DataMinds - Análise de Discursos da ONU

Anderson Nogueira Silva - 2126516
Eric Yutaka Fukuyama - 2126567

[1]Departamento Acadêmico de Informática (DAINF) –
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 – 80230-901 – Curitiba – PR – Brasil

1. Introdução

O tema "Análise dos discursos de diferentes nações no Debate Geral da Assembleia Geral da ONU", com ênfase nos países-membros do G7 e do bloco BRICS, é relevante pois permite compreender como cada país se posiciona política e economicamente ao longo da história. Além disso, possibilita traçar paralelos entre os países desses blocos para entender as convergências e divergências em suas agendas globais. Dentre os principais objetivos do estudo, destacam-se:

- Identificar quais tópicos globais foram mais discutidos por cada bloco (majoritariamente G7 e BRICS) em diferentes períodos.
- Analisar como mudanças geopolíticas, crises econômicas ou políticas impactam a retórica de cada grupo.
- Mapear os principais temas econômicos ao longo do tempo, evidenciando semelhanças e contrastes entre os blocos.
- Investigar se os sentimentos dos discursos de países aliados convergem ou divergem ao longo do tempo.

2. Processamento dos Dados

Nesta seção, são apresentados: a descrição do corpus de modo geral e sua composição, bem como o pré processamento dos dados realizado para permitir uma análise adequada de suas informações.

2.1. Descrição Geral

O corpus utilizado nesta pesquisa é o UN General Debate Corpus (UNGDC) [Harvard Dataverse 2024], uma coleção abrangente de mais de 10.000 discursos de 202 países, cobrindo o período de 1946 a 2023, cuja última atualização ocorreu em agosto de 2024.

Os dados incluem o texto completo dos discursos — assuntos de relevância política, econômica e social — e metadados com informações sobre o país, ano do discurso, número da sessão, orador e o cargo ocupado por ele.

Ao todo, o UNGDC contabiliza 10.760 discursos, sendo que 857 pertencem a países do G7 e do BRICS. Brasil, Estados Unidos, Canadá e Grã-Bretanha comparecem em todas as sessões, enquanto China, Índia e Rússia ficam ausentes em apenas uma das 78 sessões. Os demais países desses grupos possuem menos discursos no total. Em média, cada discurso contém 2.931 palavras, com o maior possuindo 22.000 palavras e o menor apenas 423.

2.2. Processamento

O processamento realizado foi feito baseado na seguinte sequência de passos descrita a seguir

1. **Obtenção dos discursos**: cada pasta, ao extrair o arquivo compactado do corpus UNGDC, corresponde a um ano de conferência da Assembleia Geral da ONU, no qual os arquivos de texto, dentro de cada pasta, possuem a nomenclatura padronizada com a sigla do país, de acordo com [International Organization for Standardization 2020], o número da sessão e o ano do discurso. Portanto, itera-se sobre essas pastas extraindo o ano, país e o texto do discurso.
2. **Remoção de símbolos**: realiza-se uma limpeza genérica de símbolos indesejados, descritos a seguir —, –, “, ”, ’, ‘, . . . , ,, ., !, ?, ;, :, (,), [,], , , ¡, ¿, /, —, @, #, $, %, &, *, +, ’s, ”, ; ”, ´.
3. **Remoção de *stopwords***: efetua-se uma extração de todas as *stopwords* do idioma dos discursos (inglês).
4. **Lematização**: pré-processamento do texto com lematização para reduzir as palavras à sua forma base (lema).
5. **Agrupamento dos discursos por bloco econômico**: após a etapa anterior os dados já estão aptos para serem trabalhados, entretanto, realiza-se uma etapa adiciona de agrupamento dos discursos dos países de acordo com seus blocos econômicos (e.g. BRICS, G7, NATO etc).

3. Resultados

Nesta seção, estão descritos os resultados obtidos, sendo apresentados na seção correspondente ao objetivo ao qual eles atendem.

3.1. Identificar quais tópicos globais foram mais discutidos

Para alcançarmos os resultados deste objetivo, a biblioteca [LDA Over Time 2024] foi utilizada extraiu tópicos bastante relevantes. Para caracterização de um tópico extraído, optou-se por utilizar as 100 palavras mais frequentes de cada um dos tópicos em uma solicitação ao [OpenAI 2025], para uma descrição geral do tópico com até 2 palavras. Após uma análise geral das palavras mais frequentes, as descrições obtidas na solicitação foram satisfatórias e bastante coerentes.

Os resultados obtidos para este objetivo, consistem na análise da evolução da relevância dos tópicos extraídos pelo LDA Over Time nos discursos ao longo de 1946-2023, portanto: a primeira parte, consiste na representação global, isto é, considerando todos os países; para a segunda parte, extraiu-se apenas os tópicos dos países pertencentes ao grupo econômico G7 e, para a terceira parte, apenas os países do BRICS.

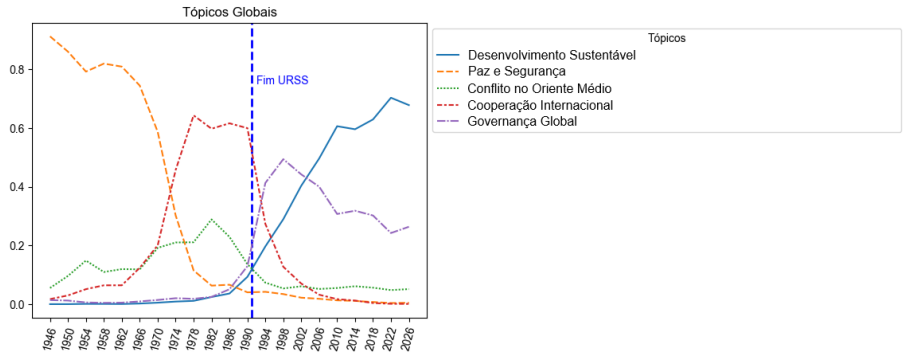
3.1.1. Tópicos Globais

Na figura 1, temos a evolução da relevância dos tópicos extraídos pelo LDA Over Time, considerando discursos de todos os países, para o período de 1946-2023. Também nesta figura, observamos a categorização atribuída a cada um deles. Pode-se visualizar

que durante determinados períodos existem tópicos que são mais relevantes, como durante o período da Guerra Fria (1947-1991) destaca-se mais o tópico de Cooperação Internacional e logo após o fim da Segunda Guerra Mundial (1939-1945), há uma relevância maior no tópico de Paz e Segurança.

Por extrair tópicos dos discursos de todos os países, podemos visualizar um tópico específico "Conflito no Oriente Médio", relevante naquele período. Por fim, tópicos de "Desenvolvimento Sustentável" e de "Governança Global" surgem com maior relevância próximo da dissolução da União das Repúblicas Socialistas Soviéticas - URSS (1991), fator relevante para alterações nas temáticas nos debates gerais da da ONU.

Figura 1. LDA Over Time: tópicos globais.

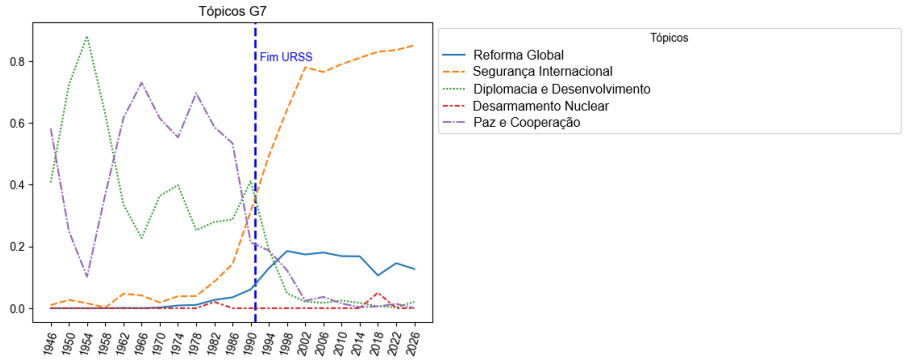


Fonte: Autores.

3.1.2. Tópicos G7

Na figura 2 temos os tópicos extraídos para o G7. Os tópicos relevantes conseguimos associar historicamente e fazem sentido. São eles: "Diplomacia e Desenvolvimento" e "Paz e Cooperação", pois aparecem no período da Guerra Fria. Com o fim deste conflito, nota-se a relevância desses tópicos serem atribuídas a outros como Segurança Internacional e Reforma Global. Em 2018 há a aparição de um tópico intitulado por Desarmamento Nuclear que muito provavelmente está associado com testes de mísseis nucleares realizados pela Coreia do Norte naquele período.

Figura 2. LDA Over Time: tópicos G7.

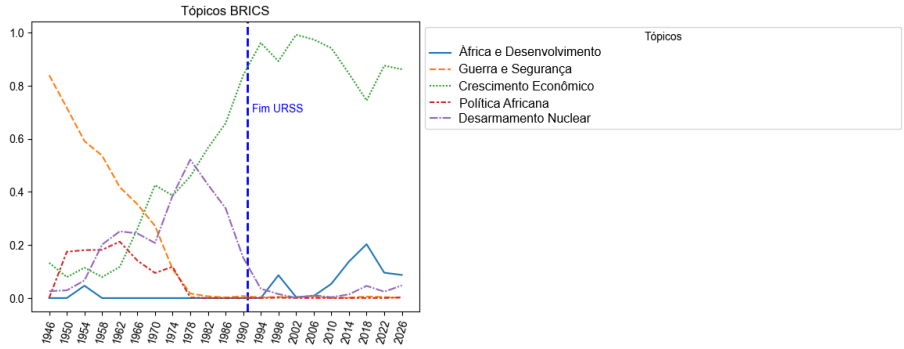


Fonte: Autores.

3.1.3. Tópicos BRICS

Os tópicos extraídos para o BRICS, podem ser vistos na figura 3. Talvez por se tratarem de países em desenvolvimento econômico, o tópico de "Crescimento Econômico" sempre esteve presente nos discursos e aumentou sua relevância ao longo dos anos. Além disso, os tópicos de "Guerra e Segurança" e "Desarmamento Nuclear" diminuem a sua relevância com o fim da Segunda Guerra Mundial e com o fim da URSS (1991), respectivamente.

Figura 3. LDA Over Time: tópicos BRICS.



Fonte: Autores.

3.2. Analisar como mudanças geopolíticas impactam os discursos

Neste objetivo, visamos entender como acontecimentos geopolíticos impactaram os discursos dos países. Para isso, procuramos verificar este impacto na semelhança (utilizando a similaridade do cosseno) dos discursos, entre países de um mesmo bloco econômico, e como ela evoluiu ao longo do tempo, no período de 1946-2023.

Como visualizamos nas figuras 1, 2 e 3, o fim da URSS (1991) foi impactante na temática dos discursos dos países, sendo que os tópicos de maior relevância, após esse acontecimento, tratam de desenvolvimento, segurança e reformas globais, demonstrando maior sintonia dos tópicos abordados entre os países. Além disso, é possível perceber picos nas relevâncias de determinados tópicos em anos que aconteceram eventos impactantes, como em 2018 com os testes nucleares da Coreia do Norte (figura 2).

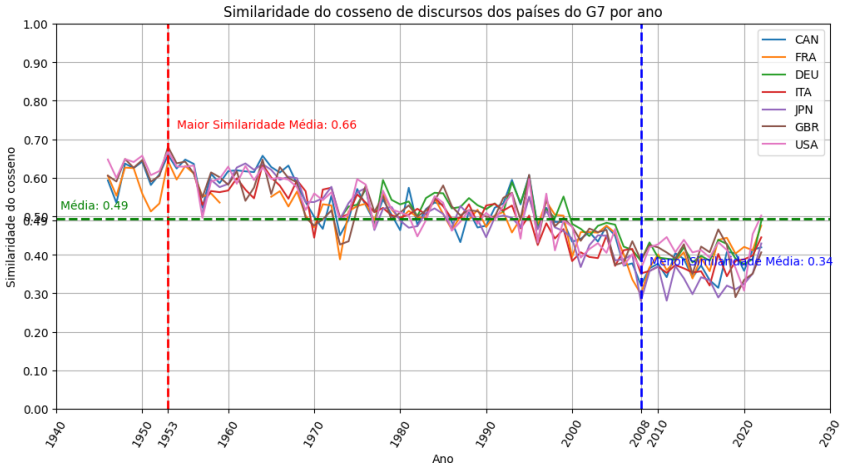
Desta forma, calculou-se para cada discurso de cada país, por ano, sua similaridade, com relação aos discursos dos outros países daquele bloco econômico, para o mesmo ano. Por fim, calculou-se a média dessas similaridades para cada país.

3.2.1. Similaridade G7

Na figura 4, evidentemente há uma similaridade maior nas primeiras assembleias gerais da ONU entre países do G7, a qual decai até sofrer um aumento após 2020, provavelmente devido ao advento da pandemia do COVID-19. Esse comportamento, talvez possa ser explicado pelo fim da Segunda Guerra Mundial (1945), no qual os países estavam mais alinhados em suas temáticas, para demonstrar um senso de aliança pós guerra.

Passado alguns anos, essas temáticas se alteraram aos poucos, até atingir uma pontuação de similaridade mínima em 2008.

Figura 4. Similaridade Cosseno: discursos G7.

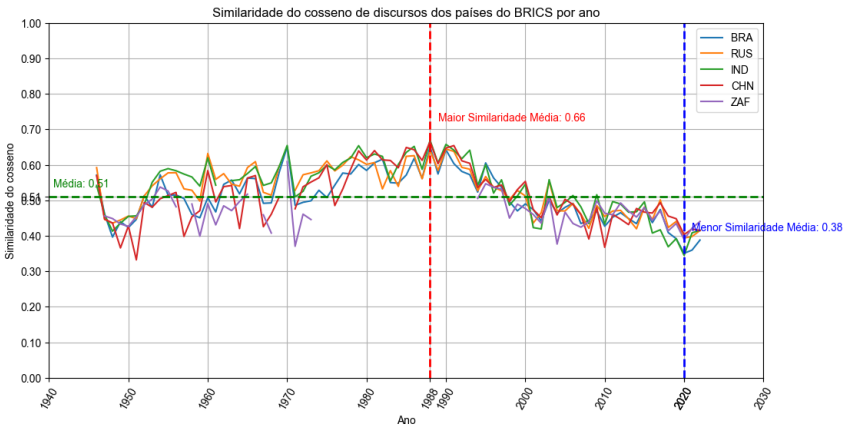


Fonte: Autores.

3.2.2. Similaridade BRICS

A similaridade para os países do BRICS está disposto na figura 5. Neste caso, verifica-se um efeito distinto daquele da figura 4. Temos a similaridade evoluindo conforme as assembleias acontecem, atingindo seu pico máximo em 1988. Supomos que por se tratar de países em desenvolvimento, foi necessário alinharem os temas e objetivos comuns para maior colaboração. Entretanto, um contraponto a esta suposição, é o fato de que a África do Sul (ZAF) não possui discursos no período de 1974-1994, comprometendo a medida de similaridade do cosseno da maneira utilizada, além disso, em anos mais recentes, a similaridade diminuiu. Por fim, percebe-se após 2020, um leve aumento que também supomos ser explicado ao advento da pandemia do COVID-19.

Figura 5. Similaridade Cosseno: discursos BRICS.



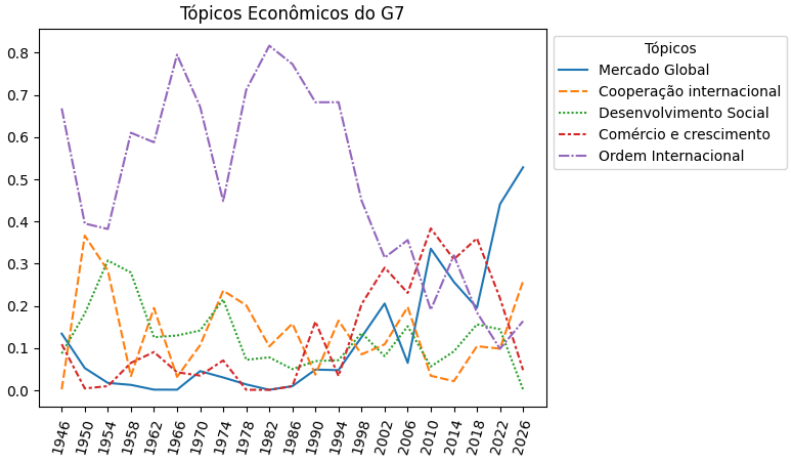
Fonte: Autores.

3.3. Mapear temas econômicos ao longo do tempo

Para compreender os tópicos econômicos ao longo do tempo para ambos os blocos também foi usado o LDA Over Time [LDA Over Time 2024]. No entanto, foi feita uma limpeza prévia do Corpus para capturar informações apenas sobre assuntos econômicos. Tal limpeza foi feita buscando palavras relacionadas a "economy"usando o conceito de similaridade. Dessa forma, palavras consideradas semelhantes foram selecionadas. A partir dessa lista, no Corpus original sem limpeza foi buscado esses tópicos junto com as suas variações(inflections) e manteve-se apenas as sentenças que tinham essas palavras. Depois, foi aplicado o LDA Over Time nesse Corpus tratado. Por fim, foi pedido para o ChatGPT [OpenAI 2025] encontrar os tópicos relacionados as 100 palavras mais frequentes.

Dessa forma, o resultado pode ser visto nas figuras 6 e 7. Importante notar que tópicos como "Comércio e Crescimento"e "Cooperação Internacional"aparecem nos dois blocos. Inclusive, "Cooperação Internacional"possui uma curva parecida mostrando a importância que países desses dois blocos tendem a focar nesse tópico ao longo do tempo. Ainda, a curva de "Comércio e Crescimento"apresentou uma crescente nos dois blocos após o fim da guerra fria. Todavia, no BRICS o crescimento foi ainda maior por ser um grupo de países emergentes. Em adição a isso, o tópico de "Desenvolvimento Econômico", sempre presente no BRICS não é visto nos tópicos do G7, provavelmente por já serem países desenvolvidos. Ainda, "Ordem Internacional"esteve em uma decrescência para o G7 depois do fim da guerra fria, mostrando então que era um tópico prezando a paz naquele período.

Figura 6. Tópicos Econômicos do G7

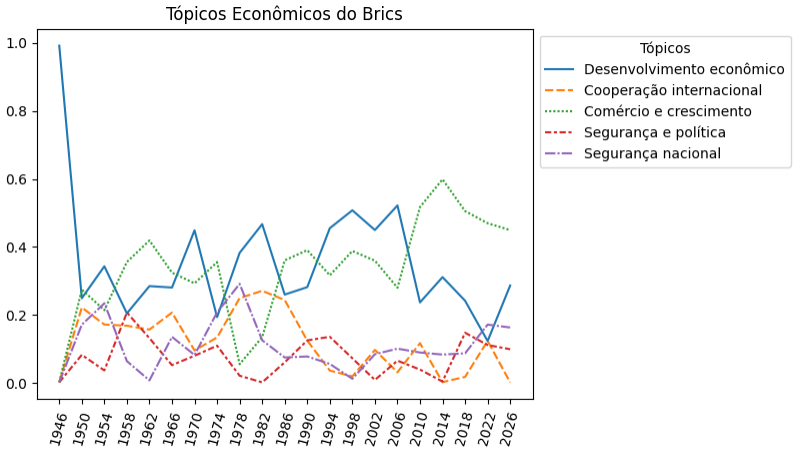


Fonte: Autores.

3.4. Investigar sentimentos nos discursos de países aliados

Para a análise de sentimentos, foi utilizado o modelo BERT disponibilizado pela biblioteca Hugging Face Transformers[Hugging Face 2025c]. Como os discursos da ONU frequentemente possuem textos extensos, foi adotada a estratégia de segmentar os discursos em "chunks", respeitando a limitação de 512 tokens do modelo. Para isso, o discurso foi segmentado em sentenças e então analisado separadamente pelo BERT. Cada

Figura 7. Tópicos Econômicos do BRICS



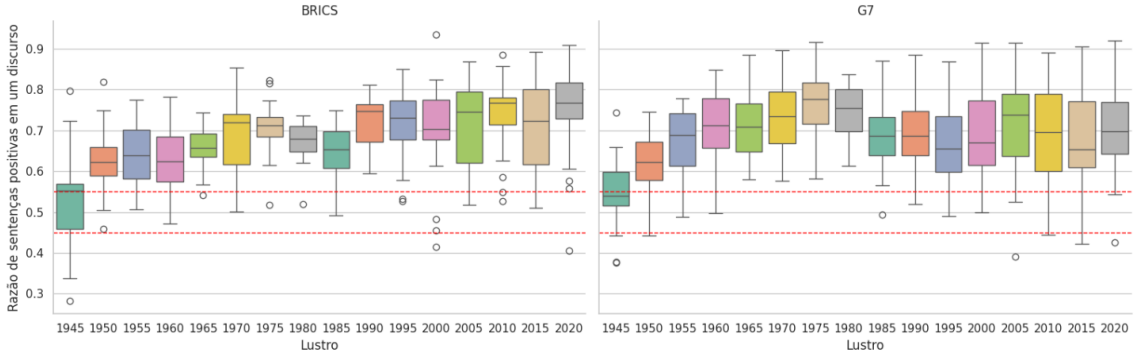
Fonte: Autores.

saída do modelo retorna dois valores: (i) um rótulo indicando se o sentimento da sentença era predominante positivo ou negativo e (ii) uma porcentagem, representando o grau de confiança do modelo na classificação atribuída.

Após essa etapa, atribuiu-se o valor +1 para cada chunk classificado como positivo, -1 para cada chunk classificado como negativo e 0 caso a porcentagem de certeza da classificação fosse menor que 0.6. A soma média desses valores resultou em uma pontuação final que representava a predominância do sentimento positivo. Dessa forma, a pontuação foi mapeada para ficar entre os valores 0 e 1 para representar o porcentual de sentenças positivas naquele discurso. Se o percentual de sentenças positivas fosse maior que 55% o discurso inteiro seria classificado como positivo. Caso fosse menor que 45% o discurso seria classificado como negativo e entre esses valores o discurso foi considerado como neutro.

Dessa forma, a figura 8 representa o *box plot* agrupados por lustro(período de 5 anos) e grupo com o *threshold* imposto para definir o discurso como neutro. Vale notar que a maior parte dos discursos é positivo e neutro, o que faz sentido visto que os discursos da ONU são discursos diplomáticos que visam a esperança e cooperação internacional.

Figura 8. Análise de sentimento: discursos BRICS e G7.



Fonte: Autores.

No entanto, entender os discursos negativos foi visto como importante. Dessa forma, 7 dos 14 discursos negativos encontrados foram logo após a segunda guerra mundial. Os 7 discursos restantes com as possíveis explicações sobre o sentimento negativo segue abaixo:

- Canadá(1951): Guerra da Coreia e a presença forte do Canadá nesse contexto;
- India(2002): Tensões com o Paquistão com ataque ao parlamento indiano;
- França(2009 e 2011): Crise financeira e intervenção internacional na Líbia;
- Japão(2017): Testes nucleares feitos pela Coreia do Norte;
- Grã-Bretanha(2020): Brexit;
- Russia(2022): Invasão na Ucrânia;

Por fim, é notável que a convergência de discursos por grupo acontece para um lado mais positivo observando o resultado. Todavia, tal convergência parece acontecer em um âmbito mais geral devido a natureza dos discursos da ONU e não somente da relação entre os países do mesmo grupo.

4. Limitações e Trabalhos Futuros

O trabalho apresenta limitações como o foco em apenas dois blocos econômicos. Além da não aplicação de um modelo pré-treinado especificamente para discursos políticos.

Portanto, para futuros trabalhos poderia ser feito um estudo com outros blocos econômicos além de outros tipos de associações como as geográficas ou países aliados com o mesmo governo político. Ainda, para o objetivo de análise de sentimento poderia ser útil pré treinar o modelo do Bert com discursos políticos. Ainda, utilizar ferramentas que possibilitem uma maior quantidade de leitura de tokens de uma vez como o longformer [Hugging Face 2025a] ou o reformer [Hugging Face 2025b]. Ainda, procurar por sentimentos em assuntos em específico poderia ser útil para entender o posicionamento de cada país para tópicos adversos.

Referências

Harvard Dataverse (2024). UNGeneral Debate Corpus (UNGDC), 1946-2023. Atualizado em agosto de 2024. Acesso em 24 out. 2024.

Hugging Face (2025a). Longformer: Efficient Attention for Long Documents. Acesso em: 17 fev. 2025.

Hugging Face (2025b). Reformer: Efficient Transformer with Reversible Layers. Acesso em: 17 fev. 2025.

Hugging Face (2025c). Transformers: State-of-the-Art Natural Language Processing for PyTorch, TensorFlow, and JAX. Acesso em: 17 fev. 2025.

International Organization for Standardization (2020). ISO 3166-1 alpha-3: Codes for the representation of names of countries and their subdivisions – Part 1: Alpha-3 code. ISO Standard, Geneva, Switzerland.

LDA Over Time (2024). LDA Over Time: A GitHub Repository. Acesso em: 24 nov. 2024.

OpenAI (2025). ChatGPT: A Large Language Model for Conversational AI. Acesso em: 17 fev. 2025.