UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ (UTFPR)

ERIC YUTAKA FUKUYAMA
ERICK JOSE TELES DE ANDRADE
ANDERSON NOGUEIRA SILVA

ANÁLISE EXPLORATÓRIA

PROCESSAMENTO DE LINGUAGEM NATURAL

CURITIBA

2024

SUMÁRIO

1 INTRODUÇÃO

1.1 TEMA

Este trabalho tem como objetivo realizar o estudo e análise dos discursos de diferentes nações no debate geral da Assembleia Geral da ONU.

Buscando analisar e identificar os tópicos discutidos nestas assembleias, em diferentes períodos, esperamos encontrar relações e semelhanças entre discursos de países de blocos aliados, assim como entender tendências globais através do discurso ao longo do tempo.

Para isso, serão aplicados conceitos de processamento de linguagem natural em um Corpus com uma coleção completa de mais de 10000 discursos de 202 países, cobrindo as assembleias realizadas entre 1946 a 2023 (**??**).

1.2 EQUIPE

A equipe denominada DataMinds, com repositório do trabalho presente no GitLab, possui 3 integrantes e tem como objetivo desenvolver o trabalho para a disciplina Processamento Natural de Linguagem. Integrantes:

- Anderson Nogueira Silva
- Eric Yutaka Fukuyama
- Erick Jose Teles de Andrade

2 DESENVOLVIMENTO

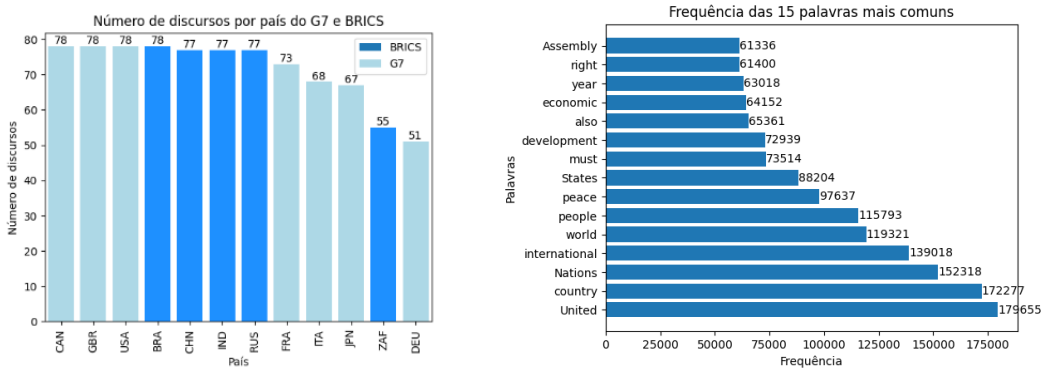
Para a etapa de desenvolvimento foram definidas as bases de dados que serão utilizadas no decorrer do trabalho, realizando diversos métodos de limpeza e ajuste dos dados presentes nas bases utilizadas. Estes procedimentos, assim como os resultados obtidos através das análises destes dados serão descritos a seguir.

2.1 OBTENÇÃO E PROCESSAMENTO DE DADOS

Definição das ferramentas e métodos utilizados para tratar cada base de dados utilizada na análise exploratória.

2.2 COBERTURA E CARACTERÍSTICAS DOS DADOS

As análises iniciais do Corpus foram focadas em analisar o número de discursos realizados por país e a frequência das palavras mais comuns. Esta análise foi fundamental para a definição dos objetivos do trabalho. Como demonstrado na Figura 1, temos mais de 50 discursos para cada pais pertencente aos grupos principais de análise no trabalho, o BRICS e o G7, além disso são apresentadas as 15 palavras mais comuns no Corpus como um todo.

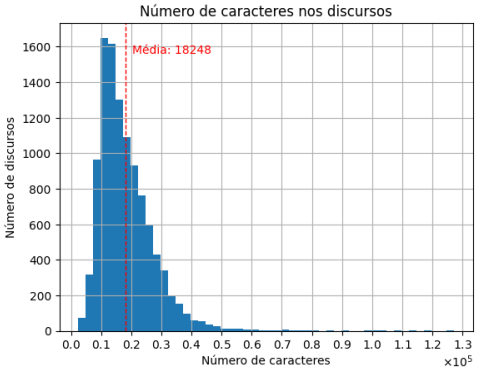


(a) Número de discursos por país.

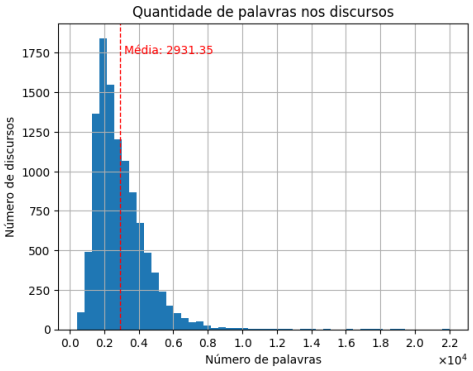
(b) 15 palavras mais comuns.

Figura 1: Gráficos de número de discursos por país e frequência de palavras.

Outra análise das características dos dados presentes no Corpus foi a contagem do número de caractere por discurso e quantidade de palavras por discurso, com o intuito de analisar os valores médios e se teríamos material suficiente para trabalhar durante o desenvolvimento. Dados são apresentados via Histograma na Figura 2.



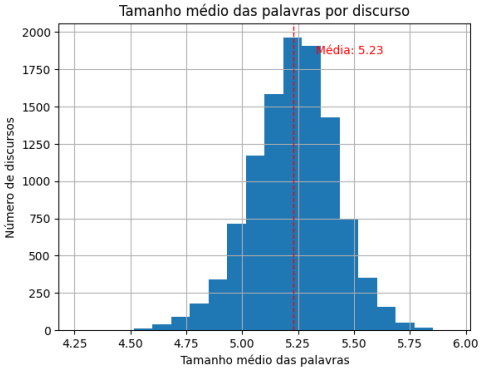
(a) Histograma do número de caracteres por discurso.



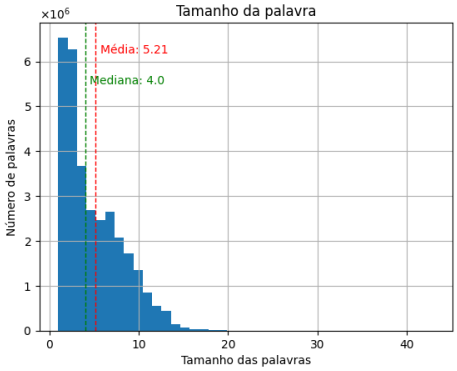
(b) Histograma da quantidade de palavras por discurso.

Figura 2: Dados dos caracteres dos discursos.

De forma semelhante, foram analisados os dados de tamanho médio das palavras, como demonstrado na Figura 3.



(a) Histograma do tamanho médio das palavras por discurso.



(b) Histograma da quantidade de palavras por tamanho de palavra.

Figura 3: Dados do tamanho das palavras.

A Figura 4 apresenta um histograma com a contagem de stopwords em inglês, tópico extremamente necessário para remover as palavras com pouco significado dos discursos, focando em analisar palavras que realmente impactam no sentido e relevância do discurso.

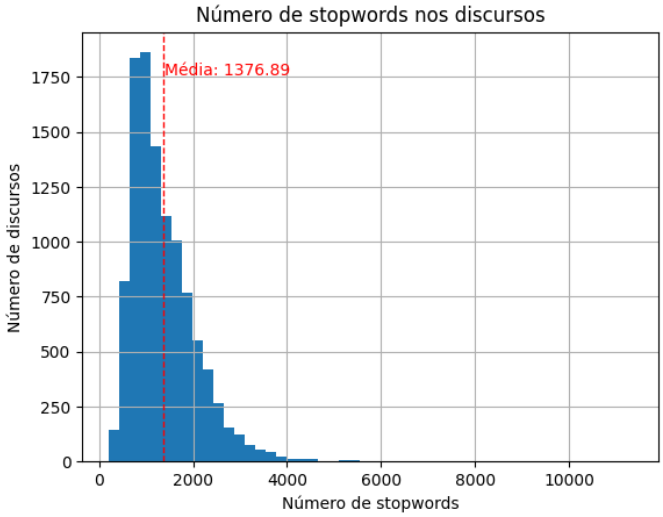
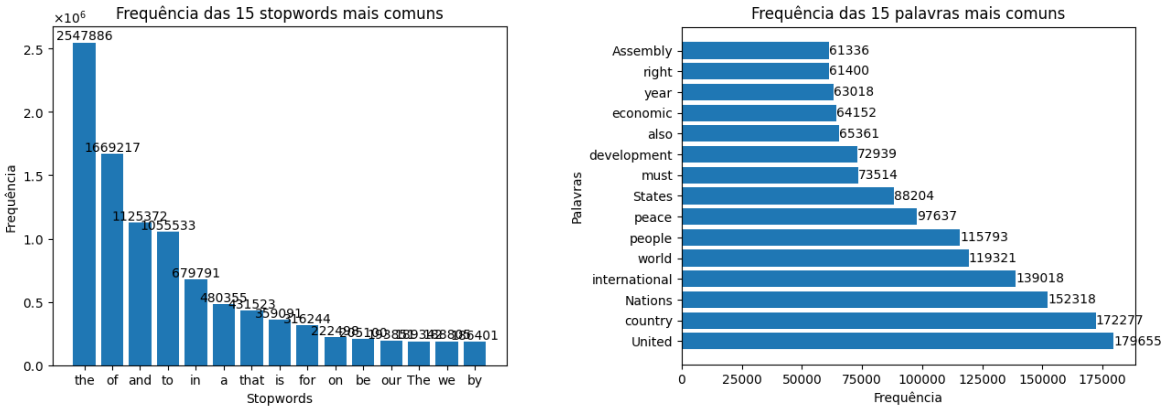


Figura 4: Histograma com a contagem de stopwords em inglês.

Após esta análise foi realizada uma verificação da frequência das palavras mais comuns e das stopwords, como demonstrado na Figura 5.

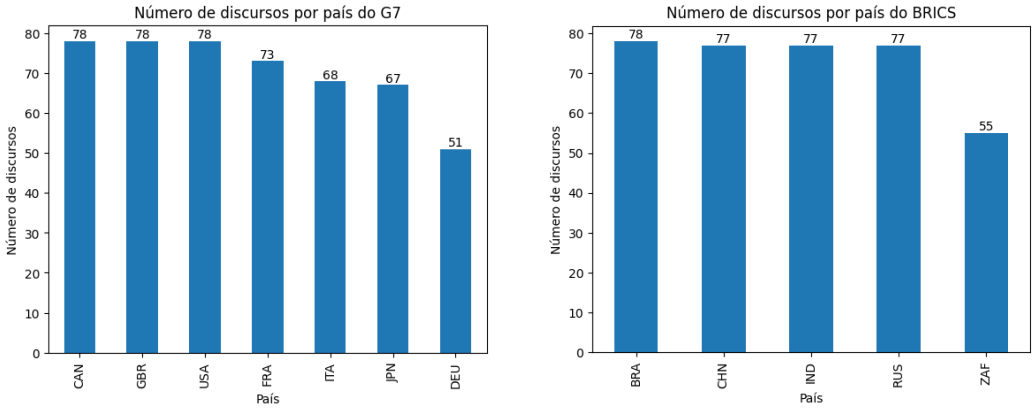


(a) Gráfico de barras da frequência das 15 stopwords mais comuns.

(b) Gráfico de barras da frequência das 15 palavras mais comuns.

Figura 5: Frequência das palavras mais comuns.

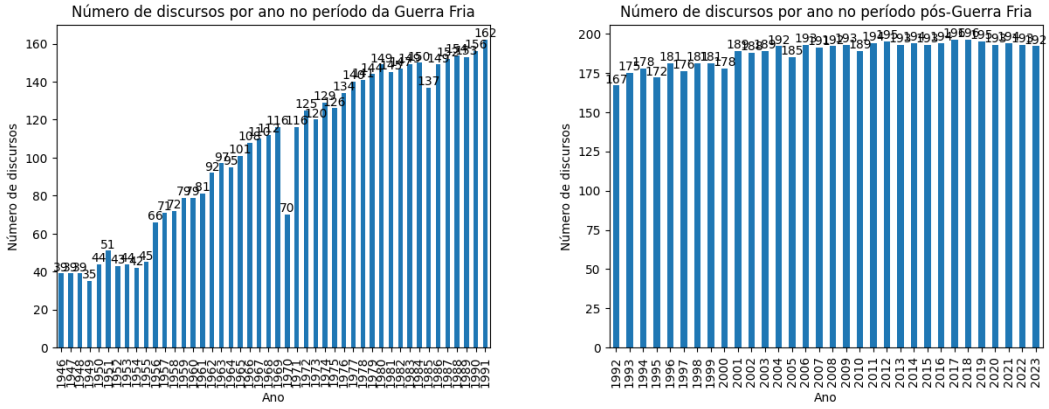
Como demonstrado nas Figura 6, temos uma comparação entre o número de discursos por país dos grupos de maior interesse (BRICS e G7), o que permite averiguar que existe material suficiente para análise comparando ambos os blocos econômicos.



(a) G7. (b) BRICS.

Figura 6: Gráfico de barras do número de discursos por país.

Por fim, também foram analisados, para o contexto global, a quantidade de discursos existentes no corpus em cada ano. Na Figura 7 a seguir, existe a separação do período de análise em dois momentos, pré e pós Guerra Fria.



(a) Período da guerra fria (1946 - 1991). (b) Período pós-Guerra Fria (1992 - 2023).

Figura 7: Gráfico de barras do número de discursos por ano.

Estes foram os processamentos realizados visando analisar a cobertura e características dos dados presentes no Corpus analisado neste trabalho.

2.3 ANÁLISE EXPLORATÓRIA

Para a análise exploratória, foram aplicados conceitos vistos em aula, além de ferramentas e aplicações apresentadas nos tutoriais usados como referência para o desenvolvimento deste trabalho. A primeira análise realizada foi a análise da nuvem de palavras. A nuvem de palavras consiste na representação visual de texto que destaca as palavras mais frequentes ou relevantes em um conjunto de dados. Esta analise é usada como uma ferramenta para explorar, visualizar e resumir o conteúdo de textos e pode ser visualizada na Figura 8, a nuvem para o Corpus usado.

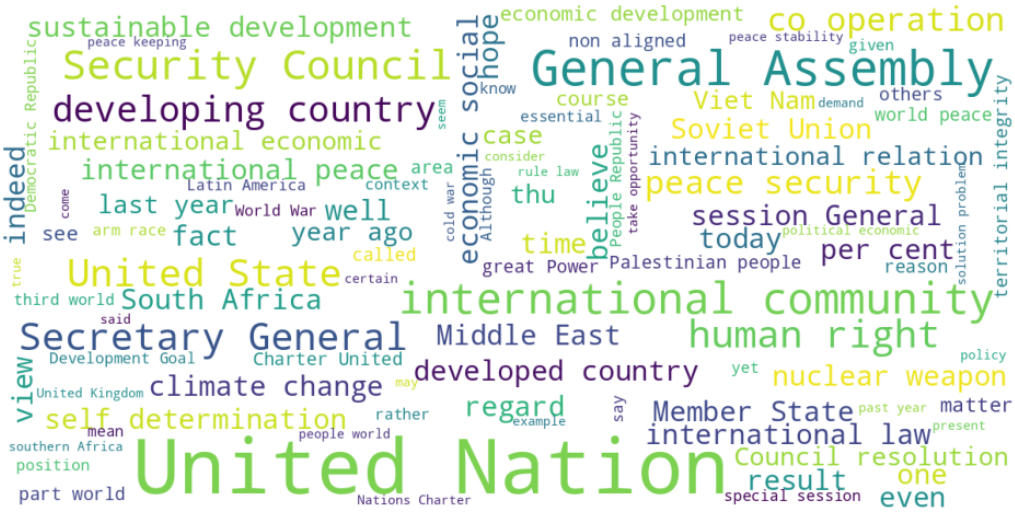


Figura 8: Nuvem de palavras presentes no Corpus.

Outra análise realizada foi focada na contagem de palavras por ano. Esta análise tinha como objetivo verificar e relacionar entre grupos como G7 e BRICS a ocorrência das palavras mais usadas nos discursos. Esta representação pode ser vista na Figura 9

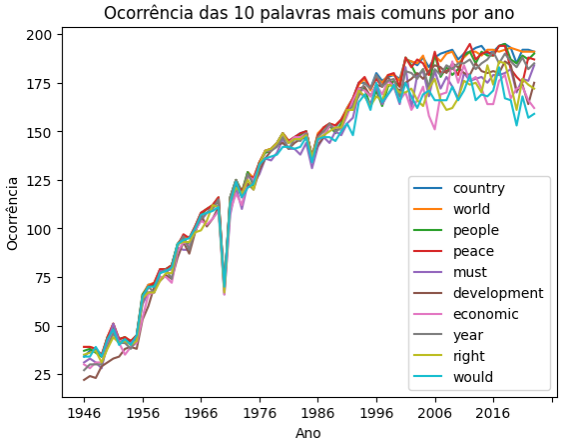
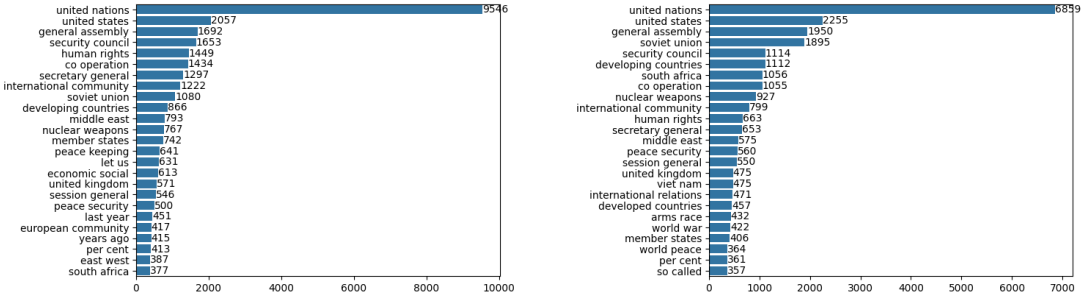


Figura 9: Contagem de palavras por ano.

Por fim, foi realizada a análise de n-grams do Corpus, sempre descartando as *stopwords*. Sabendo que os n-gramas são sequências de palavras consecutivas em um texto, esperava-se identificar padrões, frases comuns e coocorrências no texto, oferecendo *insights* sobre a estrutura e os temas principais do corpus.

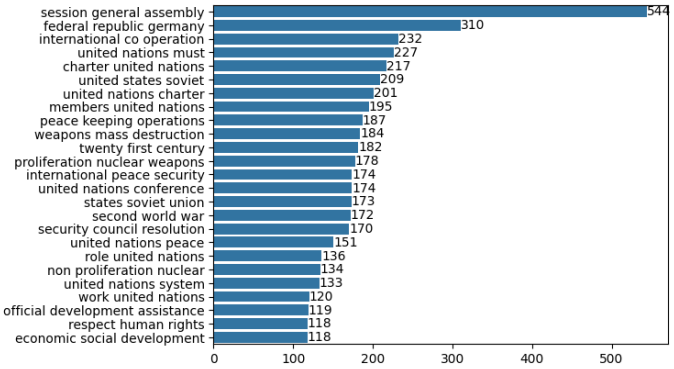
Na Figura 24 abaixo, podemos averiguar os bigramas mais frequentes.



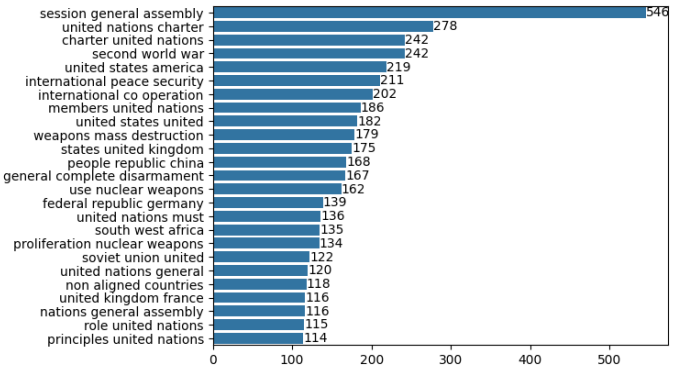
(a) G7. (b) BRICS.

Figura 10: 2Gram.

Os gráficos da Figura 11 representam os trigramas mais frequentes para os blocos econômicos de interesse.



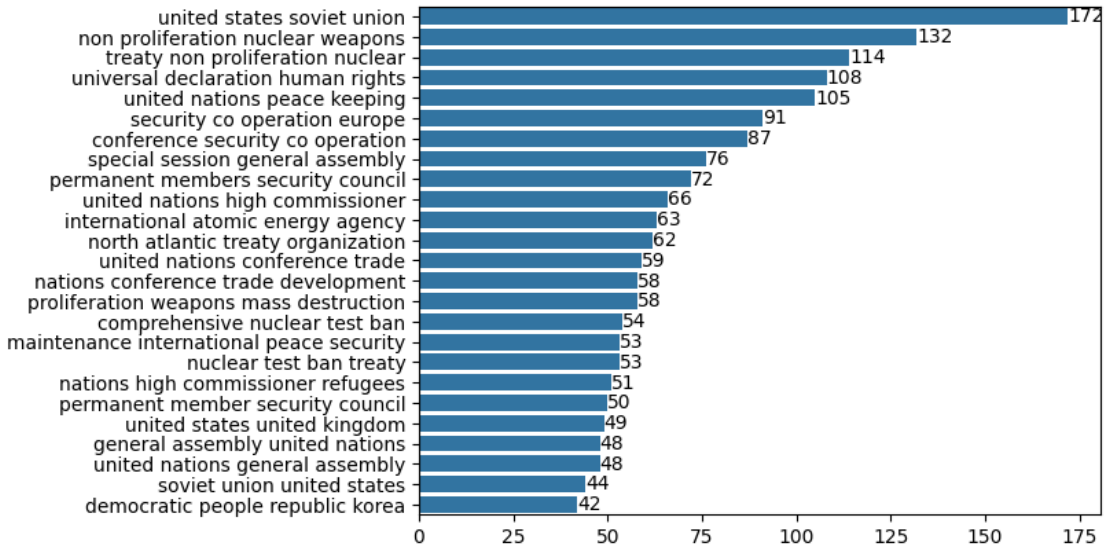
(a) G7.



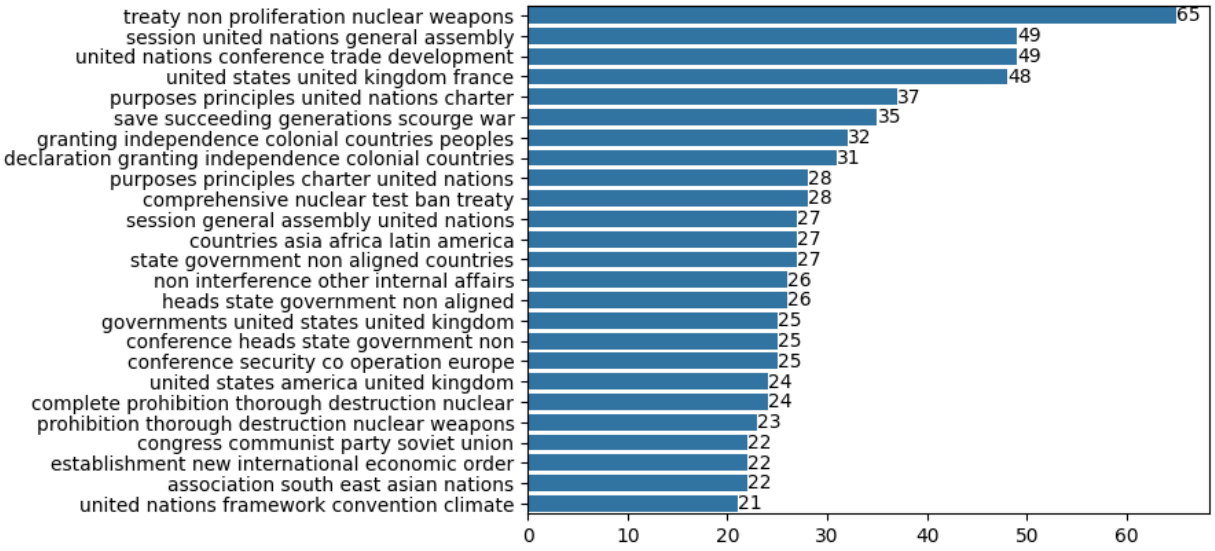
(b) BRICS.

Figura 11: 3Gram.

Na Figura 12 temos a apresentação dos quadrigramas mais frequentes para os blocos econômicos.



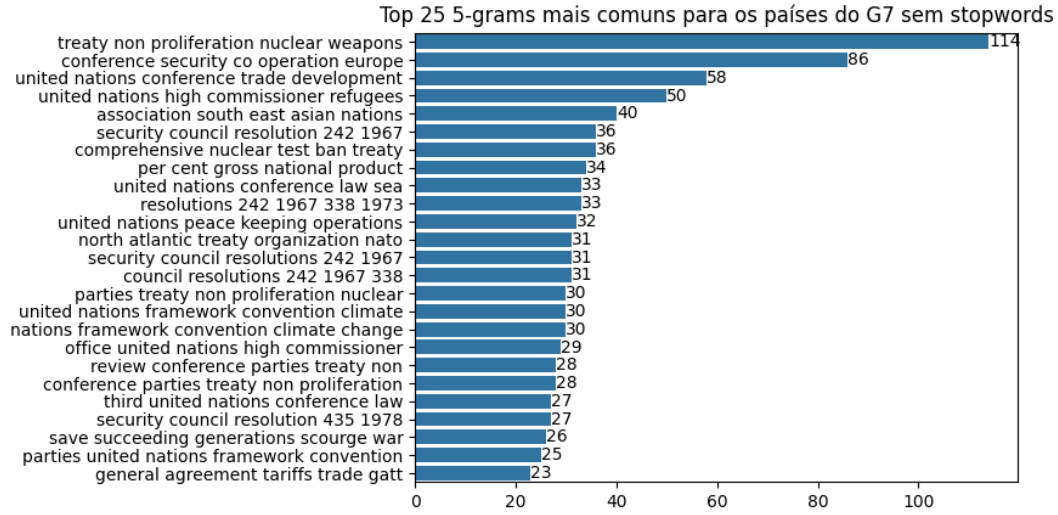
(a) G7.



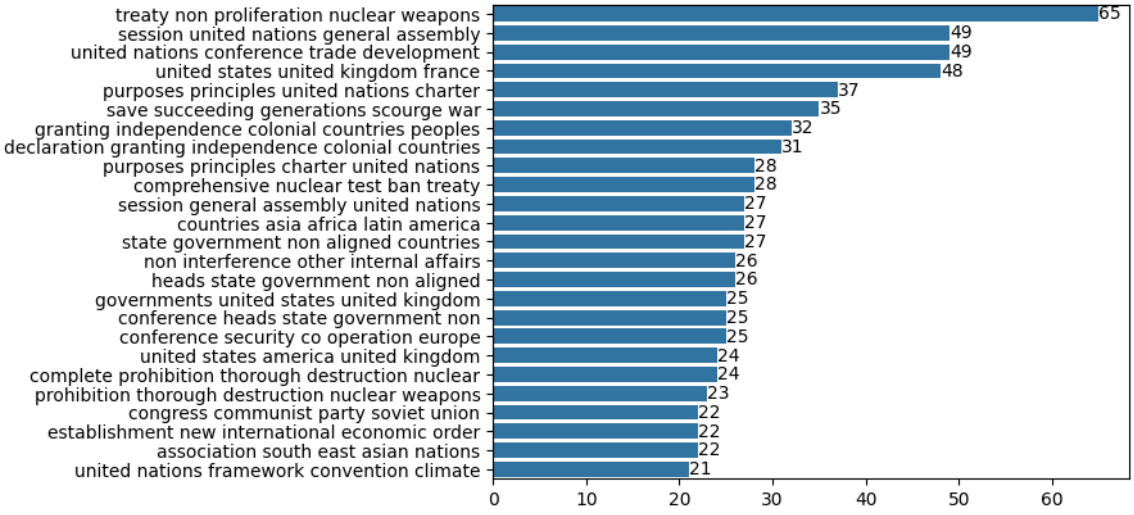
(b) BRICS.

Figura 12: 4Gram.

Por fim, na Figura 13 é apresenta os 5-gramas mais frequentes.



(a) G7.



(b) BRICS.

Figura 13: 5Gram.

2.4 PERGUNTAS DE PESQUISA E EXPLORAÇÕES INICIAIS

Esta seção tem como objetivo apresentar as primeiras análises realizadas com a intenção de concretizar os objetivos propostos na definição de tema do projeto. Portanto, para cada um dos objetivos, será apresentado aquilo obtido até esta etapa.

2.4.1 IDENTIFICAR QUAIS TÓPICOS GLOBAIS FORAM MAIS DISCUTIDOS POR CADA BLOCO (MAJORITARIAMENTE G7 E BRICS) EM DIFERENTES PERÍODOS

Neste objetivo, visamos verificar quais os tópicos são mais relevantes para cada bloco econômico ao longo do tempo. Para fazer esta análise, e também a da seção seguinte, utilizamos a biblioteca (**??**). De modo geral, o LDA Overtime possibilita agrupar os documentos analisados em intervalos de tempo que podemos definir. Para nossas analises, o intervalo de tempo utilizado foi de 4 anos.

Foi possível perceber, conforme o esperado, que eventos geopolíticos impactam os tópicos abordados por cada bloco econômico. A figura 14 apresenta a evolução dos tópicos para os discursos dos países pertencentes ao G7.

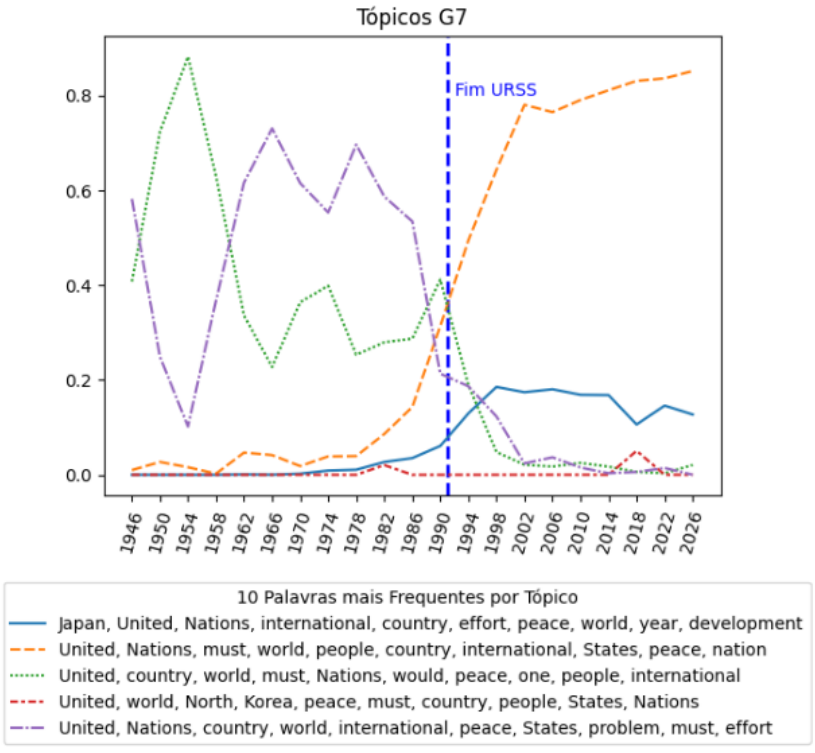


Figura 14: LDA Temporal para países do G7.

Basicamente, portanto, podemos visualizar que os tópicos mais abordados ao longo do tempo para países do G7 envolvem a o esforço para solucionar um problema, como pode ser visto pelo tópico representado pela linha roxa e verde, e que reduz bruscamente após o fim da União Soviética em 1991.

Também podemos inferir que o Japão é um dos países que está associado, de acordo com as 10 palavras mais frequentes nos tópicos apresentas, em abordar o tópico de desenvolvimento, como pode-se se visualizar pela linha azul do primeiro tópico listado.

A linha vermelha não tem uma boa representação ao longo de todo o período analisado, entretanto é bem nítido que o tópico envolvendo Korea e North tem uma leve aparição nos discursos dos países do G7, a alguns anos, muito provavelmente envolvendo quebras em acordos de não uso ou teste de armas nucleares.

Para países do BRICS, a figura 15 retrata a evolução dos tópicos abordados pelos países pertencentes a esse bloco econômico ao longo do tempo.

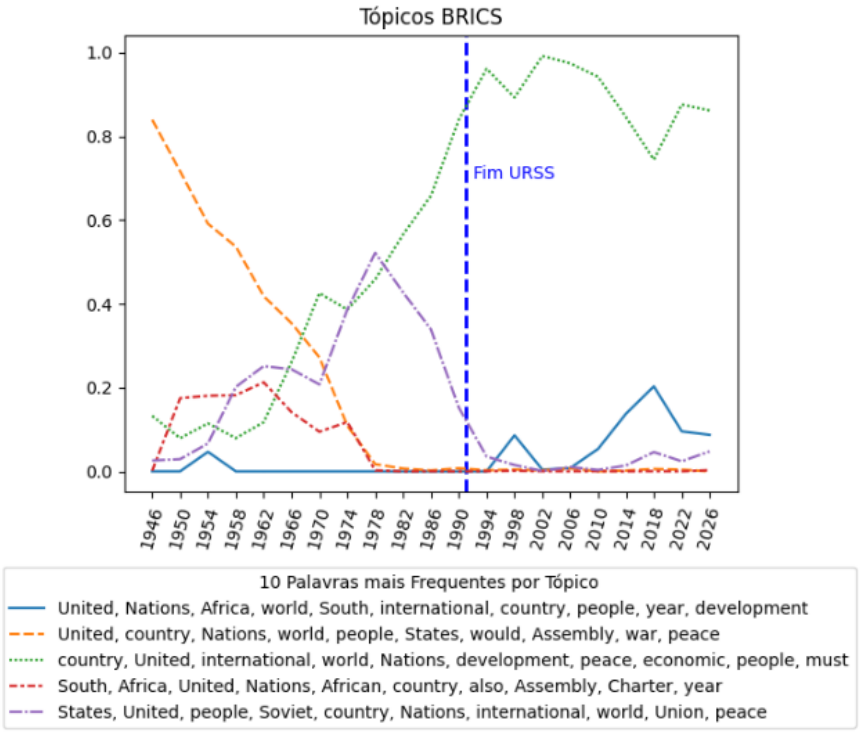


Figura 15: LDA Temporal para países do BRICS.

O primeiro ponto interessante é observar como a palavra *development* aparece em muito mais tópicos do que em países do G7, tendo em vista que este é bloco econômico de países que estão em plena fase de desenvolvimento, sendo economias emergentes.

Também é interessante observar como o fim da união soviética provoca a extinção de um tópico que envolvia diretamente a palavra *Soviet* como uma das mais frequentes.

Outro ponto interessante é que a linha vermelha retrata, com as palavras mais frequentes sendo *Africa*, *South*, *African*, como os países do eixo africano estavam sendo descolonizados, fato histórico que aconteceu entre as décadas de 1950 e 1970. Portanto, o LDA Overtime foi capaz de capturar esse tópico que provavelmente está diretamente associado a esse contexto histórico.

Para se ter um panorama geral de como se distribuiu os tópicos globais ao longo do tempo, também analisamos a evolução dos tópicos para todos os países participantes das assembléias gerais da ONU. Portanto, podemos visualizar esta evolução na Figura 16 a seguir.

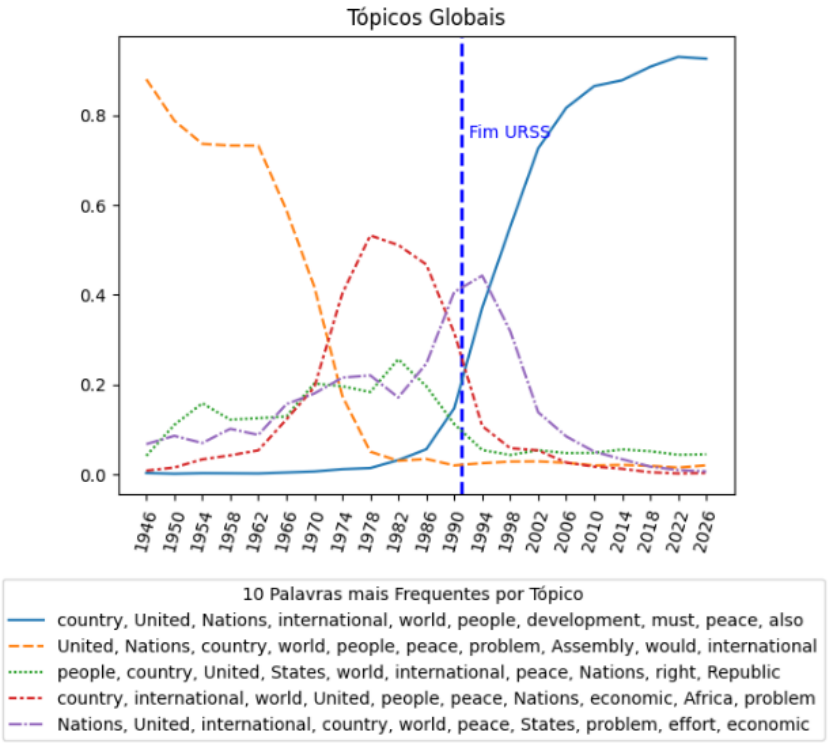


Figura 16: LDA Temporal global.

Podemos compreender portanto que países de todo o mundo, de modo geral, possuem com a queda da União Soviética, passaram a abordar temas associados diretamente a desenvolvimento. Além de que, antes deste evento, anteriormente os tópicos mais predominantes eram aqueles que envolviam as palavras *effort* e *problem*.

2.4.2 ANALISAR COMO MUDANÇAS GEOPOLÍTICAS, CRISES ECONÔMICAS OU POLÍTICAS IMPACTAM A RETÓRICA DE CADA GRUPO

Ao analisar as Figuras 14 e 15 apresentadas na seção anterior, e também a Figura 16, que representa a evolução global dos tópicos, podemos, ao menos inicialmente, verificar que eventos geopolíticos de fato influenciam na relevância dos tópicos abordados pelos países nos debates gerais da ONU, como, nitidamente, pode ser visto pela linha azul do tópico contém a palavra *development* entre as 10 mais frequentes, na Figura 16.

Além disso, uma análise individual foi realizada com alguns países de interesse específico, com o intuito de analisar como grandes eventos internos/externos e geopolíticos podem impactar nos tópicos abordados no Debate Geral da ONU. Portanto, nos gráficos de evolução temporal a seguir, alguns eventos geopolíticos chaves estarão indicados com uma linha vertical tracejada no ano de seu fim.

Portanto, a aplicação do LDA Temporal foi feita individualmente para os discursos de alguns países do BRICS (Brasil, China e Rússia) e também o EUA, pois atualmente é considerado a maior potência econômica mundial. Não serão discutidos muitos detalhes sobre os tópicos dos gráficos abaixo, apenas a sua apresentação para corroborar com a ideia inicial que tínhamos de nosso objetivo.

Sendo assim, na Figura 17, podemos visualizar a evolução temporal dos tópicos para o Brasil.

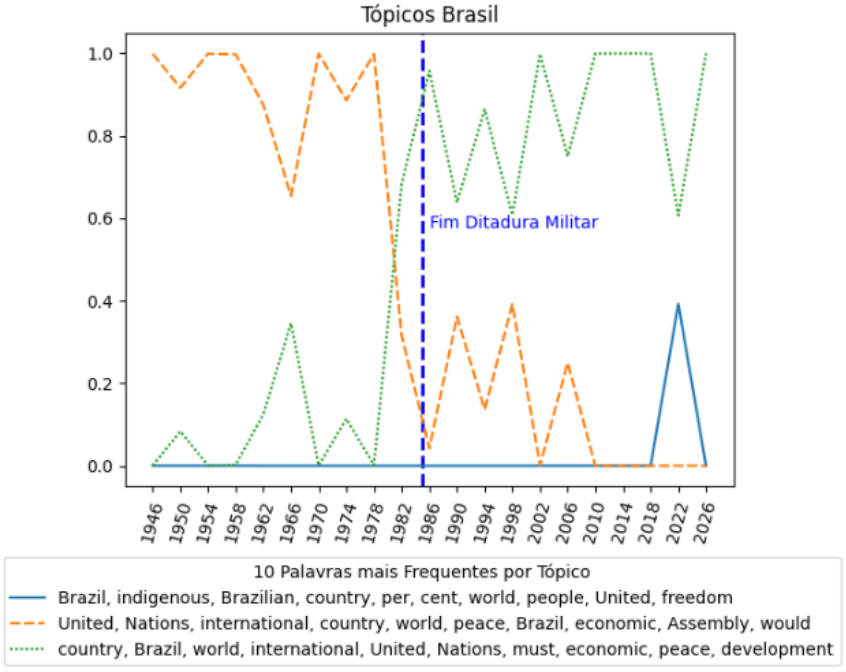


Figura 17: LDA Temporal Brasil.

A seguir, na Figura 18, temos evolução temporal dos tópicos para a China.

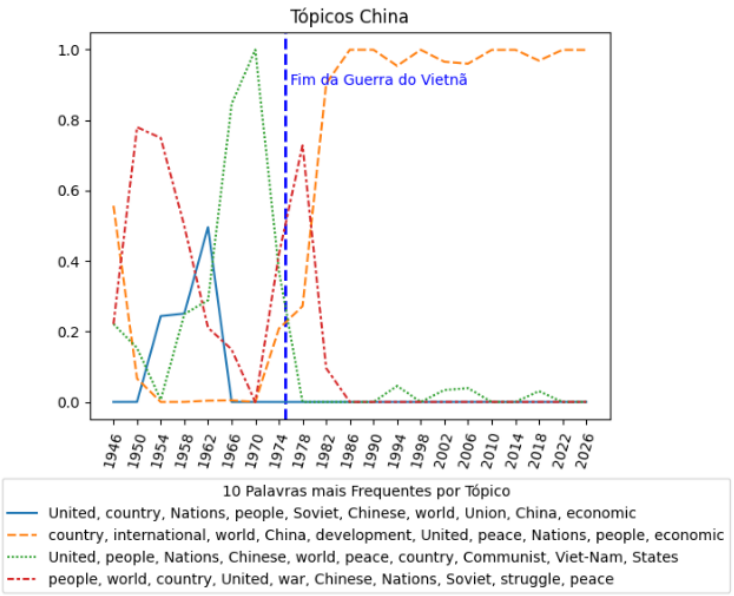


Figura 18: LDA Temporal China.

O próximo país em que aplicamos o LDA Temporal foi a Rússia, na Figura 19 podemos visualizar o LDA Temporal aplicado aos seus discursos.

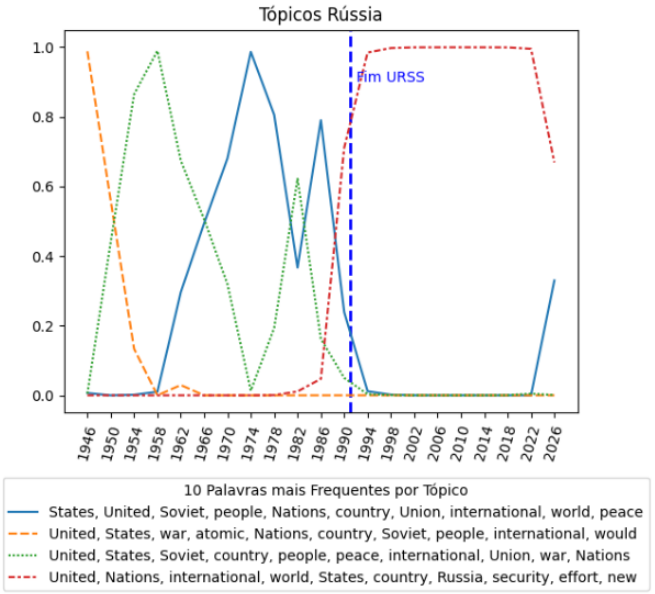


Figura 19: LDA Temporal Russia.

Por fim, para concluirmos a apresentação dos resultados para este objetivo, na Figura 20, está apresentado a evolução temporal dos tópicos abordados pelo EUA.

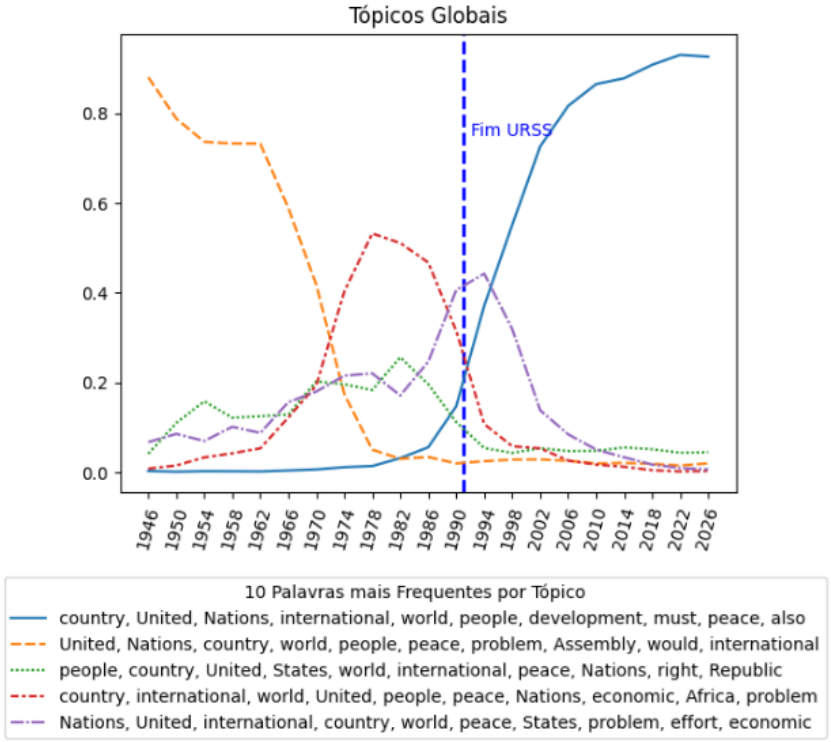


Figura 20: LDA Temporal EUA.

2.4.3 MAPEAR OS PRINCIPAIS TEMAS ECONÔMICOS AO LONGO DO TEMPO PARA DIFERENTES BLOCOS ECONÔMICOS

Para tentar responder esta perguntas, as explorações iniciais relacionadas foram feitas na tentativa de conseguir definir e encontrar temas econômicos de fato nos discursos analisados. O primeiro passo foi tentar encontrar palavras que fossem vinculadas a estes possíveis tópicos, desta forma, foi criada uma lista, com base em pesquisas, com as 10 palavras que poderiam estar vinculadas a temas econômicos quando utilizadas no discurso. Com isso, foi plotado os valores de ocorrência dessas palavras entre os grupos analisados, como apresentado na Figura 21 abaixo.

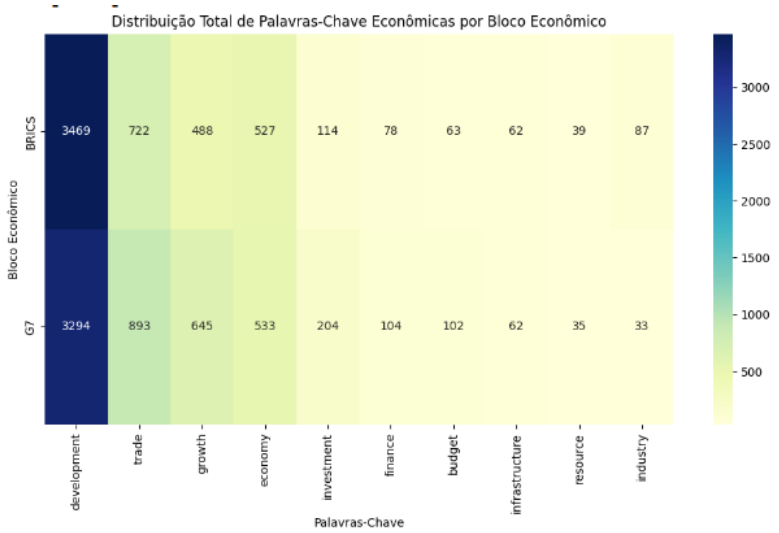
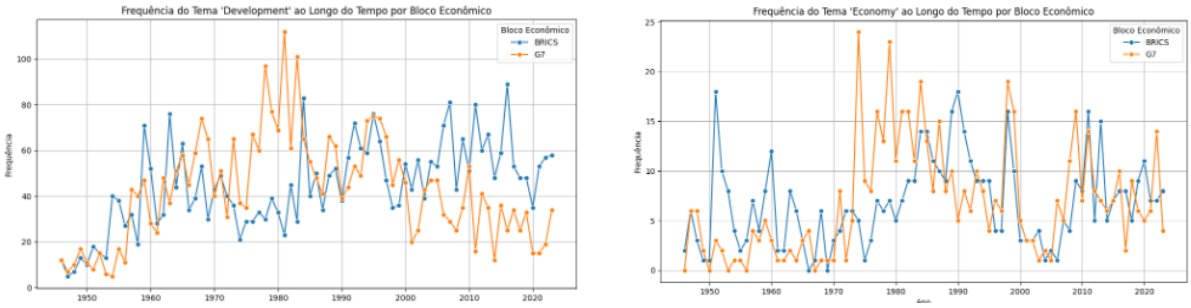


Figura 21: Ocorrência palavras tópico econômico.

Após esta primeira análise, foi verificado a frequência dessas palavras nos discursos dos países aliados ao longo do tempo, para garantir que eram tópicos utilizados amplamente. Isto pode ser visualizado na Figura 22 a seguir.



(a) Palavra ”Development”. (b) Palavra ”Economy”.

Figura 22: Frequência das palavras de tópico econômico ao longo do tempo.

Após isso, com o auxilio de n-grams foi buscado encontrar tópicos relacionados a temas econômicos, entretanto, não foi possível completar esta tarefa de forma a principio. O resultado pode ser visualizado na Figura 23.

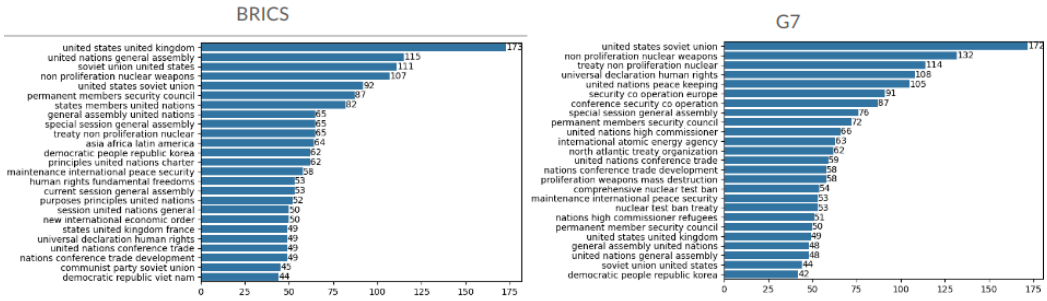


Figura 23: N-gram tópico econômico.

2.4.4 ANALISAR SE OS SENTIMENTOS DE DISCURSOS DE PAÍSES ALIADOS CONVERGEM OU DIVERGEM AO LONGO DO TEMPO

Para analisar esse problema, foram pesquisados sobre 3 modelos. Ao analisar o textBlob foi visto que ele pega a média da polaridade de frases e palavras, mas não tinha bons resultados com discursos. Além disso, o modelo Flair, todavia tal modelo foi treinado por avaliações de filmes no IMDB, tópico bem diferente ao desejado no estudo. Dessa forma, foi escolhido o BERT que é um modelo mais complexo pois é baseado em transformers, tornando uma ferramenta boa para entender o contexto do discurso. Ainda, o BERT não consegue processar mais de 512 palavras de uma vez. Dessa forma, textos maiores que isso foram separados em parte e o score de sentimento foi feito uma média. Logo, foi obtido um score de análise de sentimento para todos os países do BRICS e foi agrupado os discursos desses países por década. De forma análoga, foi feito o mesmo estudo para países do G7. Isso pode ser visto na Figura 24.

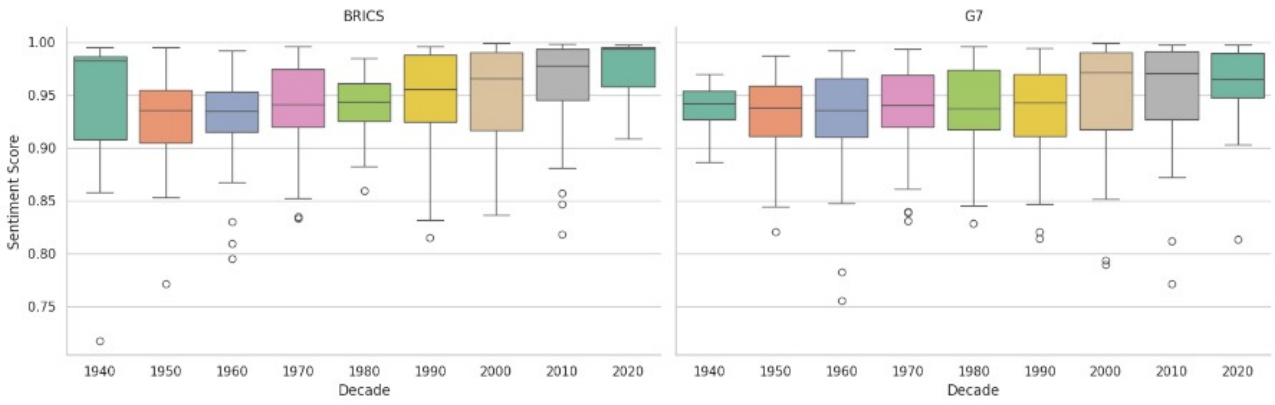


Figura 24: Bert para os grupos Brics e G7.

Dessa forma, nota-se que 50% do score de sentimentos dos grupos nas décadas se encontram em um intervalo curto, uma vez que metade dos dados encontram-se nas ”caixas”. Ainda, é notável que há poucos outliers como é visto no gráfico.

3 DISCUSSÃO E PRÓXIMOS PASSOS

A equipe encontrou dificuldades em alguns tópicos. Os principais são:

- Identificar qual modelo usar para uma melhor análise de sentimentos;
- Compreender qual é o ponto central do tópico abordado através das palavras-chaves;
- Definir tópicos econômicos em relação ao contexto;

Ainda, para os próximos passos foram definidos os seguintes itens:

- Estudo de similaridade entre tópicos de países do mesmo grupo para ver quais pontos são comuns entre eles;
- Analisar como definir os tópicos através das palavras-chaves encontradas;
- Analisar o motivo de existir décadas em que há boxplots menores;
- Analisar o comportamento dos outliers na análise de sentimentos;
- Definir eventos geopolíticos chaves para a analisar o efeito no discurso;

REFERÊNCIAS

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

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