

**UNIVERSIDADE FEDERAL DE MINAS GERAIS  
FACULDADE DE LETRAS**

**PAULO VICTOR CANTALICE DE SOUZA**

**GERAÇÃO DE LÍNGUA NATURAL:  
DESENVOLVIMENTO DE UM ROBÔ JORNALISTA RELACIONADO  
AO MERCADO EDITORIAL BRASILEIRO**

**BELO HORIZONTE  
2022**

**PAULO VICTOR CANTALICE DE SOUZA**

**GERAÇÃO DE LÍNGUA NATURAL:  
DESENVOLVIMENTO DE UM ROBÔ JORNALISTA RELACIONADO  
AO MERCADO EDITORIAL BRASILEIRO**

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de bacharel em Letras, habilitação em Edição, pela Universidade Federal de Minas Gerais.

Orientador: Evandro Landulfo Teixeira Paradela Cunha

## RESUMO

Este trabalho pretende analisar algumas questões ligadas ao campo da Inteligência Artificial, mais especificamente o Processamento de Língua Natural (PLN). O PLN tem como objetivo ampliar as possibilidades das máquinas em sua manipulação da língua natural, seja criando ou compreendendo um *corpus* linguístico. Por meio da demonstração de uma ferramenta prática, produzida por nós, vamos discutir o estado da arte da tecnologia envolvida na criação de robôs na rede social *Twitter*. Esses robôs muitas vezes usam o PLN para conseguir transmitir mensagens de maneira autônoma e escalonável, com a possibilidade de lidar com muitos dados e processá-los de maneira que se possa fazer sua divulgação nos meios digitais — são os chamados robôs jornalistas. No Brasil, essa ferramenta é utilizada para a divulgação de informações para um público amplo, muitas vezes sem o controle de qualquer órgão fiscalizador, o que demonstra, portanto, a importância de compreender melhor o funcionamento e a aplicação desse mecanismo. Nosso robô jornalista tem um formato educativo e lida com a divulgação de dados relacionados ao mercado editorial brasileiro. E, para que possamos compreender melhor a atuação desse robô, vamos discutir sua metodologia e as aplicações possíveis nesse mercado editorial. Os dados são coletados de um órgão da própria Câmara Brasileira do Livro, entidade de maior relevância para o mercado editorial em nosso país.

**Palavras-chave:** Processamento de Língua Natural. Inteligência artificial. Robôs jornalistas.

## ABSTRACT

This work aims to analyze some issues related to the field of Artificial Intelligence, more specifically Natural Language Processing (NLP). NLP aims to expand the possibilities of machines in their manipulation of natural language, whether creating or understanding a linguistic corpus. Through the demonstration of a practical tool, created by us, we will discuss the state of the art technology involved in creating robots on the social network *Twitter*. These robots often use NLP to be able to transmit messages autonomously and scalably, with the possibility of handling a lot of data and processing them in a way that can be disseminated in digital media — they are the so-called robot-reporters. In Brazil, this tool is used to disseminate information to a wide audience, often without the control of any supervisory body, which therefore demonstrates the importance of better understanding the functioning and application of this mechanism. Our robot-reporter has an educational format and deals with the dissemination of data related to the Brazilian publishing market. So we can better understand the performance of this robot, we will discuss its methodology and possible applications in this publishing market. The data are collected from a supervisory body of the Brazilian Book Chamber itself, an entity of greater relevance for the publishing market in our country.

**Keywords:** Natural Language Processing. Artificial intelligence. Robot-reporters.

## LISTA DE ILUSTRAÇÕES

<b>Figura 1.</b> <i>Screenshot</i> do portal <i>PublishNews</i>	15
<b>Figura 2.</b> <i>Screenshot</i> do ambiente Google Colab durante a execução do <i>script</i>	18
<b>Figura 3.</b> <i>Screenshot</i> do ambiente Google Colab durante a execução do <i>script</i>	19
<b>Figura 4.</b> <i>Screenshot</i> do ambiente Google Colab durante a execução do <i>script</i>	20
<b>Figura 5.</b> <i>Screenshot</i> do ambiente Google Colab durante a execução do <i>script</i>	20
<b>Figura 6.</b> <i>Screenshot</i> do ambiente Google Colab durante a execução do <i>script</i>	21
<b>Figura 7.</b> <i>Screenshot</i> do ambiente Google Colab durante a execução do <i>script</i>	21
<b>Figura 8.</b> <i>Screenshot</i> do código que captura um livro aleatoriamente	22
<b>Figura 9.</b> <i>Screenshot</i> de uma postagem do robô jornalista no Twitter	24
<b>Figura 10.</b> <i>Screenshot</i> de uma postagem do robô jornalista no Twitter	24
<b>Figura 10.</b> <i>Screenshot</i> de uma postagem do robô jornalista no Twitter	24
<b>Figura 11.</b> <i>Screenshot</i> de uma postagem do robô jornalista no Twitter	24

## SUMÁRIO

<b>INTRODUÇÃO</b>	<b>7</b>
<b>1 ROBÔS JORNALISTAS</b>	<b>9</b>
1.1 Robôs jornalistas e o <i>Twitter</i> brasileiro	9
<b>2 GERAÇÃO DE LÍNGUA NATURAL</b>	<b>10</b>
2.1 Geração de Língua Natural e Aprendizagem de Máquina	13
<b>3 DESENVOLVIMENTO DE UM ROBÔ JORNALISTA RELACIONADO AO MERCADO EDITORIAL BRASILEIRO</b>	<b>14</b>
3.1 Metodologia de coleta das listas de livros mais vendidos do brasil	15
3.2 O <i>framework Scrapy</i> e a coleta de dados do robô	15
3.3 A conversão dos dados extraídos em língua natural	17
3.4 Postagem dos resultados no <i>Twitter</i>	23
<b>CONCLUSÃO</b>	<b>23</b>
<b>REFERÊNCIAS</b>	<b>25</b>

## INTRODUÇÃO

O Processamento de Língua Natural (PLN) está presente em muitas aplicações que envolvem linguagem e tecnologia, como as ferramentas de tradução automática, de sumarização de textos e de reconhecimento de voz, entre tantas outras (GATT; KRAHMER, 2018). De maneira geral, um dos objetivos da pesquisa e do desenvolvimento em PLN é lidar com uma grande quantidade de dados linguísticos e não linguísticos para transmiti-los em formato de língua natural.

A Inteligência Artificial é formada por um conjunto de tecnologias que têm em comum a intenção de simular a capacidade humana de tomar decisões. E é por meio de uma rede neural que as ferramentas da Geração de Língua Natural podem produzir textos compreensíveis de maneira autônoma. Essas redes neurais funcionam como um sistema interconectado que usa algoritmos para que se possa reconhecer padrões e estabelecer analogias entre dados e textos. Essas analogias são treinadas e podem ser melhoradas com o tempo por meio da própria inteligência artificial.

Então, é por meio dessas ferramentas que um robô jornalista pode ser criado e adquirir a capacidade de transformar dados em textos legíveis de maneira autônoma. Sobre esses robôs jornalistas, nós também precisamos delimitar melhor sua definição e seus usos. Para isso, vamos recorrer a uma palestra de Ramón Salaverría, pesquisador de mídias digitais e professor de Jornalismo na Universidade de Navarra, na Espanha. Salaverría foi convidado para abrir o 42º congresso da Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, a Intercom, de 2019, no Pará (RAMÓN, 2019). O tema de 2019, intitulado “Fluxos comunicacionais e crise da democracia”, abrangeu, entre outros tópicos, questões ligadas às tecnologias de Inteligência Artificial e a assimilação de seus produtos por um público numeroso presente nas redes sociais e em outros espaços da Internet. Na ocasião, o pesquisador defendeu o uso de novos modelos de comunicação que integrem a robotização ao trabalho dos jornalistas. A ideia é que se possa aproveitar de alguns benefícios que essas tecnologias oferecem, como a possibilidade de lidar com trabalhos mecânicos que tomam um tempo dos profissionais da comunicação e que podem ser automatizados. Com a adesão de veículos de mídia a tecnologias de

Inteligência Artificial, os profissionais podem dedicar uma atenção maior à apuração e a outras rotinas envolvidas no trabalho de um jornalista, que demandem necessariamente sua atenção.

Essa adesão permitiria a esses profissionais adotar tecnologias que por enquanto são utilizadas, em sua maior parte, por grupos que buscam assumir o espaço livre das redes sociais para a propagação de dados enviesados ou inautênticos. Segundo Salaverría, grande parte das informações que circulam nas redes sociais são falsas e mesmo as reais têm como objetivo responder a interesses ocultos (RAMÓN, 2019). Um profissional que assuma as possibilidades oferecidas por um robô jornalista, então, utilizaria essa tecnologia — hoje majoritariamente adotada por grupos que não se identificam —, para divulgar informações, com a diferença de assumir publicamente sua identidade, estando ou não vinculado a um veículo maior, e ser submetido à regulação de conselhos de classe e mesmo à justiça comum.

Ainda no congresso citado, Salaverría falou sobre como as redes sociais contribuem para moldar as decisões de seus usuários sem que haja uma legislação que incida diretamente sobre essas redes, contribuindo com a ilusão de que a falta de uma intervenção superior significaria liberdade na mensagem (RAMÓN, 2019).

## **2 ROBÔS JORNALISTAS**

Os robôs jornalistas começaram a ser utilizados muito recentemente. O primeiro, segundo o jornal *BBC*, foi criado pelo profissional da comunicação e programador estadunidense Ken Schwencke. Sua intenção com o robô foi divulgar notícias sobre terremotos locais no jornal *The Los Angeles Times*. O robô começou a ser programado em meados de 2012 e em 2014 escreveu sua primeira notícia. Segundo o jornal *Slate*, que também noticiou o acontecimento, foram necessários três minutos para redigir o texto e publicá-lo no jornal (SLATE, 2014). Isso porque quando o terremoto ocorreu, o robô checkou as informações divulgadas pelo *U.S. Geological Survey*, órgão estadunidense responsável por dados que envolvem fenômenos naturais, e fez a redação da matéria. Em seguida, a aprovação para publicação foi feita pelo editor responsável pelo caderno de notícias.



No Brasil, a Inteligência Artificial também é usada por grandes conglomerados de mídia. Em 2020, o portal de notícias *g1*, do grupo Globo, usou essa ferramenta para publicar textos sobre a posse de prefeitos e vereadores em todos os municípios brasileiros. E essa é uma das características de um robô jornalista: por ser escalonável, a máquina pode lidar com todos os dados divulgados pelo Tribunal Superior Eleitoral, elaborar os textos e torná-los acessíveis para a revisão em poucos minutos.

## 2.1 ROBÔS JORNALISTAS E O TWITTER BRASILEIRO

Em 2021, dados divulgados pela Pesquisa Nacional por Amostra de Domicílios mostraram que 90% das casas no Brasil têm acesso à Internet. O número foi divulgado pelo Governo Federal e tem apoio do Ministério da Comunicações — sendo talvez necessária uma reavaliação no futuro —, mas ainda assim demonstra que o número é alto. Levando em consideração essa propagação, as redes sociais se tornam um terreno fértil para a transmissão de informações para um público numeroso. E isso é valorizado por vários setores da sociedade que investem na propagação de notícias pelas redes sociais, inclusive em ferramentas que tornam essa transmissão automática e escalonável.

Durante a pandemia, um robô jornalista cumpriu o papel de divulgar informações sobre os casos de COVID-19 em território nacional. O *Corona Repórter*<sup>1</sup> buscava na Internet novos dados sobre a doença e redigia os textos para a divulgação na rede social. A mesma equipe desenvolveu também o *Da Mata Repórter*<sup>2</sup>, robô que cobria desmatamentos na Amazônia Legal.

Um robô jornalista em atividade e que recebe uma atenção considerável no *Twitter* é o *Robotox*<sup>3</sup>, que acompanha o Diário Oficial da União para divulgar informações sobre a aprovação de novos agrotóxicos no Brasil — atualmente o *Robotox* tem mais de 22 mil seguidores na rede social. Outro registro interessante é o *Amazônia Minada*, que faz novas publicações sempre que encontra um pedido de mineração em terras indígenas. Por meio de ferramentas da Geração de Língua Natural, então,

---

<sup>1</sup> [www.twitter.com/coronareporter](https://www.twitter.com/coronareporter)

<sup>2</sup> [www.twitter.com/damatareporter](https://www.twitter.com/damatareporter)

<sup>3</sup> [www.twitter.com/orobotox](https://www.twitter.com/orobotox)

esses robôs divulgam notícias de forma automática, com base em dados públicos. E é preciso falar também que esses dados são divulgados em um formato quase ilegível ao público geral que porventura tentasse acessar essas informações. O *Corona Repórter*, por exemplo, acessava um banco de dados mundial, divulgado pela Organização Mundial da Saúde (OMS), e selecionava automaticamente os dados ligados ao Brasil. Com os números processados, redigia então os textos e divulgava os *tweets* na rede social.

Há também robôs criados essencialmente para alterar o debate público, em geral com postagens que envolvem temas ligados à atividade política — matéria muito presente na rede social. A utilização e o funcionamento desses robôs não são muito conhecidos porque funcionam na ilegalidade, apesar de não haver uma legislação inteiramente dedicada ao tema. A regulação da mídia serviria como essa legislação e é um assunto cada vez mais discutido pela sociedade em geral.

### **3 GERAÇÃO DE LÍNGUA NATURAL**

A Geração de Língua Natural, então, é o campo da computação que oferece as ferramentas para que se estabeleça a relação entre dados não linguísticos (em geral, organizados em formato de tabela ou banco de dados) e língua humana. Para que possamos falar sobre o nosso projeto de robô jornalista, vamos descrever as três abordagens do campo de Geração de Língua Natural utilizadas para a criação de plataformas de comunicação automática, incluindo os robôs jornalistas.

A primeira delas é a abordagem baseada em *templates*. Neste caso, os dados são transpostos para a língua natural de forma mais simplificada, com um modelo de frases preestabelecido. Assim, a máquina recebe os dados não linguísticos e aplica essas informações a uma modelo linguístico gerado anteriormente, por meio de uma substituição direta de valores. Esse procedimento torna a comunicação engessada e não permite a variação possibilitada pela inteligência artificial. Para que possamos compreender melhor esta abordagem, podemos utilizar um exemplo de frase a ser gerada de acordo com um robô jornalista que divulgue a previsão do tempo. Aconteceria assim: “Em [NOME DA CIDADE], a previsão é de [PREVISÃO

METEOROLÓGICA] durante [TEMPO DE DURAÇÃO DA PREVISÃO]”. Assim, as marcações entre chaves seriam trocadas pelas informações coletadas pelo robô.

A segunda abordagem é conhecida como *pipeline* e funciona de modo a gerar textos de acordo com as representações semânticas de entrada, mas em geral tende a gerar mais erros nas saídas de informação e não há uma variedade de escolha entre as possibilidades. Vamos falar mais sobre a abordagem *pipeline* adiante.

A abordagem mais avançada é denominada *ponta-a-ponta*. Neste caso há uma representação semântica fluente, com variação textual e utilização de métodos da Inteligência Artificial. No entanto, esta abordagem pode alucinar informações, o que significa uma confusão de entendimento dos dados de entrada para sua conversão em texto. Essa alucinação pode representar grandes problemas, se estivermos falando de um robô que lida com informações sensíveis.

Castro Ferreira et al. (2019) diferenciaram os procedimentos envolvidos na abordagem de *pipeline* e a de *ponta-a-ponta*, em suas características e procedimentos de funcionamento. A pesquisa foi feita com o objetivo de conhecer melhor as diferenças efetivas no funcionamento das duas abordagens e das possibilidades de cada uma delas. Em primeiro lugar, os pesquisadores chamam a atenção para a possibilidade de erros em cascata no sistema *pipeline*. Isso acontece quando um módulo apresenta alguma anomalia e isso é transposto para os procedimentos seguintes em um efeito que impacta todos os resultados. Esse mesmo erro não acontece na abordagem de *ponta-a-ponta*. Ao mesmo tempo, o modelo de *pipeline* pode oferecer uma performance mais significativa ao lidar com o *corpus* e até mesmo apresentar resultados de saída potencialmente melhores, se bem estruturados e livres de erros.

Os passos envolvidos na arquitetura de *pipeline*, de acordo com Castro Ferreira et al. (2019), é composto por cinco etapas. A primeira é a ordenação do discurso, quando o programa determina a ordem em que o resultado comunicativo deve ser verbalizado; a segunda é a estruturação do texto, quando o conjunto textual gera o segmento verbal com base nos *tokens* disponíveis, ou seja, no caso das pesquisas, as palavras do *corpus*. Depois disso é feita a *tokenização*, para que se possa encontrar as palavras que expressam o sentido a ser dado a cada sentença e,

depois disso, como quarto passo, tem início o processo denominado *Referring Expression Generation* (REG). Aqui, com o *template* já criado na etapa anterior, o programa codifica o texto em duas partes, em um processo de rede neural denominado *Bidirectional LSTMs* — para então sequenciar o texto nas direções de frente para trás e de trás para frente. E, por fim, resta apenas converter de fato os dados não linguísticos em textos legíveis.

A abordagem de *ponta-a-ponta* exclui esses processos intermediários e passa a utilizar as técnicas de processamento de tradução neural, com suas variações entre as ferramentas escolhidas. A conceituação estabelecida para diferenciar as duas abordagens não apresenta uma bibliografia ampla na literatura da Ciência da Computação. O artigo citado se coloca como o primeiro a estabelecer parâmetros de mensuração que possibilitam entender melhor os aspectos envolvidos em ambas as abordagens. Destacamos esta informação para dizer que os procedimentos não são simples e sua compreensão exige uma pesquisa mais dedicada ao tema. Mesmo assim, tratamos em linhas gerais para que possamos compreender o que interessa em nosso trabalho.

### 3.1 GERAÇÃO DE LÍNGUA NATURAL E APRENDIZAGEM DE MÁQUINA

Assim como o PLN, a Aprendizagem de Máquina é um campo da Ciência da Computação e da Inteligência Artificial. Essa área estuda a possibilidade de dar aos computadores a capacidade de aprender sem serem explicitamente programados, com um objetivo comunicacional específico (CHATZILYGEROUDIS, 2021).

Há alguns modelos principais de funcionamento que determinam como uma máquina pode aprender. Como dissemos, nossa pesquisa inclui a criação de um robô jornalista, no qual nos aprofundaremos mais adiante, mas, para que possamos compreender melhor nossa criação neste ponto, vamos falar sobre a sua capacidade de aprendizagem. Seu treinamento funciona de uma maneira que muitos dados são recebidos na entrada do programa e muitos dados também são retornados na saída. O robô, então, funciona com um conjunto variável de informações posteriormente convertidas na verbalização de um texto. As máquinas de tradução também funcionam desta forma. É a abordagem conhecida como de

*muitos para muitos*. Para que se entenda melhor esse modelo de aprendizagem, podemos falar sobre a possibilidade de uma máquina trabalhar no modelo de *um para um*, em que uma informação é dada como entrada e uma como saída. Há também o modelo de *muitos para um*, que é como funciona a tarefa de PLN conhecida como *análise de sentimentos/polaridade*. O modelo de *muitos para um* também possibilita identificar discursos de ódio na Internet, por exemplo, porque recebe um conjunto de informações e retorna se aquele texto está identificado como discurso de ódio ou não. Em nosso caso, no entanto, vamos lidar com a abordagem de *muitos para muitos*.

Uma das tecnologias potencialmente utilizadas para gerar essa rede neural de *muitos para muitos* é chamada *Transformers*. Criada pelo Google, a ferramenta foi lançada em 2017. Sua diferença para outras redes neurais é a possibilidade que ela oferece de analisar as frases de trás para frente e de frente para trás. Explicamos: as redes neurais recorrentes eram as mais utilizadas antes do lançamento da *Transformers* e, ao analisar uma frase, partia sempre de uma leitura da esquerda para a direita. Esse tipo de leitura gera a possibilidade de muitos erros, porque não há uma análise do contexto geral. Já quando as palavras são lidas de forma a serem comparadas com os vocábulos anteriores e posteriores, como é o caso no modelo proporcionado pela *Transformers*, há uma compreensão melhor do sentido da frase. Sobre essa compreensão, há pesquisas que sugerem a leitura humana fazendo o mesmo procedimento: tendemos a direcionar fisicamente os olhos a palavras anteriores e posteriores para que possamos compreender um texto.

Além dessa compreensão maior envolvendo *corpora* linguísticos, a rede neural *Transformers* permite também a classificação de imagens, a detecção de objetos nessas imagens e suas catalogações (USZKOREIT, 2017). Ela também é capaz de extrair informações de documentos escaneados, por exemplo, além de classificar vídeos e possibilitar a execução de muitos outros procedimentos de Inteligência Artificial ligados ao reconhecimento de áudios e outras mídias (LI, 2018).

Então, em nosso projeto, utilizamos essa capacidade da Inteligência Artificial para que a máquina consiga identificar o título de uma obra, a posição do livro no ranking anual, o nome do autor, o ano em que a obra alcançou a marca de mais vendidos, a

editora responsável pela publicação e o número de vendas alcançadas naquele período. O programa processa todas essas informações contidas em um banco de dados para que possa transformá-las em um texto legível. Sobre o funcionamento do robô em si, vamos tratar em mais profundidade adiante em nosso trabalho.

## **4 DESENVOLVIMENTO DE UM ROBÔ JORNALISTA RELACIONADO AO MERCADO EDITORIAL BRASILEIRO**

Neste ponto, passaremos a falar sobre a própria criação de um robô jornalista e como essa ferramenta pode ser utilizada em um contexto educacional para a divulgação de dados do mercado editorial brasileiro. Para isso, criamos uma máquina que, por meio do *framework Scrapy*, pudesse coletar a lista de livros mais vendidos desde 2010 até o ano de 2021. O programa foi escrito em linguagem de programação *Python* e é dividido em duas partes. A primeira está justamente na coleta dos dados do portal *PublishNews* e a segunda envolve a leitura desses dados e a conversão dessas informações em frases legíveis para que sejam publicadas no *Twitter*. Para que o funcionamento do robô jornalista fique mais compreensível, vamos dividir também a explicação de seu funcionamento nessas duas partes. Antes, porém, vamos explicar a metodologia de seu banco de dados.

### **4.1 METODOLOGIA DE COLETA DAS LISTAS DE LIVROS MAIS VENDIDOS DO BRASIL**












O portal *PublishNews*<sup>4</sup> é a fonte utilizada no robô jornalista criado neste trabalho. O portal faz parte da Câmara Brasileira do Livro, o órgão mais importante e representativo do campo da edição de livros no Brasil. Com a divulgação feita pelo *PublishNews*, pudemos ter acesso a uma lista de livros unificada, com os dados garantidos pelo órgão. É importante destacar essa informação porque até pouco tempo atrás era comum que veículos de mídia divulgassem listas de livros mais vendidos, mas nem sempre essa relação estava em conformidade entre os próprios veículos. Era comum que uma lista destoasse de outra. Hoje, então, podemos

---

<sup>4</sup> [www.publishnews.com.br](http://www.publishnews.com.br)

recorrer a um órgão que representa de maneira mais completa o campo da edição de livros em nosso país.

No site, há duas opções de listas de livros mais vendidos. Uma delas é exclusivamente de obras brasileiras. A outra engloba a venda de livros de todo o mercado editorial, inclusive os livros estrangeiros vendidos aqui. Nós escolhemos seguir pela segunda opção e obter a listagem de todos os livros comercializados no mercado brasileiro. Essa lista de livros gerais do site envolve 23 livrarias. É, portanto, uma amostra do mercado editorial como um todo. Para a divulgação, é considerada a soma simples de todas as edições vendidas nos estabelecimentos consultados.

Lista de Mais Vendidos Geral de 2020				     			
Geral   Ficção   Não ficção   Autoajuda   Infantojuvenil   Negócios				Semanal   Mensal   Anual			
Livros				Editoras			
1		<b>Mais esperto que o diabo</b> Napoleon Hill Citadel	113.041	1	<b>Sextante</b>	57	
					Sextante	40	
					Arqueiro	15	
					Estação Brasil	2	
2		<b>A sutil arte de ligar o foda-se</b> Mark Manson Intrínseca	104.649	2	<b>Gente</b>	48	
					Gente	47	
					Única	1	
					<b>Grupo Companhia das Letras</b>	48	
					Companhia das Letras	23	
					Seguinte	7	
					Objetiva	6	
					Suma de Letras	5	
					Paralela	4	
					Cia das Letrinhas	1	
					Companhia de Bolso	1	
					Quadrinhos na Cia	1	
					<b>Intrínseca</b>	48	
3		<b>Do mil ao milhão</b> 🇧🇷 Thiago Nigro HarperCollins	77.862				
4		<b>O milagre da manhã</b> Hal Elrod BestSeller	62.235				
5		<b>Box Harry Potter</b> J. K. Rowling Rocco	55.124	5	<b>Grupo Editorial Alta Books</b>	41	
					Alta Books Editora	36	

**Figura 1.** Screenshot do portal *PublishNews*

A partir da listagem feita pelo portal *PublishNews*, nosso programa faz a coleta dessas informações para que se possa gerar um banco de dados sobre as vendas do mercado editorial. Esses arquivos, então, ficam salvos para a posterior

transformação dessas informações em frases legíveis por meio da Inteligência Artificial.

#### 4.2 O FRAMEWORK SCRAPY E A COLETA DE DADOS DO ROBÔ

A ferramenta *Scrapy*<sup>5</sup> oferece a funcionalidade de fazer a coleta de dados de um site ou de uma API. No caso de um site, isso é feito por meio da própria marcação em CSS na página requerida. O processo envolve a identificação de uma informação na página por meio das marcações feitas pelo programador responsável pela criação do site. Por exemplo, para que possamos identificar o título dos livros no portal *PublishNews*, usamos o identificador “.pn-ranking-livro-nome”. E assim fazemos também com outros dados, como o nome do autor, a posição do livro no ranking anual e assim por diante. O programa, então, capta essas informações e faz a ordenação para que tudo seja lido na segunda parte do programa, quando há efetivamente o processamento de língua natural.

Por padrão, a coleta utilizando o *Scrapy* acontece de maneira aleatória — e isso exigiu nossa atenção, já que a ferramenta deveria estruturar as informações em ordem crescente, de 2010 até o presente. Isso acontece porque há uma priorização em captar dados que estão mais acessíveis. Ou seja, os dados que forem encontrados primeiro serão inscritos no banco de dados, sem que haja a linearidade pretendida, porque, funcionando de maneira linear e crescente, a máquina estaria atuando de maneira mais lenta e com maior utilização de recursos. Isso foi contornado e conseguimos salvar as informações no banco de dados partindo do ano inicial até o atual.

A partir da coleta, então, as listas são salvas de forma que a segunda parte de nosso programa possa compreendê-las e transformá-las em frases legíveis. Para que se saiba exatamente o formato em que o banco de dados está composto, vamos transcrever e explicar o funcionamento de uma de suas linhas aqui. O conjunto de dados formado por “2018, 19, Outros jeitos de usar a boca, Rupi Kaur, Planeta do Brasil , 76.069” será possivelmente convertido na frase: “Outros jeitos de

---

<sup>5</sup> [www.scrapy.org](http://www.scrapy.org)



usar a boca, de Rupi Kaur, foi o 19º livro mais vendido de 2018. Publicado pela Planeta do Brasil, as vendas da edição chegaram a 76.069 unidades.”

Dizemos que *possivelmente* ela será convertida nessa frase porque não é possível prever o resultado dado pela máquina. A partir dos modelos de treinamento dados ao robô, o resultado gerado pela inteligência artificial será feito de acordo com sua compreensão dos modelos treinados. Assim, com informações lineares e sem indicação exata de onde cada termo deve estar colocado, a tarefa da inteligência artificial é encontrar, por meio dos exemplos, a correta alocação de cada informação. Sobre isso, tratamos brevemente no tópico sobre a Geração de Língua Natural e suas abordagens mais modernas.

#### 4.3 A CONVERSÃO DOS DADOS EXTRAÍDOS EM LÍNGUA NATURAL

Com a lista convertida em formato de banco de dados, passamos então ao segundo passo do robô: a conversão das informações em língua natural. Para isso, usamos a ferramenta *Transformers*, discutida anteriormente, e montamos o modelo de treinamento com 37 frases que serviram de exemplos para que a inteligência artificial pudesse montar sua rede neural. Além do conjunto de treinamento, fizemos 33 modelos de frases para o desenvolvimento do sistema e outras 11 para os testes aplicados à inteligência artificial. Todo o conjunto de códigos e de processamento deste passo foi feito na ferramenta *Colaboratory*<sup>6</sup>, disponibilizada pelo Google.

O tempo gasto para a execução do processo completo envolve uma média de dez minutos e, destes, nove são dedicados ao treinamento da máquina de inteligência artificial. Todo o processo de treinamento foi feito com base na ferramenta *Data2Text*<sup>7</sup>, disponibilizada por Thiago Castro Ferreira. O programa utiliza o modelo criado pelo *Facebook* intitulado BART (*Bayesian Analogy with Relational Transformations*). Por meio desta tecnologia, a máquina é treinada de modo a corromper o texto com um ruído arbitrário, dificultando seu entendimento; ao mesmo tempo, o programa tenta reconstruir o modelo original com os ruídos na frase, para reencontrar seu sentido primário (RAJAPAKSE, 2020). Todo o resultado é testado

---

<sup>6</sup> <https://colab.research.google.com/drive/1CIB73HMJ9WBUKG7x51EyRaq3i757kSX8?usp=sharing>

<sup>7</sup> [www.github.com/ThiagoCF05/Any2Some](https://www.github.com/ThiagoCF05/Any2Some)

com outro *corpus* linguístico, também disponibilizado por nós, para que a máquina consiga reconhecer se os dados estão nos lugares corretos e qual a reconstrução mais adequada para chegar ao sentido da frase inicial. Ou seja, como criar outras frases, com a inteligência artificial, mantendo o mesmo sentido do *corpus* inicial. Como dissemos anteriormente, a capacidade de ler uma frase de maneira não apenas linear, mas retornando às palavras anteriores é fundamental para garantir o efeito de compreensão que esta ferramenta de inteligência artificial alcança.

O tempo necessário para que a ferramenta perpassasse todo o *corpus* linguístico criado por nós envolve também o número de *épocas* que delimitamos na programação do treinamento. Essas *épocas* são o número de vezes que o programa estuda cada uma das frases criadas. Ao todo, escolhemos colocar o número de 30 *épocas*, com a exceção de que a máquina pare de rodar os teste caso os resultados não avancem em um número de cinco *épocas*. Ou seja, caso haja recorrências em que não se mostre um progresso na compreensão do *corpus*, o programa deixa de ser executado automaticamente. Isso porque a cada volta aos exemplos, a máquina salva o avanço que teve na compreensão dos textos e há um momento em que, devido à própria limitação do número de frases e do programa, não há mais progresso na compreensão do *corpus*. Isso sempre vai acontecer em algum momento de qualquer ferramenta de inteligência artificial. Assim, quando a máquina reconhece que não há mais progresso, o programa é terminado e o conjunto de *épocas* que obteve o melhor resultado é utilizado para a criação da rede neural.

Enquanto a máquina processa o *corpus* linguístico, é possível acompanhar o desenvolvimento de sua aprendizagem. Para que se consiga compreender melhor esse progresso, vamos analisar algumas mensagens emitidas pelo programa durante o processo de leitura dos dados. A seguir, examinaremos a primeira época de treinamento, para que possamos compará-la com o avanço das seguintes.

Train Epoch: 0	[2/19 (5%) ]	Loss: 7.208451	Total Loss: 8.678090
Train Epoch: 0	[4/19 (16%) ]	Loss: 5.965469	Total Loss: 6.877240
Train Epoch: 0	[6/19 (26%) ]	Loss: 3.996363	Total Loss: 5.888760
Train Epoch: 0	[8/19 (37%) ]	Loss: 3.947131	Total Loss: 5.453610
Train Epoch: 0	[10/19 (47%) ]	Loss: 5.242540	Total Loss: 5.280830
Train Epoch: 0	[12/19 (58%) ]	Loss: 3.506931	Total Loss: 5.162830
Train Epoch: 0	[14/19 (68%) ]	Loss: 3.009919	Total Loss: 4.884760
Train Epoch: 0	[16/19 (79%) ]	Loss: 4.866531	Total Loss: 4.819320
Train Epoch: 0	[18/19 (89%) ]	Loss: 2.811140	Total Loss: 4.641040

**Figura 2.** *Screenshot* do ambiente Google Colab durante a execução do *script*

As colunas intituladas *Loss* e *Total Loss* demonstram o quanto a máquina está absorvendo daquele *corpus*. Por isso, quanto menor o número apresentado nessas informações, melhor será a compreensão da máquina em relação ao modelo estudado. Para que se possa comparar também as próprias frases geradas pela inteligência artificial, vamos demonstrar aqui o modo como elas aparecem nas saídas da máquina nessa primeira *época*:

```
Real: O livro mais vendido do ano de 2010 foi o Ágape, do autor Padre
Marcelo Rossi. A editora responsável pela publicação é a Globo Livros.
Ao todo, 337.520 livros foram vendidos naquele ano.
Pred: 2010, 1, Ágape, Padre Marcelo Rossi, Globo Livros, 337.520

Real: Em 2010, o 2º livro mais vendido do ano foi o '1822', de
Laurentino Gomes. A casa editorial responsável é a Nova Fronteira.
115.546 livros foram vendidos ao todo naquele ano.
Pred: 2010, 2, 1822, Laurentino Gomes, Nova Fronteira, 115.546

Real: Comer, rezar, amar, da autora Elizabeth Gilbert, foi o 3º livro
mais vendido de 2010. Publicado pela Objetiva, as vendas atingiram
83.160 unidades naquele ano.
Pred: Comer, rezar, amar', Elizabeth Gilbert, Objetiva, 83.160.2010,
3, 3
```

**Figura 3.** *Screenshot* do ambiente Google Colab durante a execução do *script*

Atente-se para o fato de que as frases estão divididas em duas categorias: a primeira é o *corpus* que demos à máquina, escrito por nós, intitulada pelo programa de *Real*; a segunda, intitulada *Pred*, de *prediction*, é gerada pela inteligência artificial após compreender o que pôde com o estudo de uma *época* do *corpus* dado como exemplo.

A seguir, vamos tratar da terceira *época* rodada pelo programa. A perda indicada é menor e a compreensão do que está sendo tratado nas frases aumenta:

```
Train Epoch: 3 [2/19 (5%)] Loss: 2.270333 Total Loss: 1.992820
Train Epoch: 3 [4/19 (16%)] Loss: 2.685991 Total Loss: 1.993720
Train Epoch: 3 [6/19 (26%)] Loss: 1.576469 Total Loss: 1.764500
Train Epoch: 3 [8/19 (37%)] Loss: 1.287836 Total Loss: 1.759860
Train Epoch: 3 [10/19 (47%)] Loss: 2.641802 Total Loss: 1.853040
Train Epoch: 3 [12/19 (58%)] Loss: 1.469749 Total Loss: 1.923560
Train Epoch: 3 [14/19 (68%)] Loss: 1.317377 Total Loss: 1.844380
```

```
Train Epoch: 3 [16/19 (79%)] Loss: 2.200881    Total Loss: 1.885430
Train Epoch: 3 [18/19 (89%)] Loss: 1.227752    Total Loss: 1.848510
```

**Figura 4.** *Screenshot* do ambiente Google Colab durante a execução do *script*

Então, vemos que essa perda já está menor do que 2, enquanto a primeira variava de 4 a 8. As frases geradas a partir dessa compreensão são as seguintes:

```
Real:  Os sete maridos de Evelyn Hugo, escrito por Taylor Jenkins Red,
foi o 11° livro mais vendido de 2021, com 45.943 unidades
comercializadas. A obra foi publicada pela editora Paralela.
Pred:  Os sete maridos de Evelyn Hugo, de Taylor Jenkins Reid, foi o
11° livro mais vendido de 2017. Publicado pela Paralela, as vendas
chegaram a 45.943 unidades.

Real:  Os segredos da mente milionária foi o 12° livro mais vendido de
2021. Escrito por T. Harv Eker, a publicação é da Sextante e alcançou a
marca de 43.819 unidades vendidas
Pred:  Os segredos da mente milionária, de T. Harv Eker, foi o 12°
livro mais vendido de 2015. Publicado pela Sextante, as vendas chegaram
a 43.819 unidades.

Real:  Em 2021, o 16° livro mais vendido foi O homem mais rico da
Babilônia. Escrito por George S. Clason e publicado pela HarperCollins,
a edição alcançou a marca de 40.608 volumes comercializados.
Pred:  O homem mais rico da Babilônia, George S. Clason, foi o 16°
livro mais vendido de 2010. Publicado pela HarperCollins, a obra chegou
a 40.608 unidades.
```

**Figura 5.** *Screenshot* do ambiente Google Colab durante a execução do *script*

Assim, é possível observar que o resultado ainda não está próximo do que se espera, já que o programa troca os anos em que o livro apareceu na lista de *best sellers* e mesmo a sintaxe apresenta algum problema, ainda que menor.

Agora, chegamos à parte em que o programa roda a sua 20ª época. Poderemos ver que os dados indicados em *Loss* são menores do que 1. É esperado, então, que as frases geradas se aproximem ainda mais dos modelos de exemplos dados à máquina. Vamos acompanhar:

```
Train Epoch: 20 [2/19 (5%)]      Loss: 0.496239    Total Loss: 0.461540
Train Epoch: 20 [4/19 (16%)]     Loss: 0.550890    Total Loss: 0.432920
Train Epoch: 20 [6/19 (26%)]     Loss: 0.304494    Total Loss: 0.380850
Train Epoch: 20 [8/19 (37%)]     Loss: 0.307412    Total Loss: 0.404000
Train Epoch: 20 [10/19 (47%)]    Loss: 0.703889    Total Loss: 0.436150
```

Train Epoch: 20 [12/19 (58%)]	Loss: 0.379698	Total Loss: 0.447270
Train Epoch: 20 [14/19 (68%)]	Loss: 0.279560	Total Loss: 0.424300
Train Epoch: 20 [16/19 (79%)]	Loss: 0.492593	Total Loss: 0.436540
Train Epoch: 20 [18/19 (89%)]	Loss: 0.298554	Total Loss: 0.428670

**Figura 6.** Screenshot do ambiente Google Colab durante a execução do *script*

Agora, as frases geradas nesta época:

```
Real:  A coragem de ser imperfeito foi o 13° livro mais vendido de
2021. Publicado por Brené Brown, as vendas chegaram a 43.740 unidades.
A editora responsável pela edição é a Sextante.
Pred:  A Sextante foi a editora responsável pelo livro A coragem de ser
imperfeito, de Brené Brown. O livro alcançou a marca de 43.740 unidades
vendidas.

Real:  O 14° livro mais vendido de 2021 foi A sutil arte de ligar o
foda-se. Publicado por Mark Manson e editado pela Intrínseca, a obra
alcançou as 42.052 unidades vendidas.
Pred:  A sutil arte de ligar o foda-se foi o 14° livro mais vendido de
2021. Escrito por Mark Manson e publicado pela Intrínseca, a obra
alcançou a marca de 42.052 unidades vendidas.

Real:  O poder do hábito, de Charles Duhigg, foi o 15° livro mais
vendido de 2021. Com 41.884 edições vendidas, a publicação é da
Objetiva.
Pred:  O poder do hábito foi o 15° livro mais vendido de 2021. Escrito
por Charles Duhigg e publicado pela Objetiva, as vendas chegaram a
41.884 unidades.

Real:  Em 2021, o 16° livro mais vendido foi O homem mais rico da
Babilônia. Escrito por George S. Clason e publicado pela HarperCollins,
a edição alcançou a marca de 40.608 volumes comercializados.
Pred:  O homem mais rico da Babilônia foi o 16° livro mais vendido de
2021. Escrito por George S. Clason e publicado pela HarperCollins, a
obra alcançou a marca de 40.608 unidades vendidas.
```

**Figura 7.** Screenshot do ambiente Google Colab durante a execução do *script*

Como podemos observar, o resultado está muito mais próximo do que pretendemos com os exemplos dados, e as frases estão completamente legíveis. Neste ponto, então, a máquina percebe que não será capaz de melhorar muito sua capacidade de dominar o *corpus* linguístico e termina o programa.

Agora chegamos ao ponto em que todo esse resultado permite que a rede neural seja criada. Até aqui, ainda não entramos efetivamente na criação de textos inéditos para a máquina. Isso só é feito quando o programa acessa o banco de dados e,

utilizando a rede neural criada por meio dos exemplos dados, seleciona aleatoriamente uma linha e, aí sim, começa a elaborar um novo texto. Antes de partir para a seleção aleatória no banco de dados, porém, devemos dar os testes à máquina para que ela crie efetivamente a rede neural e termine o processo de aprendizagem. Fazemos isso com um novo *corpus*, que ainda não foi analisado pelo programa. Em nosso robô, usamos 11 exemplos para que a rede neural possa executar suas provas e conseguir efetivamente ser criada. Aqui é possível analisar outro indicativo importante de checagem: o quesito BLEU, acrônimo para Bilingual Evaluation Understudy, mostra se a inteligência artificial está dominando as informações dadas à máquina. Por meio dele é possível analisar o quão próxima a frase gerada pela inteligência artificial está da resposta que se espera dela. Neste caso, quanto mais próxima de 1, melhor. Após as 20 épocas e depois das provas, o resultado foi de 0.48855. Para que se possa fazer uma comparação, a primeira execução dos estudos do programa, a primeira época, dá o seguinte resultado neste quesito 0.13084.

A partir daqui chegamos ao ponto em que, após o processamento das informações, o programa acessa o banco de dados aleatoriamente e dá início à geração de língua natural por meio de dados ainda não vistos pela máquina. Para que isso possa ser feito, usamos uma variável de randomização de números, que programamos para que passasse de 2010 a 2021, e selecionamos os arquivos com a ferramenta *Pandas*. O código ficou assim:

```
aleat = random.randint(2010, 2021)

with open(f'ano {aleat}.csv', 'r') as file:
    reader = csv.reader(file, delimiter = '\t')
    n = 20 #Número de linhas do arquivo
    s = 1 #Número de linhas desejadas
    filename = f'ano {aleat}.csv'
    skip = sorted(random.sample(range(n),n-s))
    df = pandas.read_csv(file, skiprows=skip)
    df.to_csv('temp.csv')

with open('temp.csv', 'r') as file:
    reader = csv.reader(file, delimiter = '\t')
    for row in reader:
        output = generator(row)
```

**Figura 8.** Screenshot do código que captura um livro aleatoriamente

Após todo o processo de treinamento da máquina, então, a rede neural fica salva e é possível acessá-la com os comandos dados ao programa. Assim, para que seja gerada uma frase é necessário apenas recorrer à rede neural gerada. O tempo de processamento passa então a ser de em média dez segundos. Nesse breve período o programa acessa o banco de dados, seleciona aleatoriamente um ano em que os livros estão salvos e, após selecionar o ano, encontra, também aleatoriamente, uma linha presente nesse banco de dados. A partir dessa linha é que o programa acessa a rede neural e cria uma nova frase com a inteligência artificial. Assim, a máquina está programada para gerar uma frase a cada vez que rodamos o comando de sua inicialização. Isso também torna mais fácil a posterior postagem do texto no *Twitter*.

Todas as frases geradas podem ser consultadas no *Twitter* do programa, pelo endereço [www.twitter.com/libestsellers](http://www.twitter.com/libestsellers). Um exemplo de texto gerado pela inteligência artificial após todo esse processo é o seguinte: “O poder da ação foi o 14º livro mais vendido de 2016. Escrito por Paulo Vieira e publicado pela editora Gente, a obra alcançou a marca de 98.298 unidades vendidas”. A seguir vamos demonstrar outros resultados do robô jornalista no Twitter.

#### 4.4 POSTAGEM DOS RESULTADOS NO TWITTER

Após essa seleção aleatória e a conversão dos dados em texto, é necessário apenas fazer a publicação no Twitter<sup>8</sup>. Para isso, instalamos a biblioteca da rede social na máquina do *Google Colab*.

A seguir, instalamos as chaves para acesso à conta do jornalista robô geradas pelo *Twitter* — e concedidas após a autorização da equipe da rede social para que a conta seja usada para este propósito —, e então é possível postar, por meio de uma variável, a frase gerada na conta desejada.

A seguir, vamos compartilhar, em imagens, frases representativas do nosso robô jornalista e seus resultados.

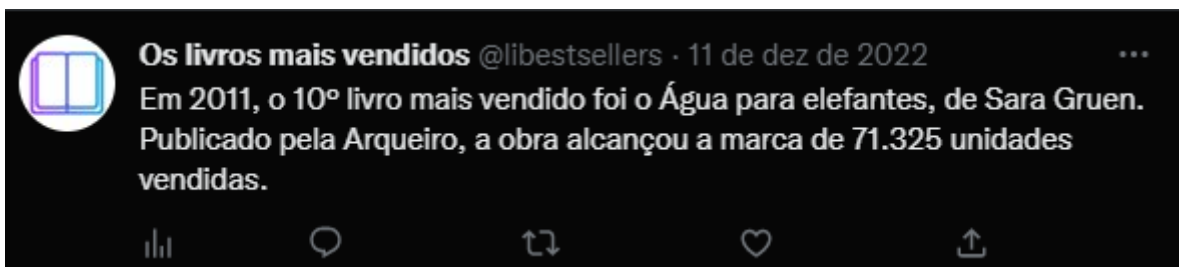
---

<sup>8</sup> [www.twitter.com/libestsellers](http://www.twitter.com/libestsellers)



**Figura 9.** *Screenshot* de uma postagem do robô jornalista no Twitter.

No caso acima, o robô criou uma frase a partir de um livro intitulado *A garota do lago*. As informações estão completas e não há qualquer falha ou alucinação no resultado.



**Figura 10.** *Screenshot* de uma postagem do robô jornalista no Twitter.

Neste caso, há uma variação na frase, que tem início pelo ano em que a obra foi publicada. No exemplo também podemos ver que as informações estão completas e não há alucinação por parte da máquina.



**Figura 11.** *Screenshot* de uma postagem do robô jornalista no Twitter.

E, por último, temos mais um exemplo da execução do robô jornalista. Neste caso, a frase tem início com o nome da obra e as outras informações seguem também sem alucinação.



Após as postagens, a máquina ficará em *standby* ou, como pretendemos, podemos colocar um temporizador, para que em alguns horários do dia o programa seja novamente rodado e uma nova divulgação seja feita.

## CONCLUSÃO

Neste trabalho, pretendemos criar um modelo de inteligência artificial que demonstrasse o estado da arte dessas ferramentas atualmente. Pretendemos também fazer uma revisão do uso feito por máquinas de geração de língua natural na rede social *Twitter* e das possibilidades que essa tecnologia apresenta à divulgação de dados, inclusive em relação ao mercado editorial brasileiro. Muito do que foi discutido neste texto tem relação com as aulas ministradas pelo pesquisador Thiago Castro Ferreira, na disciplina *Aplicações linguísticas inteligentes: processamento de língua natural para comunicação de dados abertos e jornalismo computacional*, ofertada no Programa de Pós-Graduação em Estudos Linguísticos (POSLIN) da UFMG, no 1º semestre de 2021. Os modelos de geração de língua natural e as abordagens utilizadas neste trabalho foram influenciadas diretamente pela exposição nas aulas. Em relação à metodologia de construção do nosso robô jornalista, pretendemos manter as recomendações de reportar o tempo de treinamento do modelo e os recursos computacionais utilizados, incluindo também a sensibilidade dos hiperparâmetros envolvidos no processo.

A intenção, então, foi estudar as possibilidades das ferramentas de geração de língua natural e sua aplicação à mídia brasileira. Essas possibilidades são variadas e devem ser melhor compreendidas pela sociedade em geral, inclusive para que se estabeleça regras para sua regulação, como é o caso discutido sobre a regulamentação das mídias. Com o aprofundamento do conhecimento dessas ferramentas, com certeza teremos um ambiente digital mais democrático e inclusivo. E esse avanço demanda necessariamente uma predominância maior de uma comunidade de pesquisadores, programadores e, enfim, toda a sociedade civil em torno de uma atenção maior a ferramentas de uso tão alastrado como é o caso das redes sociais.

E, por fim, o robô jornalista criado para este trabalho apresentou uma boa fluência com os exemplos criados e demonstrou capacidade de lidar com o *corpus* linguístico para a criação de frases autônomas. O percurso de programação foi muito proveitoso e ilustrativo para que possamos entender melhor o que é possível fazer com as ferramentas de inteligência artificial, que passam não só pela geração de língua natural mas também por outros caminhos, como a identificação de discursos de ódio, possibilidade citada brevemente neste trabalho mas que pode também motivar por si só pesquisas mais complexas e necessárias, tendo em vista os caminhos seguidos por diversas redes sociais.

Com nossa pesquisa, esperamos poder dar início a mais ideias que sejam desenvolvidas adiante, em outras oportunidades de pesquisa. Pretendemos continuar a desenvolver um trabalho que envolva a Geração de Língua Natural e que amplie a compreensão deste campo da Inteligência Artificial e de suas possibilidades para os mais variados campos de conhecimento da sociedade.

## REFERÊNCIAS

CASTRO FERREIRA, Thiago; VAN DER LEE, Chris; VAN MILTENBURG, Emiel; KRAHMER, Emiel. Neural Data-to-Text Generation: A Comparison between Pipeline and End-to-End Architectures. Proceedings of the 2019 **Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019. <https://doi.org/10.18653/v1/d19-1052>

CHATZILYGEROUDIS, Konstantinos; HATZILYGEROUDIS, Ioannis; PERIKOS, Isidoros. Machine learning basics. In: Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice. 2021. p. 143-193.

GATT, Albert; KRAHMER, Emiel. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research**, n. 61, p. 65–170, 2018.

LI, Hang, Deep learning for natural language processing: advantages and challenges, **National Science Review**, Volume 5, Issue 1, January 2018, Pages 24–26, <https://doi.org/10.1093/nsr/nwx110>

OREMUS, Will. The First News Report on the L.A. Earthquake Was Written by a Robot. **Slate**, [S. l.], p. 1, 17 mar. 2014. Disponível em: <https://slate.com/technology/2014/03/quakebot-los-angeles-times-robot-journalist-writes-article-on-la-earthquake.html>. Acesso em: 11 dez. 2022.

RAJAPAKSE, Thilina. BART for Paraphrasing with Simple Transformers. [S. l.], 5 ago. 2020. Disponível em: <https://towardsdatascience.com/bart-for-paraphrasing-with-simple-transformers-7c9ea3dfdd8c>. Acesso em: 11 dez. 2022.

RAMÓN SALAVERRÍA CONVOCA ACADEMIA A CONTRIBUIR COM SOLUÇÕES PARA A CRISE DO JORNALISMO. **Jornal Intercom**, [S. l.], p. 1, 5 set. 2019. Disponível em: <https://www.portalintercom.org.br/publicacoes/jornal-intercom/2019-2/09-2-2-2-2/ano-15-n-480-sao-paulo-05-de-setembro-de-2019-issn-1982-372/chamadas-1557/intercom-2019-ramon-salaverria-convoca-academia-a-contribuir-com-solucoes-para-a-crise-do-jornalismo>. Acesso em: 11 dez. 2022.

ROBOT writes LA Times earthquake breaking news article. **BBC**, [S. l.], p. 1, 18 mar. 2014. Disponível em: <https://www.bbc.com/news/technology-26614051>. Acesso em: 11 dez. 2022.

SALAVERRÍA, Ramón; BUSLÓN, Nataly; LÓPEZ-PAN, Fernando; LEÓN, Bienvenido; LÓPEZ-GOÑI, Ignacio; ERVITI, María-Carmen (2020). “Desinformación en tiempos de pandemia: tipología de los bulos sobre la Covid-19”. **El profesional de la información**, v. 29, n. 3, e290315.

USZKOREIT, Jakob. Transformer: A Novel Neural Network Architecture for Language Understanding. **Google Research**, [s. l.], 31 ago. 2017. Disponível em: <https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>. Acesso em: 11 dez. 2022.