Dataset para avaliação de sumarização de longos documentos em português

Leandro Carísio Fernandes
Guilherme Zeferino Rodrigues Dobins

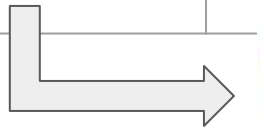
Descrição do Projeto

Criação de dataset contendo:

- (TEXTO): transcrição de audiências públicas realizadas na Câmara dos Deputados
- (SUMÁRIO): notícias sobre essas audiências
- (SUMÁRIO ESTRUTURADO): metadados extraídos das notícias indicando opiniões:quem disse/apoia o quê

Comparação com outros datasets em português

Dataset	Tamanho	Características
TeMário 2003	250	Artigos de jornais com sumários escritos por um especialista. Artigos com ~600 palavras e sumário com 25-30% desse tamanho.
CSTNews 2011	140	Artigos de jornais com anotação manual. ~330 palavras por entrada
RulingBR 2018	10k	Decisões do STF (ementa, voto, acórdão e relatório). Considera a ementa como o sumário. A tarefa é tentar extrair a decisão do texto completo. ~1200 palavras por entrada
RecognaSumm 2024	130k	Artigos de jornais. Considera que o sumário é o título + subtítulo do artigo.
Cem Mil Podcasts 2022	+100k	Transcrição de podcasts em português. Considera que o sumário é a descrição do episódio fornecida pelo criador.

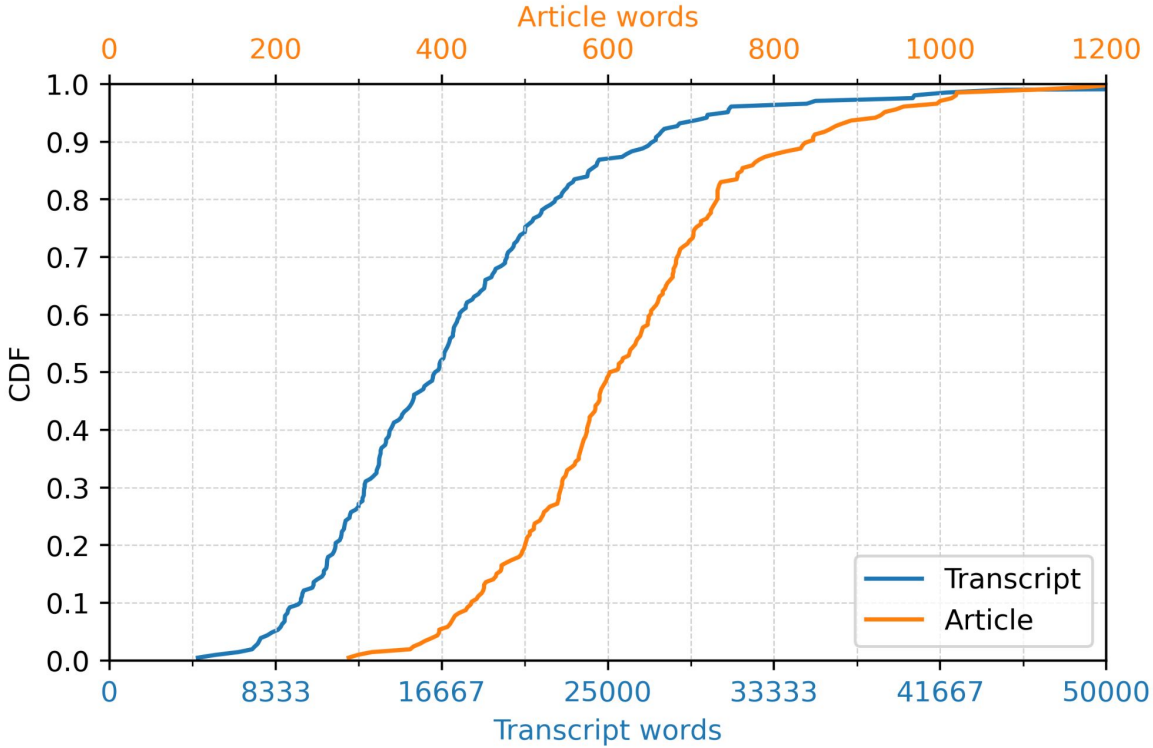


Unfortunately, due to shifting priorities, we no longer maintain the dataset. As of December 2023, we no longer take requests to access it.

Características do dataset proposto

Statistic	Transcript (words)	Article (words)	Relative size of article compared to the transcription (%)	Individuals (count)	Opinions (count)
Min	4 437	288	0.75	2	3
Q (5%)	8 495	399	1.76	3	5
Q (25%)	12 079	522	2.75	4	8
Median	16 424	607	3,97	5	10
Q (75%)	20 858	706	5.24	6	13
Q (95%)	30 761	932	7.21	8.75	18.75
Max	147 728	12 15	15.07	12	31

Características do dataset proposto



Avaliação da extração do GPT

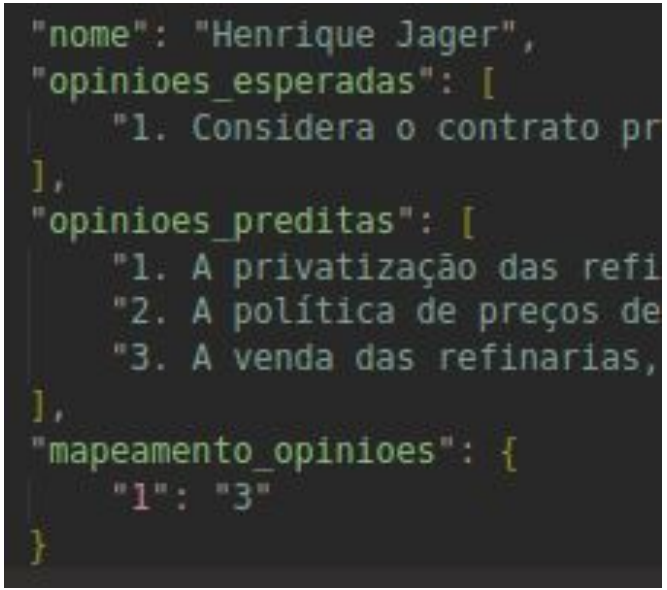
- Concluímos a avaliação da extração de opiniões pelo ChatGPT.

- Devido ao fato de o processo de avaliação ser parcialmente realizado por LLMs, observamos o efeito de alucinações nesse pipeline.

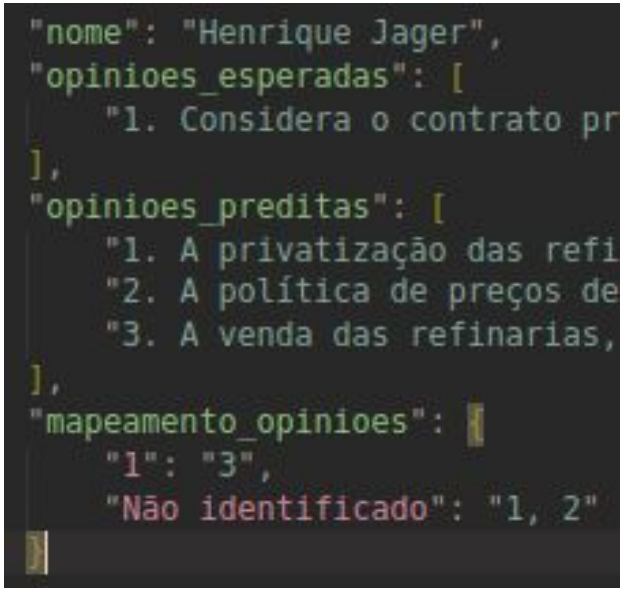
- Realizamos um pós-processamento dos resultados da avaliação para corrigir os erros, o que resolveu esse problema.

Alucinações na avaliação

Esperado:

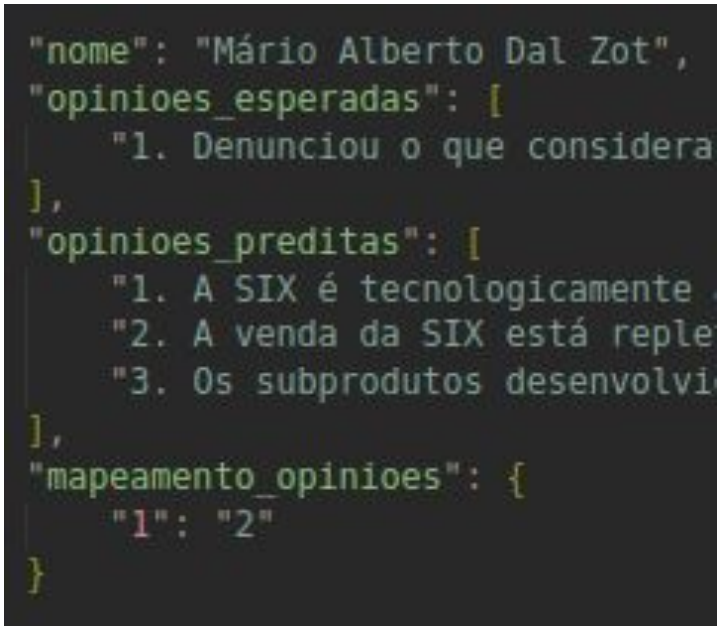


Obtido:

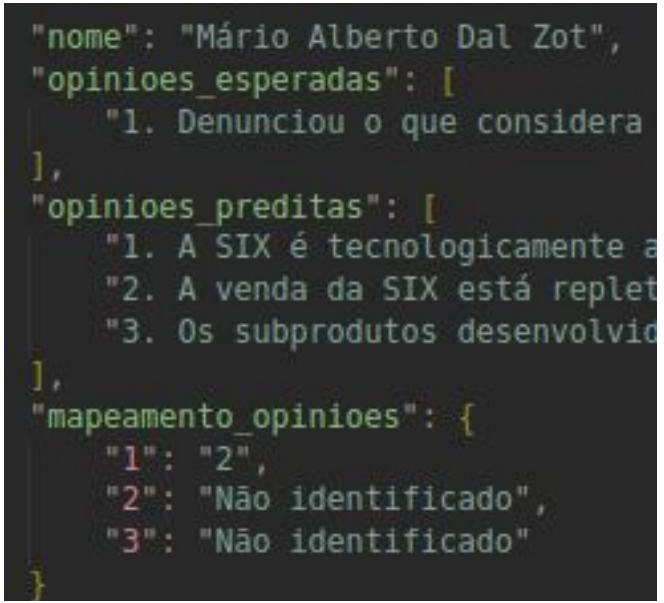


Alucinações na avaliação

Esperado:



Obtido:



Resultados - recall e precisão

Após o pós-processamento dos resultados da avaliação, foi possível calcular o recall e precisão médios obtidos pelo ChatGPT:

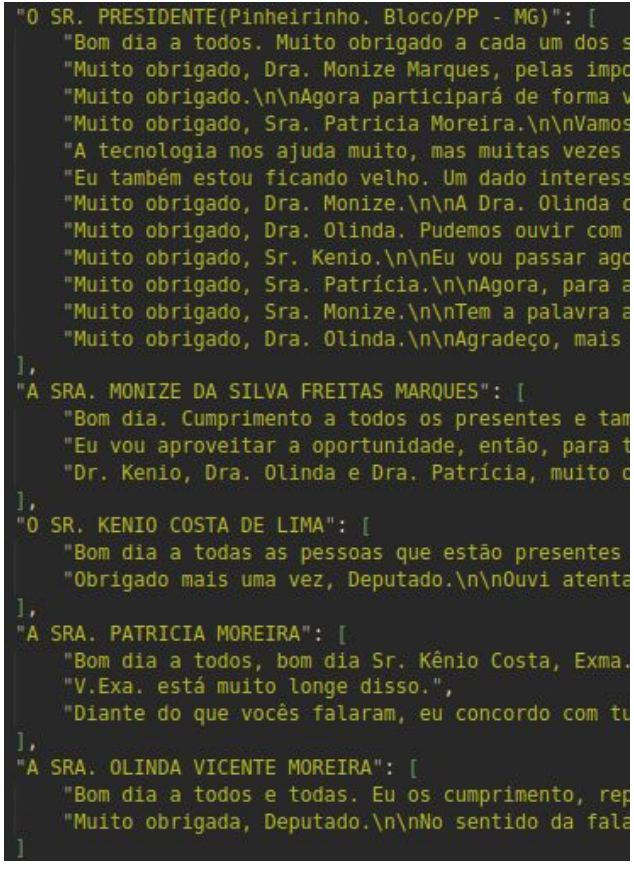
- Recall médio: 0.45
- Precisão média: 0.25

Verificação de alucinação nas opiniões extraídas

Proposta para verificar presença de alucinações: RAG

1. Aplica regex na transcrição para separar falas por pessoa
2. Para cada pessoa de cada transcrição, separa a fala em chunks e indexa utilizando a ferramenta NeuralSearchX
3. Para cada opinião retornada da pessoa, busca os top-k chunks mais relacionados à opinião no índice específico
4. Envia para um LLM a opinião e os top-k chunks e pede para informar se a opinião pode ser inferida a partir dos top-k chunks. Se não puder, considera que houve alucinação

Segmentação das transcrições



Verificação de alucinação nas opiniões extraídas

Próximos passos:

- Concluir a indexação dessas falas
- Criar o pipeline de RAG para verificar alucinações.