UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE RELAÇÕES INTERNACIONAIS

Luan Azeredo Andreata

Entre frames e fronteiras: o poder de atração do cinema como ferramenta de "soft power"

São Paulo
2025

LUAN AZEREDO ANDREATA

Entre frames e fronteiras: o poder de atração do cinema como ferramenta de "soft power"

Trabalho de Conclusão de Curso apresentado ao Curso de Relações Internacionais do Instituto de Relações Internacionais da Universidade de São Paulo como parte dos requisitos para a obtenção do título de Bacharel em Relações Internacionais.

Orientadora: Prof. ª Dr. ª Marislei Nishijima

São Paulo

2025

Catalogação na publicação Seção
Técnica de Biblioteca
Instituto de Relações Internacinais da Universidade de São Paulo
Dados inseridos pelo(a) Autor(a)

Andreata, Luan Azeredo
Entre frames e fronteiras: o poder de atração do cinema como ferramenta de “soft power” / Luan Azeredo
Andreata; orientador: Marislei Nishijima, -- São Paulo, 2025.
30 p.

Trabalho de Conclusão de Curso (Graduação) - Instituto de Relações Internacinais, Universidade de São Paulo, São Paulo.

1. Soft Power 2. Cinema 3. Joseph Nye 4. Análise de sentimentos 5. Processamento de Linguagem Natural I. Nishijima, Marislei, orient. II. Título.

Responsável: Giseli Adornato de Aguiar - CRB-8/6813

AGRADECIMENTOS

À minha família. Obrigado, Patricia e Anselmo, por acreditarem nos meus sonhos e por me oferecerem o suporte necessário para realizá-los. Tudo o que sou e conquistei tem a marca do amor e da dedicação de vocês. Ao Higor, agradeço por sua perene companhia e torcida.

À Prof.ª Dr.ª Marislei Nishijima, minha profunda gratidão pela orientação, pela paciência e pela confiança depositada neste projeto.

Aos meus amigos do curso e da vida – em especial à Cissa – obrigado por estarem presentes, por me motivarem e por serem parte essencial desta caminhada.

Ao meu namorado, Davi. Obrigado por todo o carinho, debates, revisões e, acima de tudo, por sempre acreditar em mim, mesmo nos dias difíceis.

ANDREATA, Luan Azeredo. **Entre frames e fronteiras:** o poder de atração do cinema como ferramenta de “soft power”. 2025. 30 p. Trabalho de Conclusão do Curso (Bacharelado em Relações Internacionais). – Instituto de Relações Internacionais, Universidade de São Paulo, São Paulo, 2025.

RESUMO

A teoria do *soft power* desenvolvida por Joseph Nye enfatiza o papel do poder não coercitivo na busca pela capacidade de atrair e cooptar outros Estados para atingir os resultados pretendidos pelo ator. As produções cinematográficas são capazes de gerar atração na medida em que são produtos simbólicos, nos quais são incutidos determinados valores, identidades e narrativas capazes de influenciar percepções. Neste contexto, esse trabalho propõe uma investigação empírica desse fenômeno, explorando a base de dados de filmes disponível na plataforma Kaggle, do *Internet Movie Database* (IMDb). A base conta com aproximadamente cinco mil filmes lançados de 2000 até 2020, abrangendo várias informações, tais como diferentes gêneros, mercados de bilheteria e faixas orçamentárias – para definir uma medida proxy de atração de outras populações. Com este intuito, o estudo instrumentaliza a técnica de análise de sentimento utilizando o modelo DistilBERT, correlacionando esses achados com variáveis materiais (orçamento, receita e país de origem) e investigando a hipótese de que a capacidade de produção de narrativas de apelo global está ligada não só ao poderio econômico, como também ao poderio político por influência. Com isso, são calculadas métricas para cada filme de cada gênero. O objetivo é verificar se diferentes tipos de gênero apresentam níveis distintos de sentimentos e, portanto, teriam níveis distintos de apelo em termos de *soft power*. Essa distinção é relevante porque gêneros como ação, aventura e fantasia têm custos de produção muito mais altos do que dramas e comédias, por exemplo, além da necessidade de conhecimentos tecnológicos específicos, não dominados por todos os países.

Palavras-chave: Soft Power. Cinema. Joseph Nye. Análise de sentimentos. Processamento de Linguagem Natural. DistilBERT.

ANDREATA, Luan Azeredo. **Between frames and borders:** the attractiveness of cinema as a "soft power" tool. 2025. 30 p. Trabalho de Conclusão do Curso (Bacharelado em Relações Internacionais). – Instituto de Relações Internacionais, Universidade de São Paulo, São Paulo, 2025.

ABSTRACT

The soft power theory developed by Joseph Nye emphasizes the role of non-coercive power in the pursuit of the ability to attract and co-opt other states in order to achieve the outcomes desired by the actor. Cinematic productions are capable of generating attraction insofar as they are symbolic products into which certain values, identities, and narratives are embedded, enabling them to influence perceptions. In this context, this study proposes an empirical investigation of this phenomenon by exploring the film dataset available on the Kaggle platform, from the Internet Movie Database (IMDb). The dataset contains approximately five thousand films released between 2000 and 2020, covering various information such as different genres, box office markets, and budget ranges—used here to define a proxy measure for the attraction of foreign populations. To this end, the study employs sentiment analysis techniques using the DistilBERT model, correlating these findings with material variables (budget, revenue, and country of origin) and examining the hypothesis that the capacity to produce narratives with global appeal is tied not only to economic power but also to political power through influence. Based on this approach, metrics are calculated for each film within each genre. The objective is to determine whether different genres exhibit distinct sentiment levels and, therefore, whether they have different levels of appeal in terms of soft power. This distinction is relevant because genres such as action, adventure, and fantasy typically entail much higher production costs than dramas and comedies, for example, in addition to requiring specific technological expertise that is not mastered by all countries.

Keywords: Soft Power; Cinema; Joseph Nye; Sentiment Analysis; Natural Language Processing; DistilBERT.

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

> Médico da Marinha: Você ficaria surpreso se eu dissesse que a Marinha credita a você mais de 160 mortes? Você já pensou que pode ter visto ou feito coisas lá... das quais se arrependa?
>
> Chris Kyle: Eu só estava protegendo meus companheiros. Eles estavam tentando matar nossos soldados, e eu... estou disposto a me encontrar com meu Criador e responder por cada disparo que dei. O que me assombra... são todos aqueles que eu não consegui salvar. (Sniper Americano, 2014, tradução própria)

A citação acima é retirada do filme Sniper Americano, que narra a trajetória de Chris Kyle, atirador de elite condecorado pela Marinha dos Estados Unidos no contexto da invasão dos Estados Unidos ao Iraque em 2003. No destrinchar da trama, a obra mobiliza uma série de valores como o patriotismo incondicional, a bravura individual e o dever moral de proteger a "nação" – mesmo que isso ocorra em territórios estrangeiros e sob justificativas geopolíticas pouco problematizadas. Uma possível interpretação da lógica que sustenta a narrativa é binária e simplificadora: coloca *vis-à-vis* heróis norte-americanos contra inimigos desumanizados, coragem contra ameaça, sacrifício contra fanatismo. O filme não apenas parece construir um personagem; poderia, então, estar reafirmando uma identidade nacional baseada na superioridade moral e militar dos Estados Unidos? A narrativa do filme parece ter esvaziado o contexto histórico da guerra e reproduzido, de maneiras sutis ou explícitas, a lógica do excepcionalismo americano (Nelson, 2022).

O diálogo supracitado, aparentemente uma justificativa moral despida de qualquer contexto sociopolítico e movida por valor pessoal, pode ser considerado um exemplo de como um longa-metragem hollywoodiano é capaz de incutir em sua narrativa um vetor de afirmação identitária e de reforço ideológico. A relevância desse fenômeno para as Relações Internacionais emerge quando se assume que o entretenimento de massa é um espaço de exerção e de disputa de poder. O mecanismo hipotético seria o seguinte: se o filme alcança audiências globais e sua narrativa é percebida como atraente, ele poderia funcionar como um veículo de *soft power*, moldando sutilmente as preferências e a percepção sobre o ator estatal. Partindo dessa premissa, é possível interpretar Sniper Americano como ilustração da forma como o entretenimento audiovisual pode operar como instrumento simbólico, naturalizando determinadas visões de mundo ao transformá-las em drama e heroísmo. Sendo assim, este estudo parte da suposição de que filmes operam como instrumento de *soft power* (Guan; Chagas-Bastos; Nishijima, 2023) e foca numa investigação empírica.

Partindo desse exemplo, este estudo busca se debruçar sobre a relação das produções cinematográficas como instrumento de poder para moldar as relações dos países. Para essa finalidade, utiliza-se o conceito de *soft power,* desenvolvido por Nye (2004, 2021), e ressalta-se o cinema como recurso cultural capaz de exercer atração positiva sobre um Estado.

A baliza temporal para a análise aqui proposta se inicia em 2000 e vai até 2020; trata-se de uma fase marcada por transformações significativas na indústria do cinema global, com a consolidação do modelo de franquias e blockbusters e a intensificação da circulação internacional de filmes. Além disso, corresponde ao período para o qual há ampla disponibilidade de dados estruturados sobre produções cinematográficas, o que permite análises comparativas e sistemáticas dos longas-metragens.

O tema encontra sua relevância na medida em que o cinema é entendido – por parte majoritária dos Estados – como um bem cultural capaz de englobar identidades, valores e significados para além do valor de mercado (Convention on the Protection and Promotion of the Diversity of Cultural Expressions - UNESCO Digital Library, 2005). Ao entender tal conjunto de significados como um recurso envolvido na produção de soft power, compreende-se que ele é capaz de moldar a percepção positiva das pessoas sobre um país. Esse encapsulamento de ideias e identidades promovido por filmes são objetos de atenção dos Estados, que definem políticas públicas que podem variar entre o fomento e a repressão de determinados valores – por exemplo, via subsídios a produções alinhadas à agenda nacional ou censura de conteúdos considerados contrários aos interesses do Estado.

Contudo, as produções cinematográficas não são produtos homogêneos em conteúdo e, portanto, é suposto que também difiram na capacidade de gerar atração. Ao entender os principais sentimentos associados a determinados filmes, é passo lógico pensar na sua capacidade de gerar maior ou menor atração. Nesse sentido, estabelecer padrões que relacionem essas características narrativas a variáveis indicadoras de diferenças no sucesso e na capacidade de produção oferece um insumo estratégico para que países maximizem o potencial de produção de *soft power* de sua indústria audiovisual.

Com base nessa lógica, propõe-se uma investigação que busca responder à seguinte pergunta: como as características de uma produção cinematográfica – em particular seu gênero principal, sua receita de bilheteria mundial e seu orçamento de produção – influenciam na sua capacidade de gerar *soft power*?

Como metodologia básica, realiza-se uma revisão da literatura sobre o tema e uma investigação usando dados de filmes disponíveis na base da Kaggle, que são tratados e investigados por meio de linguagem natural e estatísticas de palavras e de testes para verificar suas diferenças.

2. REVISÃO BIBLIOGRÁFICA

A bibliografia basilar que formula o conceito de *soft power* tem como principal referência os trabalhos de Joseph Nye. Na busca por uma ampliação da compreensão tradicional de poder dominante nas correntes realistas das Relações Internacionais, Nye chama a atenção para uma compreensão de poder que não se baseia em coerção e compensação (i.e, *hard power*), mas na capacidade de atrair e cooptar para atingir os resultados pretendidos (Nye, 1990, 2021). Para ele, a definição de poder se trata da capacidade de fazer coisas e, no caso de situações sociais, da habilidade de afetar outros para obter os resultados desejados (Nye, 2021, p. 197). Nye enfatiza que o sucesso de um país em obter cooperação internacional não depende apenas de seus recursos militares ou econômicos, mas também de sua habilidade de projetar uma imagem positiva, inspirar admiração e gerar identificação. Desse modo, ao comprimir conjuntos de valores em recursos geradores de *soft power*, os Estados fomentam atração na medida em que promovem lógicas que aproximem os receptores de buscar aquilo que o agente tem como objetivo. Apesar de falhas e divergências conceituais[1] serem apontadas nos estudos sobre *soft power* e na conceitualizarão fornecida por Nye, muitas das ideias sobre seu funcionamento são amplamente difundidas e adotadas (Bakalov, 2019).

Novas bibliografias surgiram a partir das obras de Joseph Nye na busca por complementar lacunas teóricas e aplicar a teoria em estudos de caso para desenvolver a compreensão do *soft power*, podendo ser divididas em cinco principais correntes que buscam entender diferentes objetos: (I) quem exerce o *soft power*, (II) quem é a audiência alvo, (III) quais são os recursos envolvidos na produção do *soft power*, (IV) qual é o objetivo ao exercer o *soft power* e, por fim, (V) qual é o mecanismo causal pelo qual recursos se transformam em *soft power* e auxiliam na consecução de objetivos políticos (Novelli; Pereira, 2023a). Os recursos, em particular, não devem ser considerados sinônimos de poder[2], mas sim elementos que dependem do contexto em que eles existem e da capacidade do agente em convertê-los em alterações de comportamento dos Estados (Nye, 2011, 2021).

Para contextualizar as produções cinematográficas neste estudo, este trabalho alinha-se à terceira corrente, analisando o cinema enquanto recurso cultural envolvido na produção de

---

[1] Bakalov aponta existir um entendimento pela maioria dos autores críticos do *framework* de Nye de lacunas conceituais relacionadas – sobretudo – na consideração dos recursos como medidas de diferenciação, utilização de uma diferenciação em degrau entre o *soft power* e o *hard power*, e conceituação sobre o modo de funcionamento do *soft power* (Bakalov, 2019, p. 14, tradução nossa).

[2] Em contraste com as escolas realistas das relações internacionais que majoritariamente definem o poder em capacidade material.

soft power. Os produtos podem se manifestar em diversos recursos que são estrategicamente dominados por diferentes nações, sendo exemplos notáveis os Estados Unidos com Hollywood, o Japão com os animes e a Coreia do Sul com o *Korean Pop* (Guan; Chagas-Bastos; Nishijima, 2023, p. 9). Desse modo, o cinema se enquadra como uma parte de um conjunto mais amplo de produtos capazes de gerar atração. Estudos que estabelecem a relevância do objeto passam pelas evidências de que filmes afetam a percepção pública e são capazes de gerar *soft power* (Guan; Chagas-Bastos; Nishijima, 2023). Entender as diferentes especificidades desse objeto e como ele pode produzir atração em maior ou menor medida são contribuições necessárias ainda pouco presentes na literatura sobre o tema.

Uma das principais limitações metodológicas dos estudos empíricos que buscam operacionalizar esse conceito é que, ao se concentrarem nos recursos que produzem *soft power*, incorrem em uma falta de distinção entre poder e recursos (Novelli; Pereira, 2023b, p. 132). Assume-se que a mera existência ou difusão global de um recurso gera, *per se*, *soft power*, desconsiderando fatores como o contexto – capaz de determinar a efetividade do uso dos recursos[3]. Tal falta de distinção é um desafio metodológico, uma vez que o contexto em que o recurso é investido envolve diretamente as percepções subjetivas da audiência focal. Este trabalho reconhece essa lacuna; contudo, opta deliberadamente por um recorte metodológico estratégico. O uso da análise de sentimento como *proxy* metodológico realiza um *trade-off* da análise contextual para ganhar amplitude sistêmica: a capacidade de mensurar empiricamente o apelo narrativo em larga escala.

No plano econômico-estratégico, algumas obras focam no papel da globalização para a indústria cinematográfica norte americana – sobretudo para a resposta de Hollywood para uma expansão da exportação de filmes – focando na produção de filmes que atendam à preferência dos consumidores internacionais (Leung; Qi, 2023). Esta linha é relevante, pois sugere que os agentes envolvidos nas produções cinematográficas selecionam certos tipos de narrativa como estratégia para criar produtos de apelo global que podem ser agrupadas em determinados gêneros. Já no plano geopolítico, a bibliografia aponta que filmes não apenas são recursos promotores de discursos e ideias envolvidos na produção de *soft power*, mas também são objetos de atenção e controle por órgãos de Estado, que oferecem incentivos e restrições no intento de moldar ou controlar aquilo que é projetado para a percepção do público (Aydemir,

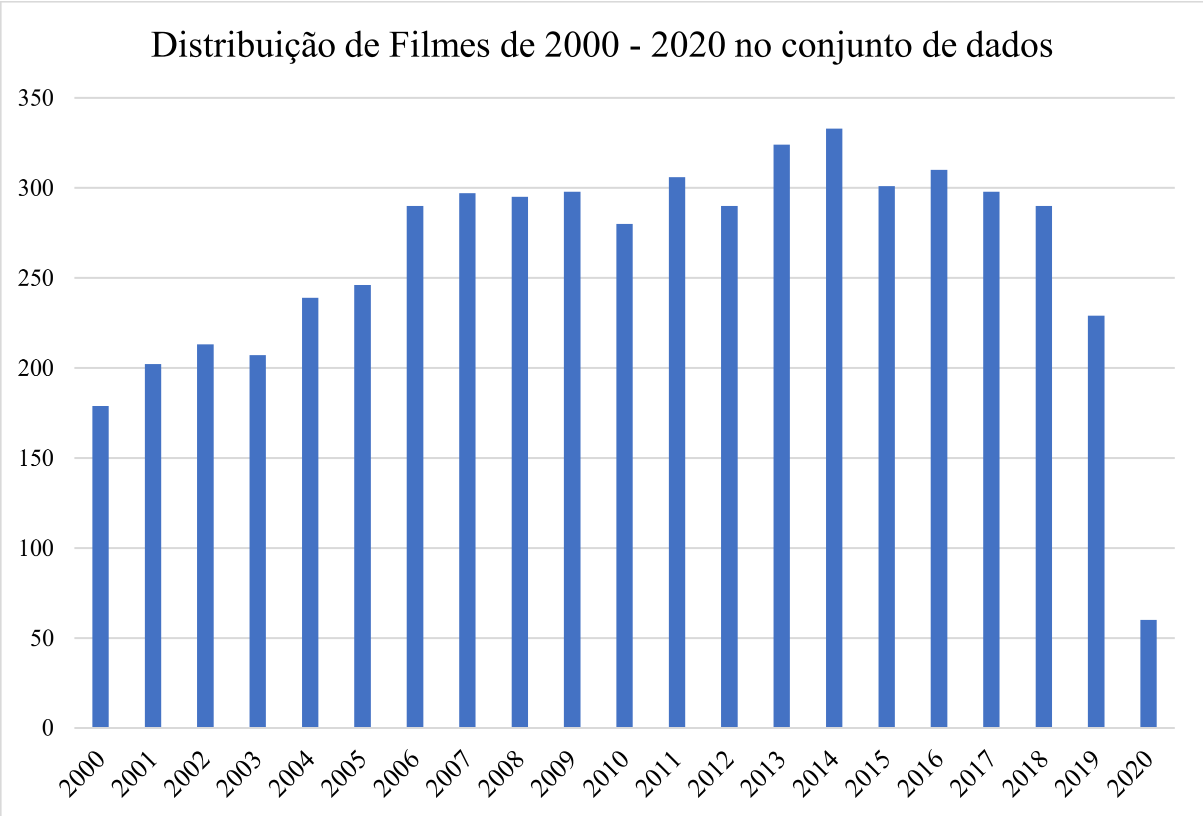
[3] Um dos principais problemas identificados nos textos são as explicações vagas do mecanismo causal em ação, ou seja, como os recursos produzem *soft power* e como esse poder pode ser utilizado para atingir determinados fins políticos (Novelli; Pereira, 2023b, p. 138).

2017; Güzelipek, 2018). Esta constatação é fundamental, pois estabelece a premissa de que o cinema é um vetor de interesse nacional e, portanto, a investigação sobre quem detém os meios materiais para produzir e disseminar esses discursos em escala global torna-se um ponto central.

3. DADOS E MÉTODOS

No intuito de enquadrar a teoria de *soft power* no caso das produções cinematográficas, o recorte temporal corresponde ao período de 2000 até 2020, abrangendo duas décadas de produção cinematográfica marcadas por transformações significativas na indústria do cinema global, com ênfase na consolidação do modelo de franquias e *blockbusters*[4] e intensificação da circulação internacional de filmes. A base de dados utilizada é proveniente do repositório Kaggle, com informações de aproximadamente cinco mil títulos extraídas do site Internet Movie Database (IMDb) (Leone, 2021), como observado na Figura 1.

Figura 1 - Distribuição por ano (2000 - 2020) dos filmes no conjunto de dados



Fonte: Próprio Autor, 2025.

O estudo tem fundamentalmente caráter exploratório, uma vez que ainda são poucas as investigações que estudam a relação entre o cinema e *soft power* de maneira sistemática. Em razão disso, a metodologia utilizada para essa investigação é o estudo de caso, com o intuito de averiguar como o caso pode contribuir para o arcabouço teórico em que se fundamenta (Seha;

[4] Uma das características dos *blockbusters* é que, mesmo que em momentos anteriores tenham tido prevalência da audiência doméstica, no recorte temporal selecionado eles apresentam a maior da audiência internacional (Nelson, 2022, p. 2).

Müller-Rommel, 2016). Dentro dessa metodologia, esse se configura como um *hypothesis-generating case study* (Lijphart, 1971, p. 692) que almeja aplicar a hipótese da relação entre os sentimentos e os gêneros em um vasto número de casos. O trabalho, então, tem como objetivo específico contribuir para o arcabouço teórico do *soft power* ao explorar empiricamente variáveis ainda pouco discutidas dentro do tema.

A operacionalização dos dados será feita por meio da aplicação de técnicas Aprendizado de Máquina[5], especificamente na vertente de Processamento de Linguagem Natural (PLN). O aprendizado de máquina é uma ramificação dentro do campo da inteligência artificial que permite modelos aprenderem a partir de um conjunto de dados e, em passo seguinte, fazerem predições, destacando-se ao lidarem com situações complexas e ambíguas ao analisar os dados e tomar decisões probabilísticas (Cerulli, 2023, p. 1). A literatura separa os tipos de aprendizado de máquina em: (I) supervisionado, (II) sem supervisão e (III) aprendizado reforçado (Cerulli, 2023, p. 2) Inserindo-se na primeira categoria, este estudo destaca metodologicamente o uso de Aprendizado Profundo (Deep Learning) na modalidade supervisionada. Diferente das abordagens estatísticas tradicionais, o Aprendizado Profundo faz parte de uma família de métodos baseado em várias arquiteturas de Redes Neurais Artificiais (RNAs) incorporadas no aprendizado (Cerulli, 2023, p. 3). Existem uma variedade de tipos de RNAs, mas em sua maioria são compostas por componentes que se assemelham aos do cérebro humano, como neurônios, sinapses, pesos, vieses e funções (Cerulli, 2023, p. 3). Já o PLN se trata de um campo interdisciplinar que combina técnicas de inteligência artificial, linguística, e ciência da computação para compreensão, classificação e interpretação de linguagem humana (Oliveira, 2024, p. 30). Assim, a metodologia proposta busca aplicar um modelo classificatório de Aprendizado Profundo supervisionado, utilizando-se do PLN para classificar os textos em determinadas categorias.

Especificamente, o modelo adotado baseia-se na arquitetura dos Transformers, RNAs de processamento de texto que permitem uma análise contextual bidirecional das palavras. O modelo que serve de base é o DistilBERT (*Distilled Bidirectional Encoder Representations from Transformers*), modelo de linguagem bidirecional de interpretação de texto com otimizações para melhoria de desempenho (Sanh *et al.*, 2020). Este modelo passa por um processo denominado "destilação de conhecimento" (Hinton; Vinyals; Dean, 2015), no qual

[5] Também pode ser referido como aprendizado estatístico (Cerulli, 2023, p. 2).

uma rede neural menor (DistilBERT) é treinada para reproduzir o comportamento de uma rede maior (BERT), resultando em um modelo 40% menor e 60% mais rápido, ao mesmo tempo que retêm 97% da capacidade de entendimento da linguagem (Sanh *et al.*, 2020, p. 1). Essa relação custo-benefício torna o DistilBERT a ferramenta adequada para esta investigação, viabilizando a análise de sentimentos em larga escala proposta pelo estudo.

Assim, este estudo utiliza a análise de sentimentos, um processo que classifica a polaridade emocional (positiva, negativa ou neutra) de um texto, usada aqui como *proxy* para medir a "capacidade de atração". Para essa tarefa, foi empregado o *Multilingual Sentiment Classification Model* (Samuel Gyamfi; Vadim Borisov; Richard H. Schreiber, 2025), uma versão ajustada do DistilBERT para análise de sentimento multilíngue. Ele utiliza dados sintéticos provenientes dados multilíngues sintéticos gerados por modelos de linguagem avançados, garantindo ampla cobertura de expressões de sentimento em diversos idiomas e contextos culturais. A partir disso, o modelo opera classificando as sinopses em inglês da base de dados nas classes de sentimento (I) Muito positivo, (II) Positivo, (III) Neutro, (IV) Negativo e (V) Muito negativo. Para o registro do código, tratamento dos dados, execução do modelo e tratamento pós-resultado foi selecionada a plataforma do Google Colab. Nela, a codificação foi escrita utilizando a linguagem de programação Python e bibliotecas do ecossistema *Hugging Face*.

Estabelecidos esses pontos, o gênero cinematográfico principal das produções é fixado como o principal objeto da análise. Um desafio metodológico no uso da base do IMDb é a hibridez dos gêneros cinematográficos, em que um filme pode ser classificado simultaneamente com mais de um gênero diferente. Para endereçar essa complexidade foi necessário um critério de seleção taxonômica: adotou-se a regra de priorização da própria fonte, em que o primeiro gênero listado é tratado como a categoria principal do filme e foi o único utilizado nesta investigação. Embora essa abordagem simplifique a complexidade narrativa de obras híbridas, ela garante um modo de classificação que permite a análise comparativa de sentimentos entre os gêneros em larga escala. Além disso, para a análise dos resultados, o foco é estritamente nos oito principais gêneros em termos de frequência na amostra. Este recorte metodológico é fundamental para a validade da análise comparativa, uma vez que gêneros com baixa representatividade não ofereceriam volume de dados suficiente para inferências estatisticamente significativas sobre suas características de produção ou de sentimento. Dessa maneira, os oito gêneros com representatividade substancial na base de dados são: (I) Comédia,

(II) Ação, (III) Drama, (IV) Crime, (V) Biografia, (VI) Animação, (VII) Terror e (VIII) Aventura[6].

Variáveis adicionais são incorporadas para aprimorar a compreensão da questão das diferenças por gênero, como a (I) receita de bilheteria mundial, (II) orçamento de produção e o (III) país de origem do filme. A receita de bilheteria mundial funciona como *proxy* de sucesso comercial e difusão global; já o orçamento de produção indica o grau de investimento, elemento que pode potencializar a visibilidade internacional e viabilizar a produção de determinados gêneros. Para os valores de orçamento e receita, os dados foram padronizados no Dólar americano, sendo então todos os valores disponibilizados em outra moeda convertidos pela média anual do câmbio correspondente (Banco Mundial, 2025). Além disso, os valores nominais foram corrigidos pela inflação utilizando o índice de preços do consumidor para o período de 2000 a 2020 (U.S Bureau of Labor Statistics, 2025). Por fim, definir o país de produção do filme como umas das variáveis de interesse ajuda a entender como se dá a distribuição internacional da produção que, quando associada às outras variáveis, é capaz de revelar assimetrias entre os Estados. Para os casos de produções cinematográficas que listam mais de um país de origem, o critério de seleção adotado foi semelhante ao do gênero principal, em que a base de dados já prioriza o primeiro da listagem como o país principal.

Desse modo, o objetivo central desta metodologia é verificar empiricamente se o potencial de atração difere sistematicamente entre os gêneros cinematográficos. Para isso, o estudo supera a mera catalogação de recursos – uma limitação metodológica recorrente na literatura – ao propor um proxy mensurável para a "capacidade de atração", buscando verificar empiricamente se a análise de sentimentos difere sistematicamente entre os gêneros cinematográficos. Para tal, propõe-se quantificar a valência sentimental das descrições de filmes como um proxy mensurável para essa "capacidade de atração". Essa métrica de sentimento será, subsequentemente, correlacionada com as variáveis materiais (orçamento, país de origem) e de difusão (receita mundial), permitindo uma análise sistêmica da infraestrutura por trás da capacidade de atração, estabelecendo assim um nexo metodológico fundamental para testar a hipótese de que a produção de narrativas de apelo global está intrinsecamente ligada ao poderio econômico do ator estatal.

---

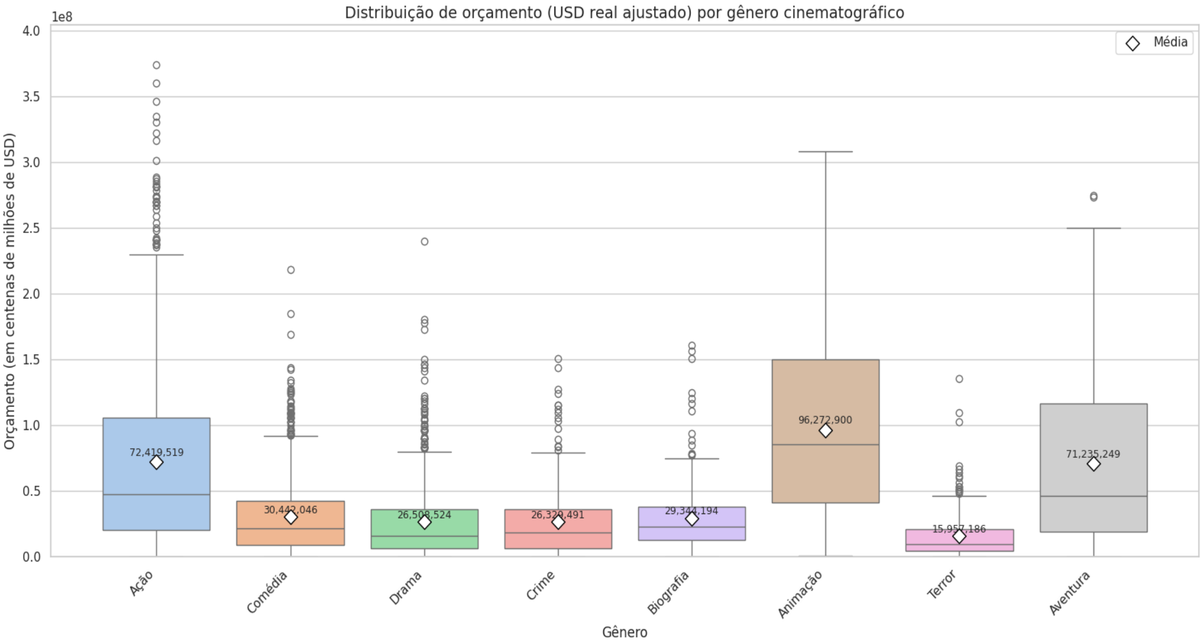
[6] Apesar de não ser a interpretação utilizada para esse estudo, é notável que alguns autores consideram filmes de Aventura também como filmes de Ação, uma vez que identificam que eles em uma quantidade significante de vezes se sobrepõem (Leung; Qi, 2023, p. 1).

4. DESENVOLVIMENTO E RESULTADOS

4.1 DISTRIBUIÇÃO DE ORÇAMENTO, RENDA E PAÍS DE ORIGEM

O primeiro ponto a ser observado são as diferenças dos custos de produção por gênero cinematográfico. A análise da Figura 2 revela uma notável disparidade no custo de produção entre os gêneros. Os três que demandam maior orçamento (Ação, Aventura e Animação) destacam-se ao apresentar tanto a média quanto a mediana de custo superiores ao dobro dos valores registrados pelo gênero seguinte, evidenciando uma elevada necessidade de capital para a produção desses gêneros. A disparidade nos custos de produção entre os gêneros é um fator analítico crucial na medida em que a escala de investimento demandada por certos filmes pode atuar como uma barreira econômica, limitando ou até mesmo inviabilizando sua realização por indústrias cinematográficas com menor aporte de capital.

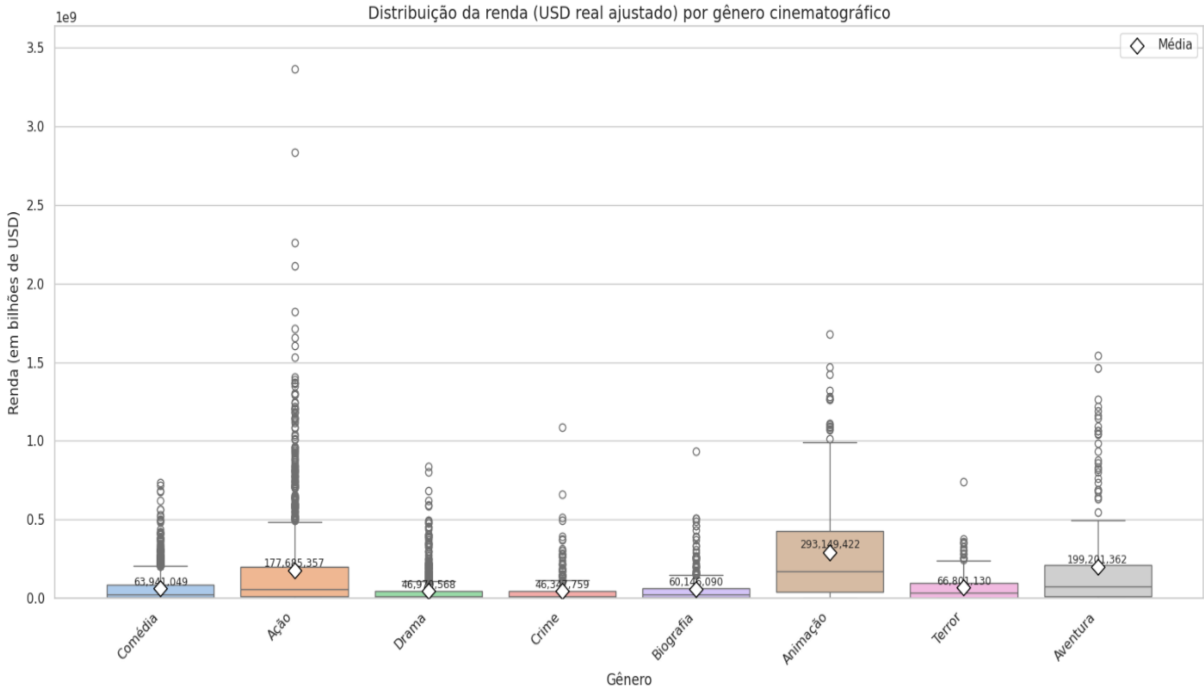
Figura 2 - *Boxplot* com a distribuição do orçamento por gênero cinematográfico



Fonte: Próprio Autor, 2025.

O mesmo padrão dos custos pode ser observado nos rendimentos, conforme a Figura 3. Esses mesmos três gêneros apresentam ampla disparidade nas métricas de distribuição em relação aos outros gêneros. O sucesso financeiro desses filmes garante que essas narrativas sejam as mais vistas globalmente, disseminando em massa produtos culturais incutidos com determinadas visões de mundo. A questão que se coloca, portanto, é: qual ator nacional domina a produção desses gêneros e, por consequência, essa disseminação?

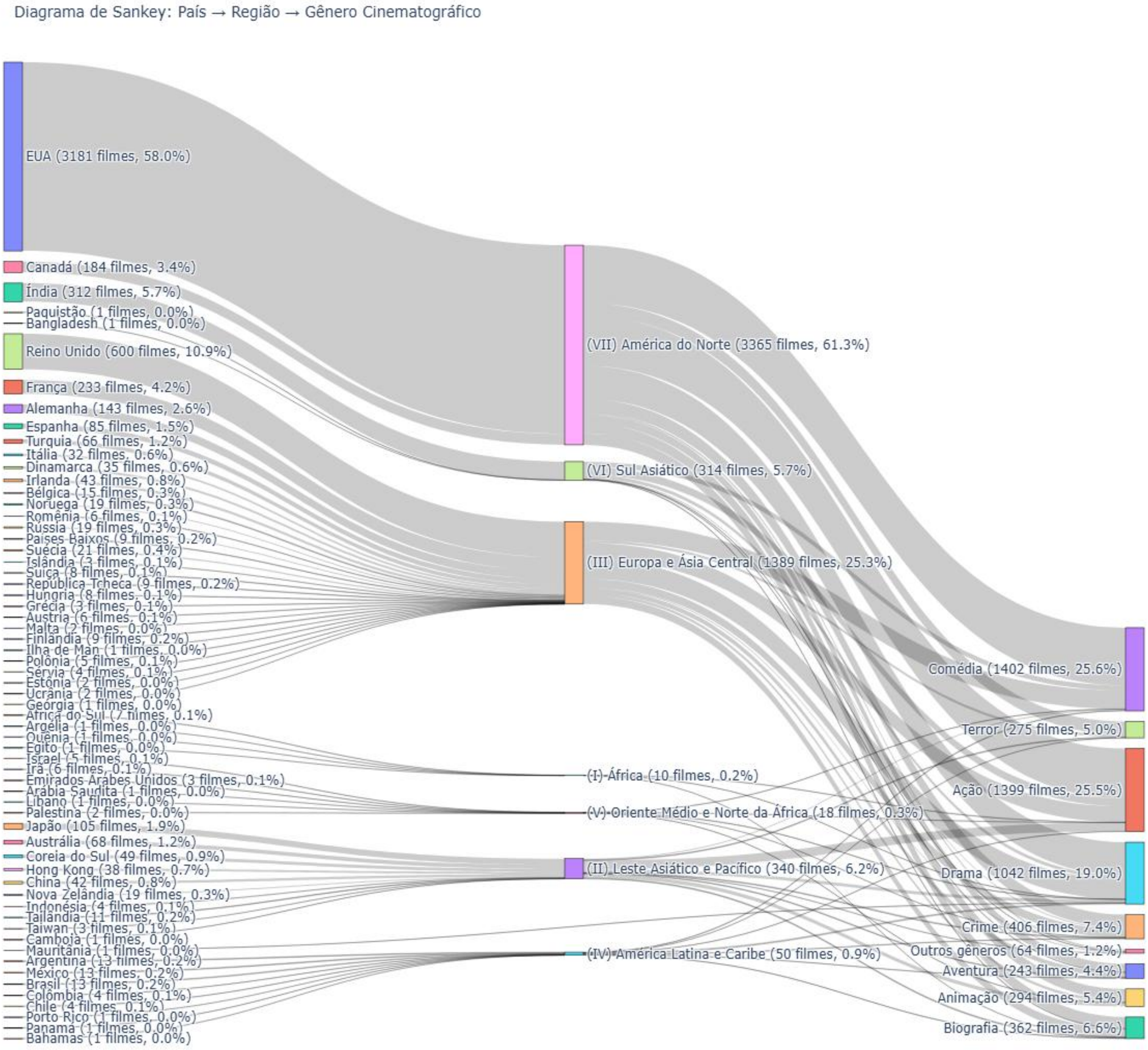
Figura 3 - *Boxplot* com a distribuição da renda por gênero cinematográfico



Fonte: Próprio Autor, 2025.

Para responder quem domina a disseminação dessas 'visões de mundo', a análise da distribuição internacional da produção, conforme ilustrado na Figura 4, revela uma hegemonia dos Estados Unidos. O país é a origem de 3.181 filmes, representando 58% de toda a amostra. Embora a base de dados seja uma amostra e não um censo total, a participação americana é representativa da realidade do mercado, um fato amplamente corroborado pela literatura (Crane, 2014; Guan; Chagas-Bastos; Nishijima, 2023; Nelson, 2022). Esta dominância não é apenas em volume agregado; ela se mantém de forma sistêmica por todos os oito gêneros de maior representatividade, com o cinema estadunidense figurando como o principal produtor em cada uma dessas categorias. A literatura menciona que esse domínio de mercado em ampla vantagem pode ser explicado por uma arquitetura de poder dual: ele se apoia tanto em uma capacidade produtiva e competitiva superior quanto em ações estratégicas de desincentivo à concorrência (Guan; Chagas-Bastos; Nishijima, 2023, p. 3).

Figura 4 - Diagrama de Sankey de produções cinematográficas, com os eixos de país de origem, região e gênero cinematográfico



Fonte: Próprio Autor, 2025.

Essa correlação direta entre os gêneros de maior custo (Figura 2) e maior rendimento (Figura 3) induz à hipótese de que é evidência de um ciclo de retroalimentação econômica impulsionado pela globalização, que define a estratégia do *blockbuster* global. A suposição central é que a expectativa de um rendimento massivo – especialmente no mercado internacional – é o que justifica o investimento inicial em níveis tão elevados. Essa suposição encontra embasamento na literatura, uma vez que os estúdios de Hollywood demonstram responder à expansão do mercado de exportação, adaptando seus produtos às preferências dos consumidores internacionais (Leung; Qi, 2023, p. 32). É possível que gêneros com menos barreiras culturais e de idioma atinjam a preferência de um maior número de consumidores e, portanto, possuem maior possibilidade de sucesso internacional. Dessa maneira, o retorno percebido de um dólar adicional investido nesses gêneros de alto retorno é significativamente maior nesses mercados do que o mesmo dólar investido em outros gêneros. Assim, é notável que esse alto investimento não se trata apenas de um custo de produção, mas um investimento estratégico. Cria-se, assim, um sistema fechado no qual apenas os estúdios com capacidade material para arcar com essa aposta inicial colhem as maiores recompensas.

4.2 APLICAÇÃO DA ANÁLISE DE SENTIMENTOS

Implementado o modelo de análise de sentimentos, a distribuição das classes é apresentada na Tabela 01. Conforme os dados, a classe amplamente predominante para as descrições dos filmes é a Neutra, representando 72% do total. O grupo de sentimentos negativos (agregando "Negativo" e "Muito Negativo") representa 23% da amostra, em forte contraste com o grupo positivo (agregando "Positivo" e "Muito Positivo"), que constitui apenas 4%.

Tabela 01 - Dados por Sentimentos na base de filmes de 2000 - 2020

	Percentual	Renda Média	Orçamento Médio
Muito positivo	1%	$ 116.332.976,87	$ 49.778.408,31
Positivo	3%	$ 106.933.245,41	$ 46.974.476,66
Neutro	72%	$ 114.570.083,11	$ 49.191.532,75
Negativo	17%	$ 91.787.292,80	$ 39.217.007,38
Muito negativo	6%	$ 58.899.917,64	$ 30.763.539,04

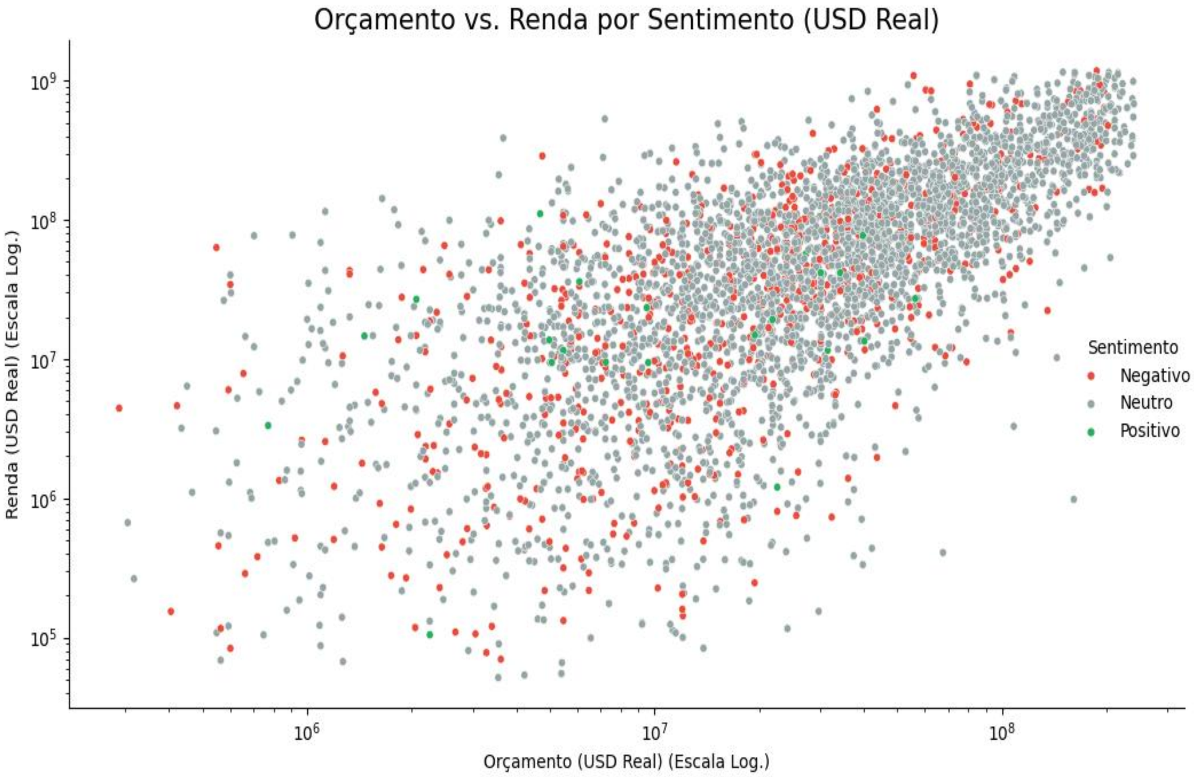
Fonte: Próprio Autor, 2025.

Essa alta incidência de neutralidade (72%) é um achado metodológico esperado e se justifica por uma confluência de fatores que definem o rigor desta análise. Primeiramente, a própria natureza da fonte de dados – sinopses do IMDb – é, por definição, de caráter descritivo, priorizando o enredo em detrimento da emoção. Fatores secundários, como o tamanho relativamente curto das descrições e o uso de um modelo de PLN treinado com dados gerais – e não especificamente para sinopses de filmes – reforçam essa tendência: o modelo conservadoramente classifica como "Neutro" qualquer texto que não contenha uma linguagem explícita e inequivocamente emocional. Apesar das limitações, essa distribuição não se trata necessariamente de uma falha, mas sim de uma definição da linha de base. Isso torna os resultados "Positivos" (4%) e "Negativos" (23%) estatisticamente ainda mais relevantes: eles representam os casos em que a carga emocional da narrativa é intensa o suficiente para superar a natureza descritiva do texto e o viés conservador do modelo. É a análise dessas variações minoritárias – porém significativas – que permite identificar as assinaturas emocionais de cada gênero.

O passo seguinte é entender como a renda e o orçamento médio podem estar relacionados com as classificações fornecidas pelo modelo. A observação da Tabela 01 aponta para uma relação entre as variáveis, onde a "capacidade de atração" (quantificada pelas classes da análise de sentimento) varia substancialmente de acordo com as variáveis de produção (orçamento médio) e sucesso (renda média). Os dados demonstram que orçamentos e rendas médias não estão associados apenas a sentimentos positivos, mas sim à ausência de sentimentos negativos. As classes Neutra, Positiva e Muito Positiva operam em um patamar de investimento e retorno financeiro similar e significativamente elevado, enquanto as classes Negativa e Muito Negativa estão associadas a valores drasticamente inferiores.

Essa ideia encontra mais reforços pela análise da Figura 5. O neutro é o sentimento predominante: ele demonstra alta elasticidade, estando presente desde produções de baixo custo e baixa renda até os filmes de maior arrecadação e investimento do conjunto de dados. Em contraste, os filmes classificados como "Negativo" concentram-se predominantemente nos estratos de orçamento e renda baixos a médios. É aparente uma espécie de "teto" financeiro para esses filmes; eles raramente penetram no escalão superior de arrecadação. A partir disso, levanta-se a hipótese de que a "capacidade de atração" está ligada a narrativas que evitam a polaridade negativa, o que os dados sugerem ser uma característica de produções de alto investimento.

Figura 5 - *Scatter Plot* do Orçamento pela Renda

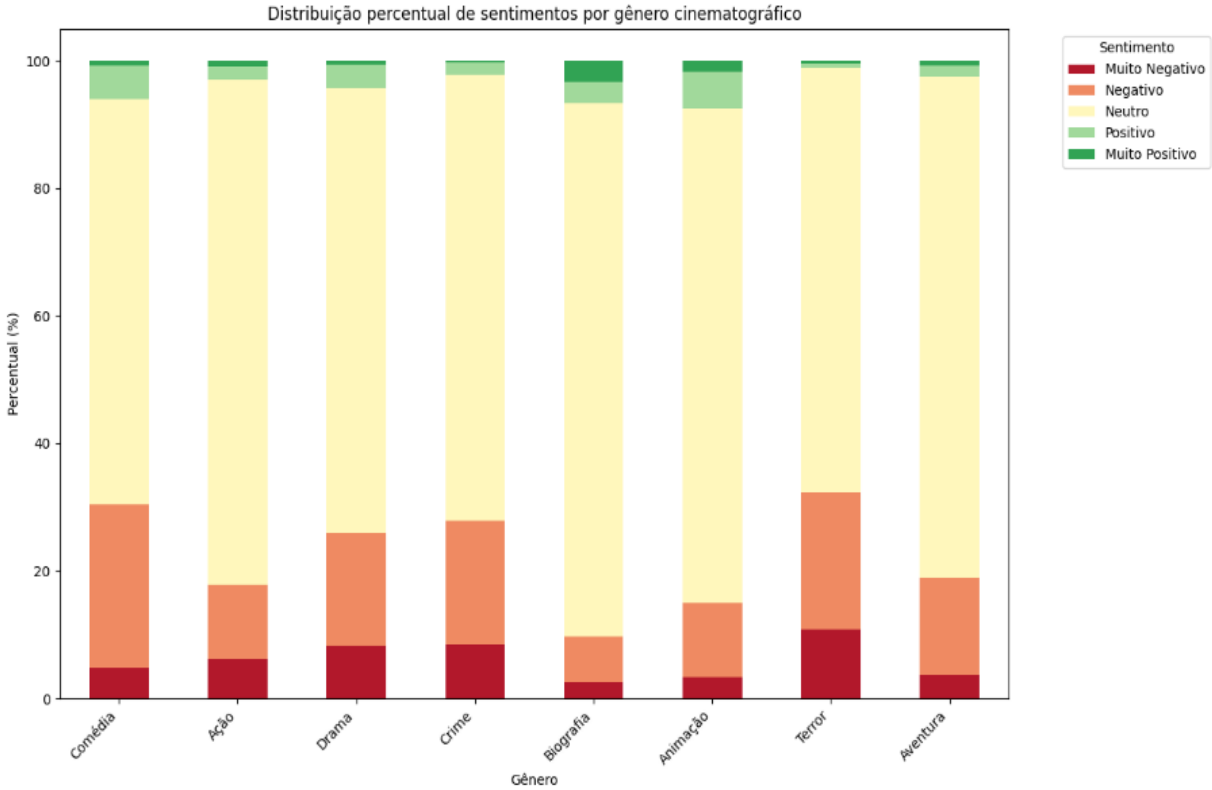


Fonte: Próprio Autor, 2025.

A distribuição de sentimentos por gênero apresentada na Figura 6 oferece a evidência empírica mais clara da tese. Os dados confirmam que os gêneros de maior custo[7], maior receita[8] e maior predominância americana[9] (Ação, Animação, Aventura) são precisamente aqueles que, na análise de suas sinopses, registram a menor incidência de polaridade negativa. Enquanto gêneros como Biografia se destacam por um pico de sentimentos "Muito Positivos" e o Terror se confirma como o polo de maior negatividade, Ação e Aventura consolidam-se por sua massiva neutralidade. Esse achado é relevante na medida em que reforça a hipótese levantada previamente, sugerindo que o investimento financeiro massivo não necessariamente opera para criar positividade explícita, mas sim para garantir a ausência de negatividade. É essa "não-negatividade", ou neutralidade focada na trama, que parece definir o perfil narrativo do *blockbuster* global.

[7] Figura 02.
[8] Figura 03.
[9] Figura 04.

Figura 6 - Distribuição percentual dos sentimentos por gênero cinematográfico



Fonte: Próprio Autor, 2025.

A análise empírica sustenta que a geração de *soft power* cinematográfico não é um campo de disputa simétrico. A questão que se coloca é: o que permite a esses gêneros alcançar esses resultados narrativos positivos em uma escala de impacto global? A resposta aparenta residir na capacidade material de produção. O orçamento, nesta ótica, não é apenas um custo, mas a condição de possibilidade para forjar um produto final que seja inequivocamente positivo ou neutro, ou, mais precisamente, que evite os sentimentos negativos. Filmes com orçamentos menores, que frequentemente exploram temas mais ambíguos ou de nicho, podem não se traduzir em uma sinopse de polaridade tão marcadamente "não-negativa". Consequentemente, se a análise de sentimento do produto final indica quais gêneros são os "motores" mais eficazes de atração e se a produção desses motores é extremamente custosa, então a capacidade de gerar *soft power* via cinema está estruturalmente ligada ao poderio econômico de uma nação. A dificuldade não está em conceber ideias atrativas, mas sim em possuir os meios materiais para executá-las e transformá-las em produtos culturais de alcance global. Isso estabelece uma clara hierarquia, na qual nações com maior capital – como os EUA, que dominam a produção em todos os gêneros analisados (Figura 4) – têm uma vantagem sistêmica na exerção de influência e prestígio no cenário internacional.

Partindo da análise empírica na qual a geração de *soft power* cinematográfico está atrelada ao poderio econômico, é cabível uma discussão das implicações diretas para os atores estatais fora do eixo hegemônico de produção. Para nações que não possuem essa capacidade material, a tentativa de replicar o modelo *blockbuster* pode ser ineficaz ou contraproducente. A alternativa estratégica, portanto, parece residir na assimetria: seja pela exploração de nichos narrativos de alta positividade – como o gênero Biografia, que se destacou positivamente na análise –, seja pela consolidação de esferas de influência regionais, nas quais a proximidade cultural pode suplantar a necessidade de uma atração universalizante.

5. CONSIDERAÇÕES FINAIS

Este trabalho se propôs a analisar uma das dimensões do poder nas relações internacionais, investigando a relação entre as produções cinematográficas e o *soft power*, conforme conceitualizado por Joseph Nye. O estudo partiu de uma lacuna empírica, buscando responder uma questão fundamental: como as características de uma produção cinematográfica – em particular seu gênero principal, sua receita de bilheteria mundial e seu orçamento de produção – influenciam sua capacidade de gerar *soft power.* Para navegar nesta questão, o estudo propôs uma operacionalização quantitativa utilizando uma vasta base de dados (2000-2020) e instrumentalizando técnicas de Processamento de Linguagem Natural. A "capacidade de atração" foi mensurada através de um proxy: a valência sentimental extraída das sinopses dos filmes. Os resultados dessa operacionalização retornaram não apenas uma resposta à pergunta de pesquisa, mas um retrato da infraestrutura material que sustenta a influência por meio do cinema no cenário internacional.

A análise empírica revelou, primeiramente, que o mercado cinematográfico global opera sob uma lógica de disparidade econômica extrema. Gêneros como Ação, Animação e Aventura não apenas demandam orçamentos de produção que eclipsam os demais, mas também são aqueles que capturam a maior fatia da receita global. Esta dinâmica confirma a estratégia do *blockbuster* global: um ciclo de retroalimentação onde o investimento massivo é justificado pela expectativa de um retorno igualmente massivo.

Em segundo lugar, a investigação demonstrou que este poderio econômico-produtivo não é difuso; ele é geopoliticamente concentrado. A análise da origem das produções revelou uma hegemonia sistêmica dos Estados Unidos, que figuram como o principal produtor em todos os oito gêneros de maior representatividade analisados. Este domínio, como aponta a literatura, configura uma arquitetura de poder dual, baseada tanto na capacidade competitiva superior quanto em estratégias de desincentivo à concorrência.

O achado mais significativo deste trabalho, no entanto, foi a conexão entre essa estrutura material (custo e origem) e o tipo de narrativa produzida. Ao cruzar as variáveis financeiras com a análise de sentimentos (Tabela 01, Figura 6), a descoberta central é que a atração cultural de maior escala não está ligada a narrativas explicitamente "positivas", mas sim a uma evasão da polaridade negativa. Enquanto gêneros como Biografia se destacaram pela positividade e o Terror pela negatividade, os gêneros "motores" da produção cinematográfica internacional

(Ação, Aventura, Animação) consolidaram-se por sua massiva neutralidade e sua notável ausência de sentimentos adversos. O orçamento elevado funciona como a condição de possibilidade para forjar um produto final majoritariamente não-ambíguo ou "não-negativo". É o capital intensivo que permite a criação de mundos imersivos e efeitos visuais admiráveis, sustentando narrativas que evitam a complexidade, o nicho ou a controvérsia que poderiam limitar seu alcance global. Uma das possíveis implicações para as nações fora do eixo hegemônico é que a tentativa de replicar o modelo de *blockbuster* pode ser não apenas ineficaz, mas contraproducente. A alternativa estratégica talvez resida não na competição direta, mas na exploração de nichos de alta positividade (como as Biografias) ou na consolidação de esferas de influência regionais, reconhecendo o desenho fornecido pelas necessidades materiais predefinidas.

Portanto, esta pesquisa sugere que a capacidade de gerar *soft power* via cinema está estruturalmente ligada ao poderio econômico de uma nação. Mais do que uma simples aplicação do conceito de Joseph Nye, este trabalho oferece uma investigação da capacidade de atração do produto. A dificuldade não reside em conceber ideias atrativas, mas em possuir os meios materiais para executá-las e transformá-las em produtos culturais de alcance global. Isso estabelece uma clara hierarquia, na qual nações com maior capital – como os EUA – detêm uma vantagem sistêmica e estrutural na competição por influência e prestígio no cenário internacional.

A robustez destas conclusões deve ser ponderada à luz das limitações inerentes ao desenho metodológico. Primeiramente, o proxy utilizado – análise de sentimentos de sinopses – mede o artefato cultural em sua descrição, e não a recepção ou o impacto real na audiência. A própria natureza descritiva das sinopses, aliada ao uso de um modelo de PLN generalista, influi na alta incidência de "neutralidade" e deve ser considerada. Além disso, a base de dados, embora extensa, representa uma amostra do mercado (IMDb), e não um censo total das produções cinematográficas mundiais.

Ainda são poucos os estudos que exploram a transformação de recursos como obras cinematográficas em *soft power*. Futuros estudos podem analisar quais valores, ideias e visões de mundo são, de fato, incutidas nessas narrativas "não-negativas" de alto orçamento, indo além da métrica de sentimento na descrição e analisando o discurso. Além disso, a literatura ainda carece de iteratura obras que foquem na dimensão contextual do cinema como produto capaz de gerar *soft power*. Se faz necessário mais estudos que passem pela recepção de audiências reais em diferentes contextos nacionais, e como elas interpretam e são, ou não, influenciadas por esses produtos culturais, validando a conexão entre o artefato e a atração. Em última análise, a pesquisa revela que a arquitetura do *soft power* internacional é desenhada primordialmente com as mesmas ferramentas do poder material, como capital, infraestrutura e escala industrial.

REFERÊNCIAS

AYDEMIR, Emrah. Use of Hollywood as a Soft Power Tool in Foreign Policy Strategy of the United States of America. [*s. l.*], 2017.

BAKALOV, Ivan. Whither soft power? Divisions, milestones, and prospects of a research programme in the making. **Journal of Political Power**, [*s. l.*], v. 12, n. 1, p. 129–151, 2019.

BANCO MUNDIAL. Official exchange rate (LCU per US$ period average). , 2025. Disponível em: https://databank.worldbank.org/embed/Official-exchange-rate-%28LCU-per-US%24-period-average%29/id/bbfa60bb?utm_source=chatgpt.com. Acesso em: 7 out. 2025.

CERULLI, Giovanni. The Basics of Machine Learning. *In*: CERULLI, Giovanni (org.). **Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata**. Cham: Springer International Publishing, 2023. p. 1–17. Disponível em: https://doi.org/10.1007/978-3-031-41337-7_1. Acesso em: 17 nov. 2025.

CRANE, Diana. Cultural globalization and the dominance of the American film industry: cultural policies, national film industries, and transnational film. **International Journal of Cultural Policy**, [*s. l.*], v. 20, n. 4, p. 365–382, 2014.

GUAN, Miaofang; CHAGAS-BASTOS, Fabrício H; NISHIJIMA, Marislei. Winning Hearts and Minds: Soft Power, Cinema, and Public Perceptions of the United States and China in Brazil. **Global Studies Quarterly**, [*s. l.*], v. 3, n. 2, 2023. Disponível em: https://academic.oup.com/isagsq/article/doi/10.1093/isagsq/ksad029/7186191. Acesso em: 23 maio 2025.

GÜZELIPEK, Yiğit Anıl. THE IMPLEMENTATION OF USA'S SOFT POWER VIA HOLLYWOOD: LOOKING BACK TO COLD WAR. **Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü Dergisi**, [*s. l.*], v. 1, n. 32, p. 228–240, 2018.

HINTON, Geoffrey; VINYALS, Oriol; DEAN, Jeff. **Distilling the Knowledge in a Neural Network**. [*S. l.*]: arXiv, 2015. Disponível em: http://arxiv.org/abs/1503.02531. Acesso em: 26 nov. 2025.

LEONE, Stefano. IMDb Movies Dataset from 2000–2020 [dataset]. Kaggle, , 2021. Disponível em: https://www.kaggle.com/datasets/stefanoleone992/imdb-extensive-dataset. Acesso em: 19 set. 2025.

LEUNG, Tin Cheuk; QI, Shi. Globalization and the rise of action movies in hollywood. **Journal of Cultural Economics**, [*s. l.*], v. 47, n. 1, p. 31–69, 2023.

LIJPHART, Arend. Comparative Politics and the Comparative Method. **American Political Science Review**, [*s. l.*], v. 65, n. 3, p. 682–693, 1971.

NELSON, Travis. Captain America? On the relationship between Hollywood blockbusters and American soft power. **Globalizations**, [*s. l.*], v. 19, n. 1, p. 139–151, 2022.

NOVELLI, Douglas Henrique; PEREIRA, Alexsandro Eugenio. The Use of the Soft Power Concept in Empirical Studies. *In*: THE ROUTLEDGE HANDBOOK OF SOFT POWER. 2. ed. [*S. l.*]: Routledge, 2023a.

NOVELLI, Douglas Henrique; PEREIRA, Alexsandro Eugenio. The Use of the Soft Power Concept in Empirical Studies. *In*: THE ROUTLEDGE HANDBOOK OF SOFT POWER. 2. ed. [*S. l.*]: Routledge, 2023b.

NYE, Joseph S. Soft Power. **Foreign Policy**, [*s. l.*], n. 80, p. 153–171, 1990.

NYE, Joseph S. Soft power: the evolution of a concept. **Journal of Political Power**, [*s. l.*], v. 14, n. 1, p. 196–208, 2021.

NYE, Joseph S. The future of power. [*s. l.*], 2011.

OLIVEIRA, Brenno Ruschioni De. Analisando a Influência do Twitter na Criação de Sequências de Filmes de Terror: Uma Abordagem Baseada em Dados. [*s. l.*], 2024.

SAMUEL GYAMFI; VADIM BORISOV; RICHARD H. SCHREIBER. multilingual-sentiment-analysis. [*s. l.*], 2025. Disponível em: https://huggingface.co/tabularisai/multilingual-sentiment-analysis. Acesso em: 30 out. 2025.

SANH, Victor *et al.* **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. [*S. l.*]: arXiv, 2020. Disponível em: http://arxiv.org/abs/1910.01108. Acesso em: 30 out. 2025.

SEHA, Esther; MÜLLER-ROMMEL, Ferdinand. Case study analysis. *In*: HANDBOOK OF RESEARCH METHODS AND APPLICATIONS IN POLITICAL SCIENCE. [*S. l.*]: Edward Elgar Publishing, 2016. p. 419–429. Disponível em: https://china.elgaronline.com/display/edcoll/9781784710811/9781784710811.00037.xml. Acesso em: 18 abr. 2024.

SNIPER AMERICANO. Estados Unidos: Warner Bros, 2014.

U.S BUREAU OF LABOR STATISTICS. Consumer Price Index for All Urban Consumers (CPI-U). , 2025. Disponível em: https://data.bls.gov/timeseries/CUUR0000SA0. Acesso em: 13 out. 2025.

APÊNDICE – CÓDIGOS DAS CÉLULAS EM PYTHON

```
# ----------------------------------------------------------

# Célula 1 — Montagem do Drive e importação de bibliotecas

# (Breve: monta o Google Drive e importa bibliotecas para manipulação

#  de dados, visualização, pré-processamento, ML e NLP)

# ----------------------------------------------------------


from google.colab import drive

drive.mount('/content/drive')


import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import random

import re


from sklearn.linear_model import LogisticRegression

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

from sklearn.feature_extraction.text import CountVectorizer

from sklearn.model_selection import LeaveOneOut

from sklearn.datasets import make_blobs

from sklearn.model_selection import cross_val_score

from sklearn.model_selection import KFold
```

```
from sklearn.metrics import precision_recall_fscore_support
from transformers import pipeline

from nltk import word_tokenize
import nltk
nltk.download('punkt')
from nltk.corpus import stopwords
nltk.download('stopwords')
nltk.download('punkt_tab')

# ----------------------------------------------------------
# Célula 2 — Carregamento da base IMDB
# (Breve: carrega o arquivo Excel com os dados e mostra as primeiras linhas)
# ----------------------------------------------------------

file_path = '/content/drive/MyDrive/TCC/IMDB_analise_sentimento_usd_deflacionado.xlsx'
df = pd.read_excel(file_path)
df.head()

# ----------------------------------------------------------
# Célula 3 — Modelo de Análise de Sentimento (tokenizer, modelo, função e aplicação)
# (Breve: carrega tokenizer e modelo, define mapeamento de classes,
#  cria função que retorna sentimento, score e probabilidades, aplica ao DF e salva)
# ----------------------------------------------------------
```

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
import torch.nn.functional as F
import pandas as pd

# === Carregar modelo e tokenizer ===
model_name = "tabularisai/multilingual-sentiment-analysis"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSequenceClassification.from_pretrained(model_name)

# === Mapeamento de classes ===
sentiment_map = {
    0: "VERY_NEGATIVE",
    1: "NEGATIVE",
    2: "NEUTRAL",
    3: "POSITIVE",
    4: "VERY_POSITIVE"
}

# === Função de análise ===
def analisar_texto_com_score(texto):
    if not isinstance(texto, str) or texto.strip() == "":
        return pd.Series(["", 0.0, 0.0, 0.0, 0.0, 0.0, 0.0])

    inputs = tokenizer(texto, return_tensors="pt", truncation=True, padding=True,
max_length=512)
```

```
with torch.no_grad():
    outputs = model(**inputs)
    probs = F.softmax(outputs.logits, dim=-1).squeeze(0)

probs_np = probs.detach().cpu().numpy()

# Probabilidades individuais
prob_very_negative = float(probs_np[0])
prob_negative      = float(probs_np[1])
prob_neutral       = float(probs_np[2])
prob_positive      = float(probs_np[3])
prob_very_positive = float(probs_np[4])

# Classe mais provável
sentiment = sentiment_map[int(torch.argmax(probs))]

# Score contínuo em [-1, +1] (média ponderada)
weights = torch.linspace(-1.0, 1.0, steps=model.config.num_labels)
score = float(torch.sum(probs * weights))

return pd.Series([
    sentiment,
    score,
    prob_very_negative,
    prob_negative,
```

```
        prob_neutral,

        prob_positive,

        prob_very_positive

    ])


# === Aplicar no DataFrame ===

df[['sentiment', 'score', 'prob_very_negative', 'prob_negative', 'prob_neutral', 'prob_positive', 'prob_very_positive']] = \

    df['description'].apply(analisar_texto_com_score)


# === Criar coluna main_genre (se ainda não existir) ===

if 'main_genre' not in df.columns:

    df['main_genre'] = df['genre'].apply(lambda x: x.split(',')[0] if isinstance(x, str) else x)


# === Visualizar amostra ===

print(df[['description', 'sentiment', 'score']].head())


# Salvar resultados

output_path = '/content/drive/MyDrive/TCC/IMDB_analise_sentimento_usd_deflacionado_final.xlsx'

df.to_excel(output_path, index=False)

print(f"Arquivo salvo em: {output_path}")


# ----------------------------------------------------------

# Célula 4 — Recarregar arquivo final

# (Breve: reabre o arquivo gerado para usos posteriores)
```

```
# ---------------------------------------------------------

file_path =
'/content/drive/MyDrive/TCC/IMDB_analise_sentimento_usd_deflacionado_final.xlsx'

df = pd.read_excel(file_path)

df.head()


# ---------------------------------------------------------

# Célula 5 — Estatísticas descritivas e histograma do score

# (Breve: calcula distribuição de sentimentos, médias financeiras por sentimento,

#  exibe tabelas e plota histograma do score)

# ---------------------------------------------------------


import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

import numpy as np


# Calcula a contagem de amostras para cada sentimento

sentiment_counts = df['sentiment'].value_counts().reset_index()

sentiment_counts.columns = ['Sentiment', 'Count']


# Calcula a porcentagem de cada sentimento

total_count = sentiment_counts['Count'].sum()

sentiment_counts['Percentage'] = (sentiment_counts['Count'] / total_count) * 100
```

```
# Exibe a tabela

print("Distribution of Sentiment:")

display(sentiment_counts)


# Calcula e exibe as métricas financeiras médias por sentimento

sentiment_finance = df.groupby('sentiment')[['worldwide_gross_usd_real', 'budget_usd_real']].mean().reset_index()

print("Average Worldwide Gross and Budget by Sentiment (Real USD):")

pd.options.display.float_format = '{:,.2f}'.format

display(sentiment_finance)


# Distribuição do 'sentiment_score'

plt.figure(figsize=(10, 6))

sns.histplot(df['score'].dropna(), kde=True, bins=30)

plt.title('Distribution of Sentiment Score')

plt.xlabel('Sentiment Score')

plt.ylabel('Frequency')

plt.show()


# ----------------------------------------------------------

# Célula 6 — Tradução de sentimentos e gêneros + gráficos por gênero

# (Breve: traduz rótulos, identifica top8 gêneros e plota distribuição absoluta e percentual)

# ----------------------------------------------------------


import matplotlib.pyplot as plt

import seaborn as sns
```

```
import pandas as pd

import numpy as np


# Traduções

sentiment_translation = {

    'VERY_NEGATIVE': 'Muito Negativo',

    'NEGATIVE': 'Negativo',

    'NEUTRAL': 'Neutro',

    'POSITIVE': 'Positivo',

    'VERY_POSITIVE': 'Muito Positivo'

}


genre_translation = {

    'Action': 'Ação',

    'Comedy': 'Comédia',

    'Drama': 'Drama',

    'Crime': 'Crime',

    'Biography': 'Biografia',

    'Animation': 'Animação',

    'Horror': 'Terror',

    'Adventure': 'Aventura',

    'Sci-Fi': 'Ficção Científica',

    'Mystery': 'Mistério',

    'Family': 'Família',

    'Fantasy': 'Fantasia',
```

```
    'Thriller': 'Suspense',
    'History': 'História',
    'Music': 'Música',
    'War': 'Guerra',
    'Western': 'Western',
    'Sport': 'Esporte',
    'Musical': 'Musical',
    'Film-Noir': 'Film-Noir',
    'Documentary': 'Documentary'
}

# Aplicar traduções
df['sentiment_translated'] = df['sentiment'].map(sentiment_translation).fillna(df['sentiment'])
df['main_genre_translated'] = df['main_genre'].map(genre_translation).fillna(df['main_genre'])

# Define uma paleta personalizada para sentimentos (neutral com amarelo levemente mais escuro)
sentiment_palette = {
    'Muito Negativo': '#b2182b', # Vermelho escuro
    'Negativo': '#ef8a62',       # Vermelho mais claro
    'Neutro': '#fff7bc',         # Amarelo claro
    'Positivo': '#a1d99b',       # Verde claro
    'Muito Positivo': '#31a354'  # Verde escuro
}

# Ordem desejada dos sentimentos para plot
```

```
sentiment_order = ['Muito Negativo', 'Negativo', 'Neutro', 'Positivo', 'Muito Positivo']



# Identifica os top 8 gêneros por contagem (usando nomes traduzidos)

genre_counts = df['main_genre_translated'].value_counts().reset_index()

genre_counts.columns = ['main_genre_translated', 'total_count']

top_8_genres_translated = genre_counts['main_genre_translated'].head(8).tolist()


# Filtra o dataframe para incluir somente os top 8 gêneros

df_top8_genres = df[df['main_genre_translated'].isin(top_8_genres_translated)].copy()


# Tabela pivô de contagens de sentimento dentro dos top 8 gêneros

# Garante que todas as categorias de sentimento estejam presentes nas colunas

pivot_sentiment_top8 = df_top8_genres.groupby(['main_genre_translated', 'sentiment_translated']).size().unstack(fill_value=0).reindex(columns=sentiment_order, fill_value=0)


# Ordena gêneros pela lista top_8_genres_translated

pivot_sentiment_top8 = pivot_sentiment_top8.loc[top_8_genres_translated]


# --- Gráfico 1: Distribuição de Sentimentos por Gênero (Contagem) ---

plt.figure(figsize=(13, 8))

ax1 = pivot_sentiment_top8.plot(kind='bar', stacked=True, ax=plt.gca(), color=[sentiment_palette[s] for s in pivot_sentiment_top8.columns])

plt.title('Distribuição de sentimentos por gênero cinematográfico')

plt.xlabel('Gênero')

plt.ylabel('Quantidade de Filmes')
```

```
plt.legend(title='Sentimento', bbox_to_anchor=(1.05, 1), loc='upper left')

plt.xticks(rotation=45, ha='right')

plt.tight_layout()

plt.show()




# Calcula a distribuição percentual para o segundo gráfico

pivot_sentiment_top8_percent = pivot_sentiment_top8.divide(pivot_sentiment_top8.sum(axis=1), axis=0) * 100



# --- Gráfico 2: Distribuição Percentual de Sentimentos por Gênero ---

plt.figure(figsize=(13, 8))

ax2 = pivot_sentiment_top8_percent.plot(kind='bar', stacked=True, ax=plt.gca(), color=[sentiment_palette[s] for s in pivot_sentiment_top8_percent.columns])

plt.title('Distribuição percentual de sentimentos por gênero cinematográfico')

plt.xlabel('Gênero')

plt.ylabel('Percentual (%)')

plt.legend(title='Sentimento', bbox_to_anchor=(1.05, 1), loc='upper left')

plt.xticks(rotation=45, ha='right')

plt.tight_layout()

plt.show()



# ------------------------------------------------------------

# Célula 7 — Boxplot do orçamento por gênero (top 8)

# (Breve: converte orçamento para numérico, seleciona top8 e plota boxplot com médias)

# ------------------------------------------------------------
```

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np

# --- Preparação dos dados ---
# Garante que budget_usd_real seja numérico
df['budget_usd_real'] = pd.to_numeric(df['budget_usd_real'], errors='coerce')

# Remove linhas sem gênero ou sem budget
df_clean = df.dropna(subset=['main_genre', 'budget_usd_real']).copy()

# Tradução dos gêneros para português
genre_translation = {
    'Action': 'Ação',
    'Adventure': 'Aventura',
    'Animation': 'Animação',
    'Comedy': 'Comédia',
    'Crime': 'Crime',
    'Documentary': 'Documentário',
    'Drama': 'Drama',
    'Family': 'Família',
    'Fantasy': 'Fantasia',
    'History': 'História',
```

```
    'Horror': 'Terror',
    'Music': 'Música',
    'Mystery': 'Mistério',
    'Romance': 'Romance',
    'Science Fiction': 'Ficção Científica',
    'Thriller': 'Suspense',
    'Biography': 'Biografia',
    'War': 'Guerra',
    'Western': 'Faroeste'
}

# Aplica a tradução, mantendo o original caso não exista tradução
df_clean['main_genre_pt'] = df_clean['main_genre'].map(genre_translation).fillna(df_clean['main_genre'])

# Seleciona os 8 gêneros mais frequentes (em português)
top8_pt = df_clean['main_genre_pt'].value_counts().nlargest(8).index.tolist()

# Filtra o dataframe e mantém a ordem dos gêneros no plot
df_top8 = df_clean[df_clean['main_genre_pt'].isin(top8_pt)].copy()
df_top8['main_genre_pt'] = pd.Categorical(df_top8['main_genre_pt'], categories=top8_pt, ordered=True)

# --- Plot ---
sns.set(style="whitegrid")
plt.figure(figsize=(16, 8))
```

```
ax = sns.boxplot(
    data=df_top8,
    x='main_genre_pt',
    y='budget_usd_real',
    order=top8_pt,
    showfliers=True,
    showmeans=False,
    palette='pastel'
)

# Calcula as médias por gênero (na mesma ordem)
means = df_top8.groupby('main_genre_pt')['budget_usd_real'].mean().reindex(top8_pt)

# Desenha as médias como diamantes
x_positions = np.arange(len(top8_pt))
ax.scatter(x_positions, means.values, marker='D', s=80, edgecolor='black', facecolor='white', zorder=10, label='Média')

# Anota os valores médios acima dos diamantes
for i, m in enumerate(means.values):
    ax.text(i, m * 1.03, f'{m:,.0f}', ha='center', va='bottom', fontsize=9)

# Ajustes de eixos e título
plt.title('Distribuição de orçamento (USD real ajustado) por gênero cinematográfico', fontsize=14)
```

```
plt.xlabel('Gênero', fontsize=12)
plt.ylabel('Orçamento (em centenas de milhões de USD)', fontsize=12)
plt.xticks(rotation=45, ha='right')

# Define limite inferior do eixo Y
y_max = df_top8['budget_usd_real'].max()
if pd.notna(y_max) and y_max > 0:
    plt.ylim(bottom=0, top=y_max * 1.08)

# Legenda
ax.legend(loc='upper right')

plt.tight_layout()
plt.show()

# ----------------------------------------------------------
# Célula 8 — Boxplot da renda mundial por gênero (top 8)
# (Breve: procedimento análogo ao anterior, aplicado à receita mundial ajustada)
# ----------------------------------------------------------

import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
```

```
# --- Preparação dos dados ---

# Garante que worldwide_gross_usd_real seja numérico

df['worldwide_gross_usd_real'] = pd.to_numeric(df['worldwide_gross_usd_real'], errors='coerce')


# Remove linhas sem gênero ou sem renda

df_clean = df.dropna(subset=['main_genre', 'worldwide_gross_usd_real']).copy()


# Tradução dos gêneros para português

genre_translation = {

    'Action': 'Ação',

    'Adventure': 'Aventura',

    'Animation': 'Animação',

    'Comedy': 'Comédia',

    'Crime': 'Crime',

    'Documentary': 'Documentário',

    'Drama': 'Drama',

    'Family': 'Família',

    'Fantasy': 'Fantasia',

    'History': 'História',

    'Horror': 'Terror',

    'Music': 'Música',

    'Mystery': 'Mistério',

    'Romance': 'Romance',

    'Science Fiction': 'Ficção Científica',

    'Thriller': 'Suspense',
```

```
    'Biography': 'Biografia',

    'War': 'Guerra',

    'Western': 'Faroeste'

}


# Aplica a tradução, mantendo o original caso não exista tradução

df_clean['main_genre_pt'] =
df_clean['main_genre'].map(genre_translation).fillna(df_clean['main_genre'])


# Seleciona os 8 gêneros mais frequentes (em português)

top8_pt = df_clean['main_genre_pt'].value_counts().nlargest(8).index.tolist()


# Filtra o dataframe e mantém a ordem dos gêneros no plot

df_top8 = df_clean[df_clean['main_genre_pt'].isin(top8_pt)].copy()

df_top8['main_genre_pt'] = pd.Categorical(df_top8['main_genre_pt'], categories=top8_pt,
ordered=True)


# --- Plot ---

sns.set(style="whitegrid")

plt.figure(figsize=(16, 8))


ax = sns.boxplot(

    data=df_top8,

    x='main_genre_pt',

    y='worldwide_gross_usd_real',

    order=top8_pt,
```

```
    showfliers=True,

    showmeans=False,

    palette='pastel'

)


# Calcula as médias por gênero (na mesma ordem)

means = df_top8.groupby('main_genre_pt')['worldwide_gross_usd_real'].mean().reindex(top8_pt)


# Desenha as médias como diamantes

x_positions = np.arange(len(top8_pt))

ax.scatter(x_positions, means.values, marker='D', s=80, edgecolor='black', facecolor='white', zorder=10, label='Média')


# Anota os valores médios acima dos diamantes

for i, m in enumerate(means.values):

    ax.text(i, m * 1.03, f'{m:,.0f}', ha='center', va='bottom', fontsize=9)


# Ajustes de eixos e título

plt.title('Distribuição da renda (USD real ajustado) por gênero cinematográfico', fontsize=14)

plt.xlabel('Gênero', fontsize=12)

plt.ylabel('Renda (em bilhões de USD)', fontsize=12)

plt.xticks(rotation=45, ha='right')


# Define limite inferior do eixo Y

y_max = df_top8['worldwide_gross_usd_real'].max()
```

```
if pd.notna(y_max) and y_max > 0:
    plt.ylim(bottom=0, top=y_max * 1.08)

# Legenda
ax.legend(loc='upper right')

plt.tight_layout()
plt.show()

# ---------------------------------------------------------
# Célula 9 — Tradução de países, mapeamento de regiões e Sankey
# (Breve: dicionários de tradução/regionamento, funções auxiliares,
#  agregações e criação de diagrama de Sankey: País → Região → Gênero)
# ---------------------------------------------------------

# Requisitos: pip install pandas plotly
import pandas as pd
import plotly.graph_objects as go

REGION_LABELS = [
    "(I) África",
    "(II) Leste Asiático e Pacífico",
    "(III) Europa e Ásia Central",
    "(IV) América Latina e Caribe",
    "(V) Oriente Médio e Norte da África",
```

"(VI) Sul Asiático",

"(VII) América do Norte"

]

Note: não colocamos "Outros" nas regiões — o usuário pediu isso.

--- DICIONÁRIO DE TRADUÇÃO (INGLÊS -> PORTUGUÊS) ---

Cobertura ampla com muitas variações comuns; adicione itens extras se o relatório final mostrar nomes não mapeados.

COUNTRY_TRANSLATION = {

América do Norte

"United States": "Estados Unidos", "United States of America": "Estados Unidos", "US": "Estados Unidos", "USA": "EUA",

"Canada": "Canadá",

América Central / Caribe / América do Sul

"Mexico": "México", "Mexico, USA": "México",

"Guatemala": "Guatemala", "Belize": "Belize", "Honduras": "Honduras", "El Salvador": "El Salvador",

"Nicaragua": "Nicarágua", "Costa Rica": "Costa Rica", "Panama": "Panamá",

"Cuba": "Cuba", "Dominican Republic": "República Dominicana", "Puerto Rico": "Porto Rico",

"Jamaica": "Jamaica", "Bahamas": "Bahamas", "Barbados": "Barbados",

"Trinidad and Tobago": "Trinidad e Tobago",

"Colombia": "Colômbia", "Venezuela": "Venezuela", "Ecuador": "Equador", "Peru": "Peru",

"Bolivia": "Bolívia", "Paraguay": "Paraguai", "Chile": "Chile", "Argentina": "Argentina",

"Uruguay": "Uruguai", "Brazil": "Brasil", "Brasil": "Brasil",

Europa Ocidental / Central / Norte / Sul

"United Kingdom": "Reino Unido", "UK": "Reino Unido", "England": "Reino Unido",

"Scotland": "Reino Unido", "Wales": "Reino Unido", "Northern Ireland": "Reino Unido",

"Ireland": "Irlanda", "France": "França", "Germany": "Alemanha", "Spain": "Espanha", "Italy": "Itália",

"Portugal": "Portugal", "Netherlands": "Países Baixos", "Holland": "Países Baixos", "Belgium": "Bélgica",

"Switzerland": "Suíça", "Austria": "Áustria", "Luxembourg": "Luxemburgo",

"Denmark": "Dinamarca", "Sweden": "Suécia", "Norway": "Noruega", "Finland": "Finlândia",

"Iceland": "Islândia", "Poland": "Polônia", "Czech Republic": "República Tcheca",

"Czechia": "Tchéquia", "Hungary": "Hungria", "Romania": "Romênia", "Bulgaria": "Bulgária",

"Slovakia": "Eslováquia", "Slovenia": "Eslovênia", "Croatia": "Croácia", "Serbia": "Sérvia",

"Montenegro": "Montenegro", "Bosnia and Herzegovina": "Bósnia e Herzegovina",

"North Macedonia": "Macedônia do Norte", "Macedonia": "Macedônia do Norte", "Albania": "Albânia",

"Greece": "Grécia", "Turkey": "Turquia", "Cyprus": "Chipre", "Malta": "Malta",

"Monaco": "Mônaco", "Andorra": "Andorra", "San Marino": "San Marino",

"Vatican": "Vaticano", "Liechtenstein": "Liechtenstein",

"Estonia": "Estônia", "Latvia": "Letônia", "Lithuania": "Lituânia",

Europa Oriental / ex-URSS / Cáucaso / Ásia Central

"Russia": "Rússia", "Russian Federation": "Rússia", "Ukraine": "Ucrânia", "Belarus": "Bielorrússia",

"Moldova": "Moldávia", "Georgia": "Geórgia", "Armenia": "Armênia", "Azerbaijan": "Azerbaijão",

"Kazakhstan": "Cazaquistão", "Uzbekistan": "Uzbequistão", "Turkmenistan": "Turcomenistão", "Kyrgyzstan": "Quirguistão", "Tajikistan": "Tajiquistão",

África

"Nigeria": "Nigéria", "Egypt": "Egito", "South Africa": "África do Sul", "Kenya": "Quênia",

"Morocco": "Marrocos", "Algeria": "Argélia", "Tunisia": "Tunísia", "Ghana": "Gana",

"Mauritania": "Mauritânia", "Ethiopia": "Etiópia", "Sudan": "Sudão", "Angola": "Angola",

"Mozambique": "Moçambique", "Madagascar": "Madagáscar", "Cameroon": "Camarões",

"Ivory Coast": "Costa do Marfim", "Cote d'Ivoire": "Costa do Marfim", "Senegal": "Senegal",

"Uganda": "Uganda", "Tanzania": "Tanzânia", "Zimbabwe": "Zimbábue", "Zambia": "Zâmbia",

"Namibia": "Namíbia", "Botswana": "Botsuana", "Libya": "Líbia", "Somalia": "Somália",

"Burkina Faso": "Burkina Faso", "Niger": "Níger", "Benin": "Benim",

Médio Oriente / Norte da África (MENA)

"Saudi Arabia": "Arábia Saudita", "United Arab Emirates": "Emirados Árabes Unidos", "UAE": "Emirados Árabes Unidos",

"Iran": "Irã", "Iraq": "Iraque", "Israel": "Israel", "Jordan": "Jordânia", "Lebanon": "Líbano",

"Syria": "Síria", "Oman": "Omã", "Yemen": "Iémen", "Qatar": "Catar", "Bahrain": "Bahrein", "Kuwait": "Kuwait",

"Palestine": "Palestina", "State of Palestine": "Palestina", "Palestinian": "Palestina",

Sul da Ásia

"India": "Índia", "Pakistan": "Paquistão", "Bangladesh": "Bangladesh", "Sri Lanka": "Sri Lanka",

"Nepal": "Nepal", "Bhutan": "Butão", "Maldives": "Maldivas", "Afghanistan": "Afeganistão",

```
    # Leste Asiático e Pacífico / Sudeste Asiático / Oceania

    "China": "China", "Hong Kong": "Hong Kong", "Macau": "Macau", "Japan": "Japão",

    "South Korea": "Coreia do Sul", "North Korea": "Coreia do Norte", "Korea": "Coreia",

    "Taiwan": "Taiwan", "Mongolia": "Mongólia",

    "Indonesia": "Indonésia", "Malaysia": "Malásia", "Philippines": "Filipinas",

    "Vietnam": "Vietnã", "Thailand": "Tailândia", "Singapore": "Singapura",

    "Cambodia": "Camboja", "Laos": "Laos", "Brunei": "Brunei", "Myanmar": "Mianmar",

    "Australia": "Austrália", "New Zealand": "Nova Zelândia", "Fiji": "Fiji", "Papua New Guinea": "Papua Nova Guiné",



    # Territórios/variações comuns em datasets

    "Hong Kong SAR": "Hong Kong", "Isle Of Man": "Ilha de Man", "Isle of Man": "Ilha de Man",

    "Sao Tome and Principe": "São Tomé e Príncipe", "Curaçao": "Curaçau", "Aruba": "Aruba",

    "Bermuda": "Bermuda", "Guernsey": "Guernsey", "Jersey": "Jersey",



    # Ruídos e padrões

    "Other": "Outro", "Others": "Outro", "Unknown": "Outro", "New Line Cinema": "New Line Cinema",

    # Garantias explícitas pedidas:

    "Estonia": "Estônia", "Isle Of Man": "Ilha de Man", "Palestine": "Palestina", "Serbia": "Sérvia",

    # (adicione/edite conforme o relatório final mostrar novos nomes)

}


# --- MAPEAMENTO (inglês) país -> região (valores em português) ---

# Versão em inglês que mapeia para as sete regiões em português.
```

```
REGION_MAP_EN = {
    # África
    "Nigeria": "(I) África", "Egypt": "(I) África", "South Africa": "(I) África", "Kenya": "(I) África",
    "Morocco": "(I) África", "Algeria": "(I) África", "Tunisia": "(I) África", "Ghana": "(I) África",

    # Leste Asiático e Pacífico
    "China": "(II) Leste Asiático e Pacífico", "Japan": "(II) Leste Asiático e Pacífico",
    "South Korea": "(II) Leste Asiático e Pacífico", "North Korea": "(II) Leste Asiático e Pacífico",
    "Australia": "(II) Leste Asiático e Pacífico", "New Zealand": "(II) Leste Asiático e Pacífico",
    "Philippines": "(II) Leste Asiático e Pacífico", "Indonesia": "(II) Leste Asiático e Pacífico",
    "Singapore": "(II) Leste Asiático e Pacífico", "Thailand": "(II) Leste Asiático e Pacífico",
    "Vietnam": "(II) Leste Asiático e Pacífico", "Hong Kong": "(II) Leste Asiático e Pacífico",
    "Malaysia": "(II) Leste Asiático e Pacífico", "Taiwan": "(II) Leste Asiático e Pacífico",
    "Cambodia": "(II) Leste Asiático e Pacífico", "Mongolia": "(II) Leste Asiático e Pacífico",
    "Laos": "(II) Leste Asiático e Pacífico", "Brunei": "(II) Leste Asiático e Pacífico",

    # Europa e Ásia Central
    "United Kingdom": "(III) Europa e Ásia Central", "France": "(III) Europa e Ásia Central",
    "Germany": "(III) Europa e Ásia Central", "Italy": "(III) Europa e Ásia Central",
    "Spain": "(III) Europa e Ásia Central", "Russia": "(III) Europa e Ásia Central",
    "Poland": "(III) Europa e Ásia Central", "Netherlands": "(III) Europa e Ásia Central",
    "Sweden": "(III) Europa e Ásia Central", "Switzerland": "(III) Europa e Ásia Central",
    "Norway": "(III) Europa e Ásia Central", "Denmark": "(III) Europa e Ásia Central",
    "Finland": "(III) Europa e Ásia Central", "Ireland": "(III) Europa e Ásia Central",
```

"Belgium": "(III) Europa e Ásia Central", "Austria": "(III) Europa e Ásia Central",

"Greece": "(III) Europa e Ásia Central", "Iceland": "(III) Europa e Ásia Central",

"Malta": "(III) Europa e Ásia Central", "Serbia": "(III) Europa e Ásia Central",

"Georgia": "(III) Europa e Ásia Central", "Ukraine": "(III) Europa e Ásia Central",

"Kazakhstan": "(III) Europa e Ásia Central", "Czech Republic": "(III) Europa e Ásia Central",

"Hungary": "(III) Europa e Ásia Central", "Slovakia": "(III) Europa e Ásia Central",

"Romania": "(III) Europa e Ásia Central", "Luxembourg": "(III) Europa e Ásia Central",

"Croatia": "(III) Europa e Ásia Central", "Slovenia": "(III) Europa e Ásia Central",

"Turkey": "(III) Europa e Ásia Central", "Monaco": "(III) Europa e Ásia Central",

"Bulgaria": "(III) Europa e Ásia Central", "Portugal": "(III) Europa e Ásia Central",

"Estonia": "(III) Europa e Ásia Central", "Latvia": "(III) Europa e Ásia Central", "Lithuania": "(III) Europa e Ásia Central",

"Isle Of Man": "(III) Europa e Ásia Central", "Bosnia and Herzegovina": "(III) Europa e Ásia Central",

"North Macedonia": "(III) Europa e Ásia Central", "Montenegro": "(III) Europa e Ásia Central",

América Latina e Caribe

"Brazil": "(IV) América Latina e Caribe", "Argentina": "(IV) América Latina e Caribe", "Mexico": "(IV) América Latina e Caribe",

"Colombia": "(IV) América Latina e Caribe", "Chile": "(IV) América Latina e Caribe", "Peru": "(IV) América Latina e Caribe",

"Cuba": "(IV) América Latina e Caribe", "Uruguay": "(IV) América Latina e Caribe", "Aruba": "(IV) América Latina e Caribe",

"Bahamas": "(IV) América Latina e Caribe", "Puerto Rico": "(IV) América Latina e Caribe", "Panama": "(IV) América Latina e Caribe",

"Ecuador": "(IV) América Latina e Caribe", "Dominican Republic": "(IV) América Latina e Caribe",

Oriente Médio e Norte da África

"Saudi Arabia": "(V) Oriente Médio e Norte da África", "United Arab Emirates": "(V) Oriente Médio e Norte da África",

"Iran": "(V) Oriente Médio e Norte da África", "Iraq": "(V) Oriente Médio e Norte da África",

"Israel": "(V) Oriente Médio e Norte da África", "Jordan": "(V) Oriente Médio e Norte da África",

"Lebanon": "(V) Oriente Médio e Norte da África", "Syria": "(V) Oriente Médio e Norte da África",

"Oman": "(V) Oriente Médio e Norte da África", "Yemen": "(V) Oriente Médio e Norte da África",

"Qatar": "(V) Oriente Médio e Norte da África", "Bahrain": "(V) Oriente Médio e Norte da África",

"Kuwait": "(V) Oriente Médio e Norte da África", "Palestine": "(V) Oriente Médio e Norte da África",

Sul Asiático

"India": "(VI) Sul Asiático", "Pakistan": "(VI) Sul Asiático", "Bangladesh": "(VI) Sul Asiático",

"Sri Lanka": "(VI) Sul Asiático", "Nepal": "(VI) Sul Asiático", "Bhutan": "(VI) Sul Asiático",

"Maldives": "(VI) Sul Asiático", "Afghanistan": "(VI) Sul Asiático",

América do Norte (territórios)

"Bermuda": "(VII) América do Norte", "Greenland": "(VII) América do Norte"

}

--- construir REGION_MAP em português usando COUNTRY_TRANSLATION e REGION_MAP_EN ---

REGION_MAP = {}

for en_name, pt_name in COUNTRY_TRANSLATION.items():

```
    region = REGION_MAP_EN.get(en_name)
    if region:
        REGION_MAP[pt_name] = region

# adições diretas (garantias)
REGION_MAP.update({
    "Estados Unidos": "(VII) América do Norte",
    "EUA": "(VII) América do Norte",
    "Canadá": "(VII) América do Norte",
    "Brasil": "(IV) América Latina e Caribe",
    "Argentina": "(IV) América Latina e Caribe",
    "México": "(IV) América Latina e Caribe",
    "Portugal": "(III) Europa e Ásia Central",
    "França": "(III) Europa e Ásia Central",
    "Alemanha": "(III) Europa e Ásia Central",
    "China": "(II) Leste Asiático e Pacífico",
    "Japão": "(II) Leste Asiático e Pacífico",
    "Índia": "(VI) Sul Asiático",
    "Emirados Árabes Unidos": "(V) Oriente Médio e Norte da África",
    "Arábia Saudita": "(V) Oriente Médio e Norte da África",
    "Egito": "(I) África",
    "Nigéria": "(I) África",
    "África do Sul": "(I) África",
    "Palestina": "(V) Oriente Médio e Norte da África",
    "Ilha de Man": "(III) Europa e Ásia Central",
```

```
    "Estônia": "(III) Europa e Ásia Central",
})

# --- TRADUÇÃO DE GÊNEROS (inglês -> português) ---
GENRE_TRANSLATION = {
    "Drama": "Drama", "Comedy": "Comédia", "Romance": "Romance", "Action": "Ação",
    "Documentary": "Documentário", "Horror": "Terror", "Thriller": "Suspense",
    "Adventure": "Aventura", "Animation": "Animação", "Sci-Fi": "Ficção Científica",
    "Science Fiction": "Ficção Científica", "Fantasy": "Fantasia", "Family": "Família",
    "Biography": "Biografia", "Music": "Musical", "Mystery": "Mistério", "Crime": "Crime",
    "History": "História", "Sport": "Esporte", "Western": "Faroeste", "Short": "Curta-metragem",
    "War": "Guerra", "Musical": "Musical", "Talk-Show": "Talk Show"
}

# --- FUNÇÕES AUXILIARES ---
def traduzir_pais(pais_raw):
    if pd.isna(pais_raw):
        return pais_raw
    s = str(pais_raw).strip()
    if ',' in s:
        s = s.split(',')[0].strip()
    # se já estiver em português e mapeado por região, retorna
    if s in REGION_MAP:
        return s
    # tradução direta
```

```
    if s in COUNTRY_TRANSLATION:
        return COUNTRY_TRANSLATION[s]
    # tentar formas title/upper
    t = s.title()
    if t in COUNTRY_TRANSLATION:
        return COUNTRY_TRANSLATION[t]
    u = s.upper()
    if u in COUNTRY_TRANSLATION:
        return COUNTRY_TRANSLATION[u]
    # se nada, retorna original (será listado no relatório "não mapeados")
    return s


def mapear_pais_para_regiao(pais_trad):
    if pd.isna(pais_trad):
        return None
    s = str(pais_trad).strip()
    # mapeamento direto
    if s in REGION_MAP:
        return REGION_MAP[s]
    # equivalências curtas
    equi = {"EUA": "Estados Unidos", "USA": "Estados Unidos", "US": "Estados Unidos",
            "UK": "Reino Unido", "U.K.": "Reino Unido"}
    if s in equi:
        return REGION_MAP.get(equi[s])
    return None  # sem região mapeada
```

```
def traduzir_genero(g_raw):
    if pd.isna(g_raw):
        return g_raw
    s = str(g_raw).strip()
    # strings compostas (ex.: "Drama,Comedy")
    if ',' in s:
        parts = [p.strip() for p in s.split(',')]
        translated = [GENRE_TRANSLATION.get(p, p) for p in parts]
        # juntar por "/" — depois o agrupamento top8 tratará a string resultante
        return "/".join(translated)
    if s in GENRE_TRANSLATION:
        return GENRE_TRANSLATION[s]
    t = s.title()
    if t in GENRE_TRANSLATION:
        return GENRE_TRANSLATION[t]
    return s

# --- CARREGAR DADOS ---
df = pd.read_excel(file_path)

# --- TRADUZIR PAÍS E GÊNERO ---
df["pais_trad"] = df["main_country"].apply(traduzir_pais)
df["genero_trad"] = df["main_genre"].apply(traduzir_genero)
```

```
# --- MAPEAR REGIÃO (None se não mapeado) ---

df["regiao"] = df["pais_trad"].apply(mapear_pais_para_regiao)


# --- REMOVER LINHAS COM GÊNERO/NOME DE PAÍS NULOS (mas manter países sem região) ---

df_limpo = df.dropna(subset=["pais_trad", "genero_trad"]).copy()

total_filmes = len(df_limpo)


# --- DETERMINAR OS 8 MAIORES GÊNEROS (após tradução) e AGRUPAR RESTO em "Outros gêneros" ---

top8 = df_limpo["genero_trad"].value_counts().nlargest(8)

top8_list = top8.index.tolist()

print("Top 8 gêneros (após tradução):")

for g, cnt in top8.items():

    print(f"  {g}: {cnt}")


df_limpo["genero_top"] = df_limpo["genero_trad"].apply(lambda g: g if g in top8_list else "Outros gêneros")

outros_count = (df_limpo["genero_top"] == "Outros gêneros").sum()

print(f"\nFilmes agrupados em 'Outros gêneros': {outros_count} de {total_filmes}")


# --- AGREGAR DADOS ---

# 1) Contagens país -> região (apenas para países com regiao)

pais_regiao = df_limpo.dropna(subset=["regiao"]).groupby(["pais_trad", "regiao"]).size().reset_index(name="count")

# 2) Contagens região -> gênero_top (apenas para linhas com região)

regiao_genero = df_limpo.dropna(subset=["regiao"]).groupby(["regiao", "genero_top"]).size().reset_index(name="count")
```

```
# 3) Contagens país -> gênero_top para países SEM região (fluxo direto)

pais_genero_sem_regiao = df_limpo[df_limpo["regiao"].isna()].groupby(["pais_trad", "genero_top"]).size().reset_index(name="count")


# --- CRIAR NÓS ---

# países (tanto com região quanto sem), regiões (apenas as 7), gêneros (top8 + "Outros gêneros")

paises_presentes = df_limpo["pais_trad"].astype(str).unique().tolist()

regioes_presentes = sorted(list(set(REGION_LABELS)))  # só as 7

generos_presentes = sorted(df_limpo["genero_top"].unique().tolist())


nodes = list(paises_presentes) + regioes_presentes + generos_presentes

node_idx = {label: i for i, label in enumerate(nodes)}


# --- RÓTULOS COM TOTAIS E PERCENTUAIS ---

pais_totais = df_limpo["pais_trad"].value_counts().to_dict()

regiao_totais = {r: int(df_limpo[df_limpo["regiao"] == r].shape[0]) for r in regioes_presentes}

genero_totais = df_limpo["genero_top"].value_counts().to_dict()


node_labels = []

for n in nodes:

    if n in pais_totais:

        total = pais_totais[n]; pct = (total / total_filmes) * 100 if total_filmes else 0

        node_labels.append(f"{n}\n({total} filmes, {pct:.1f}%)")

    elif n in regiao_totais:

        total = regiao_totais[n]; pct = (total / total_filmes) * 100 if total_filmes else 0
```

```
        node_labels.append(f"{n}\n({total} filmes, {pct:.1f}%)")
    elif n in genero_totais:
        total = genero_totais[n]; pct = (total / total_filmes) * 100 if total_filmes else 0
        node_labels.append(f"{n}\n({total} filmes, {pct:.1f}%)")
    else:
        node_labels.append(n)

# --- LINKS ---
sources, targets, values, link_labels, customdata = [], [], [], [], []

# Links País -> Região (somente quando região existe)
for _, row in pais_regiao.iterrows():
    pais = str(row["pais_trad"]); reg = row["regiao"]; cnt = int(row["count"])
    if pais in node_idx and reg in node_idx:
        sources.append(node_idx[pais]); targets.append(node_idx[reg]); values.append(cnt)
        link_labels.append(f"{pais} → {reg}: {cnt}"); customdata.append(f"{cnt} filmes")

# Links Região -> Gênero (apenas para países com região)
for _, row in regiao_genero.iterrows():
    reg = row["regiao"]; gen = str(row["genero_top"]); cnt = int(row["count"])
    if reg in node_idx and gen in node_idx:
        sources.append(node_idx[reg]); targets.append(node_idx[gen]); values.append(cnt)
        link_labels.append(f"{reg} → {gen}: {cnt}"); customdata.append(f"{cnt} filmes")

# Links País -> Gênero direto (para países sem região mapeada)
```

```
for _, row in pais_genero_sem_regiao.iterrows():
    pais = str(row["pais_trad"]); gen = str(row["genero_top"]); cnt = int(row["count"])
    if pais in node_idx and gen in node_idx:
        sources.append(node_idx[pais]); targets.append(node_idx[gen]); values.append(cnt)
        link_labels.append(f"{pais} → {gen}: {cnt}"); customdata.append(f"{cnt} filmes")

# --- CRIAR SANKEY ---
sankey = go.Sankey(
    node=dict(
        pad=10, thickness=18, line=dict(color="black", width=0.5),
        label=node_labels,
        hovertemplate="%{label}<extra></extra>"
    ),
    link=dict(
        source=sources, target=targets, value=values, label=link_labels,
        customdata=customdata,
        hovertemplate="%{label}<br>%{customdata}<extra></extra>"
    )
)

fig = go.Figure(sankey)
fig.update_layout(
    title=dict(text="Diagrama de Sankey: País → Região → Gênero Cinematográfico", font=dict(size=14)),
    autosize=False,
    width=1200,
```

```
    height=1000,
    margin=dict(l=50, r=50, t=80, b=50)
)


# --- MOSTRAR (sem salvar) ---
fig.show()

# ----------------------------------------------------------
# Célula 10 — Scatter Orçamento vs Renda (ajustes, remoção de outliers, escala log)
# (Breve: carrega DF v7, limpa outliers, traduz sentimentos e plota scatter em escala log)
# ----------------------------------------------------------

import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np

# Explicitamente carrega o dataframe para garantir que tenha as colunas de sentimento mais recentes
file_path = '/content/drive/MyDrive/TCC/IMDB_analise_sentimento_usd_deflacionado_v7.xlsx'
df = pd.read_excel(file_path)

# Garante que as colunas financeiras sejam numéricas e lida com NaNs para plotagem
df_plot = df.dropna(subset=['budget_usd_real', 'worldwide_gross_usd_real', 'sentiment']).copy()
```

```
df_plot['budget_usd_real'] = pd.to_numeric(df_plot['budget_usd_real'], errors='coerce')

df_plot['worldwide_gross_usd_real'] = pd.to_numeric(df_plot['worldwide_gross_usd_real'], errors='coerce')


# Remove linhas onde a conversão para numérico resultou em NaN

df_plot.dropna(subset=['budget_usd_real', 'worldwide_gross_usd_real'], inplace=True)


# Remove valores <= 0 para compatibilidade com escala logarítmica

df_plot = df_plot[(df_plot['budget_usd_real'] > 0) & (df_plot['worldwide_gross_usd_real'] > 0)].copy()


# --- NOVO: Remover outliers extremos usando percentis ---

# Calcular os percentis para Orçamento e Renda

budget_lower_bound = df_plot['budget_usd_real'].quantile(0.01)

budget_upper_bound = df_plot['budget_usd_real'].quantile(0.99)

gross_lower_bound = df_plot['worldwide_gross_usd_real'].quantile(0.01)

gross_upper_bound = df_plot['worldwide_gross_usd_real'].quantile(0.99)


df_plot = df_plot[

    (df_plot['budget_usd_real'] >= budget_lower_bound) &

    (df_plot['budget_usd_real'] <= budget_upper_bound) &

    (df_plot['worldwide_gross_usd_real'] >= gross_lower_bound) &

    (df_plot['worldwide_gross_usd_real'] <= gross_upper_bound)

].copy()


# Tradução dos rótulos de sentimento para melhores títulos e legendas
```

```
sentiment_translation = {
    'NEGATIVE': 'Negativo',
    'NEUTRAL': 'Neutro',
    'POSITIVE': 'Positivo'
}
df_plot['sentiment_translated'] = df_plot['sentiment'].map(sentiment_translation)

# Define uma ordem consistente para cores e legenda
sentiment_order = ['Negativo', 'Neutro', 'Positivo']
# Filtra sentimentos não presentes nos dados ou não na ordem
sentiment_order_present = [s for s in sentiment_order if s in df_plot['sentiment_translated'].unique()]

# Define paleta de cores customizada com saturação aumentada
sentiment_palette_custom = {
    'Negativo': '#e74c3c',  # Vermelho mais saturado
    'Neutro': '#95a5a6',    # Cinza mais distinto
    'Positivo': '#27ae60'   # Verde mais saturado
}

# Cria o scatter plot com todos os sentimentos em um único gráfico
g = sns.relplot(
    data=df_plot,
    x='budget_usd_real',
    y='worldwide_gross_usd_real',
    hue='sentiment_translated',
```

```
    hue_order=sentiment_order_present,

    palette=sentiment_palette_custom,

    height=6, aspect=1.5,

    s=20, alpha=1.0 # Transparência removida (alpha = 1.0), tamanho dos pontos reduzido

)


# Define títulos e rótulos e formata ticks dos eixos

g.set_axis_labels('Orçamento (USD Real) (Escala Log.)', 'Renda (USD Real) (Escala Log.)') # Indicador de escala log

g.set_titles('') # Remove títulos de facetas (não há facets)


# Aplica escala logarítmica aos eixos x e y

g.ax.set(xscale="log", yscale="log")


# Altera o título da legenda

g.legend.set_title('Sentimento')


plt.title('Orçamento vs. Renda por Sentimento (USD Real)', fontsize=16)

plt.tight_layout()

plt.show()
```