# Automated Text Analysis

## Summer Institute in Computational Social Science (SICSS) 2025

Felipe Lamarca

Instituto de Estudos Sociais e Políticos (IESP-UERJ)

2025-07-02

# Motivação

Nas ciências sociais, em particular, e em várias outras ciências, em geral, textos são fontes de **evidência empírica**:

- Discursos de políticos
- Publicações em redes sociais
- Respostas a perguntas abertas em surveys
- Documentos históricos
- ...

Intuitivamente, a ideia do "texto como dado" é natural, para não dizer óbvia.

# Text as data

Podemos analisar textos de várias formas. Nas ciências sociais, a análise de texto por muito tempo esteve restrita a métodos qualitativos: análise de conteúdo, de discurso e assim por diante.

A análise quantitativa de texto é uma oportunidade de analisar corpus textuais **maiores** e de forma mais sistemática.

# A evolução do text as data

Até recentemente, a análise quantitativa de texto era estritamente baseada em métodos como:

- Bag of Words
- Contagem de palavras
- Métodos baseados em dicionário
- Modelagem de tópico
- N-gramas

# A evolução do text as data

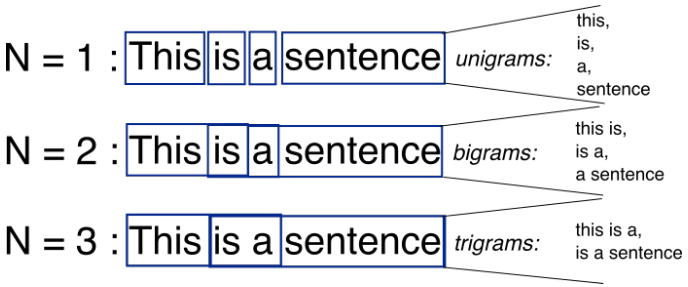


Figure 1: N-grams

| | about | bird | heard | is | the | word | you |
|---|---|---|---|---|---|---|---|
| About the bird, the bird, bird bird bird | 1 | 5 | 0 | 0 | 2 | 0 | 0 |
| You heard about the bird | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| The bird is the word | 0 | 1 | 0 | 1 | 2 | 1 | 0 |

Figure 2: Bag of Words

São, é claro, métodos muito úteis. Mas são incapazes de incorporar **contexto** e **significado** de frases e palavras.

# Embeddings: textos $\rightarrow$ vetores semânticos

O que diferencia um banco de um banco?


Figure 3: Banco

Figure 4: Banco

# Embeddings: textos $\rightarrow$ vetores semânticos

Modelos de linguagem precisam representar palavras e frases como vetores numéricos para processá-las. Esses vetores são chamados de **embeddings**.

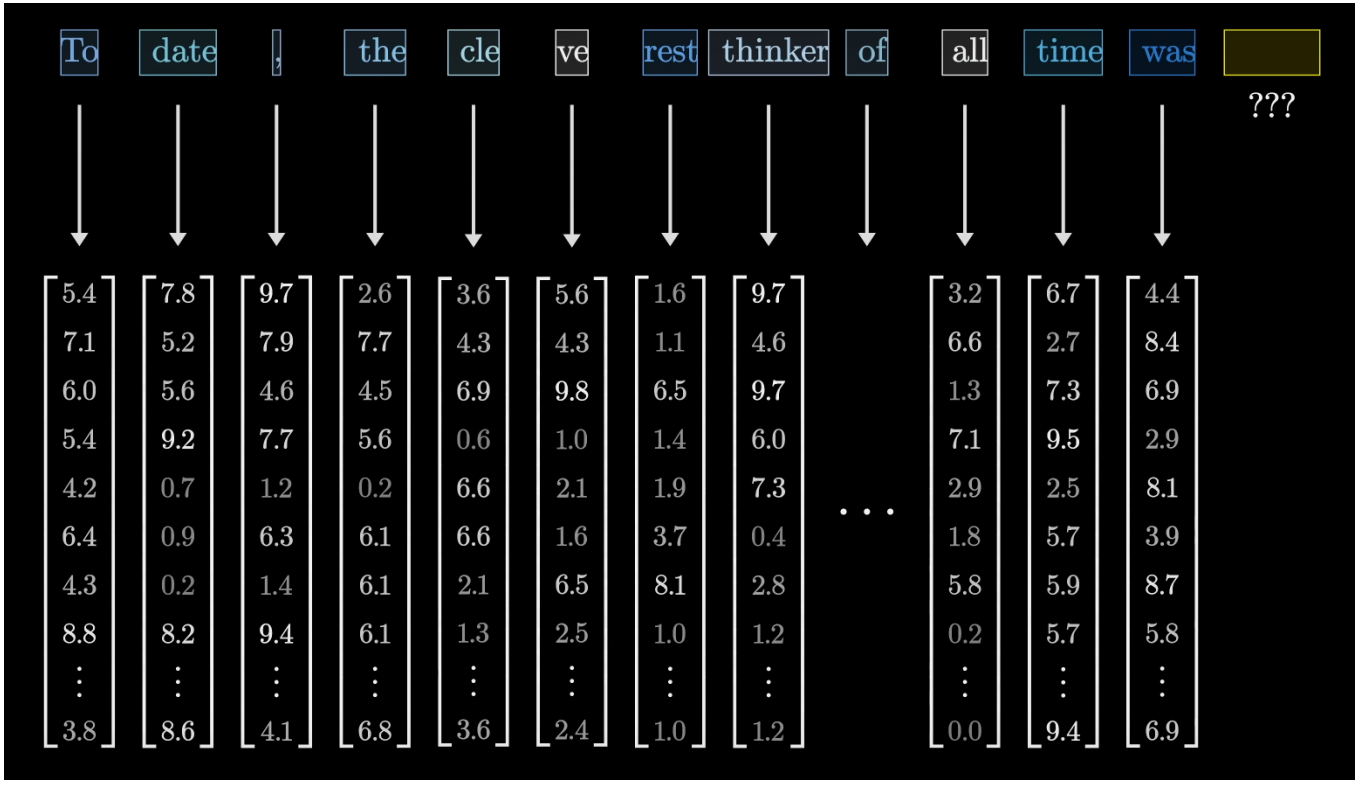


Figure 5: Word Embeddings – Fonte: 3b1b

# Embeddings: textos $\rightarrow$ vetores semânticos

A ideia é que, ao treinarmos um modelo de linguagem, a representação vetorial de palavras e frases passa a capturar algum significado semântico.

Embedding Projector

# Transformer: o que é?

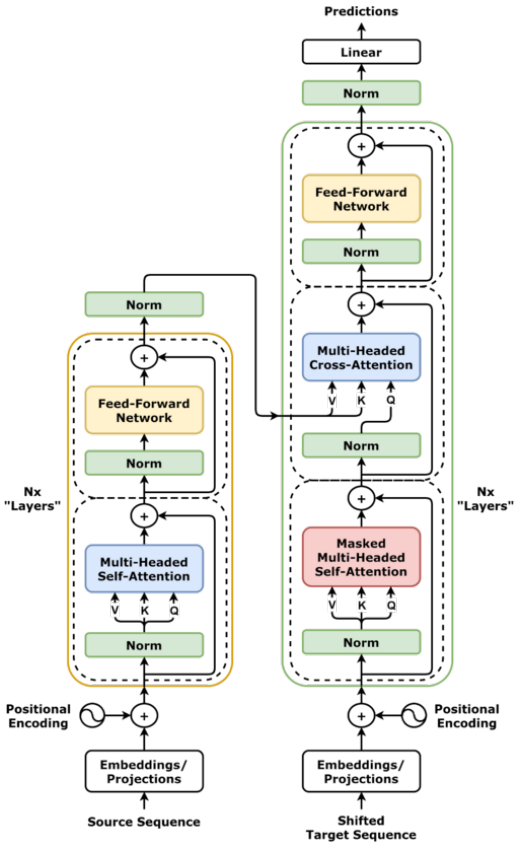


Figure 6: Arquitetura de um Transformer

*Transformers* são uma arquitetura de rede neural que revolucionou o campo de NLP ao permitir que os modelos aprendam relações contextuais entre palavras em uma frase a partir do mecanismo de *attention*. [link]

# Transformer: o que faz?

Na prática, os *transformers* são capazes de prever a próxima palavra em uma frase, levando em consideração o contexto de todas as palavras anteriores.

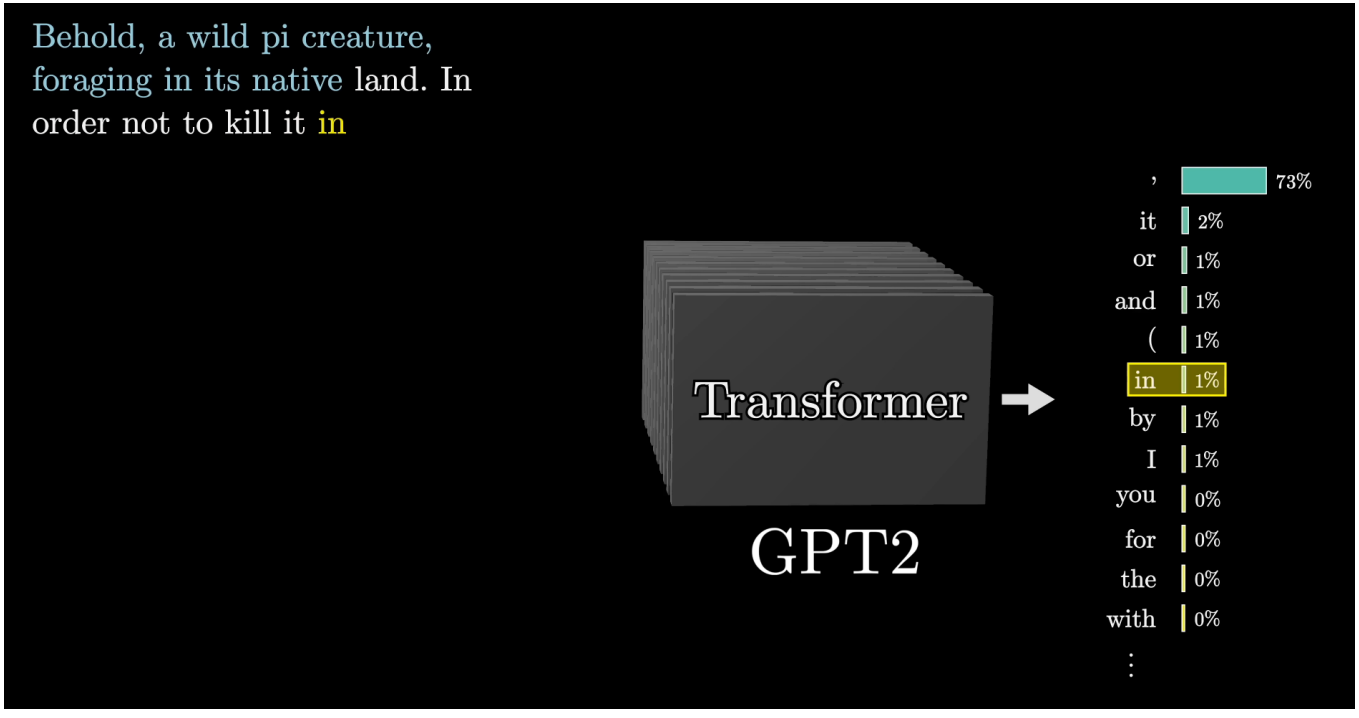


Figure 7: Transformer – Fonte: 3b1b

# Fine-tuning: como adaptar modelos para tarefas específicas

A técnica de fine-tuning envolve ajustar um modelo pré-treinado em um conjunto de dados específico para melhorar seu desempenho em uma tarefa particular. Isso é especialmente útil quando se tem um conjunto de dados limitado, mas ainda assim se deseja aproveitar o conhecimento adquirido pelo modelo durante o treinamento inicial.

# Modelos “grandes” e modelos “pequenos”

Em geral, modelos de linguagem podem ser “grandes” ou “pequenos”:

- **Modelos grandes:** treinados com grandes quantidades de dados, geralmente em várias línguas, e com muitos parâmetros. Exemplo: `llama3.1:405b`.
- **Modelos pequenos:** treinados com menos dados e com menos parâmetros. Menor capacidade de generalização. Exemplo: `llama3.2:1b`.

# Modelos “grandes” e modelos “pequenos”

Em geral, modelos de linguagem podem ser “grandes” ou “pequenos”:

- **Modelos grandes:** treinados com grandes quantidades de dados, geralmente em várias línguas, e com muitos parâmetros. Exemplo: `llama3.1:405b`.
- **Modelos pequenos:** treinados com menos dados e com menos parâmetros. Menor capacidade de generalização. Exemplo: `llama3.2:1b`.

**Saiba**

Em geral, quanto maior o modelo, melhor o desempenho; no entanto, menores são as chances de que você consiga rodar o modelo no seu próprio computador.

# No que usar?

O céu é o limite! Alguns exemplos:

- *Tradução* de textos
- *Transcrição* de textos
- *Classificação* de textos
- *Cálculo de distância* (i.e., similaridade) entre textos
- Reconhecimento de entidades nomeadas
- Sistemas de recomendação
- E por aí vai...

# Como usar?

Beleza, Felipe, já entendi. Mas como eu uso esse negócio?

Opções open-source

- HuggingFace
- Ollama
- Meta

Opções pagas

- Modelos open-source via Groq (com limite gratuito diário)
- ChatGPT, da OpenAI
- Grok, da xAI
- Claude, da Anthropic
- Sabiá, da Maritaca AI
- ...

# Indicações de material

[link] Can Large Language Models Transform Computational Social Science? (2024)

[link] Do 'texto como texto' ao 'texto como dado': o potencial das pesquisas em Relações Internacionais (2022)

[link] Série de vídeos do 3Blue1Brown sobre redes neurais e *transformers*.

[link] How Much Does It Cost to Train Frontier AI Models? (2024)

# Contato

**Email:** felipe.lamarca@hotmail.com
**Site:** felipelamarca.com
**LinkedIn:** felipe-lamarca