina do site deles? Não, a gente tinha esse roteamento para cada Regional receber somente os textos que eram dedicados àquela regional. Então a região que você falou da Grande Matão recebia os textos da Regional que era responsável. Para não ter essa bagunça, porque seria bem difícil dos Jornalistas encontrarem os textos lá para depois fazer esse essa revisão e publicação,.

[51:57]
Era um roteamento interno que o pessoal fez. Como era feito lá dentro do sistema deles, eu não faço a menor ideia também.

[52:15]
Faltou alguma ferramenta digital que a gente não abordou ainda?

[52:52]
Você trabalhou no código em Python? Sim, eu que escrevi. Eu trabalhei nesse projeto meio que como o gerente. Eu cuidava de toda a equipe de coleta de dados da Globo, eu que falava com a equipe de distribuição do CMA e o que fazia toda essa parte de implementação do código em Python. Para depois fazer o tratamento da informação, a geração do texto e a criação do arquivo posterior. Então toda a parte da codificação ali pra geração dos textos e dos arquivos, ela foi feita por mim.

[53:34]
Certo, então na hora de fazer esse código, você encontrou muita coisa já?

[53:39]
A gente reaproveitou aquela solução que a equipe de experimentação da Globo havia feito. Então a gente utilizou o formato que eles haviam feito. Eu posso até te dar uma outra informação. Esse mesmo sistema de geração de texto em templates também foi utilizado em outro projeto na Globo. Eu não lembro exatamente o nome da página, mas era alguma coisa como "Espião Estatístico". Ele era usado em estatísticas de futebol pro campeonato brasileiro. O projeto foi feito para isso, esse era o core inicial do projeto.

[54:25]
Por exemplo, eles coletavam as informações da partida entre Corinthians e Palmeiras. Então tinha lá a posse de bola do time, os chutes a gol, os cartões amarelos e os cartões vermelhos. Eles utilizaram essa solução inicialmente para fazer esse tipo de publicação. Eles tinham uma página que eu vou procurar para tentar encontrar, mas acho que era alguma coisa como "Espião Estatístico". No mesmo modo que a gente fazia, só que eles faziam com informações referentes ao futebol. Então o texto entre Corinthians e Palmeiras ficava como: "o jogo terminou empatado", ou" vitória do Corinthians por 2 a 0 e os gols foram marcados por Fulano", "durante a partida tiveram tantas faltas e tantos cartões amarelos", "a posse de bola do time foi essa". Aí, a gente pegou esse mesmo código e aproveitou.

[55:25]
Pelo menos lógica que tinha nele, mas fazendo as adaptações necessárias. Colocamos os sinônimos que a gente precisava para publicação do texto da eleição, criando os templates e definido com o pessoal do jornalismo, inserindo as informações com o fill the blanks para o preenchimento das lacunas, com as informações coletadas, anteriormente e também com o sistema do TSE em tempo real.

[55:54]
Como eu expliquei para você no começo, a ideia toda do projeto surgiu desse projeto piloto. A experimentação que o pessoal da Globo fez para geração de textos para o espião estatístico, a gente reaproveitou a ideia e utilizou para a publicação dos textos da eleição também.

[56:20]
A gente fez também um outro uso, que não foi divulgado, não foi publicado, mas por exemplo só para te te contextualizar,.

[56:32]
Esse mesmo formato de template foi utilizado para fazer a geração dos textos em off, quanto a divulgação das pesquisas eleitorais que são feitas durante o pleito das eleições. Então esse mesmo template a gente utilizou para fazer a geração desses textos em off, que é lido pelo ãncora do jornal quando tá fazendo a divulgação dos textos das pesquisas.

[57:03]
É uma informação nova, sei que não é o foco, mas para você ver como é esse formato de geração de texto ele pode ser bastante mutável. Você pode utilizar ele para publicação de

qualquer tipo de texto, de forma em tempo real, ou massiva, ou quando você precisa ter muita agilidade. Por exemplo no caso da pesquisa e da eleição, os resultados da pesquisa chegam em torno de 45 minutos a 30 minutos antes do jornal ir ao ar. Então você precisa fazer o texto nesses 45 a 30 minutos. Você precisa fazer todo aquele esforço de geração de gráficos, de geração de texto, de publicação para leitura no jornal em tempo real. Então essa era uma das reclamações que o pessoal do jornalismo tinha.

[58:03]
O uso disso foi bem fundamental no dia a dia publicações das pesquisas, porque a gente recebia os arquivos dos institutos de pesquisa, tipo o Ibope e outros mais. A gente fazia a coleta desse dado e fazia também a geração do texto off, também distribuia esse dado para fazer a geração dos gráficos das pesquisas em tempo real. Então para você vê como o sistema é Coringa. Você pode ter múltiplos usos a partir do mesmo código e a partir da mesma raiz. Desde que você tenha uma base de dados confiável.

[58:43]
Se você tiver uma ideia boa, dados robustos e uma plataforma para execução desse código, dá para aplicar isso em qualquer coisa, Você consegue fazer qualquer tipo de projeto que trabalhe com a geração de textos.

[59:55]
Qual a semelhança desse sistema com o Chat GPT?

[01:00:46]
Eles são modelos diferentes. Por exemplo, no Chat você têm o uso de Deep Learning e redes neuraus. Você também tem um treinamento prévio do modelo para ele entender a linguagem que você tá passando para ele. Quais são os sinônimos, as conexões entre as palavras, para fazer a geração posterior do texto, ou responder perguntas e tudo mais, o uso dele também é bastante vasto. É bastante diferente do formato de template, em que você não precisa fazer um treinamento desse modelo anteriormente. Você não precisa dar um input para ele aprender por meio desse texto. Você simplesmente tem um formato ali mais quadradinho, com algumas variações mínimas.

[01:01:46]
Mas você nâo tem uma rede neural treinada para fazer isso, você não tem um Deep Learning por trás para fazer isso. Então é um nível bem mais simplificado, bem mais rústico de geração de texto, ele se encaixa dentro do NLP, mas é um outro tipo. Não é o Hype do momento, que nem o Chat GPT e outros mais que tem por aí.

[01:02:08]

Por ele se encaixar dentro do NLP. Pelo o que eu estudei da teoria, você tem a Geração de Linguagem Natural, o Processamento e a Compreensão,. Três campos dentro do que se chama de Computação Linguística.

[01:02:32]
Esse modelo de template, ele é uma inteligência artificial ou uma automação por regras? É meio que uma automação por regras, porque como eu falei para você, você não tem uma rede neural treinada, ou um input para esse modelo aprender. Você simplesmente tem um código quase estático, que faz a sua geração do texto com algumas pequenas variações no meio. Você não tem uma inteligência artificial por trás disso.

[01:03:11]
Qual seria a distância de um automação por regras e uma inteligência artificial assim? São modelos são diferentes. Primeiramente você tem a toda codificação dele. Tem a questão do treinamento que você precisa fazer anteriormente. Acho que a grande diferença é a flexibilidade. Quando você está trabalhando com templates em formato estático, você tem uma caixinha sem muita surpresa. Quando você está trabalhando com geração autônoma, por Inteligência Artificial, você não tem um formato.

[01:03:59]
Você tem um certa variabilidade, mas não é muito mais dinâmico, muito mais flexível do que o formato por template, no caso. A distância é bem considerável.

[01:04:13]
Não tem o treinamento que você mencionou, mas tem a fase de testes, correto? Isso, só para ver se a solução está funcionando, se a solução está de acordo com o treinamento.

[01:04:44]
No dia da apuração você fez algum tipo de operação? Você fez algum tipo de acompanhamento? Sim, ao longo daquela semana a gente fez uma batelada de testes, em todas as pontas. A gente testou desde o recebimento, ao processamento e entrega para o pessoal do CMA, para ver se eles estavam recebendo o arquivo no formato especificado. Para que eles fizessem o roteamento da informação e a distribuição interna. A gente fez todo o teste com uma quantidade massiva de arquivos. Se eu não me engano uma ou duas semanas antes, a gente teve um último teste com o TSE. Que era o teste final e oficial.

[01:05:39]
Depois daquilo não teria mais nenhum teste com o TSE, ele seria o formato final que todo mundo receberia e tudo mais. No dia a gente começou a acompanhar a partir das 4:00 ou 5 horas da tarde, quando começou a chegar aos primeiros arquivos. A gente detectou um pequeno erro num dos formatos de template, que a gente tinha feito e testado, mas passou batido por todo mundo que estava acompanhando o desenvolvimento. A gente fez uma alteração rápida ali no código e corrigiu isso, mas no dia mesmo foi mais uma questão de

monitoramento. A gente teve um grande problema no dia que não foi culpa nossa, foi uma culpa do TSE. Não sei se você lembra, se você pegar as notícias ali de 2020, você vai ver que teve um problema de porque o TSE começou a utilizar um outro sistema para contagem e apuração dos votos.

[01:06:39]
Isso gerou um gargalo e a distribuição dos arquivos por um formato ficou impactado. Aí foi feito um outro tipo de distribuição que o sistema da Globo também foi impactado. Isso atrasou a publicação dos textos. A nossa ideia era publicar os textos até às 9:00, 10:00 horas da noite. Mas a gente só terminou de publicar os textos na segunda-feira pela manhã, porque teve esse gargalo do sistema do TSE, que ocasionou um problema no sistema da Globo. Isso não estava previsto, mas impactou a entrega dos arquivos para nós. Mas no diaa gente acompanhou a chegada dos arquivos. Para ver se os arquivos estavam sendo salvos no Big Querry, como eram pra ser salvos. A gente conseguia acompanhar dentro do Pipe Line cada um dos passos, para ver como estava a execução, se tava demorando mais do que o planejado, se não tinha algum impacto na fila, a gente não detectou nada.

[01:07:36]
A gente acompanhou na ponta com o pessoal do G1, para ver se estava sendo roteado tudo certo.

[01:07:48]
Mas foi mais assim para cuidar. Para ver se não tinha nenhum desvio, nenhum mau comportamento ali no meio. Esse gargalo do TSE não apareceu no dia do teste?

[01:08:10]
Não e ficou muito mal explicado, porque eles não revelaram. Se você procurar as notícias você vai encontrar assim: foi um sistema novo que não tinha sido treinado ainda o suficiente e que não estava contabilizando da forma correta. Só que isso não apareceu nos testes e foi uma reclamação que a gente fez para o pessoal que tinha essa interface com eles. Por que foi uma falha bem considerável. Foi nesse episódio que surgiu aquele burburinho de que o sistema não funciona e de que não é confiável. Aí que começou a surgir essas fofocas e essas balelas, mas até hoje eu não sei exatamente o que aconteceu.

[01:09:10]
Nos dois ou três testes anteriores que a gente já havia feito com eles, não tinha aparecido esse comportamento. Foi no dia da eleição mesmo.

[01:09:24]
Ou seja, mesmo com testes recorrentes e mesmo com um dia de teste oficial, ainda existem fatores externos que podem aparecer? É, porque na verdade a nossa solução não era independente. A nossa solução tinha várias etapas, em que ela era totalmente dependente. Então se o sistema do TSE demorar para processar o arquivo e não entregar a tempo, o sistema de coleta da Globo também não vai coletar esse arquivo e consequentemente o nosso

sistema não vai gerar os textos. Então tem todas essas dependências, por que assim você testa, você imagina todos os cenários possíveis, mas às vezes podem ter alguns cenários que não são previstos.

[01:10:15]
Aí você tem um problema que ninguém imagino que poderia acontecer. Isso aí foi mais uma falha com o TSE e o sistema de captação da Globo, que não trataram essa possível falha.

[01:10:42]
Para fechar, na verdade era uma pergunta que eu ia fazer no início, mas a gente acabou já entrando nas questões mais técnicas. Como é que foi o teu contato com os jornalistas ao longo do projeto? Você recebia os templates de texto deles ali por e-mail, dai rebatia com comentários? A gente tinha uma reunião semanal com jornalismo, porque a gente não trabalhava só nesse projeto, mas com alguns outros projetos também. Principalmente nessa época que foi a época da Covid-19, você tinha a distribuição dos dados do consórcio de jornalismo. Então a gente tinha uma reunião semanal que a gente discutia vários temas. O tema da eleição sempre era recorrente, porque tinha tanto esse projeto da publicação dos textos automatizados, quanto também a distribuição dos dados da pesquisa. Então tinha esses dois projetos.

[01:11:42]
Então era uma troca recorrente. O Thiago Reis fazia um template e a gente contestava: "poxa, mas isso aqui talvez quando a gente fizer as variações do sinônimos, não vai ficar legal". Eles trabalhavam de volta e nos enviavam novamente. Aí a gente testava e mandava para eles alguns exemplos do texto. Eles retornavam falando o que não ficou legal, ou o que teria que mudar. A gente trabalhava novamente e enviava para ele. Então tinha essa troca constante, durante todo o desenvolvimento. Por que eram várias etapas. A primeira foi a definição do template. Antes veio a definição da ideia. Foi a definição de que seria um formato mais pré-definido sem muita variação, por causa da questão da revisão. Posteriormente a gente foi para o template em si, definir os formatos as variações de sinônimo, quais nformações iam ser colocadas no texto e tudo mais.

[01:12:42]
Depois você têm a questão dos testes, para ver se estava tudo sendo gerado da forma adequada, com todas as variações possíveis e com todas as profissões possíveis, com todos as idades possíveis, com candidatos homens e mulheres. A gente sempre estava trocando, fazendo essas validações. Posteriormente teve a questão da publicação junto com o pessoal do CMA, que eles também acompanharam para ver se estava recebendo no formato correto e se o texto estava vindo com a formatação que era esperada, com a quebra de linha que era esperada e com os espaçamentos que eram esperados, para por lá dentro do CMA e para a publicação ser feita manualmente.

[01:13:31]

As questões que apareceram na definição do template de formatação, elas se repetiram depois na fase do CMA?

[01:13:41]
Não as mesmas, mas algumas. Por que o formato que a gente gerava o texto era no formato bruto, de texto corrido com os espaçamentos e as as definições. Quando ia para o CMA a gente precisava passar para eles em outro formato, com outras tags e com outros caracteres especiais ali dentro do texto, para fazer a quebra de linha, para fazer a quebra de parágrafos, para fazer a divisão do título, do subtítulo e do corpo do texto mesmo. Então eles eram similares, mas não eram iguais.

[01:14:18]
Ali dentro do aquivo você não conseguia determinar todas as quebra de linha, parágrafo e coisas assim?

[01:14:35]
A limitação não era nem pelo aquivo json, era por como o CMA conseguia tratar essas informações que a gente colocava lá dentro. Porque o CMA não conseguia fazer um processamento de qualquer coisa que a gente colocava lá. Por exemplo, a nossa ideia que a gente queria era colocar imagens. A imagem do prefeito, a foto dele. A gente queria colocar o mapa da cidade, algum mapa do estado, a gente queria colocar o gráfico da porcentagem da eleição. Mas o CMA não era compatível para fazer esse tratamento. A gente queria colocar algumas partes do texto em negrito e em Itálico, com uma fonte maior ou menor. Mas o CMA também não era compatível para fazer esse tipo de tratamento. Então tinha algumas limitações que o sistema deles impunha e que a gente não conseguia fazer qualquer coisa dentro do json. O json também não permite fazer qualquer coisa, mas permite fazer bastante.

[01:20:58]
Maravilha era só isso mesmo. Quer dizer só isso não, né? Foi uma hora e meia de conversa. Espero que eu tenha colaborado para você.

[01:21:37]
Incrível, muito obrigado, viu? Hector bacana Mateus. Obrigado. Espero que tenha colaborado. Espero que seja produtiva.

[01:21:48]
Até mais.

Entrevista decupada com Tiago Reis

[00:00]

Quando o projeto começou assim nas primeiras conversas, vocês tinham algum exemplo de um veículo nacional ou do exterior? Então, a gente tinha visto um um projeto da BBC. A BBC tinha feito um projeto não igual, mas eles tinham usado Processamento de Linguagem natural e alguma coisa automatizada para fazer alguns textos numa numa eleição lá do Reino Unido. A gente viu aquilo ali, mas não era para todas as séries, era para algumas localidades só e e enfim era um pouco diferente.

[00:44]
Tinha muito muito da ação humana ali porque não temos que era feito com ajuda de uma parte automatizada, mas uma boa parte que era que feita por um jornalista. E aí a gente começou a pensar que a gente poderia fazer eh de forma similar aqui no Brasil. E aí uma das ideias foi ter um resultado da cada uma das cidades nas Eleições Municipais. Então esse foi o ponto de partida a gente tinha uma referência que era BBC é, mas a gente queria fazer uma coisa maior e era uma coisa que a gente então a gente já pensou desde o começo de fazer algo que poderia marcar todas as 5.568 cidades. Porque Fernando de Noronha e Brasília não tem eleição Municipal. Mas a gente queria algo que fosse um pouco maior e que exigisse menos a participação do jornalista, por que a gente não ia ter como entrar em 5.500 textos pra ficar mexendo ou melhorando o texto enfim.

[01:47]
Então foi no início de 2020 mais ou menos que vocês começaram a idealizar esse projeto? Quantos meses de trabalho prévio, vocês precisaram antes dos primeiros testes para ver se a coisa estava realmente funcionando e pronto para ir para o ar? A gente começou a fazer isso no começo de 2020. Eu não lembro que que data exata tá confesso que a minha memória não é das melhores. Eu já vou te dizendo tudo que tiver que lembrar datas exatas eu vou ficar devendo, mas a gente começou a fazer um planejamento ali pra pra eleição no comecinho do ano. Então assim, a gente tinha vários projetos que a gente já sabia que ia tocar, como o jogo eleitoral que a gente sempre faz, que é um jogo que você consegue identificar ali, de acordo com a resposta que você dá alguma quantidade de pessoas que se identifica e vários outros ali que a gente já tinha programado e esse (projeto) seria uma

[02:46]
novidade que a gente nunca tinha feito até então. Essa era uma nossas apostas para fazer durante o ano. Aí a gente tinha, uma coisa que é importante colocar, a gente tinha desde 2019 um contato em reuniões periódicas com a área de Tecnologia lá da Globo. Foi em uma dessas reuniões em que a gente foi apresentar o nosso projeto de eleição. A gente falou: queremos muito fazer esse tipo de coisa com texto automatizado, vocês acham que é possível? E aí em um primeiro momento eles falaram que precisam estudar para pensar se é possível, como que eles fariam e tal. Então esses primeiros papos começaram ali numa dessas reuniões periódicas que a gente tinha com esses setor de Tecnologia e no começo do ano e a gente foi desenvolvendo ao longo desse período uma tentativa de protótipo do que seria esse texto.

[03:44]

Então a primeira ação na verdade foi eu mandar para eles quais eram as informações que a gente ia precisar e de que fontes de dados. Então assim, a gente vai precisou olhar todas as informações possíveis que estão lá no TSE do candidato, a gente queria também acessar as informações do IBGE. Eu fui lisitando ali algumas fontes que a gente poderia usar. Depois, você vai ver que no fim a gente basicamente usou só as informações do TSE. A gente resolveu simplificar porque ia ser é muito complicado ter várias fontes ali ao mesmo tempo, então a gente ficou só com o TSE, mas no começo a gente começou a fazer um trabalho meio que de

[04:29]
pensar, em um brainstorm mesmo, do que poderia ser. Então a gente começou a pensar nas fontes que poderiam ser usadas. E ali eu comecei a rascunhar um texto padrão, como um rascunho. Antes de começar a redigir o texto padrão, seu ponto de partida foram as bases de dados do TSE e do IBGE? No começo a gente queria mostrar em quais bases públicas a gente poderia conseguir as informações, isso aconteceu depois do texto. Então assim, vamos fazer um texto padrão do que seria esse rascunho desse texto.

[05:29]
Na verdade, foram quatro rascunhos. Um para a vitória em primeiro turno. Empate com decisão só no segundo turno. Candidatura impugnada e uma para aguardando decisão do TSE. Dai eu fiz os modelos de texto basicamente colocando espaços para serem preenchido. Então assim, vou dar um exemplo: o prefeito da cidade de..., dai tinha um espaço, foi eleito neste domingo ... para um mandato de quatro anos, dai tinha o negócio de ensino completo ou não. Dai aí a gente foi colocando um texto com um monte de lacunas que eu já sabia que eu queria preencher.

[06:30]
Que tinha as informações que tinha sido eleito, as informações de quem tinha sido derrotado, uma listinha com o percentual de quanto cada um tinha recebido. E aí tinha uma lista com várias informações do Legislativo, que são coisas que a gente depois não conseguiu colocar tudo. Porque essa questão do Legislativo, a gente sabe que tem um coeficiente, então tem uma conta que precisa ser feita e que não sai ali automaticamente. A gente achou que ia ser muito difícil colocar isso logo no quente da eleição, então a gente acabou também diminuindo essa parte do Legislativo e usando uma parte muito maior do resultado do executivo. A dificuldade era ter uma lista exata de quem de fato ia tomar posse, porque tem que ter aquele calculo do coeficiente eleitoral e Isso muda muito

[07:30]
dependendo do julgamento algumas de algumas candidaturas. Então a gente tinha essa opção também sendo jogados ali e eles podiam perder esse mandato depois de serem julgados inelegíveis ali e e perder. Tinha uma opção lá que a gente fazia quando isso acontecia, mas no Legislativo isso era muito mais imprevisível e dificil de controlar. A gente não ia ter como controlar. Eu vi agora, por exemplo, que mudou a lista de eleitos pra Assembleia Legislativa

de São Paulo, faz uma semana. O cara teve um novo julgamento e mudou um cara que tinha sido eleito, dai mudou para outro. Isso é natural que aconteça depois da eleição. Mas assim, mesmo logo depois da eleição, a gente já teve uma uma dificuldade de que talvez o resultado não fosse fiel.

[08:30]
A saída seria colocar os mais votados, que seria ruim também porque a pessoa iria acham que o cara mais votado tinha sido eleito e não foi. Então a gente preferiu não colocar nada do Legislativo nesse nesse sentido.

[08:43]
Eu tenho mais algumas dúvidas sobre o template. Foi você, o editor de Dados da Globo na época que fez o esses dois modelos de texto, já enquanto você estava em tratativos com o Setor de Tecnologia. Então eles já tinham ele já tinham te dado algum tipo de recomendação sobre como esse modelo deveria ficar? Não, isso partiu de mim eu que fiz desse modelo de texto e mandei para eles uma versão.

[09:20]
E a gente foi conversando a partir desse modelo que eu mandei, desse template. Então eu mandei esse rascunho no e-mail mesmo, com essas várias lacunas do que eu achava que se deveria preencher.

[09:26]
Então no começo, como eu falei, era um texto bem maior assim com várias informações. Eu não vou lembrar exatamente, mas haviam informação do IBGE, como o número de habitantes e o IDH da cidade. Tinha algumas informações a mais que a gente acabou não usando. E aí a gente foi burilando isso junto com o pessoal da tecnologia para ver o que fazia sentido e o que era mais fácil fazer. Então a gente já chegou a conclusão de que o ideal era ter apenas dados do TSE. Em algum momento, não lembro qual momento. Mas a gente chegou a essa conclusão. A partir dali a houve mais uma preocupação de como que funcionaria esse Processamento de Linguagem Natural no texto ali para mudar feminino e masculino, para mudar algumas algumas construções do texto, então por exemplo:

[10:27]
O Giovani do PT de 54 anos tem superior completo. A gente percebeu que lá na base não tinha só se era fundamental e superior completo. Tinha também se só lê e escreve. Dai tinha que mudar a palavra que acompanha, por que não dá para ficar "tem lê e escreve". Então tinham várias informações que a gente tinha que ter noção de tudo que poderia vir da base do TSE. Por exemplo, a informação da profissão foi bem complicado. A gente teve que ver todas as profissões possíveis para ver se tinha alguma ali que fugia do padrão e ficava estranho. Tanto o grau de instrução como todas as outras informações tiveram que avaliar as possibilidades dentro do Processamento de Linguagem Natural. Então gente até teve essa discussão com um dos engenheiros da Tecnologia se seria uma solução por

[11:20]
Inteligência Artificial. Ele falava que é Inteligência Artificial por que tudo é Inteligência Artificial, quando você mexe tipo de coisa, mas assim não

[11:37]
chega a ser avançado. Basicamente uma preenchimento de lacuna e o que a gente fez é um Processamento de Linguagem Natural para entender se é masculino ou feminino, para entender uma construção sintática dentro da frase. Mas não tem uma coisa tão avançada ponto de construir um texto sozinho, sabe esse tipo de coisa que a gente vê como assim sabe? Sim sim, já aproveitando esse assunto, pelo o que a teoria me mostrou existem três técnicas de Processamento de Linguagem Natural. Existe uma que é estatística, o Machine Learning. Uma que é o Sentence Embeding, que já é Deep Learning; E uma que é mais simples, a Automação por Regras e Eurística. Eu entendo que é essa que tenha sido usada.

[12:37]
Só para te falar, eu não sei se vocês chegaram a falar com alguém la da Tecnologia. A pessoa que mais ficou a frente disso, eu acho que ela também não está mais na Globo e que ela está no exterior, se chama Hector. Ele foi a pessoa que ficou responsável, por pensar nessa parte de Processamento de Linguagem Natural ele era meu ponto focal na Tecnologia. Então boa parte desse processo era

[13:37]
basicamente eu e ele conversando Então durante todo esse processo do início da ideia até boa parte da concretização era um eu e ele. Eu era o ponto focal ali no jornalismo, ele era um pouco focal na tecnologia, dai a gente conseguir fazer acontecer.

[14:13]
A gente fez até uma apresentação juntos na Campus Party. Uma apresentação falando sobre como é que foi e ele tem várias telas assim que poderia te ajudar de como foi toda a construção do negócio. Ele fez todo um organograma de como ele trabalhou e de como que foi feito essa Processamento de Linguagem Natural, como que foi usado. Então seria legal você falar com ele, porque ele tem todo esse mapeamento de como que foi feito. tempoEsse organograma que você tá falando, provavelmente alguma coisa numa estrutura de uma árvore de decisão, correto?

[15:13]
Sim, mas também um esquema de como foi montado o Processamento de Linguagem Natural, que tipo de método e toda a parte teórica, com as referências e tal.

[16:35]
Entre a definição do template, desses dois modelos que depois viraram micro-modelos. Pelo que você falou, vocês tiveram que criar mais variabilidade ali de possibilidades de frases dentro do texto, né? E você foi foi fazendo isso junto com Hector, né? Ao mesmo tempo vocês foram ali esmiuçando a base de dados do TSE para ver o que que dava para entrar. Em

algum momento você teve uma discussão sobre em qual ordem as informações entrariam no texto ou Isso aí foi bem assim padrão do lide jornalístico? Isso foi bem de boa, eu tinha basicamente o lide ali das informações principais.

[17:19]
Quem foi eleito, o percentual de votos que teve. Isso não mudou muito desde o começo com o rascunho do primeiro texto. Acho que não teve muito muito problema nisso, foi bem de boa e a gente tinha bem definido. Foi feito com base no que a gente escreveria em um texto normal, como se fosse um humano escrevendo. Então a minha ideia foi basicamente fazer um texto padrão que a gente mesmo; Só que pensando no lugar das informações escritas, a gente tinha puxar elas pela base do TSE. Mas é o texto não teve muito problema desde o começo. O rascunho definitivo ficou bem parecido com o rascunho inicial, em relação ao que foi publicado depois.O que estava disponível ali na base do TSE que vocês conseguiram ir avaliando ao longo do ano foi basicamente o que delimitou o texto? Exato, aí tinha as informações que iriam sair em tempo real ali como o percentual de

[18:19]
votos, número de brancos e nulos, enfim. Também tinham as informações que já estavam lá. Inclusive eles baixaram todos esses dados para não precisarem baixar isso em tempo real ali do TSE, pois não fazia sentido. Então informações de perfil das pessoas como profissão, idade, grau de instrução e todo o resto que já tava ali no DivulgaCand desde o início, meses antes ali da da apuração. Então todo esse trabalho prévio de baixar todos esses dados, de ter uma base ali toda estruturada foi feito pelo pessoal lá de Tecnologia. A gente ficou muito preocupado só com a parte da apuração. Para os veículos que de fato acompanham a apuração, o pessoal do TSE abre

[19:19]
algumas semanas antes para um período de teste, que eles fazem com dados fake. É uma simulação da apuração para ver se os dados estão sendo transmitidos está chegando. A gente colocou essa equipe de tecnologia junto com a equipe que da Globo que faz essa parte da apuração em si. É porque é uma outra equipe que faz toda essa parte da apuração que vai pra TV e que vai para todos os outros lugares. Existe toda uma tecnologia já feita e preparado ali pra pra receber esses dados. Então existe uma pessoa ali só para ter um backup no TSE, caso tenha algum um problema.

[20:19]
São feitas várias calls com o pessoal do TSE e vários testes. O pessoal da Tecnologia participaram dos testes pensando só na apuração para os textos automatizados. Então no dia que foi feito o teste eles acompanharam para ver se estavam chegando os dados e se os dados estavam nos textos. Enfim, então teve toda essa preocupação de entender se esse negócio estava funcionando ou não e se estavam precisando dos dados, se estava tudo certinho. Eu imagino que esse medo tinha a ver com suponho eu, como o volume de textos era muito grande, seriam 5.568 erros, né? A preocupação é que se os dados entrassem errado, iriam ser 5.568 erros?

[21:18]
Exatamento, então o que a gente fez foi para evitar isso, inclusive. No dia da apuração a gente fez uma operação de guerra na redação. Nós tinhamos muito receio de publicar os textos diretamente no ar. Depois que a gente estruturou o projeto e fez o processamento de linguagem natural, nós vimos que estava funcionando. Tinha um outro desafio ainda que era a gente ter um publicador. O G1 tem um publicador que se chama CMA, onde você coloca o texto lá, aperta o botão e publica. A gente precisava de uma interface ainda de como que esse texto ia entrar no publicador. Nunca tinhamos feito nada parecido. O pessoal da tecnologia ainda teve que se envolver com o pessoal do publicador. Foram várias áreas envolvidas. Tem uma área que só mexe com essa parte de estruturação do publicador o setor quer fica totalmente afastado da tecnologia.

[22:18]
Esse pessoal da tecnologia então foi colocado em contato com esse pessoal do publicador pra entender como que eles conseguiam colocar esse texto automaticamente, deixando ele lá como rascunho, sem publicar.

[22:29]
Então o pessoal da tecnologia teve que fazer um esquema para deixar esse texto lá como um rascunho para um jornalista entrar e ver se tá tudo certo para ir publicando. Como a gente não ia ter como publicar esse texto direto, a gente não tinha a segurança de publicar ele direto. Na verdade por parte da direção ali do G1 , por mim a gente publicaria, mas enfim, seria mais fácil publicar e depois dar uma olhadinha para ver se estava tudo certo. Ninguém ia ficar olhando tudo para ver se está tudo certo. Então teve que fazer uma operação de guerra envolvendo todas as afiliadas do G1. A gente envolveu mais de 200 jornalistas pelo Brasil, porque temos afiliadas em todos os estados e alguns estados mais uma afilhada. Por exemplo, São Paulo, a gente tem em Sorocaba em Ribeirão Preto em Campinas e Itapetininga. No Pará a gente.

[23:29]
tem em Santarém, a gente tem em Belém. Em Minas, a gente tem no triangulo Mineiro. São umas 55 afiliadas se eu não me engano. A gente envolveu todas elas, ficando responsável pelos textos da área delas. Em algumas áreas, como eram muitas cidades, na Bahia por exemplo tem muitas cidades. Eles não iam conseguir dar conta dos textos que eles tinham lá. A gente pegou uma um que fica em São Paulo.

[23:55]
Que ia subir alguns textos nesses estados onde tinha muitos e muitos textos. Então tinha um pessoal lá que estava responsável pelos da Bahia algumas vezes do Rio Grande do Sul, que também tinha muito texto. Tinha alguns de Minas. Se não me engano. Enfim, então a gente pensou e fizemos uma conta do que cada uma dessas afiliadas ia ter que publicar. A gente fez uma planilha de controle em que a gente tinha os textos que chegavam e a URL do CMA. E

logo que a matéria era publicada, tinha a pessoa que publico, quuantas publicou e o link da matéria publicada. Nessa planilha a gente teve o controle.

[24:37]
Então realmente é uma operação com muitas mãos mexendo numa planilha e mexendo nas matérias. Então eh, a gente chegou no final e a gente tinha que ver se tinham 5.568 links lá, lembrando que também eh… Só uma dúvidazinha? O trabalho dos mais de 200 jornalistas de todas as afiliadas, pelo que você falou, era de bater o olho no texto e ver se estava certo, de uma maneira mecânica? eu entrava no texto no senado estava certo. O cara que foi Prefeito que foi eleito foi eleito mesmo, ele chama uma coisa muito estranha no texto e mandava ver o que a gente fez a gente fez no no começo. A gente como eu entrar uns cinco mil vezes no CNA. A gente fez junto com uma combinação de tecnologia pra entrar num centro primeiro e a gente vê se tava tipo entrou sei lá uns 50 textos logo. Começou assim e a gente viu? Como que teve um problema estava tendo Tava tendo problema com o número de débitos e

[25:37]
Tive uma informação que tava errada. Tava puxando errado. E aí cara foi o caos você pode imaginar. E aí eles começaram a trabalhar no problema, porque senão ia baixar todas as notícias erradas. Então na verdade acho que a gente nem chegou a fazer uma trava que percebeu isso não tava começando a baixar. E aí eu falei meu interrompe até agora senão vai chegar tipo 50, cem textos no máximo, errados. Eu falei ó, esse sentença a gente mexe na mão aqui a gente vai entrar lá para ver o percentual e mete na mão. mas os próximos tem que vir certo, se não vai ficar muito grosso esse trabalho. A gente não vai ter como pedir para cada um das pessoas que estão ao mesmo tempo entrar mudar aquele dado chegarem muito tempo nisso. E aí vocês conseguiram ajustar rápido até e a gente seguiu na toada o que a gente combinou. E aí é importante você saber também é que assim como é um volume muito grande de textos.

[26:37]
Esses textos só iam entrar no nosso sistema quando a eleição estivesse fechado 100% dos votos.

[26:44]
Por quê? Porque a gente não queria que entrasse uma cacetada de texto, toda vez que tivesse ficado eleito depois. E aí tivesse que entrar nesse texto depois para mexer dos percentuais finais de como que foi a votação, mas a gente fez uma exceção para ser mais ou menos cento e pouca cidades que eram as cidades mais importantes do Brasil. Então a gente fez um recorte ali que era um tipo era uma substância. Eu sei que entre 100 e 200 cidades nessa cidade a gente tinha ela ia entrar duas vezes no texto do CNA. E então tivesse matematicamente eleito.

[27:19]
Porque nessas cidades, como são cidades mais importantes, eram cidades onde a gente precisava publicar logo. Assim que tivesse matematicamente eleito, a gente precisava desse texto tudo pronto no CMA, antes de de totalizar 100%. Então por exemplo, eu vou dar o

exemplo São Paulo, a gente não tinha como esperar estar 100% das zonas apuradas para publicar o texto dizendo que o prefeito de São Paulo foi eleito, nem do Rio de Janeiro, nem de nenhuma das filiais onde tinha cidades importantes, tipo Campinas. por exemplo.

Esse recorte aí de cidade importante seria talvez cidades médias e metrópoles?

[28:03]
É, basicamente. Que são sempre poucas cidades, talvez a gente tenha feito um recorte um pouquinho maior, se eu não me engano. A gente pegou algumas cidades que tinha mais eleitores e habitantes, mesmo que não tivesse segundo turno. Então tinha um recorte, mas não consigo lembrar qual foi essa régua, mas assim eh, a gente fez uma lista e o acesso à lista lá pro pessoal de tecnologia. Cara, a gente precisa dessa cidade aqui a gente precisa ter o resultado quando tiver matematicamente eleito. São Paulo, por exemplo a gente nem usou o texto automatizado porque o pessoal de São Paulo da editoria de São Paulo, estava preparando um texto super completo com várias informações com histórico da eleição com muito mais elementos desses básicos de olhar foto, por exemplo.

O que eu reparei na análise de conteúdo dos textos que foram publicados, é que geralmente textos de metrópoles e capitais trazem um pouco da história do político e da cidade. Então a minha dúvida é se foi um texto automatizado, ou se foi complementado por um jornalista? Depois que subimos os textos no CMA, eles estavam prontos, mas podiam ser alterados. Via de regra, nas principais cidades, eu digo nas capitais, quase todos devem ter feito isso manualmente. Eu recomendava para eles que os textos automatizados estariam ali, mas é claro que a filial tinha liberdade para preparar mais alguma coisa na véspera.

[29:33]
Vou dar um exemplo vai pessoal de Campinas e tem outras cidades ali grandes ali da turma da turma é uma cidade que não não consigo dizer com certeza, mas talvez tenha segundo turno, na verdade. Talvez eles não tenham conseguido fazer um teste de idade. Então assim eu sei que algumas praças usaram desde automatizado das necessidades grandes para poder aproveitar mais essa principais cidades da praça, por exemplo. Campinas. Deve ter feito isso de Campinas é sentarem provavelmente publicou o texto na mão de Santarém. Então é até porque esse texto ele já estava esse texto maior de de lugares mais importantes, eles são meio preparados na véspera assim, né? Você coloca ali, ele já isso que você falou um histórico da eleição, como é que foi fotos links saiba mais assim, isso a gente não conseguiu é é aliás. Isso é uma outra coisa assim é que é importante falar.

[30:33]
A gente gostaria de ter tido vários elementos a mais né? Mas que a gente não conseguiu nesse primeiro momento. Por exemplo, colocar foto foi impossível. A gente conseguiu trabalhar nesse espaço que ele tem para pensar e como foi uma foto lá dentro ou então um elemento de sei lá de link, não sei o quê. Então a gente optou por ter só um texto direto ali mesmo, nesse primeiro momento. Sendo que nessa segunda não participei desse em 2022. Agora já não tava

mais no G1, mas eu tava participando do planejamento no finalzinho do ano de 2021 do que que a gente faria nessa eleição e uma das coisas que eu tinha idealizadamente era ter um vídeo. Eu acho que eles fizeram pelo o que eu vi. Porque a gente tinha muita demanda por vídeos e a gente queria fazer um vídeo que fosse muito simples com o resultado da eleição também em vídeo para as pessoas postarem. Então essa foi uma coisa que a gente falou em deixar essa pra próxima. A gente fez ali o que era possível fazer na primeira vez e já era muita coisa, já era um esforço de inovação. A gente se enfiou

[31:33]
para fazer o texto e a gente já tinha esse problema que eu falei de ter colocado no sistema para ter a aprovação de jornalista, tinha muito muitos processos envolvidos, então o que aconteceu era que mesmo de algumas cidades pequenas a afiliada ia e colocava uma foto lá, colocava um saiba mais, colocavam alguma coisa que não estava previsto no template do texto automatizado. As vezes a afiliada entrava lá dentro pra checar e ela colocava um elemento a mais ali, entendeu? E isso tudo bem, tava liberado, não tinha que todos os textos serem iguais é se ela quisesse entrar e mudar alguma coisa no texto, tudo bem também. A nossa ideia era conseguir ter um texto para cada uma das cidades do Brasil. O objetivo não era ter uma coisa pasteurizada, não era ter uma coisa que tipo fosse igual para todos. Até o próprio o pessoal da tecnologia mudaram algumas frases para todas não serem iguais sabe.

[32:33]
Então alguns textos começavam com "o prefeito da cidade" na outra começava com "a cidade", algumas ordens em coisas bem pequenas eram mudadas. Mas não eram mudanças tão absurdas na construção do texto, mas tinham umas frases um pouquinho diferentes, não eram todos 5.568 textos iguais. Toda questão de estrutura de texto passava pela equipe editorial e pelos jornalistas? Sim, a gente combinou antes e a gente fala galera, vamos mudar isso daqui aquilo dali pode ser ? Isso foi combinado bem antes e aí a gente foi com algumas mudanças bem pequenas para não ficar todos os 5.500 textos iguais. Mas só mudança bem pequenas, como a descrição do prefeito eleito ou da cidade. Eram pequenas mudanças. mas com a estrutura dos parágrafos igual em todas. Então tinha ali o eleito, quanto recebeu de voto, o perfil do eleito e isso não mudava

[33:30]
Que mais. Então essas questões de estilo do texto, formatação, como que o número entraria, se era dividido por ponto ou dividido por vírgula, o tamanho das frases, o número de caracteres por frases, tudo passou por você? Tudo passou por mim, se tinha zero depois da vírgula, se era caixa alta ou baixa, se iria usar o cifrão, se o patrimônio viria com vírgula mais x 's zeros depois. De certa forma, você transferiu o seu manual de redação mental aí para equipes de tecnologia? Exato, então tinha algumas coisas que para eles não fariam sentido, de formatação de texto mesmo. A gente teve essa conversa para poder chegar num texto no padrão do G1 e no padrão jornalístico, que fizesse sentido. Então foram várias conversas e várias trocas antes de eles começarem a fazer esse texto

[34:34]
Dai antes da eleição, claro, eles fizeram umas simulações e mandaram para a gente dar uma olhada. Eu dividi com a minha equipe que era o Grandin, a Clara e a Gabi. Eu tinha quatro pessoas ali na equipe, mas uma estava muito focada no fato & fake. Os três eram mais de jornalismo de dados mesmo. Então os três ali me ajudaram a revisar o texto para ver se tinha alguma coisa muito estranha e para dar sugestões. Então teve esse trabalho também. Você falou que era você e mais três nessa equipe de jornalismo de dados. Vocês eram baseados ali no Rio ou em São Paulo?

[35:18]
Então, eu a Clara e a Gabi estavamos em São Paulo e o Grandin ficava no Rio. Quantas pessoas na equipe de de tecnologia que vocês correspondiam regularmente e como era essas correspondência? A gente tinha reuniões semanais com eles, eu

[35:40]
acho eram quinzenais e depois viraram semanais. Nessas reuniões não eram só da gente do G1 tá? Eram reuniões do jornalismo com a tecnologia. Então tinha gente do G1, da TV, da Globo News e tinha gente de outros setores ali da TV de alguns programas específicos como o Fantástico. Era uma reunião em que a gente conversava com ele sobre coisas que a tecnologia estava fazendo para o jornalismo, então a gente tinha uma fila de coisas que eles faziam pra gente. Vou dar um exemplo, um dashboard de dados da covid.

[36:40]
Então tinham outros projetos que eles tocavam com a gente esse projeto de eleição era um dos que estavam sendo tocados. Inclusive na eleição, eles também fizeram dashboard de consulta com os dados de eleição para a gente ajudar no dia da eleição com a apuração. A gente tinha uma editoria chamada eleição em números em que a gente fazia vários detox com os resultados da eleição. Então sei lá, qual que é a cidade que teve a menor e maior diferença de votos entre um candidato e outro, por exemplo. Algumas curiosidades também. Qual é o perfil do deputado eleito. Esse tipo de coisa eles também nos ajudaram a conseguir os dados mais rápido. A gente pegava e fazia alguma coisa em R, ou no Excel e eles fizeram uma coisa mais intuitiva pra gente conseguir pegar mais rápido. Então você e sua equipe chegaram a programar alguma coisa foi usada no projeto?

[37:39]
Desse projeto específico não, mas a gente na equipe tanto o Grandin, quanto a Gabi, eles mexiam super em R. Então eles faziam outras coisas em R, para outros projetos, mas para esse projeto específico não. Nesse projeto ficou tudo com a tecnologia dessa parte de programação. Mas o fato de vocês terem uma literacia digital, digamos assim, de entendimento de programação e de lógica de programação, ajudou? Isso ajudou a conversar com eles e a entender tudo muito mais fácil. As conversas eram muito mais fluidas e a gente percebia isso nas próprias reuniões quinzenais, porque o pessoal que participava dos outros setores da Globo, não era da área de dados e não tinham essa expertise. Então eles tinham muito mais dificuldade de entender até as possibilidades do que podia ser feito. Na própria

conversa ali com o pessoal de tecnologia, então com certeza esse contato nosso com a linguagem de programação

[38:39]
ajudava muito nessa conversas. Vocês de certa forma dentro da organização jornalística, eram como que híbridos ali, né? Vocês conseguiam entender tanta cabeça do repórter do editor como um pouco a cabeça do técnico?

[39:00]
Era até meio natural. A Gabriela César que estava na equipe, mais não está mais lá também, ela está no no Google agora, mas ela era uma pessoa que flutuava muito entre jornalismo e programação. Ela gostava muito de programar, estudava muito e estava em um nível muito avançado.

[39:24]
E ela tinha muito interesse em tecnologia. Tanto que em algum momento ali a gente tinha alguns projeto de tecnologia que não estavam tão ligados ao jornalismo diretamente e a gente combinou de ela participar de algumas coisas e fazer alguns intercâmbios, digamos assim, lá na na tecnologia para poder entender como que era feito todo o processo desde o começo até o final de como que era que eles lidavam lá porque ela queria tinha curiosidade nisso assim. Então a gente tinha já essa simbiose com essa área (tecnologia) até pelo pelo interesse das pessoas da nossa equipe. Então a Gabi era uma pessoa que estava muito ligada a eles e tinha esse interesse de conhecer e de entender como que era o trabalho dele, entender os processos então a gente chegou a combinar de um desses projetos ela participar desde o começo lá com eles.

[40:24]
Falando um pouco de ferramentas que foram usadas ao longo desse projeto específico da cobertura automatizada? No dia a gente fez uma planilha compartilhada com todo mundo de google docs para as pessoas poderem entrar e colocar os links, mas isso era uma coisa estrutural de ter um controle.

[41:04]
Assim a parte mais de tecnologia foi feito com a tecnologia mesmo, a gente não usou nenhuma ferramenta para poder automatizar o texto, a não ser a própria ferramenta de publicação (CNA).

[41:19]
Toda essa parte ferramental foi da Tecnologia mesmo. Toda essa parte, desde que eles usaram para fazer o processamento de linguagem natural até que eles fizeram como que eles usaram para colocar dentro do CMA. Enfim é pra pegar os dados do TSE tudo isso seria mais com ele vai entender tudo que eles utilizaram essa ferramenta assim Maravilha é o último

penúltima pergunta é para ir Fechando assim, que muito se fala, né? Ainda mais com esse avanço muito rápido que teve essa área de processamento de linguagem natural teve de sei lá três uns quatro anos para cá. Um certo fantasma do desemprego estrutural. A partir da aplicação desta tecnologia e de que você trabalhou nesse projeto. Qual foi o sentimento que ficou? Primeiro que a gente brincava com isso na redação. A gente falava que os robôs iam roubar nossos empregos. Nós fazíamos muita brincadeira com isso..

[42:36]
Então primeiro que eu tinha esse projeto é um exemplo assim do jeito que ele foi construído e ele é é foi colocado no ar é um exemplo contrário, né, gente pra caramba pra publicar os textos. A gente não confiou de fato na tecnologia para poder entrar automaticamente. A gente teve que me envolver muita gente, mas muita gente mesmo pra conseguir fazer isso é eu não sei se o grande se você falou que tinha antes da eleição, depois a eleição desse ano foi depois foi depois e ele não sei como é que foi esse ano se eles usaram também uma cacetada de gente pra publicar de novo não ele falou que esse ano foi foi direto entendi. É porque assim a primeira vez é É teve muita gente que foi usado, mas aí Respondendo a sua pergunta, sobre esse projeto específico, eu acho que não porque se fosse assim a gente não teria feito esses textos. Se não tivesse sido automatizado, só haviam textos para as principais cidades, endente?

[43:36]
Agora, para as 5.568 cidades, ninguém teria feito. Eu não acho que seria um problema de vamos substituir as pessoas, porque a gente simplesmente não conseguiria fazer com as pessoas que a gente tinha. Então não é nem essa a questão. Segundo que os textos mais importantes como eu falei que são os textos que são os mais livres de fato, porque assim é tem uma componente que é importante saber aí também que é a gente queria fazer uma uma inovador era complicado isso pra cada uma das cidades, mas conseguir algum tipo de audiência com isso porque as pessoas dessas cidades iam querer olhar lá enfim e a gente gostaria de ter é esse alcance é é.

[44:16]
Como que eu posso dizer, é bem distribuído no Brasil todos a gente conseguir ter uma audiência qualificada ali em várias cidades, mas ao mesmo tempo a gente sabe que a audiência desses textos sozinhos, no final juntando todas, não se comparar a audiência de São Paulo, ou por exemplo do Rio de Janeiro, até porque esse texto inclusive eles iam ser destacados dentro do São Paulo. Eu não sei cara eu posso voltar não sei compartilhados eu finalizar nas redes sociais, então assim é a distribuição mesmo. A gente sabe que hoje tipo só você publicar perdida, você não consegue tem que chegar nele, né? Então a gente tinha ali até uma chamada veja todos os tempos todos do Brasil inteiro a pessoa.

[45:03]
Por esse caminho ir lá e chegar na cidade, mas os outros terço que estavam na rede social, tá? Não tipo sei lá colocados no Instagram e no Facebook pra você voltava é colocados nos perfis, eu precisava de G1 tava chamando ponto com tava chamado da capa do G1. Esse texto

é uma audiência muito maior que nós não precisa ser mais trabalhados e eu preciso de um humano, de um jornalista ali para conseguir fazer um texto atrativo e é substancioso , com mais informações e informações críticas, então é ao meu ver esse projeto não é um exemplo de como robôs podem substituir humanos e sim, de ajudar numa tarefa que ia ser impossível um humano fazer sozinho. Eu vejo mais nesse sentido, como uma tarefa complementar ao trabalho do jornalista, mais do que uma substituição do papel do jornalista. Nesse caso sabe, então acho que esse é um exemplo bem claro de não substituição e sim de complementaridade.

[46:01]
Complementariedade é uma palavra similar ao que o Felipe usou, que foi potencializador do trabalho do jornalismo, não ia conseguir publicar esse texto sozinho com jornalistas com mesmo tendo essa Força Tarefa e é assim Impossível a gente pegar e ele pegar e puxar essas informações do TSL e fazer por um ano seguinte começar a fazer isso tipo dias em semanas meses antes, sabe? Acho que todos os dias prontos com várias opções ali do que ele poderia ganhar, né? Porque por exemplo imprevisível não fazia sentido não teria sentido.

[46:36]
Maravilha eh para fechar assim, eu só queria ver com você outros nomes eh e e talvez até documentos assim que você tenha que possa compartilhar para eu conseguir ter uma noção melhores assim o que não fosse sigiloso É claro é mas você falou da Gabi que era da sua equipe. A Gabriela não participou tão ativamente. Eu não sei se Valeria acho que não sei se vai apenas ser ela participou mais lateralmente assim, eu não sei se foi a pena você perder um tempo assim, eu acho que a pessoa que você tem que falar tem que falar, ué, acho que vai então você tem que falar porque ele é a pessoa chave do negócio, ele que teve envolvido muito nesse nessa parte do do processamento natural, ele construiu para dentro do começo junto comigo ali é que tem sonho, como que seria Toda a modelagem que estava envolvendo ali no no caso ali do dia da eleição. Eu não sei se a gente falou isso a gente teve alguns problemas com um próprio sistema do TSE que interrompeu a transmissão. Em algum momento no meio da eleição eles interromperam o envio dos dados a gente ficou umas quatro ou cinco horas sem conseguir fazer nada. Estavam ali todas as pessoas dispostas para esperar até chegar os textos e não estavam vindo os dados. Não estava conseguindo entrar no CMA, então a gente ficou horas ali eh sem conseguir fazer nada e a gente sempre uma hora a gente falou cara. Vamos interromper aqui essas pessoas que foram embora e Vamos retomar isso de manhã e a gente fez tipo do reverter uma operação de guerra ali de mandar gente de casa e as pessoas que estavam chegando na redação no dia seguinte nesses lugares começarem a tocar porque caiu no TSE voltou muito pela manhã, então a gente demorou na manhã do dia primeiro na manhã do dia que foi do dia da eleição, né? Porque em 2020 foi dia primeiro, você tá falando do primeiro turno ou da posse no caso e aí a gente

[48:23]
Que a gente teve essa a gente ficou então a gente demorou mais de 24 horas para publicar quase todos os tempos sabe a gente teve uma um tempo bem maior do que a gente programável e planejava fazer entre o início da da apuração e o momento de publicação que

você falou, né? Que foram exato exato, até porque teve um problema de de fato na apuração nesse nesse meio tempo, então a gente dependia do TSE fazer além do Hector. Eu tinha mais alguém né? Que pediu de tecnologia que você lembra o nome mais ou menos assim, cara, tinha uma das pessoas mas acho que talvez seja a melhor pessoa e ele pode te indicar. Talvez quando você falar com ele aqui que quem mais ele acha que poderia falar sabe primeira coisa que você falar tem que ser o Hector e ele pode depois te indicar alguém que enfim.

[49:08]
Tá não maravilha, quem você tiver de nome aí para mim indicar. Eu aceito se você lembrar depois ou antes tem acho que é o principal mesmo, cara.

[49:21]
Tá beleza, eu vou tentar ver se consigo achar, mas eu não tô achando minha internet tudo no meus e-mails da Globo. Se tiver saindo eu conseguia te mandar super fácil rascunho que eu fiz, mas tudo cara, eu não guardei eh. Talvez o grande tenha isso. Se você pedir para ele porque ele tava acompanhado desses e-mails, não esqueci de te passou alguma coisa eh e também não sei se ele vai se ele vai sentir confortável passar porque as pessoas ele deve ter a situação muito chatas na Globo para atualizar. Qualquer coisa eu não sei como é que a pessoa tá com ele, mas para

[50:00]
Para ser super aberto porque normalmente é muito chato, ela tem uma burocracia muito tem que passar por você, todo mundo universidade para ele estar bem estar conversando com um pesquisador de tudo que você vai passar de documento. Enfim pensando bem chatos com essa burocracia, tá? Beleza, mas a gente eu não tenho nada que eu lembre aqui agora porque eu acabei realmente perdendo meus e-mails assim, mas talvez não tenha sido pelos iniciais. Talvez ele tenha assim e o eco. Talvez ele vai fazer alguma coisa, mas ele é tipo como ele tem todo o processo dentro tudo que ele fez se ele guardou alguma coisa, talvez ele possa te ajudar também perfeito. E esse vídeo que você falou também do do da Campus, Party, se você lembrar. Sei lá o nome do evento que eu tenho.

[50:59]
Enfim, se eu achar eu te falo sim, o o Hector ele com certeza ele deve ser porque foi ele que me mandou o vídeo quando for no quando a gente fez então talvez ele tenha o vídeo tá? Eu vou tentar esse contato dele aqui no Linkedin, você precisa falar com ele, você está no finalzinho aí da da do trabalho, mas cara, ele é uma pessoa super importante assim nesse

[51:22]
nesse vamos ver se ele é acessível, mas eu vou tentar aproveitar que ele não tá mais na Globo também fica mais fácil para ele poder falar e tal Maravilha. Obrigado aí viu, Tiago você precisar de alguma coisa me manda e-mail. Estamos estamos falando demais foi foi muito.

Entrevista decupada com Felipe Grandin

[02:29]
Eu queria saber de você, como é que foi o seu envolvimento nesse projeto?

[02:35]
Por que que você está mais interessado da posse do que do primeiro turno?

[02:41]
Traz mais informações, né do que no primeiro turno.

[03:00]
Segundo esse mesmo texto que está no anúncio do projeto, tanto a primeira publicação como a segunda são duas partes do mesmo projeto, correto?

[03:16]
Eu perguntei porque foi mais ou menos a mesma coisa.

[03:38]
A gente decidiu fazer os textos da Posse também. Que tinha tinha mais informações. E esse ano também de 2022, a gente também fez.

[03:48]
Eh no caso vocês fizeram para corrida de Presidente, senadores e deputados. isso , mas daí com percentual em cada município.

[04:02]
Não é tão não é tão bom quanto Municipal, que é mais lógico. Mas nessa eleição, principalmente, as pessoas queriam saber quem que foi mais votado em cada município. Dai agora em 2022 a gente a inovação que a gente fez foi publicar

[04:26]
um vídeo para cada município, também com os resultados tanto no primeiro turno quanto no segundo turno.

[04:36]
Quando que essa ideia apareceu primeiro para você lá em 2020. Então, lá eu não era o editor do G1 dados, mas eu participei como repórter no início.

[05:37]

Isso surgiu há vários meses antes das eleições.

[05:49]
A ideia era publicação dos resultados para o primeiro turno e originalmente era era só isso.

[05:55]

Nas discussões iniciais tinha alguma referência de um veículo internacional ou mesmo nacional?

Deixa ver se tentado algo parecido, cara, vou ter que recuperar.

[06:07]
Eu imagino que sim.

[06:31]
Quantos meses mais ou menos de trabalho prévio vocês tiveram que ter antes do primeiro turno?

Desde o início do ano, eu acho que o grosso mesmo foram uns três a quatro meses assim antes.

[06:53]
Foi quando a gente definiu o template e tudo mais.

[06:58]
E desde o início do ano essa tratativas eram entre os jornalistas e esse antigo departamento de tecnologia?

[07:31]
Sempre foi junto com o COE o projeto.

[09:15]
Em 2020 não foi eu que coordenei o projeto. Esse ano, sim. Em 2022 fui eu que coordenei o projeto da publicação dos vídeos e dos textos automatizados. Então é legal você falar como Thiago e o Hector.

[09:58]
Você tinha dua equipes basicamente: a equipe de conteúdo, ou editorial, que era a gente de jornalismo de dados e a equipe da tecnologia, que cuidava do script do algoritmo que gerava o template de acordo com os dados da base. Como vocês lidaram com questões de limpeza e

formatação da base? Inicialmente a gente (jornalismo de dados) fez uma análise do tipo de dado que tinha

[10:57]
e de como era a base, o que a gente podia usar para fazer a matéria. Então no caso dos resultados, a gente queria dizer quem foi o eleito em cada cidade e isso era o mais importante. "Foi eleito o prefeito, ou a prefeito do município tal". Acho que essa era a informação principal. Além disso a gente queria dizer quem eram os vereadores. Essa era a segunda informação mais importante. Quais são os vereadores eleitos ali para cada cidade.

[11:35]
A partir daí, a gente olhou na base. A gente já sabia mais ou menos o que que tinha porque a gente já cobria eleições , mas aí a gente viu que tipo de informação sobre cada candidato eleito podia acrescentar na matéria. Então eu indiquei qual o percentual de votos, quem vai para segundo turno. Ou quem foi para o segundo turno junto com ele. A partir dai a gente pegou os gastos de campanha, qual o perfil do candidato, qual é a profissão. Tudo isso são informações que tinham sobre os candidatos na base do TSE, que é basicamente o DivulgaCand.

[12:35]
É uma base do TSE que é onde você vê as informações dos candidatos.

[12:43]

[13:13]
Tem as outras bases que são de resultados, que daí é o principal de onde a gente puxa. Quem foi eleito e quem não foi eleito, quantos votos teve e como ficou percentual etc.

[14:02]
Basicamente a gente puxa do TSE. Inicialmente, que tipo de dados a gente vai ter no dia para colocar? Então a gente tem os dados prévios do candidato no DivulgaCand. Ele é divulgado antes da eleição, né? Então a gente já tinha esses dados. Deu para deixar a base pronta com esses dados. O que ficou faltando eram os dados do resultado de apuração.

[14:32]
Com base no resultado da apuração, dá para puxar lá quem é o candidato que ganhou e etc. Então eram basicamente duas fontes de dados: uma já estava divulgada previamente e uma que estava sendo atualizada no dia. Vocês já sabiam quais dados iam ter no dia da apuração? Isso, a gente fez o template a partir. A gente vê que dados ia ter e daí faz um template que prevê quais são as variações que a gente pode ter. Como vocês definiram a ordem em que as informações entram no texto, porque por exemplo o estado civil do candidato entra antes grau de escolaridade é que aparece antes do patrimônio declarado.

[15:28]
Então como que foi essa discussão para dizer qual dado entra antes e depois? Isso foi uma decisão editorial mesmo. No geral, foi a pirâmide invertida do jornalismo. Primeiro vai ser quem foi eleito, do mais importante para o menos importante. Mas no caso do perfil do candidato, foi uma decisão editorial. De certa forma a ordem desses dados reflete um pouco os critérios de notíciabilidade? Disso aí não tem a menor dúvida. A ideia foi fazer uma reportagem. Lógico que usando outro tipo de recurso. Mas a ideia é manter os critérios.

[16:31]
Como você falou, antes disso a base de dados serviu de forma para vocês? Sim, é então a base de dados seria a apuração. A gente tem essas informações aqui, agora com base nessas informações, a gente vai escrever um texto.

[17:00]
Foi basicamente isso. A gente vai saber se naquelas cidades os prefeitos foram eleito no primeiro turno, ou se os dois foram para o segundo turno. A gente vai saber quem são os vereadores eleitos, entende? A gente vai saber quantos votos teve. Qual percentual e etc. A partir da base a gente viu o que a gente ia ter de informação.

[17:28]
A partir disso, a gente construiu um texto jornalístico, mas que fosse obviamente adaptável

[17:40]
para cada município. Então, você tinha que prever por exemplo, se ganhou no primeiro ou no segundo turno, se é homem ou é mulher. Tudo isso é feito por uma árvore de decisão. Se ganhou no primeiro turno, vai ser esse texto aqui, se não, vai ser o outro. Aí vai ter essa preposição aqui vai ter aquela proposição ali, entende? Daí vai indo dessa forma.

[18:11]
Você participou dessa delimitação do template?

[18:17]
Sim. Tinham questões sobre se estivesse impugnada a candidatura. Como é que avisa? Tem outras variáveis que a gente tem que prever o que vai acontecer.

[18:54]
Uma candidatura pode não estar valendo por vários motivos. A gente juntou todos elas. Dai tem uma coluna lá no TSE que é apto e inapto.

[19:02]
Essas informações tem na base do TSE, aí tem uma frasezinha para cada uma. Dai o texto era publicado com a devida ressalva.Vocês tiveram uma preocupação com a extensão do texto, sobre a frase ter no máximo x caracteres e o parágrafo também, como é que foi isso?

[20:02]
Sim, a gente não tinha uma regra que seja usada no dia a dia, mas a ideia era não ficar um texto enorme com um monte de dados que as pessoas não iriam ler. Inclusive no G1 a gente não têm nenhuma restrição a textos longos.

[20:26]
A gente queria ser um mais objetivo possível, mas incluir os principais informações ali. Mas eu lembro que a gente não queria que ficasse um texto enorme com um monte de informação ali. Porque no final tinha a lista com os vereadores eleitos. O texto é um texto jornalístico, como é publicado no G1, não tem diferença, ele segue as regras e segue os parâmetros.

[21:19]
Teve muita preocupação com questão da formatação do texto? Também segue o mesmo padrão do G1.

[21:32]
A ideia é que se fosse uma matéria normal e que o resultado final não tivesse diferença.

[21:43]
Então assim, o esforço todo na elaboração do template, foi para que não houver essa diferença entre essa matéria e uma matéria que a gente fosse escrever normalmente sobre o resultado. É lógico que tem coisas que a gente talvez fosse escrever na própria matéria e que não tinham na base do TSE. Por exemplo, algum detalhes que aconteceu no dia, ou uma aspa de alguém. Isso não teria como incluir, mas o texto ali factual é uma matéria normal do G1.

[22:20]
Essas escolhas dentro da da árvore de decisão e dentro do template sobre sinônimos, que tipo de palavra inicial finaliza uma frase, como é que isso foi discutido no início do projeto?

[22:40]
Esse ano também houve a mesma discussão, mas foi um pouco mais complicado, porque em 2020 era mais objetivo assim.

[22:49]
Cada cidade tinha um resultado para cada cidade. Esse ano não foi o resultado, foi a votação, né? Então foi um pouco mais complicado, mas a ideia é que com os recursos de que a gente tinha, com o dado que a gente tinha, a forma como ele ia chegar, ele ficasse o mais próximo possível de um texto corrido normal. O mais fluido possível. Que não tivesse nenhum estranhamento. Então a gente vê como que pode chegar aquele dado, como é que aquele dado vai entrar e vai ser colocado ali e de que forma isso se encaixa. Fazer um texto em que esse dado se encaixe de forma flúida, sem ficar uma coisa telegráfica que pareça automatizada.

[23:43]
Para isso vocês foram fazendo testes? A gente foi fazendo e refazendo o template, fazendo testes com o texto, por exemplo:

[24:02]
a gente faz um template básico primeiro aí, manda para o departamento de tecnologia template, eles rodam com o script do algorítmo, pegam a base da última eleição (2018) e rodam com os dados antigos. Quando foi chegando mais perto, o TSE soltou uma base fake também. Eles criaram essa base de candidatos fake para fazer testes.

[24:43]
Aí eles usaram essa base de teste do TSE, que aí já usa a mesma API e aproveita tudo. Onde que entra o processamento de linguagem natural entre esse processo e a redação das matérias?

[25:08]
É nesse caminho entre o a base e a geração do texto.Qual parte do template que vai usar o dado e como vai usar. Não é um uso super sofisticado de NLG. Se você for pegar os mais sofisticados hoje, ele vê expectativa, compara com números da série histórica e tudo mais. O texto final seria algo na linha assim: essa foi a eleição mais disputada em x anos. Por que o texto seria gerado a partir de uma séria de comparações.

[25:55]
A árvore de decisões é bem mais complexa do que a que a gente usou. A gente uma árvore decisões mais simples, com algumas variáveis.

[26:22]
O texto automatizado mais simples é o que você tem o dado e ele tem um template fixo,

[26:37]
que bota sempre aquele dado. Seria o mais mais simples de todos. Você pode aumentar a complexidade. Aí você pode ter um template, ou dois, três template, ou um template com diferentes partes entre si. São outros tipos de complexidade de acordo com o número de variávelis. Outra possibilidade é que além daquele dado recebido, você compara ele com outros dados. Então por exemplo, se for pegar o Automated Insights, ou o Narrative Science, que fazem para a Forbes. Vai sair o resultado de fechamento das empresas, eles publicam automatizado. Então sai o resultado de alguma empresa, ele

[27:22]
faz um texto sobre aquele resultado. Sem contato humano, só no algoritmo. Ele não pega simplesmente o número do lucro da empresa naquele trimestre, ou o faturamento. Eles veem.

[27:40]
qual era o consenso de mercado do que ia ser o lucro e se o lucro superou, ou ficou abaixo daquele consenso.

[27:46]
Aí ele muda o texto de acordo com isso, por exemplo, a empresa teve um um lucro 5% maior do que o que era esperado. A empresa tal teve um lucro acima do esperado pelos analistas. A empresa está 50% maior. A empresa tal surpreendeu analistas e com um lucro muito acima do que estava esperado. Isso tudo o algoritmo faz sozinho. E aí você nivela. Você programa isso antes. Por exemplo, até 5% é "maior do que "acima de 10% é "muito maior do que", entendeu? Você vai colocando condições para o algoritmo. Isso aí é um programinha. Não tô falando não do que a gente fez, mas dio que tem hoje no mercado.

[28:45]
É um programinha que você vai ajustando ali. Então assim, a gente não chegou nesse nível. O que a gente tem é diferente templates. Você tem uma árvore de decisão

[28:58]
e com a base de dados você vai alimentando, de acordo com os dados que vão chegando, você vai usando cada um dos templates.

[29:09]
O template ele serve tanto na estrutura do parágrafo, quanto da frase, quanto de orações?

[29:19]
Sim

[29:24]
Sim, por exemplo: que ganhou no primeiro turno, vai ser lide e sublide, se for para o segundo turno vai ser outro lide e sublide. Se tiver um problema de candidatura vai ter outra frase ali. Aí no nível da palavra,

[29:50]
se for mulher vai ter uma preposição, se for home vai ter outra fome vai ter outra preposição. Pode ser prefeito ou prefeita. Daí você vai indo.

[30:02]
Essa tecnologia de geração de linguagem natural foi desenvolvida di dentro? Sim, foi internamente.

[30:35]

Então a preparação e o desenvolvimento do código e dos templates foi uma conversa constante entre as equipes? Como é que acontecia essa colaboração no dia a dia? A gente fazia reunião pelo menos uma vez por semana e conversava principalmente pelo Teams, que é a ferramenta que a gente usa no dia a dia.

[30:55]
Quando foi chegando mais perto, não tinha tanta periodicidade, a gente foi se falando e resolvendo.

[31:03]
Fiz uma alteração no template, aí o cara aí eles vão rodam e mandam de volta. A gente via e falava se não estava legal. Eu acho melhor a gente fazer assim. Ah, acho melhor a gente botar um link, acho melhor a gente colocar em negrito, melhor colocar em Bullet. E aí foi indo e voltando. De certa forma vocês estavam mais preocupados com conteúdo do resultado final e eles com o "como fazer" ? Isso, exatamente. Para você já era um repórter de dados, mas pelo que você pode observar da sua experiência, a necessidade de se familiarizar com novas ferramentas e novos conceitos para conseguir acompanhar o desenvolvimento desse projeto?

[31:51]
Com ferramentas não porque foi essa parte de desenvolver ficou toda com a tecnologia.

[32:00]
Então para você foi basicamente o texto e briefar a equipe de tecnologia sobre a base de dados do TSE.

[32:11]
Isso, acho que o conhecimento que precisa ter é de como funciona e o que o algorítmo pode fazer, mais do que escrever o script.

[32:22]
E como você desenvolveu esse conhecimento? Você já tinha algum estudo prévio? Eu já tinha porque eu gostava dessa area e nem precisou muito por que em 2020 quem foi o cara que ficou à frente foi o Tiago. Esse ano que para mim acabou sendo mais útil. Eu participei com ele no template.

[32:59]
Eu já eu já tinha pesquisado isso porque eu eu fiz um pré-projeto de doutorado que eu acabei desistindo e era sobre "robô jornalista", que era o nome na moda da época.

[33:24]

Eu conversei com o pessoal das companhias que tinham na época. Eram poucas. Acho que agora também mais.

[33:37]
Mas assim, não exigiu nada super sofisticado da parte editorial não.

[33:43]
Muita gente nem nem houveram muitas necessidades de explicação assim por parte da equipe de tecnologias de tentar mostrar para vocês o código rodando?

[33:58]
Não. Então os instrumentos que você usou para colaborar no projeto foram o Teams e as ferramentas de texto mesmo? Sim, exatamente. Isso e as bases.

[34:15]
Como que vocês passaram essas informações da base do TSE pra equipe de tecnologia?

[34:38]
Pelo Teams e por e-mail.

[34:51]
Esse ano eles tiveram um contato com o TSE, por que eles estavam atuando em outros projetos de apuração.

[35:02]

[35:06]
Esmiuçando um pouco mais, você ia indicando quais colunas da base de dados a informação estava. Sim, a gente tinha que dizer o que iria entrar. Então por exemplo, vai fazer o texto, tem que estar lá indicado no texto qual dado vai entrar.

[35:30]
Vai ser o candidato a prefeito, vai ser o mais votado. Tinha que ter uma indicação ali de qual dado vai entrar. Qual a idade, qual é a UF (Unidade da Federação), qual é o município,

[35:53]
tem que ter ao longo do texto

[36:00]
entre colchetes, qual o dado que vai entrar e quais são as possibilidades.

[36:11]

Ou seja, dentro do template você já define aonde o dado vai entrar é qual era o conteúdo escrito? O que vocês alinhavam com a equipe tecnologia era mais ou menos essa questão da árvore de decisão, né? Isso. Talvez seja melhor ver essa linha do tempo com Thiago, mas pelo o que eu me lembre

[36:46]
foi que mais perto da eleição é que os templates já estavam prontos, aí começaram a fazer os testes rodando. Mas antes o que pegou mais era mais a definição do template. A gente definiu como é que a gente queria o texto e aí depois quando foi chegando mais perto, ele já conseguiam rodar

[37:06]
para ver como ficava na página e tudo mais. Aí o que restou era fazer o ajuste fino ali, né detalhes.

[37:15]
Mas pelo que eu entendi em momento nenhum você e o Thiago tiveram que enfrentar desafios com o código? Não, a programação não ficou com a gente. O anúncio fala sobre um processo de revisão dos textos, antes dele serem publicados. Como foi isso?

[38:02]
O que aconteceu foi o seguinte, a gente nunca tinha feito nada disso. A gente testou e tudo mais, mas assim na hora que sai, pode ter erro na base do TSE, pode ser mil coisas. Então a principal preocupação era de não rodar direito o código e sair alguma coisa errada. Por que uma vez que você rodou tudo e publicou, para despublicar é uma trabalheira, sem falar que aquilo ali ia estar no ar. Se voc~e publicou alguma coisa errada, vão ter 5.568 matérias erradas para você despublicar. Não têm uma forma de depublicar todas de uma vez. Nesse novo modelo você consegue publicar várias matérias de uma vez. Mas para despublicar é manual.

[38:39]
E você ter que despublicar com elas no ar é um problema enorme, porque assim você afeta credibilidade do veículo. Você não pode publicar um monte de coisa errada, não importa se aquilo foi feito automatizado ou não. Então é

[39:23]
algo que é uma preocupação constante, de publicar só o que está correto. Então como a gente nunca tinha feito isso e não sabia se ia dar certo, se ia chegar certo os dados do TSE, se ia mandar os dados corretos, se o algoritmo ia funcionar e se não ia dar nenhum nenhum problema no meio do caminho. Daí a gente não sabia se iria dar algum erro que nós não estavamos prevendo, mas que que acontece, né? Então optou-se como era a primeira vez de fazer um trabalho braçal, de antes da publicação, de gerar os textos todos mas sem publicar.

[40:09]

E aí a gente revisou, manualmente, né? Revisou todos e olhou para ver se estava ok. Mas quem era "a gente". Era um monte de jornalista do G1 e das afiliadas.

[40:33]
Vocês priorizaram, por exemplo, revisar texto de capitais, foi uma amostra ou foram todos? Foram todos os textos.

[40:55]
Isso em 2020.

[40:58]
Quanto tempo esses jornalista levaram para fazer isso? Foram publicando até o dia seguinte. Começou

[41:06]
quando saiu os resultado no domingo foi até o dia seguinte.

[41:12]
Menos de 24 horas.

[41:15]
Então em menos de 24 horas essa equipe do G1 e das afiliadas conseguiu revisar os 5 mil textos? Isso.

[41:29]
E vocês fizeram depois algum "lições aprendidas" sobre os erros mais comuns que passaram? O bom é que não teve. Os textos saíram redondos assim como a gente estava esperando, tanto que esse ano a gente não revisou antes a gente publicou direto.

[41:57]
A gente viu que funcionava então a gente publicou direto. Só que esse ano a gente fez a mesma coisa com os vídeos automatizados. Foi a primeira vez que a gente publicou esse ano. Então a gente revisou por lote.

[42:17]
Geramos 10 vídeos, confere, gerou mais 10, gerou mais 30, confere. Depois que chegou em um número X, a gente falou "bom, agora tá tudo certo ". Se está tudo certo e não teve nada errado,

[42:36]
pode rodar e publicar. Então em 2020 vocês não tiveram grandes problemas com erro ortográfico, ou erro no código? E aí esse ano vocês foram para publicação direta, automatizada mesmo?

[42:52]
O que diferiu da experiência que você teve em 2020 e da experiência que você teve em 2022?

[43:03]
A diferença para mim pessoalmente foi que agora eu toquei a parte editorial. Em 2020 tinha outra pessoa tocando e eu estava ali dando dando apoio. E o que você sente que progrediu?

[43:26]
O principal avanço foram os vídeos. Porque além do texto, a gente fez vídeos automatizados.

[43:32]
Então assim, os textos já estavam mais ou menos dominado. O que avançou nos textos foi que a gente não precisa revisar e publicar manualmente.

[43:41]
Teve uma revisão por lote, mas depois de já publicado. Então os textos já entraram publicados. Então o que teve de grande novidade, né? Foi um salto bem grande fazer os vídeos automatizados, porque aí foi uma operação bem mais complexa do os textos.

[44:05]
Para fazer o vídeo automatizado você precisa de um template também, só que há vários outros processos envolvidos. Tem uma plataforma que a gente usa para fez vídeo.

[44:14]

[44:21]
Aí tem outra plataforma para publicar. E aí tem uma parte gráfica, né? Daí entram outras variáveis.

[44:36]
Por que entra a parte de design junto. Não só a parte editorial. Então já são três equipes trabalhando. Entra a necessidade de usar alguma base de imagem? Você tem a base de imagem, você tem de áudio, tem uma trilha, você tem texto e também a parte editorial que é definir quais dados e que tipo de foto que vai entrar. O que vai ter mais destaque, né? Informação, como é que vai aparecer e tal então é bem mais complexo sim.

[45:25]
Aí envolveu o departamento de tecnologia.

[45:30]
O editorial ficou comigo, o design com o Guilherme e a parte de vídeo com a Tati.

[45:38]

Daí tudo isso virou uma playlist. Foi bem mais complicado, mas deu tudo certo também.

[45:46]
Pelo o que eu estou entendendo. o sucesso de ter dado tudo certo e da revisão não ter sido um trabalho tão pesado, se deu muito em conta da qualidade da própria base de dados, né? Ah, sim. Se não fosse por uma base boa, a gente nem ia fazer.

[46:08]
Pensa que podem ter vários problemas. Pode ter o problema da transmissão dos dados, de não ser confiável por algum motivo em termos de recursos da manutenção.

[46:26]
Na hora de transmitir, pode sei lá. O TSE tem um sistema bem robusto, então eles vão mandando os dados e os dados vão chegando certinho, não cai a conexão e não chega pela metade. Esse tipo de coisa acontece sem erro. Então é bem confiável. Isso é uma coisa, a outra coisa é a base em si. A base do TSE é bem limpinha. Não tem coisas fora de formatação, entendeu? Não tem erros ali dentro que você não identifica na hora. Então ela é uma base bem padronizada, bem formatada. Ela é bem limpa e você pode trabalhar direto com ela.

[47:17]
Nesse ponto a qualidade da base é fundamental. Se você tiver uma base suja, você teria que fazer um outro processo. É que assim, no caso da base do TSE você já faz um processo de formatação dos dados antes. A estruturação dos dados também acontece antes. A parte da tecnologia deve explicar melhor, mas eu imagino que eles puxam os dados para um data lake, tem um processo ali de extrair os dados que a gente quer e formatar do jeito que a gente quer.

[47:58]
Para depois poder jogar para o algoritmo. Então eu imagino que seja isso. Você já ter o dado limpinho ali, você não precisa daquele processo de ter um processo de limpeza enorme entende? Deixar os nomes iguais, ver se não tem um NaN ali diferente, ou campo vazio, que não era para ter. Então com certeza ajuda muito você ter uma base limpa.Você minimiza muito o erro, né?

[48:27]
Pelo que você acabou de fala, os dois pilares de certa forma da da qualidade base de dados é tanta a questão da transmissão, quanto da padronização e da formatação?

[48:44]
Sim.

[48:53]

O mais difícil de trabalhar com bases de dados é a limpeza. Então assim, primeiro a base tem que ser confiável, a origem dos dados tem que ser confiável, você pode inventar uma base que seja toda bonitinha, mas que os dados não sejam verdadeiros, isso é uma questão. Mas nesse caso a gente parte uma base que é confiável. Partindo de uma base confiável, uma coisa que dá muito trabalho e aumenta muito a chance de erro é ter uma base suja. Então não está bem estruturada, ou ou ela foi mal formatada. Por exemplo, uma base suja é a base das vacinas. O Data Sus. É uma base bem suja por que é print da mão, feita no papel. Dai depois alguém vai lá e digita aquela base ali no computador.

[49:50]
Tem coisas absurdas como uma pessoa que tomou 10.000 doses de vacina. Porque é um erro de digitação, erros do trabalho manual ali na hora de colocar. Então assim, você tem que analisar em detalhes a base para ver o que pode estar errado, o que que pode estar pegar os outliers e o que que viveu? Da onde veio o que que aconteceu e tal até você criar um scriptzinho para limpar aquilo ali tem uma base que dê para usar então no caso das vacinas faz isso.

[50:28]
Tem coisas que você precisa apurar ligar lá para entender, tipo, eu lembro nessa de vacina. Acho que tem coisas assim em São Paulo, cara que tinha um cara que tava tudo no cpf dele por causa de um erro lá de uma base a gente ligava lá aí. Liga lá pra secretaria de saúde, ó isso aqui nessa base então isso aí vamos ver o que que é. Ah, é isso aconteceu isso aí ficava lá com você aí revisa os dados aí vai lá e faz vê outra coisa tá errada.

[51:02]
Por muito que você disse eu entendi. Que que vocês conseguiram automatizar essa cobertura justamente pela qualidade da base, né? Existe hoje no Brasil alguma outra base que vê contas assim o que vocês prospectam que poderia ser usado como fonte por uma competição, mas tem várias vozes muito boas Cara eu acho que tem tem bastante.

[51:28]
Assim transparência melhorou muito né? Assim 21, sei lá.

[51:34]
15 anos aí desde a lei de acesso à informação principalmente, né?

[51:40]
A gente tem base muito boas no próprio governo federal de hoje, nós também nos Estados, né assim e vocês pensam em automatizar algum outro tipo de cobertura, já que tem.

[51:54]

Tem alguns desafios, né? A gente pensa sim, mas é que tem um desafio. Quais são os principais desafios assim não são a base não só a base cebola, né confiável e tudo mais mas assim.

[52:09]
Ela tem que ser.

[52:13]
Vou falar só faz sentido a gente a gente automatizar. Se for para gerar uma grande quantidade de texto. Então assim eu tenho que ter uma base e tem uma desagregação ali que faça sentido eu fazer x meu texto, entendeu? Pode ser pode ser sei lá alguma coisa de município, pode ser? Não sei por exemplo, quais os casos hoje no mundo. Que que tenha automatização e que já tá bem popularizado e tal. Não sei o quê resultado de empresa você pega sei lá nos Estados Unidos 55.000 empresas lá, não tem como humanamente você fazer um texto para cada uma um veículo fazer então você tem que ser tem que ser automatizar.

[53:07]
Automatização começou com a little League americana, né? Primeiro fez aí, né? Foi na liga de beisebol infantil lá que tinha não sei quantos mil jogos e que os pais eh, colocavam os resultados só os resultados, né? Não sai e a partir desse site o se criou o primeiro.

[53:37]
Primeiro projeto aí, né de você gerar fazer matérias a partir desses dados do site que aí os pais das crianças iam ler. Então assim. Então você já tinha uma base que era o todos os pais colocando os os resultados de milhares de jogos lá dos filhos.

[53:59]
Então você tinha esses dados você podia pegar ela e gerar milhares de matérias uma para cada time de cada cada criança lá que tava jogando.

[54:10]
Então assim você tem que ter uma base que sirva de fonte para milhares de matérias e até agora isso tem alguma periodicidade não a gente tá analisando tá vendo? Ainda não temos nada e ficou limitado a política, né?

[54:30]
Tinha para fechar assim, eh que muito quando se fala de automatização implementação de algoritmo. Sempre tem aquela discussão ali do fantasma do desemprego estrutural ou de que possa de trabalho vão desaparecer ou que eles vão mudar de lugar para você que participou assim, qual foi a impressão a impressão que foi de que você trabalhou?

[54:59]
Colaboração com algoritmo. É um tipo de aplicação que não nos substitui. O trabalho que a gente fez usando o algoritmo nesse caso foi para fazer o que não dá para fazer no braço,

entendeu? A gente não tá pegando o trabalho que alguém que faz e a pessoa deixou de fazer esse trabalho. A gente está pegando uma coisa que não existe, que não dá para fazer, que não tem gente suficiente, não tem tempo suficiente, não tem recurso e a gente está usando um algoritmo fazer isso. Então o algoritmo, nesse caso, é o potencializador do nosso trabalho.

[55:40]
A gente está falando para nichos que não valeria a pena a gente fazer reportagem. Até vale a pena, mas a gente não tem recurso. Pode valer a pena em termos de relevância, dizer quem é cada Prefeito de cada cidade é super relevante, só que a gente não tem braço, não tem recurso para fazer isso, porque são 5.568 municípios então se não fosse pelo algoritmo simplesmente você não tem como fazer. Aí o que a gente faz? Antes de ter essa solução a gente tinha os dados de todas as cidades ali. Então o cara podia abrir um mapa com todos os municípios e botar o nome do município dele para ver qual foi o resultado. Isso é um recurso que a gente já tinha.

[56:29]
Só que além disso. As pessoas gostam de ler e elas querem ter um texto com o resultado daquilo que elas buscam. Mas não dá para atender isso né? Aí a gente conseguiu com ajuda do algoritmo fazer um volume que era inviável com com os recursos que a gente tinha. Então eu vejo o robô jornalismo assim. Vejo pelo lado positivo dele e acho que pode ter um lado negativo, mas eu não vi até hoje como uma substituição de mão de obra. Eu vejo como potencializador do trabalho jornalístico e que a gente escreveu para nichos que ninguém escrevia.

[57:19]
Com isso a gente pode mostrar coisas que ninguém tava vendo também.

[57:30]
Qual foi o nicho nesse caso da cobertura da eleição municipal?

[57:41]
O nicho foi o município, por que a gente estava escrevendo para municípios com 800 pessoas. Borá no interior de São Paulo tem 800 pessoas e a gente escreveu uma matéria só para ele. Falando quem era o prefeito. Quem era os vereadores e tudo mais.

[57:50]

[58:06]
Esses públicos menores são contemplados só nas notícias que afetam o Estado. Mas em geral é muito difícil que veículos nacionais façam conteúdos específicos para pequenos grupos de pessoas, até porque não tem como.

[58:50]
Então eu vejo o jornalismo automatizado como o potencializado e a gente conseguiu alcançar mais gente. Futuramente, eu acho que você pode ter um nível de personalização mais alto.

[59:02]
Porque você pode usar o perfil das pessoas como um input para os dados, na seleção dos dados dentro da árvore de decisões. E aí você pode você pode produzir reportagens que são mais personalizadas. Você diz, por exemplo, cobertura de jogo de futebol. O cara que torce para um time recebe uma notícia e o cara que torce para outro, recebe outra? No caso de futebol isso dá para fazer na mão. Por que as series do Brasileirão tem uns 20 times.

[59:37]
Você bota uma tag ali para cada time e vai o conteúdo de cada time para o outro. São produzidos uma matéria ou mais para cada um dos 20 times. Se for no Globo Esporte, vai ter várias matérias de cada um dos times da série B.

[01:00:00]
Isso é feito no braço, é feito normalmente. Estou falando que daria para fazer do time do bairro, do time de futebol da escola, entendeu? Você falou aí que você não conseguiu identificar nenhum ponto negativo da relação com o trabalho, mas a única exigência assim é o projeto teve foi esse preparo prévio bem extenso, né?

[01:00:30]
Ah, bom. Sim, principalmente por ser um um projeto inédito.

[01:00:37]
Não é uma coisa simples de ser feita, pelo menos hoje, mas eu acho que no mercado você já tem soluções, fora do Brasil principalmente que são mais de prateleira.

Entrevista decupada com Rafael Muniz

[00:00:00] - Prontinho.
[00:00:02] - Vamos lá, Rafael.
[00:00:03] - Me conta como é que esse projeto apareceu para você na primeira vez?

[00:00:08] - Foi em 2020, foi em 2019. Poxa, na verdade foi em 19. Foi o ano que eu entrei
[00:00:24] - Aí foi o ano que eu entrei na Globo. Covid começou quando?
[00:00:29] - Foi só no início de 2020.
[00:00:32] - No início, tá. Então tá, beleza. Foi em 2020 mesmo. Quando eu entrei foi o ano da Covid. Era a minha referência. Então sim.
[00:00:43] - Qual foi o primeiro contato que eu tive?

[00:00:46] - Quando eu entrei na Globo o projeto já existia em fase de planejamento, estava até na mão do Hector ali tocando. Então fui apresentado para ele quando eu entrei na equipe para atender a galera de jornalismo. Falaram ali quais os projetos que estavam no andamento e um deles era o de eleição. É basicamente foi isso o primeiro contato…

[00:01:10] - E você entrou para qual cargo?

[00:01:13] - Então, eu entrei como engenheiro de dados. Aí e a gente tinha uma… a gente tem uma divisão ainda por Squad. Aí era a Squad para atender a galera de jornalismo.

[00:01:25] - Era o mesmo departamento do Hector?

[00:01:27] - Isso, isso. O Hector era cientista de dados e eu o engenheiro de dados.

[00:01:33] - E eles se referem ao departamento como COI? Era o departamento de jornalismo de dados? O que significa COI?

[00:01:40] - Boa pergunta.
[00:01:47] - Caraca, eu esqueci.
[00:01:49] - O "Center of Excellence"
[00:01:51] - ou "Certer of Expertise", algumas dessas coisas, Centro de Excelência, centro de Expertise em um determinado assunto. No caso de análise de dados.
[00:02:02] - Então era um centro de Expertise, ou de Excelência, em análise de dados.

[00:02:16] - Beleza.
[00:02:18] - Pode continuar.

[00:02:20] - Entrei nesse centro em 2020 como Engenheiro de Dados, no Squad do Hector
[00:02:26] - Isso.
[00:02:27] - Então a área que eu entrei foi o COI.
[00:02:30] - Aí no COI tem vários Squads. As Squads eram os times menores, geralmente de 4 pessoas. Aí que eu entrei, tinham 4 pessoas e eu fui o quinto quando entrei.
[00:02:45] - Então a gente trabalhou ali em 2020 para com 5 pessoas.
[00:02:50] - Legal.

[00:02:51] - O contato que você tinha com o projeto era por meio do Hector, ou era direto com a equipe do Thiago e do Grandin?

[00:03:00] - Então, aí isso seria dividir em duas partes o projeto. A primeira parte seria do planejamento. Aí acho que o contato era só através do Hector. O Hector estava vendo, era quem estava tocando o planejamento. E depois já em 2020, vamos dizer assim, um desenvolvimento em si comigo e com o Hector

[00:04:34] - E o Thiago Reis?

[00:04:36] - E com o Tiago Reis, exato.
[00:04:38] - O contato com ele no primeiro momento foi de definir como é que ia ficar o texto, então ele ia mandando opções, que era mais ou menos o template, a ideia da solução.

[00:04:51] - Como é que foi essas primeiras trocas?

[00:04:54] - É, porque primeiro a gente definiu qual seria a solução, né? Seria essa de criar os 5.500 textos automatizados. Aí beleza. Aí a gente foi caçar e ver se era possível, né? Se a gente tinha todos os dados disponíveis no TSE. A gente conversou, tinha um time dentro da… um time dentro da Globo que era responsável para fazer essa ligação com o TSE e distribuir os arquivos dentro da Globo então a gente conversou com esse time para entender quais dados a gente poderia receber. A gente voltava, eu não vou lembrar, se era exatamente com o Grandin, ou com Thiago Reis, mas com o Jornalismo, que eram os dois que estavam envolvidos. A gente falava que deu aqui, a gente tem esses dados e dá para montar o texto
[00:05:40] - Aí começou o Thiago Reis com esse trabalho de elaborar um template onde a gente iria colocar a cara de informação ali, e como tinha questão de prefeitos sendo eleitos no primeiro, mas com alguns indo com dois candidatos para o segundo, cidades que não tinham segundo turno, cidades que tinham, então a gente trabalhou com 4 textos, 4 formatos de textos.
[00:07:11] - Isso, entendi. O Hector comentou que esses dados vinham de dois lugares. Era os dados que eram de boletim de urna na hora e os que já estavam numa base do TSE, procede?
[00:07:25] - Ah sim, é, mais ou menos, só não era boletim de urna o boletim de urna é um dado um pouco diferente seria o dado da apuração do TSE divulga. Então esse dado ali a gente coletava a quantidade de voto por porcentagem de votos brancos, votos nulos e tinha alguns dados da base do TSE que é quando candidato envia os dados dele, profissão, renda, cor…
[00:07:50] - Então tinham alguns a mais, mas a gente não colocou no texto não. Mas a profissão a gente botou, idade acho que botou. Então assim para montar o texto "o fulano foi eleito na cidade do Rio de Janeiro, ele é, engenheiro, com 42 anos. Então esses dados o candidato que informa o TSE, aí o TSE valida ali a candidatura dele e fica apto a concorrer.

[00:08:22] - Mas quem indicou quais dados pegar e como ia ficar o template?

[00:08:26] Foi o Thiago, ou o Felipe. Bom, eu não vou lembrar exatamente quem foi, acho que os dois estavam trabalhando no texto, aí teria que pegar essas conversas lá de 2020. Mas era da equipe deles e assim assim, eles montavam o texto, então assim se não me engano, teve até coisa que a gente cortou, que a gente falou isso não tem necessidade.
[00:08:50] - Então assim eles montavam o template, esse template ele já vinha com a indicação de, por exemplo, da onde esses dados estariam, algum dado que eles, colocassem no template, a gente não tivesse do TSE e não fosse receber durante a apuração
[00:09:09] - A gente falou, oh esse dado aqui não tem como entrar, porque a gente não vai ter ele no tempo de gerar o texto, mas se eu não me engano foi só um campo.
[00:09:18] - Eu não vou lembrar qual é, mas um que a gente conversou e que resolveu tirar.

[00:09:23] - Me fala um pouco mais desse processo aí de fazer o pipeline?

[00:09:51] - Sobre o pipeline, o Data Lake a gente mantém no Big Querry. Então talvez eles chamem de Data Lake e talvez eu fale aqui Big Querry, mas seria basicamente a mesma coisa. Então, esses dados referentes aos candidatos o TSE divulga previamente e vai atualizando. Então a gente tinha um fluxo ali de que era um script que coletava esses dados, se eu não me engano, eram dois CSVs do TSE, um era referente a Bens de Candidatos e outro aos candidatos porque um dos campos era o patrimônio do candidato. Então a gente tinha que somar a declaração de Bens deles nesse script, ele recolhia e atualizava.
[00:11:03] - Agente fez isso usando o Python na Cloud Functions. Então, o do TSE, se eu não me engano, a gente botou pra rodar uma vez por semana e a gente ficou com a tarefa de fazer uma execução extra na vez para a deleção, então, a gente tinha ali uma atualização semanal que, basicamente para testar, para a gente pegar um dado mais atualizado e a gente fez uma execução manual e fez um disparo manual nas vésperas.
[00:11:28] - Quando estávamos prestes a iniciar a apuração por volta das 4 horas e 30 minutos da tarde de domingo, realizamos uma execução forçada para garantir que capturaríamos a última atualização quando as urnas estivessem prestes a serem encerradas. Nesse momento, você estava realizando o scraping - ou seja, coletando dados - no site do TSE. No entanto, não era tão complicado, pois a página do TSE possuía um link de download de um arquivo CSV fixo, que nos permitia acessar diretamente o arquivo desejado. Era necessário apenas buscar na página onde o arquivo estava localizado, mas era relativamente simples. Basicamente, utilizamos Cloud Functions para coletar os dados disponibilizados previamente pelo TSE e também durante a eleição, capturando os dados que eram enviados pelo próprio TSE. Você mencionou que eram dois arquivos CSV, o que significa que tínhamos duas bases de dados para trabalhar.
[00:12:31] - Então, tínhamos duas bases de dados - uma para os candidatos e outra para o patrimônio. O trabalho de organização foi feito no Big Query. Foi uma decisão tomada pela nossa equipe, onde criamos algumas tabelas e uma visualização que capturava todos os campos que precisávamos, tanto dos dados da apuração que recebíamos quanto dos dados do TSE. Houve algum trabalho no momento de extrair esses dados e formatá-los para carregá-los no Big Query. Por exemplo, alguns campos poderiam estar em letras maiúsculas ou minúsculas, então nos scripts tínhamos que escolher como tratar e carregar esses dados, como um processo meio automatizado com um tempo de vida útil definido.

[00:13:17] TTL?

[00:13:19] Track, Transform and Load
[00:13:22] Então, tínhamos alguns campos nos quais precisávamos realizar transformações antes de carregá-los. Por exemplo, se tivéssemos um campo com valor decimal, tínhamos que substituir o ponto por vírgula. Fazíamos esse tratamento antes de carregar os dados. Essa formatação já era uma etapa para guardar os dados de forma estruturada, de modo a utilizá-los posteriormente e inseri-los no template. Não necessariamente era a maneira como os dados seriam apresentados no texto final. Se houvesse algum tratamento específico, como

mencionei antes, no CSV, onde tudo era tratado como string, realizávamos o tratamento para armazenar esse campo como numérico. No texto final, faríamos a formatação adequada, como adicionar pontos a cada três dígitos e uma vírgula para separar os decimais. Isso era tratado no código, seja na visualização (SQL) quando era mais conveniente ou no próprio código Python. Após essa etapa, o próximo passo seria estruturar os dados no BigQuery, utilizando as Cloud Functions para extrair os dados do TSE. O pipeline basicamente fazia a coleta desses dados do TSE com antecedência, antes do momento que as urnas estavam fechando, né?
[00:15:13] Então, durante a apuração, não mexíamos muito com esses dados, pois os dados da apuração eram recebidos conforme o TSE ia atualizando. O fluxo que observávamos era o seguinte: tínhamos nosso pipeline que recebia os dados e, ao final dele, verificávamos se a cidade havia alcançado 100% dos votos ou se havia alguma definição de candidato. Quando isso ocorria, iniciávamos os scripts que geravam o texto. Basicamente, era isso. Os scripts geravam o texto, provavelmente em lotes. Quando o TSE liberava uma atualização, ela não era apenas para uma única cidade, mas sim para várias. Por exemplo, a atualização poderia informar que a eleição do Brasil estava em 60% e agora estava em 65%, o que representava uma diferença de 5%. Essa atualização poderia levar dois minutos para ser concluída e poderia afetar a atualização de 3 mil cidades, por exemplo. Portanto, cada arquivo que recebíamos continha uma quantidade pequena de dados no início, aumentava durante o pico da eleição e diminuía no final, à medida que os votos eram contabilizados.
[00:16:23] Então, sobre o script que montava o texto, como conversamos com o Hector e o Tiago, eles escolheram um modelo de template que permitia ter certo controle sobre como o texto final seria gerado. Isso proporcionava uma padronização para todas as cidades, tornando o produto final confiável e previsível.
[00:17:25] Uma parte interessante era a escolha feita no template em relação à conjugação verbal e ao uso de sinônimos para evitar repetições. O desafio nesse caso era desenvolver essa lógica no script. Embora eu não tenha participado diretamente do desenvolvimento desse script, posso compartilhar o que lembro sobre os templates.
[00:17:50] Basicamente, havia quatro templates devido à questão do segundo turno. Havia cidades que tinham segundo turno, cidades que não tinham e candidatos que eram eleitos diretamente no primeiro turno. Por causa disso, a divisão resultou em mais um template. Os jornalistas definiam os templates, e o script se comportava de maneira semelhante, selecionando um dos templates com base nos dados disponíveis para compor o texto.
[00:18:03] As substituições no texto funcionavam para qualquer template, ou seja, as substituições eram feitas de acordo com os dados disponíveis, independentemente do template escolhido. Também havia considerações sobre o uso de artigos e a forma de mencionar a profissão. Por exemplo, se a base de dados indicasse "professora", mas fosse necessário escrever "professora" no texto quando se tratava de uma mulher, o script lidava com essa substituição.
[00:18:18] Sobre as decisões no script em relação às profissões com formas masculinas e femininas, o script tomava uma série de decisões nesse sentido. No entanto, para entrar nos detalhes específicos do script, a resposta do Hector seria mais adequada, já que ele foi o responsável por construí-lo.

[00:18:58] No momento, estou verificando o código, mas não acredito que conseguirei lembrar dos detalhes específicos. No entanto, acredito que havia uma combinação de abordagens, com uso de estruturas condicionais (if) em casos mais simples e o uso de bibliotecas quando a lógica era mais complexa. Também acredito que havia uso de técnicas de processamento de linguagem natural em algumas situações.

[00:19:12] Em relação ao pipeline, você mencionou algumas bibliotecas consagradas para lidar com projetos de geração de linguagem natural, como NLTK, SpaCy e Flux ?

[00:20:01]Quanto a um módulo ou repositório específico que tenha sido fundamental nesse trabalho, a resposta provavelmente seria melhor fornecida pelo Hector.
O que consigo lembrar é que dentro da Globo já havia outras iniciativas de geração automatizada de textos, então houve uma troca significativa de conhecimentos entre as equipes. No entanto, em relação a bibliotecas específicas utilizadas, isso ficou mais sob a responsabilidade do Hector, que cuidou desse script. Parece que esse aspecto estava mais ligado à área de cientista de dados.

[00:20:07] beleza, me fala um pouco mais sobre como foi a dinâmica entre você e o Hector dentro da equipe.

[00:20:13] Ah, beleza, aí foi bem tranquilo. Na verdade, o Hector tinha entrado na Globo alguns meses antes de mim, então acabou tendo uma aproximação natural. Éramos duas pessoas novas nos times. Pelo que entendi, quando entrei alguns meses depois, ele já estava encarregado do projeto de eleições desde o início e cuidou do planejamento inicial para entender o que deveria ser entregue. Quando entrei, já estava definido que seriam textos, um por cidade, mas ainda não estava claro como seria a divisão do trabalho. Não era necessário um engenheiro atuando inicialmente, pois já havia dois na equipe. Quando as eleições começaram, eu acabei assumindo o desenvolvimento, enquanto outro engenheiro estava envolvido em outra atividade. A troca com o Hector sempre foi muito boa, nos falávamos diariamente, já que éramos da mesma equipe, mesmo estando em cidades diferentes. Se precisássemos de apoio um do outro, marcávamos um papo, foi bem tranquilo durante todo o projeto.

A distinção de papéis era clara: Hector atuava como cientista de dados, e eu como engenheiro. Basicamente, eu ficava responsável pela coleta e carregamento dos dados de forma que o cientista pudesse utilizá-los. O cientista, por sua vez, era responsável pela montagem do texto no projeto em si. Se faltasse algum dado no texto, ele me avisava para que pudéssemos incluí-lo. A engenharia consistia em garantir que os dados estivessem disponíveis o mais rápido possível, enquanto a qualidade do texto ficava sob responsabilidade do cientista, que se certificava de que as substituições corretas fossem feitas para que fizessem sentido.

No que diz respeito à arquitetura, utilizamos principalmente o BigQuery, que é uma ferramenta do Google. Também usamos o Cloud Functions e o Pub/Sub. O Pub/Sub é um

serviço de mensageria que dispara eventos. Ele explicou que o disparo de eventos era o Cloud Functions
[00:23:35] Ah, beleza, aí foi bem tranquilo. Na verdade, o Hector tinha entrado na Globo alguns meses antes de mim, então acabou tendo uma aproximação natural. Éramos duas pessoas novas nos times, e pelo que entendi, quando entrei alguns meses depois, o Hector já estava tocando esse projeto de eleições desde o início. Ele cuidou do planejamento inicial para entender o que deveria ser entregue. Quando entrei, já estava definido que seriam textos, um por cidade. No início, não estava claro se precisaria de um engenheiro atuando, pois já havia dois engenheiros na equipe. A divisão do trabalho não tinha sido definida ainda.

Quando as eleições de fato começaram, eu comecei o desenvolvimento, pois outro engenheiro estava envolvido em outra atividade. Mas a troca com o Hector sempre foi muito boa, foi bem tranquilo. A gente se falava diariamente e marcava um papo caso precisasse de apoio um do outro. Durante o projeto todo, não tivemos problemas.

Quanto aos papéis, o Hector atuava como cientista de dados, enquanto eu atuava como engenheiro. Basicamente, minha responsabilidade era coletar e carregar os dados para que o cientista pudesse usá-los na montagem do texto. Se faltasse algum dado, ele me avisava para que eu pudesse disponibilizá-lo. A engenharia consistia em garantir que os dados estivessem disponíveis o mais rápido possível, para que o texto refletisse as informações atualizadas sobre as eleições.

Em termos de infraestrutura, utilizamos o BigQuery, Cloud Functions e Pub/Sub. O Cloud Functions era responsável por guardar eventos, enquanto o Pub/Sub era usado para enviar esses eventos. O desafio era disponibilizar o dado quase em tempo real, garantindo sua qualidade e fazendo as transformações necessárias para refletir o trabalho realizado pelo cientista.

O principal desafio para mim foi conseguir disponibilizar o dado o mais rápido possível, garantindo sua qualidade. Tínhamos a responsabilidade de produzir matérias confiáveis para a Globo, e qualquer erro poderia ter repercussões negativas. O projeto demandou um ano de trabalho intenso para que pudesse rodar em apenas um dia. Era crítico garantir a qualidade dos dados antes de disparar os textos, evitando informações erradas ou prematuras. Afinal, era um projeto único que não permitia ajustes contínuos após a publicação.

[00:29:19] E os testes?

[00:29:22] Sim, realizamos testes para garantir a qualidade e o bom funcionamento do sistema. Para isso, utilizamos dados fictícios fornecidos pelo próprio TSE em simulados. O TSE disponibiliza um calendário de simulados em que as equipes interessadas se candidatam e recebem arquivos com esses dados fictícios.
Com esses dados fictícios, conseguíamos testar o fluxo completo, desde o envio dos dados pelo TSE até a geração dos textos. Durante os testes, manipulávamos esses arquivos para

tentar refletir a realidade do dia da eleição o máximo possível. Em alguns casos, se um dado fictício não atendesse ao texto esperado, fazíamos pequenas alterações para adequá-lo. Após os simulados, continuávamos realizando testes internos utilizando esses dados manipulados. Dessa forma, buscávamos testar diferentes cenários e identificar possíveis problemas ou ajustes necessários no sistema. Os testes eram realizados em várias etapas para garantir a qualidade dos dados e a correta geração dos textos. Espero ter esclarecido suas dúvidas.

[00:31:11] - Sim, esse projeto demandou cerca de um ano de preparo, desde o planejamento inicial até a viabilização da solução de geração de texto. Durante esse período, foram realizadas diversas etapas, como o desenvolvimento da arquitetura do sistema, a implementação das funcionalidades necessárias, os testes com dados fictícios do TSE, entre outras atividades.

A experiência adquirida ao longo desse primeiro ano foi muito valiosa para replicar o projeto em anos subsequentes. Com base nos aprendizados e nas melhorias identificadas, foi possível otimizar o processo e aprimorar a qualidade do sistema. Essa continuidade e aperfeiçoamento são fundamentais, uma vez que o projeto é executado apenas durante um dia específico, e é crucial garantir que tudo funcione corretamente nesse período.

Assim, o primeiro ano foi essencial para estabelecer as bases do projeto e construir uma solução sólida. Nos anos seguintes, a equipe pôde se concentrar em aprimoramentos e ajustes, aproveitando a experiência anterior para obter um melhor desempenho e garantir a confiabilidade dos resultados.

[00:31:39] - Aí seria falar um pouco de 22. Mas basicamente foi evoluir a solução de 2020.

[00:31:42] Entendi, tá, maravilha. E qual que você acha que foi? Você falou do desafio né, mas não falou, por exemplo, qual que você acha que foi a conquista do projeto, qual foi o principal aprendizado por você assim, e a relevância dele na sua visão?

[00:31:58] - Então, acho que o bacana assim, acho que logo que começou as cidades terem eleição finalizada em publicar testes, já começou a ter repercussão assim legal nas redes sociais. A gente achava textos no Twitter, comentando ali de outros jornalistas, questionamento ali sobre é, matérias automatizadas e tudo mais. Então, assim, a repercussão disso foi muito bacana assim, você vê seu trabalho sendo comentado por outros profissionais sim. Então, acho que assim, a repercussão que teve de imediato assim foi muito legal.
[00:32:18] - E seu trabalho assim que a gente entregava o dado para o que nunca foi feito assim, de o cara em na menor cidade do Brasil, ele tinha um texto contando ali a história de como foi a eleição na cidade dele, assim, praticamente em tempo real, né? Então, esse cidadão ali conseguiu ter uma matéria personalizada para a cidade dele contando a história. Isso aí também foi bem bacana assim, algo que nunca tinha sido feito e foi publicado em 2022.

[00:32:34] - Legal. Aproveitando que você acabou falando um pouco de 2022 assim, mas qual que você acha que foi? O Philip também falou um pouco, né, que ele assumiu o projeto em 2022. Qual que você acha que foi o grande avanço assim vendo o projeto que aconteceu pela primeira vez e uma segunda vez?

[00:32:48] - Então, o principal foi que em 2022, além do vídeo, além do texto, a gente incluiu o vídeo, né? Então, aí foi um desafio maior ainda. Aí, acho que foi a grande evolução. Fez algumas melhorias na arquitetura, assim, coisas que a gente achou que poderiam ser melhoradas. A gente também acabou mudando assim, então é. Eu disse mais, eu destacaria o vídeo, acho que o vídeo foi a principal evolução. Legal.

[00:33:04] - Podemos ficar por aqui, Rafael, se você quiser comentar mais alguma coisa assim, mas eu acho que deu para entender legal como é que foi o seu trabalho nesse projeto. Um bobo, acho que você está satisfeito, eu também estou. E qualquer coisa e se depois se lembrar de alguma coisa ficou de posição, estou fácil de achar.

[00:33:21] - É, não é interesse mais nesses detalhes técnicos assim. Entendo que o script quem teve mais envolvimento foi o ator mesmo, né? Então, esse de montar o texto assim.
[00:33:31] - Sim, mas eu acredito que na parte do pipeline você explicou bem, né, que foi o pubsub, cloud functions, big query. É, essas duas bases do TSE que vocês estavam constantemente atualizando e essa questão do script rodando automático para vocês não terem nenhum dado desatualizado.

[00:33:45] - E imagina também que tinha alguma preocupação para não haver duplicações de arquivo também. É, então, nesse caso, era tão difícil assim porque como o TSE coloca a base completa, basicamente tinha que olhar sempre para o último arquivo deles assim.

[00:33:59] - Então, não era algo que a gente precisava do histórico, que a gente olhava sempre para a última posição, sem ser eventualmente o candidato fizesse a correção no nosso texto, no contado da história. O candidato cadastrou que era professor e agora ele colocou aqui como funcionário público, assim, então a gente só publicava do texto a última informação que o candidato deu, né?
[00:34:17] - Então, essa questão de duplicidade de arquivo assim não precisamos nos preocupar tanto, bastava olhar sempre para o último arquivo que o TSE disponibilizava. Esses arquivos são públicos, né, do TSE. Então, depois eu gostaria de olhar lá no site. Não, sim, eu estou inclusive olhando.

[00:34:32] - Beleza, podemos ficar por aqui, obrigado Rafael. Obrigado Igor por acompanhar a gente. Se eu tiver qualquer dúvida aí, se eu tiver qualquer dúvida aí, eu entro em contato.

ANEXO III

Código em Python para raspagem de notícias e análise de similitude

```
#CRIAR URL BASE DE REFERÊNCIA + O DIRETÓRIO (PASTA)

BASE_URL =
"https://g1.globo.com/ba/bahia/eleicoes/2020/noticia/2020/11/16/veja-tod
os-os-prefeitos-e-prefeitas-eleitos-as-nos-417-municipios-da-bahia.ghtml
"
DATA_DIR = "BA_Notícias"

#FAZER UMA PRIMEIRA REQUISIÃ‡ÃƒO COM A URL, QUE IRÁ RETORNAR UM HTML

response = requests.get(BASE_URL)
```

```
#INICIAR A INSTÂNCIA BEAUTIFULSOUP, A FIM DE RASPAR O QUE ESTÁ NA BASE
URL

bs = BeautifulSoup(response.text, 'html.parser')


links_noticias = bs.find_all(class_="content-unordered-list")


a_tags = links_noticias.find_all('a', href=True)  # 👉 encontrar todas
tags <a> que tem o atributo a href

#👇 fazer com que o print se transforme em uma lista

noticias_ba = []

# 👇 fazer um loop nos resultados

links_noticias = [noticias_ba.append(tag['href']) for tag in a_tags]

#transformar a lista em data frame para baixar arquivo csv dos links

df = pd.DataFrame(noticias_ba)

# salvar o data frame

df.to_csv('noticias_ba.csv')
```

```
print(len(noticias_ba))
```

```
results_name = []

for i in a_tags:
    results_name.append(i.text)

print(len(results_name))
```

```
results_title = []
results_article = []
results_soutien = []
erros_conexao = []
attr_error = []

for url in noticias_ba:
    # 1. Obter a resposta da url:
    response = requests.get(url)
    # 2. Criar o objeto beautiful soup:
    bs = BeautifulSoup(response.content, 'html.parser')
    # 3. Extrair os elementos para cada url:
    try:
       results_title.append(bs.find("title").get_text().strip())
       results_article.append(bs.find("article").get_text().strip())
       results_soutien.append(bs.find(class_="medium-centered subtitle",
attrs={"class": "meta content"}).get_text().strip())
    except AttributeError as err:
        attr_error.append(err)
        erros_conexao.append(url)
        break
```

```
#CHECAR O TAMANhO DE CADA LISTA

print(len(results_name))
print(len(results_title))
print(len(results_soutien))
print(len(results_article))
```

```
dict_noticias_ba = {}

for i in range(len(results_name)):
    dict_noticias_ba[results_name[i]] = (results_title[i],
results_soutien[i], results_article[i])
```

```
print(len(dict_noticias_ba))
```

```
#SALVAR O DICIONÁRIO EM UM ARQUIVO JSON

with open('Bahia_Notícias.json', 'w', encoding='utf-8') as f:
    json.dump(dict_noticias_ba, f, ensure_ascii=False, indent=4)
```

```
#SALVAR OS ERROS EM UM ARQUIVO CVS

df_erro = pd.DataFrame(erros_conexao + attr_error)

df_erro.to_csv('erros_conexão.csv')

#REPETIR PARA TODOS OS ESTADOS E/OU PORTAIS DE NOTÍCIA QUE VOCÊ QUEIRA
RASPAR
```

```
#UNIFICAR TODOS AS NOTICIAS RASPADAS EM UM ÚNICO ARQUIVO
def merge_dicts(dicts):
    result = {}
    for d in dicts:
        for key, value in d.items():
            if key in result and isinstance(result[key], dict) and
isinstance(value, dict):
                result[key] = merge_dicts([result[key], value])
            else:
                result[key] = value
    return result

def merge_json_files(directory):
    merged_dict = {}
    for filename in os.listdir(directory):
        if filename.endswith('.json'):
            file_path = os.path.join(directory, filename)
            with open(file_path, 'r', encoding='utf-8') as file:
                try:
                    json_data = json.load(file)
                    merged_dict = merge_dicts([merged_dict, json_data])
                except UnicodeDecodeError:
                    print(f"Error reading file: {file_path}")
    return merged_dict

def save_merged_json(data, output_file):
    with open(output_file, 'w', encoding='utf-8') as file:
```

```
        json.dump(data, file, ensure_ascii=False)

# INDICAR O NOME DO ARQUIVO E O SEU DIRETÓRIO
input_directory = r'C:\Users\mathe\Documents\Beautiful Soup G1\_BRASIL'
output_file = r'C:\Users\mathe\Documents\Beautiful Soup
G1\_BRASIL\df_brasil.json'

merged_data = merge_json_files(input_directory)
save_merged_json(merged_data, output_file)
```

```
# Abrir o arquivo json
with open('df_brasil.json') as file:
    data = json.load(file)

# Definir os padrões de texto
patterns = [
    (r"(é eleito prefeito|é eleita prefeita)", r".* Esta reportagem foi
produzida de modo automático com o apoio de um sistema de inteligência
artificial"),
    (r"vão disputar", r".* Esta reportagem foi produzida de modo
automático com o apoio de um sistema de inteligência artificial"),
    (r"sub judice", r".* Esta reportagem foi produzida de modo
automático com o apoio de um sistema de inteligência artificial"),
    (r"está indeferida", r".* Esta reportagem foi produzida de modo
automático com o apoio de um sistema de inteligência artificial")
]

# Inicializar a contagem para cada padrão
counters = {pattern[0]: 0 for pattern in patterns}

# Iterar sobre as chaves do dicionário para identificar os padrões
for city_data in data.values():
    for item in city_data:
        for pattern, ending in patterns:
            if re.search(pattern, item['text']) and re.search(ending,
item['text']):
                counters[pattern] += 1

# Printar os resultados
for pattern, count in counters.items():
    print(f"Pattern: {pattern}")
    print(f"Occurrences: {count}")
    print("---")
```

ANEXO VI

As 2.966 notícias que compõem o corpus desta pesquisa foram salvas no Google Drive. Disponível em: <https://drive.google.com/drive/folders/1WICzjUwdfGBiOF9J4U3SlwME88pf0I9Z?usp=drive_link>. Acesso feito 25 de maio de 2023.