



UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

BRENNO RUSCHIONI DE OLIVEIRA

**Analisando a Influência do Twitter na Criação de Sequências de Filmes de
Terror: Uma Abordagem Baseada em Dados**

São Paulo

2024

BRENNO RUSCHIONI DE OLIVEIRA

**Analisando a Influência do Twitter na Criação de Sequências de Filmes de
Terror: Uma Abordagem Baseada em Dados**

Versão original.

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação. Área de concentração: Metodologia e Técnicas da Computação.

Orientador: Profa. Dra. Marislei Nishijima

São Paulo

2024

Dissertação de autoria de Brenno Ruschioni de Oliveira, sob o título “**Analisando a Influência do Twitter na Criação de Sequências de Filmes de Terror: Uma Abordagem Baseada em Dados**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em ____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr.
Instituição
Presidente

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Ruschioni de Oliveira, Brenno
Analisando a Influência do Twitter na Criação de
Sequências de Filmes de Terror: Uma Abordagem
Baseada em Dados / Brenno Ruschioni de Oliveira;
orientadora, Marislei Nishijima. -- São Paulo,
2024.

101 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de
Pós-Graduação em Sistemas de Informação, Escola de
Artes, Ciências e Humanidades, Universidade de São
Paulo, 2024.

Versão original

1. e-WOM (Electronic Word-of-Mouth). 2.
Aprendizado de Máquina. 3. Análise de Sentimentos.
4. Sequências Cinematográficas. 5. Gênero Terror.
6. Twitter. I. Nishijima, Marislei, orient. II.
Título.

Dedico este trabalho a todos que estiveram ao meu lado durante esta jornada. À minha parceira, Larissa, cuja empatia e coragem inspiraram nossa mudança para Recife, longe da cidade cinza de São Paulo. Ao meu companheiro canino, Zig, meu maior amor e fiel amigo, que enfrentou comigo momentos difíceis, desde a depressão até a solidão da mudança. Aos meus pais, Marcos e Rita, que me apoiaram incondicionalmente: meu pai, que superou uma condição pós-operatória durante a pandemia, e minha mãe, sempre atenta e carinhosa, oferecendo todo tipo de auxílio. Ao meu irmão, Lucas, por nossa reconexão e amizade renovada. E à minha professora Marislei, por respeitar meu tempo, me apoiar nos momentos difíceis e me impulsionar a concluir este estudo, que foi uma imensa fonte de aprendizado e satisfação.

“A arte não é o espelho que reflete o mundo, mas o martelo que o molda”

(Bertolt Brecht)

Resumo

OLIVEIRA, Brenno Ruschioni de. **Analisando a Influência do Twitter na Criação de Sequências de Filmes de Terror: Uma Abordagem Baseada em Dados.**

2024. 104 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2024.

Este estudo explora a influência do e-WOM no Twitter sobre a criação de sequências de filmes de terror, destacando como as interações e sentimentos expressos nas redes sociais podem moldar as decisões das produtoras de cinema. Utilizando um conjunto robusto de dados de fontes como Box Office Mojo, IMDb, Wikipedia e, principalmente, o Twitter, foram identificados padrões de comportamento dos consumidores que indicam uma relação significativa entre a atividade nas redes sociais e a continuidade de franquias cinematográficas.

Os resultados revelaram uma forte correlação entre as métricas de engajamento no Twitter—como retweets, likes e replies—e a probabilidade de um filme gerar uma sequência. A aplicação de algoritmos de aprendizado de máquina, como Random Forest e XGBoost, permitiu prever com alta precisão quais filmes teriam sequências, ressaltando a importância das variáveis financeiras e do engajamento social como indicadores decisivos.

A análise de sentimentos e emoções mostrou que tweets carregados de sentimentos positivos, como alegria e antecipação, eram mais frequentes em filmes que geraram sequências. Em contraste, filmes que não receberam sequências tiveram uma predominância de sentimentos neutros ou negativos, destacando que não é apenas a quantidade de interações que importa, mas também a qualidade emocional do conteúdo.

Além disso, a precisão dos modelos diminuiu quando o balanceamento de classes não foi adequadamente tratado, especialmente no XGBoost, sublinhando a importância de um tratamento cuidadoso dos dados para garantir previsões confiáveis.

Este estudo contribui para a compreensão de como as interações nas redes sociais podem prever o sucesso futuro de filmes e influenciar decisões estratégicas na indústria cinematográfica. As conclusões sugerem que produtoras de cinema podem se beneficiar substancialmente ao aplicar modelos preditivos baseados em análises detalhadas de redes sociais, permitindo decisões mais assertivas sobre a continuidade de filmes. Pesquisas futuras poderiam expandir este estudo para outros gêneros cinematográficos e explorar novas variáveis que captem melhor as nuances culturais e emocionais dos consumidores em diferentes mercados.

Palavras-chaves: e-WOM (Electronic Word-of-Mouth), Aprendizado de Máquina, Análise de Sentimentos, Sequências Cinematográficas, Gênero Terror, Twitter.

Abstract

OLIVEIRA, Brenno Ruschioni de. **Analyzing the Influence of Twitter on the Creation of Horror Movie Sequels: A Data-Driven Approach**. 2024. 104 f.

Dissertation (Master of Science) – School of Arts, Sciences, and Humanities, University of São Paulo, São Paulo, 2024.

This study explores the influence of e-WOM on Twitter in the creation of horror movie sequels, highlighting how interactions and sentiments expressed on social media can shape the decisions of film producers. Utilizing a robust dataset from sources like Box Office Mojo, IMDb, Wikipedia, and primarily Twitter, patterns of consumer behavior were identified, indicating a significant relationship between social media activity and the continuity of film franchises.

The results revealed a strong correlation between Twitter engagement metrics—such as retweets, likes, and replies—and the likelihood of a movie generating a sequel. The application of machine learning algorithms like Random Forest and XGBoost allowed for high-accuracy predictions of which movies would have sequels, emphasizing the importance of financial variables and social engagement as decisive indicators.

Sentiment and emotion analysis showed that tweets loaded with positive sentiments, such as joy and anticipation, were more frequent in movies that generated sequels. In contrast, movies that did not receive sequels had a predominance of neutral or negative sentiments, highlighting that it is not just the quantity of interactions that matters, but also the emotional quality of the content.

Furthermore, the accuracy of the models decreased when class imbalance was not adequately addressed, particularly in XGBoost, underscoring the importance of careful data handling to ensure reliable predictions.

This study contributes to the understanding of how social media interactions can predict the future success of films and influence strategic decisions in the film industry. The findings suggest that film producers can substantially benefit by applying predictive models based on detailed social media analyses, enabling more assertive decisions about film continuity. Future research could expand this study to other film genres and explore new variables that better capture the cultural and emotional nuances of consumers in different markets.

Keywords: e-WOM (Electronic Word-of-Mouth), Machine Learning, Sentiment Analysis, Cinematic Sequels, Horror Genre, Twitter.

Lista de figuras

Figura 1 – Modelo conceitual do valor monetário da extensão de marcas	17
Figura 2 – Principais diferenças entre WOM e e-WOM pelas visões de credibilidade, privacidade, velocidade de transmissão e acessibilidade	24
Figura 3 – Filmes lançados que são sequência em comparação com séries de filmes.	30
Figura 4 – % da bilheteria bruta dos EUA por gênero, considerando o investimento em sequências ou prequelas. Dados referentes aos 100 filmes de maior bilheteria de cada ano, entre 2005 e 2014.	31
Figura 5 – (a) Quantidade de Trabalhos Encontrados e Aceitos	40
Figura 6 – Pesos para Revisão Sistemática	40
Figura 7 – Quantidade de Trabalhos por Critérios de Inclusão	41
Figura 8 – (a) Distribuição de Trabalhos por Ano	42
Figura 9 – (b) Tipos de Trabalhos Aceitos	42
Figura 10 – Dados extraídos do Twitter por Hodeghatta, 2013	47
Figura 11 – Modelo Relacional do Banco de Dados de Tweets e Filmes	53
Figura 12 – (a) Informações de Tweets dos filmes que têm sequência	58
Figura 13 – (b) Informações de Tweets dos filmes que não têm sequência	58
Figura 14 – (a) Informações de Tweets dos Top 10 filmes (Receita Mundial) na base que têm sequência	60
Figura 15 – (b) Informações de Tweets dos Top 10 filmes (Receita Mundial) na base que não têm sequência	60
Figura 16 – Fluxograma dos Modelos de Processamento de Linguagem Natural . .	62
Figura 17 – (a) Medidas de Sentimentos de filmes que têm sequência	63
Figura 18 – (b) Medidas de Sentimentos de filmes que não têm sequência	63
Figura 19 – (a) Sentimentos dos Top 10 filmes (Receita Mundial) na base que têm sequência	64
Figura 20 – (b) Sentimentos dos Top 10 filmes (Receita Mundial) na base que não têm sequência	64
Figura 21 – (a) Emoções de filmes que têm sequência	65
Figura 22 – (b) Emoções de filmes que não têm sequência	65

Figura 23 – (a) Emoções dos Top 10 filmes (Receita Mundial) na base que têm sequência	66
Figura 24 – (b) Emoções dos Top 10 filmes (Receita Mundial) na base que não têm sequência	66
Figura 25 – RF: Filmes que têm ou não sequência	67
Figura 26 – Importância das Variáveis no RF	68
Figura 27 – XGBoost: Importância das Features	70
Figura 28 – RF (Sem Balanceamento): Importância das Features	72
Figura 29 – XGBoost (Sem Balanceamento): Importância das Features	72

Lista de quadros

Quadro 1 – Lista de trabalhos aceitos na revisão sistemática (<i>Continuação</i>)	42
Quadro 1 – Lista de trabalhos aceitos na revisão sistemática	43
Quadro 1 – Lista de trabalhos aceitos na revisão sistemática (<i>Continuação</i>)	43
Quadro 2 – Lista de trabalhos inseridos manualmente na revisão	44
Quadro 3 – Lista de Features	56
Quadro 4 – Tabela de Features dos Modelos de RF e XGBoost.	56
Quadro 5 – Matriz de Confusão RF.	66
Quadro 6 – Relatório de Classificação RF.	68
Quadro 7 – Matriz de Confusão XGBoost.	69
Quadro 8 – Relatório de Classificação do XGBoost.	70
Quadro 9 – Relatório de Classificação do Random Forest (sem balanceamento de classes).	71
Quadro 10 – Relatório de Classificação do XGBoost (sem balanceamento de classes). .	71

Lista de tabelas

Tabela 1 – Palavras-chaves selecionadas para o estudo	37
Tabela 2 – Critérios de Inclusão e Exclusão	38
Tabela 3 – Strings de Busca	39

Lista de abreviaturas e siglas

WOM	Word-of-Mouth
e-WOM	Eletronic-Word-of-Mouth
VADER	Valence Aware Dictionary for sEntiment Reasoning

Sumário

1	Introdução	16
1.1	<i>Tema</i>	16
1.1.1	Sequências de Filmes	16
1.1.2	e-WOM (Eletronic Word-of-Mouth)	16
1.2	<i>Motivação</i>	18
1.3	<i>Lacuna</i>	19
1.4	<i>Hipótese</i>	19
1.5	<i>Objetivos</i>	20
1.6	<i>Justificativa</i>	21
2	Conceitos Fundamentais	22
2.1	<i>WOM e e-WOM</i>	22
2.1.1	Diferenças entre WOM e e-WOM	23
2.2	<i>Twitter</i>	25
2.2.1	Twitter e e-WoM	26
2.2.2	Tweets e cinema	27
2.3	<i>Sequências de Filmes</i>	27
2.3.1	Classificação das Sequências	28
2.3.2	Sequências como Extensão de Marca	29
2.4	<i>Sequências e Filmes de Terror</i>	29
2.5	<i>Processamento de Linguagem Natural</i>	30
2.5.1	BERT	31
2.5.2	RoBERTa	32
2.6	<i>Machine Learning ou Aprendizado de Máquina</i>	33
2.6.1	Tipos de Modelos de Machine Learning	33
2.6.2	Floresta Aleatória (Random Forest)	34
2.6.3	XGBoost	34
3	Revisão Sistemática	36
3.1	<i>Protocolo de Pesquisa</i>	36
3.1.1	Palavras-chaves e sinônimos	36

3.1.2	Critérios de seleção	37
3.1.3	Critérios de Inclusão e Exclusão de Artigos	38
3.1.4	Strings de Busca	38
3.1.5	Atualização da Revisão Bibliográfica	39
3.1.6	Tabela de Artigos Seleccionados	42
4	Trabalhos Correlatos: Tweets Influenciando Hollywood?	45
4.1	<i>Principais Trabalhos e Pontos Observados</i>	45
4.1.1	Sequências de Filmes	45
4.1.2	Análise de Dados e Twitter	46
4.1.3	Revisões de Filmes	48
5	Dados e Métodos	50
5.1	<i>Extração de Tweets</i>	50
5.2	<i>Dados complementares dos filmes</i>	52
5.3	<i>Banco de Dados de Filmes e Tweets</i>	52
5.4	<i>Plataforma Utilizada</i>	53
5.5	<i>Modelos NLP nos Tweets</i>	53
5.5.1	Seleção dos Modelos	53
5.6	<i>Modelos de Classificação de Texto nos Tweets</i>	54
5.6.1	Seleção dos Modelos	54
6	Resultados e Discussão	57
6.0.1	Medidas dos Tweets	57
6.0.2	Modelos de Análise de Sentimento e de Emoção	59
6.0.3	Aplicação do RF	65
6.0.4	Aplicação do XGBoost	68
6.0.5	Resultados sem Balanceamento de Classes	70
6.0.6	Comparativo entre os Modelos Random Forest e XGBoost	71
7	Conclusão	75
	REFERÊNCIAS	77

Appendices 83

	Apêndice A – Apêndices	84
<i>A.1</i>	<i>Códigos e Notebooks</i>	84
A.1.1	Python para coleta de Tweets	84
A.1.2	Notebook - Modelos de Emoção e Sentimento (RoBERTa)	85
A.1.3	Notebook - Random Forest	93
A.1.4	Notebook - XGBoost	98

1 Introdução

1.1 *Tema*

1.1.1 Sequências de Filmes

As sequências de filmes são componentes centrais na cultura cinematográfica, amplamente reconhecidas pela sua significativa importância, especialmente pela capacidade de replicar ou ampliar o sucesso de uma obra original. Essas sequências são caracterizadas por filmes que se baseiam em produções anteriores já estabelecidas e reconhecidas por seu êxito, seja em termos de prêmios ou de bilheteria. Geralmente, esses filmes continuam as narrativas e histórias introduzidas em suas produções predecessoras. As grandes produtoras monitoram atentamente o desempenho de seus filmes, buscando constantemente oportunidades para o desenvolvimento de continuações ou sequências, com o objetivo de mitigar riscos financeiros e fortalecer o valor da marca associada à franquia. Isso se torna ainda mais relevante, dado que muitas produções fazem parte de marcas ou franquias cinematográficas consolidadas ([HENNIG-THURAU; HOUSTON; HEITJANS, 2009](#)). O modelo proposto por esses autores para a avaliação monetária do valor de marca de um produto pode ser visualizado na Figura 1.

1.1.2 e-WOM (Eletronic Word-of-Mouth)

O word-of-mouth (WOM), ou boca-a-boca, é uma das formas mais antigas de disseminação de informações em nossa sociedade ([DELLAROCAS, 2003](#)). Com o avanço das tecnologias de informação e comunicação (TIC), especialmente das plataformas digitais, o WOM passou de uma forma física de comunicação para um formato digital, denominado electronic word-of-mouth (e-WOM). Existem algumas distinções importantes entre o WOM tradicional e o e-WOM, conforme ilustrado na Figura 2. O e-WOM tornou-se uma ferramenta de grande impacto na propagação de opiniões e avaliações sobre produtos e serviços ([HUETE-ALCOCER, 2017](#)).

O e-WOM é amplamente utilizado tanto por consumidores ativos quanto por consumidores passivos. Consumidores ativos são aqueles que geram feedback, como avaliações ou comentários sobre a qualidade de produtos ou serviços, e compartilham essas informa-

ções em plataformas online. Já os consumidores passivos são aqueles que buscam essas informações online, sem necessariamente produzir conteúdo (WANG; FESENMAIER, 2004).

Em outras palavras, o impacto do word-of-mouth (WOM), potencializado e medido por meio de tweets, exerce influência direta no comportamento de consumo, seja de forma positiva ou negativa (HU; KOH; REDDY, 2014). Essa influência pode ser expressa por meio de avaliações numéricas ou simples menções (KARNIOUCHINA, 2011).

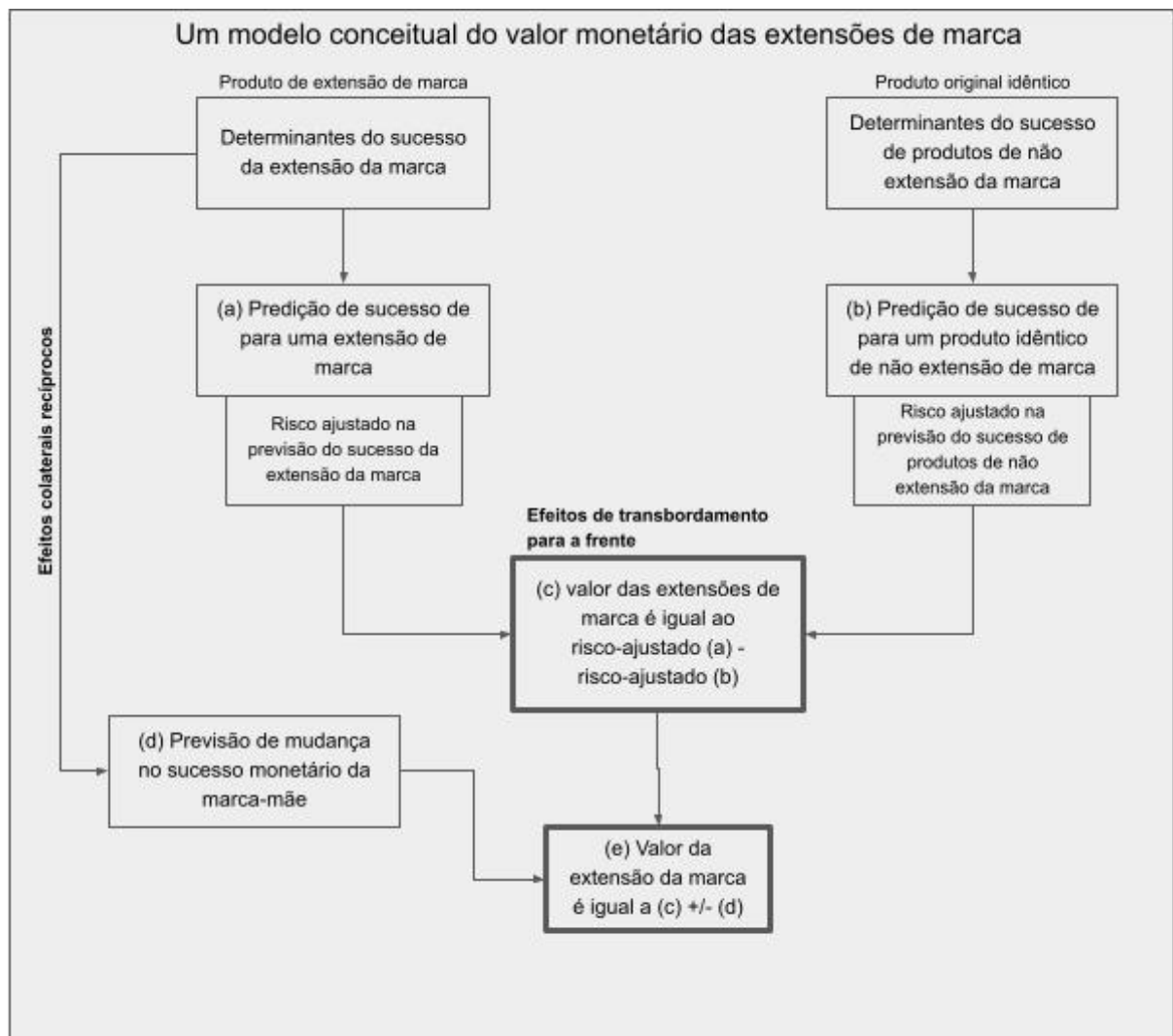


Figura 1 – Modelo conceitual do valor monetário da extensão de marcas

Fonte – Adaptado e Traduzido de Hennig-Thurau, Houston e Heitjans (2009)

1.2 Motivação

O número de filmes classificados como sequências tem crescido significativamente nos últimos anos, com sua produção aumentando de forma expressiva entre meados de 1990 e 2014, passando de 6% para 12% (OPITZ; HOFMANN, 2016). Este fenômeno reflete um comportamento mais conservador entre as produtoras de cinema, que buscam minimizar os riscos em suas atividades produtivas e garantir uma receita considerável com filmes subsequentes aos seus maiores sucessos. A avaliação crítica, ou *feedback*, desses bens culturais — que também são bens digitais de experiência (SHAPIRO; VARIAN, 2008) — é amplamente acompanhada nas redes sociais, influenciando, de maneira positiva ou negativa, o sucesso dos filmes, sejam eles sequências ou não.

Dado que os filmes são considerados bens de experiência, sua qualidade só pode ser plenamente conhecida após o consumo. Dessa forma, os consumidores tendem a consultar avaliações de produtos e serviços para obter informações sobre sua qualidade antes da aquisição, seja por meio de revisões (*reviews*) em blogs, microblogs ou até comentários em outras redes sociais (PARK; LEE, 2012). Essa prática é motivada pela assimetria de informação entre produtores e consumidores. Entre as plataformas online disponíveis, o Twitter destaca-se como uma das mais consultadas e influentes, capaz de moldar tanto as decisões de consumo quanto o comportamento individual (ALLCOTT; GENTZKOW, 2017).

As redes sociais evoluíram a ponto de serem consideradas o epicentro do compartilhamento instantâneo de informações através do *e-WOM* (eletronic word-of-mouth), alcançando rapidamente milhares de usuários (HANNA; ROHM; CRITTENDEN, 2011). Devido a essa característica, o foco do marketing mudou do ponto de vista do "fornecedor" para o ponto de vista do "usuário". Em outras palavras, os consumidores agora controlam o fluxo das informações de marketing, e não mais as empresas, pois são os próprios consumidores os responsáveis por compartilhar suas experiências e opiniões de maneira abrangente nas redes sociais (HODEGHATTA; SAHNEY, 2016).

1.3 *Lacuna*

O Twitter é amplamente reconhecido como a plataforma de microblogging mais utilizada globalmente, contando com mais de 330 milhões de usuários registrados e mais de 165 milhões de usuários ativos diariamente (Twitter Q3 2019 Earnings Report). Além disso, o Twitter caracteriza-se por ser uma rede social de acesso aberto, cujo formato de compartilhamento de mensagens facilita a disseminação do *e-WOM* (eletronic word-of-mouth). Nessa plataforma, cada usuário pode publicar mensagens de até 280 caracteres, incluindo a possibilidade de compartilhar vídeos e imagens. Essas pequenas publicações podem ser amplamente difundidas de maneira simples, atingindo milhares de usuários em questão de segundos, o que gera uma enorme quantidade e diversidade de dados a serem explorados.

A revisão bibliográfica conduzida neste estudo revela que há um volume considerável de análises de dados oriundos do Twitter, com ênfase em análises de sentimento, demonstrando a viabilidade desse tipo de abordagem. No entanto, não foram identificados estudos que estabeleçam uma correlação entre as informações extraídas do Twitter e a tomada de decisão quanto à criação de sequências de filmes no mercado cinematográfico. Diante desse cenário, o presente estudo visa preencher essa lacuna, investigando o impacto das informações oriundas do Twitter no processo decisório das produtoras em relação à criação de sequências cinematográficas.

1.4 *Hipótese*

A hipótese central deste estudo é a existência de uma relação de influência entre a quantidade de tweets e/ou o conteúdo dos tweets e a decisão sobre a possível criação de sequências de filmes. De acordo com a literatura, um grande volume de tweets exerce influência direta sobre o sucesso e a adesão ao consumo de filmes, independentemente de os comentários serem positivos ou negativos, evidenciando uma correlação entre o uso das redes sociais e o sucesso de produtos cinematográficos (VUJIC; ZHANG, 2018). Ademais, sabe-se que a criação de sequências cinematográficas é frequentemente motivada pelo desejo de expandir a marca do filme original ou pela necessidade de mitigar os riscos enfrentados pelas produtoras (VUJIC; ZHANG, 2018). No entanto, a literatura também

indica que o comportamento financeiro das sequências pode divergir significativamente em relação ao desempenho de seus filmes predecessores.

Dada a revolução tecnológica que as mídias vêm experimentando nos últimos anos (WALDFOGEL, 2017), e a alta velocidade e amplitude de disseminação de informações entre consumidores através de canais influentes como o Twitter, supõe-se que os comentários sobre filmes podem impactar de forma tanto positiva quanto negativa a decisão de produção de uma sequência, seja antes, durante ou após o lançamento do filme. Além disso, conforme sugerido por Eliashberg e Shugan (1997), o *e-WOM* (eletronic word-of-mouth) no Twitter pode, por outro lado, refletir a qualidade de um filme. Dessa forma, estabelece-se um problema de dupla causalidade: o *e-WOM* do Twitter não apenas influencia, mas também reflete a qualidade de um filme. Esses autores observaram que a influência exerce um efeito de curto prazo, enquanto o reflexo da qualidade — também denominado efeito de previsão — refere-se a um efeito de longo prazo.

1.5 Objetivos

O objetivo geral deste projeto é:

1. Investigar e apresentar, por meio da aplicação de modelos estatísticos, a relação entre diferentes métricas de tweets sobre um filme original e o potencial para a criação de uma sequência cinematográfica. Adicionalmente, diante do problema de dupla causalidade envolvendo *e-WOM* e o consumo de filmes, este estudo buscará identificar os efeitos causais utilizando dados de *e-WOM* disponíveis em curto prazo ((ELIASHBERG; SHUGAN, 1997)).

Os objetivos específicos deste estudo incluem:

1. Definir metodologias adequadas para a extração e o tratamento de tweets históricos, bem como coletar dados do *Box Office Mojo* relacionados a informações detalhadas sobre filmes lançados nos Estados Unidos.
2. Aplicar análise de sentimentos aos textos dos tweets, a fim de identificar as percepções dos usuários em relação aos filmes.
3. Calcular a correlação entre a quantidade e/ou o conteúdo dos tweets e a criação de sequências cinematográficas, testando a possível influência desses fatores com base em efeitos de curto prazo Eliashberg e Shugan (1997).

4. Avaliar se os tweets apresentam um elevado grau de manifestação da emoção "medo" e investigar se esta emoção predomina em relação a outras emoções.

1.6 *Justificativa*

Com o objetivo de proporcionar maior precisão às produtoras de Hollywood e outros estúdios cinematográficos no processo de decisão sobre a criação de sequências, este estudo busca investigar novas variáveis que possam influenciar esse processo decisório. Dada a vasta quantidade de dados gerados diariamente, o volume de informações disponível permite o reconhecimento de padrões por meio de processamento de linguagem natural, além de possibilitar a criação de novas métricas de avaliação da qualidade dos filmes.

O Twitter, em particular, oferece uma estrutura de dados concisa, com mensagens limitadas a um número máximo de caracteres, o que pode reduzir possíveis ruídos nas informações. Estudos que exploram dados provenientes do Twitter demonstram que a análise de sentimentos pode ser aplicada para classificar tweets, sendo essa uma variável potencialmente relevante no contexto de decisão sobre sequências cinematográficas.

Embora as produtoras geralmente não divulguem os fatores específicos considerados em seus processos decisórios para a criação de sequências, pesquisas anteriores citadas neste trabalho indicam que certos aspectos são comumente influentes, como a existência de obras literárias que precedem a criação de filmes, e, mais comumente, o faturamento obtido nas bilheterias. Este estudo propõe que outras variáveis possam igualmente desempenhar um papel significativo e, portanto, devem ser consideradas no processo de tomada de decisão.

2 Conceitos Fundamentais

2.1 WOM e e-WOM

O advento e a popularização da Internet resultaram no surgimento de uma nova forma de troca de experiências e informações, conhecida no Brasil como "boca a boca" ou, como referenciado na literatura, word-of-mouth (WOM). Essa troca, quando realizada no meio digital, é chamada de electronic-word-of-mouth (e-WOM), sendo reconhecida como uma das formas informais mais influentes de interação entre consumidores, empresas e a sociedade em geral ([HUETE-ALCOCER, 2017](#)).

O WOM é uma das formas mais antigas de disseminação de informação ([DELLAROCAS, 2003](#)), tendo sido amplamente discutida e definida de diversas maneiras na literatura. Em uma de suas definições, WOM é descrita como uma ferramenta de comunicação interpessoal, na qual o comunicador transmite informações para o receptor a respeito de aspectos variados, como marcas, produtos ou serviços, sendo que a fonte dessas informações é vista como independente de influências comerciais ([LITVIN; GOLDSMITH; PAN, 2008](#)). Esse processo de troca interpessoal possibilita o acesso a avaliações sobre a qualidade de produtos ou serviços feitas por consumidores, situando-se fora do escopo da publicidade formal. Dessa forma, WOM vai além das mensagens corporativas e, involuntariamente, exerce influência sobre a decisão de compra dos consumidores ([BROWN; BRODERICK; LEE, 2007](#)).

O WOM é amplamente reconhecido como uma das métricas mais influentes no processo decisório dos consumidores ([DAUGHERTY; HOFFMAN, 2014](#)). Tal influência é especialmente relevante para produtos intangíveis, como aqueles associados à experiência de consumo, como o turismo, cuja qualidade é difícil de ser avaliada antes do uso. Consequentemente, WOM se estabelece como a principal fonte de informação sobre a qualidade de bens e serviços no processo de decisão de compra dos consumidores ([LITVIN; GOLDSMITH; PAN, 2008](#)).

Os consumidores tendem a confiar mais nas opiniões de outros consumidores do que nas informações oferecidas pelos próprios vendedores ([NIETO; HERNANDEZ-MAESTRO; MUNHOZ-GALLEGO, 2014](#)). O WOM pode, ainda, impactar simultaneamente muitos receptores de informação, como observado por [Lau e Ng \(2001\)](#), sendo percebido como um canal de marketing dominado pelos consumidores, uma vez que os provedores de

informações são considerados independentes dos ofertantes, o que lhes confere maior credibilidade (BROWN; BRODERICK; LEE, 2007).

Uma das definições mais amplamente aceitas de e-WOM foi proposta por Litvin, Goldsmith e Pan (2008), que descreve essa forma de comunicação como toda e qualquer interação informal realizada via Internet, direcionada ao consumidor, e relacionada ao uso ou às características de um bem ou serviço, ou ainda, aos vendedores que os oferecem. A principal vantagem do e-WOM é sua acessibilidade, estando disponível a todos os consumidores, independentemente de suas preferências, desde que tenham acesso a plataformas online onde podem criar, compartilhar e obter opiniões sobre a qualidade de produtos com outros usuários.

No passado, os consumidores confiavam apenas no WOM de amigos e familiares. Atualmente, porém, eles recorrem às plataformas online (e-WOM) em busca de informações sobre produtos e serviços (NIETO; HERNANDEZ-MAESTRO; MUNHOZ-GALLEGO, 2014). Dessa forma, consumidores de qualquer parte do mundo podem deixar comentários e avaliações que influenciam a decisão de outros usuários. Tanto consumidores ativos quanto passivos utilizam o e-WOM. Consumidores ativos são aqueles que produzem e compartilham informações de feedback online, enquanto consumidores passivos buscam essas informações nas plataformas (WANG; FESENMAIER, 2004).

No campo do comportamento do consumidor, estudos como o de Park e Lee (2009) evidenciam que os consumidores tendem a prestar mais atenção a informações negativas do que positivas (LEE *et al.*, 2011). Um exemplo disso é o consumidor insatisfeito com o produto ou serviço oferecido que publica uma avaliação negativa online (e-WOM) Royo-Vela e Casamassima (2011), o que pode representar uma desvantagem competitiva significativa, especialmente para pequenos vendedores, que geralmente dispõem de menos recursos, incluindo os de marketing.

2.1.1 Diferenças entre WOM e e-WOM

Entre as principais características que distinguem as formas de comunicação WOM e e-WOM, a credibilidade da fonte de informação destaca-se como um fator relevante (e.g., Cheung e Thadani (2012), Hussain *et al.* (2017)). Ambas as modalidades exercem influência no comportamento do consumidor em relação a produtos e serviços; entretanto,

Figura 2 – Principais diferenças entre WOM e e-WOM pelas visões de credibilidade, privacidade, velocidade de transmissão e acessibilidade

	WOM	eWOM
Credibility	The receiver of the information knows the communicator (positive influence on credibility)	Anonymity between the communicator and the receiver of the information (negative influence on credibility)
Privacy	The conversation is private, interpersonal (via dialogs), and conducted in real time	The shared information is not private and, because it is written down, can sometimes be viewed by anyone and at any time
Diffusion speed	Messages spread slowly. Users must be present when the information is being shared	Messages are conveyed more quickly between users and, via the Internet, can be conveyed at any time
Accessibility	Less accessible	Easily accessible

Source: The author.

Fonte – [Huete-Alcocer \(2017\)](#)

o e-WOM apresenta um alcance significativamente maior [Veasna, Wu e Huang \(2013\)](#). No contexto de serviços relacionados ao turismo, considerados de alto risco, o e-WOM pode ter um impacto substancial ([SOTIRIADIS; ZYL, 2013](#)). [Luo et al. \(2013\)](#) sugerem que comentários e avaliações anônimas podem prejudicar a credibilidade das informações. Em contrapartida, outros estudos (e.g., [Hussain et al. \(2017\)](#)) apontam que os consumidores utilizam o e-WOM como uma estratégia para mitigar os riscos associados às suas decisões de consumo. Além disso, o e-WOM tende a ser mais confiável quando o autor das avaliações possui experiência anterior em realizar revisões e é devidamente identificado ([SOTIRIADIS; ZYL, 2013](#)).

A privacidade da mensagem é outro aspecto que diferencia as duas formas de comunicação. Enquanto o WOM tradicional ocorre de maneira privada, em tempo real, por meio de conversas presenciais, o e-WOM não se restringe ao ambiente privado e geralmente pode ser acessado por indivíduos anônimos, que não necessariamente se conhecem. Além disso, as revisões e avaliações de produtos e serviços no contexto do e-WOM podem ser interpretadas de maneiras diversas, dependendo de quem consome essa informação ([CHEUNG; THADANI, 2012](#)). Como essas revisões são disponibilizadas por escrito, tanto consumidores quanto empresas podem acessá-las a qualquer momento, desde que o conteúdo não tenha sido removido. Este é um fator distintivo importante em relação ao WOM

tradicional, em que a mensagem, após ser transmitida ao receptor, é absorvida e tende a desaparecer.

Uma diferença adicional entre WOM e e-WOM é a velocidade de disseminação da mensagem. O e-WOM propaga-se muito mais rapidamente, dado que é compartilhado em plataformas digitais (GUPTA; HARRIS, 2010). Mídias sociais, sites, blogs e outras plataformas online são elementos que diferenciam o e-WOM do WOM tradicional (CHEUNG; THADANI, 2012). Em primeiro lugar, essas plataformas tornam as informações mais acessíveis a um número maior de consumidores (CHEUNG; THADANI, 2012; SOTIRIADIS; ZYL, 2013). Em segundo lugar, a informação permanece disponível por tempo indeterminado, o que aumenta a sua acessibilidade ao longo do tempo (CHEUNG; THADANI, 2012; HENNIG-THURAU *et al.*, 2004).

Embora os termos WOM e e-WOM pareçam similares, eles possuem diferenças fundamentais em sua natureza. A internet transformou o WOM tradicional em e-WOM, alterando a maneira como as opiniões são comunicadas. A interação, que antes ocorria de forma interpessoal (i.e., de pessoa para pessoa ou face a face), passou a ser mediada por plataformas digitais (HUETE-ALCOCER, 2017). Muitos estudos, como os conduzidos por Brown, Broderick e Lee (2007), Daugherty e Hoffman (2014), Katz e Lazarsfeld (1966), Yang (2017), concordam que o e-WOM é o meio mais influente no comportamento do consumidor, sendo amplamente utilizado para obter informações antes, durante e após a experiência de consumo de um produto ou serviço. No campo do turismo, por exemplo, o e-WOM é considerado a fonte de informação mais influente no processo de decisão de compra, superando qualquer outra fonte de informações sobre viagens (SOTIRIADIS; ZYL, 2013).

2.2 Twitter

O Twitter, uma plataforma de microblogue (do inglês, microblogging), lançada em 2006, consolidou-se como uma das maiores redes sociais do mundo, com mais de 330 milhões de usuários registrados e mais de 165 milhões de usuários ativos diariamente (Twitter, 2019). A plataforma permite que os usuários publiquem mensagens curtas, conhecidas como *tweets*, com um limite de até 280 caracteres por postagem. Essas mensagens podem versar sobre uma variedade de temas, incluindo atividades cotidianas, novidades e opiniões,

e possuem o potencial de alcançar um vasto público na internet. O Twitter distingue-se por sua capacidade de disseminar rapidamente informações para além da rede imediata de contatos do usuário, atingindo, assim, camadas mais amplas de sua audiência. O objetivo principal da plataforma é promover a comunicação eficiente e ágil entre seus usuários.

2.2.1 Twitter e e-WoM

O conceito de *word-of-mouth* (WOM) tem sido amplamente utilizado por empresas na formulação e execução de suas estratégias de marketing. As redes sociais modernas atuam como plataformas para essa prática, transformando o conceito de WOM em *electronic-word-of-mouth* (e-WoM). Dada a assimetria de informação entre consumidores e produtos, os consumidores tendem a consultar avaliações de produtos e serviços antes de realizarem uma compra, seja por meio de *reviews* em blogs, microblogs ou comentários em redes sociais (PARK; LEE, 2012). O WOM sempre foi reconhecido como um conceito de alto valor, tanto no campo acadêmico quanto no empresarial, por ser uma métrica relevante para avaliar o alcance e a influência de um produto, serviço ou marca (LEE; KIM; KIM, 2011).

As plataformas de redes sociais, como Facebook, YouTube e Twitter, evoluíram para se tornarem centrais no compartilhamento instantâneo de informações através do e-WoM, dado o potencial de alcançar milhares de usuários em uma velocidade extremamente rápida (HANNA; ROHM; CRITTENDEN, 2011). Essa capacidade de disseminar experiências de forma quase instantânea representou um desafio significativo para as empresas, já que, independentemente da plataforma de rede social, os consumidores estão cada vez mais atentos a revisões e comentários sobre os produtos, marcas e serviços que consomem.

O crescimento no número de usuários de plataformas como Twitter e Facebook tem exposto as empresas a um público maior, ao mesmo tempo em que influencia um número crescente de consumidores (HAUBL; TRIFTS, 2000). As redes sociais mudaram o foco do marketing, deslocando-o de uma perspectiva centrada no 'fornecedor' para uma centrada no 'usuário'. Ou seja, agora são os consumidores que controlam o fluxo de informações de marketing, e não mais as empresas, pois são eles que compartilham suas experiências e opiniões de forma abrangente nas redes sociais (HODEGHATTA; SAHNEY, 2016).

Na atualidade, os consumidores conseguem acessar relatos de experiências com produtos e serviços antes de efetuarem qualquer compra (STEUER, 1992). Dessa forma, as empresas passaram a adotar as redes sociais como parte integral de suas estratégias de marketing, buscando agregar valor à marca, criar presença no mercado por meio da construção de reputação, aumentar os níveis de satisfação dos clientes e promover a retenção daqueles já adquiridos (CULNAN; MCHUGH; ZUBILLAGA, 2010). Com o crescimento contínuo dos usuários nas redes sociais, como Facebook e Twitter, a vasta quantidade de dados gerada por esses usuários tem se tornado uma fonte estratégica valiosa para o marketing nas mídias sociais (e-WoM), análises de sentimento e mineração de opiniões.

2.2.2 Tweets e cinema

Dada a influência observada do e-WoM no sucesso de produtos, diversos pesquisadores se dedicaram a analisar o impacto do e-WoM proveniente de tweets sobre o faturamento de filmes. Rui, Liu e Whinston (2013) demonstram, por meio da aplicação de algoritmos de *machine learning*, que o conteúdo veiculado no Twitter tem relevância para as vendas de filmes. No entanto, a magnitude e a direção desse efeito dependem de quem gera o e-WoM e do conteúdo discutido. O estudo evidencia que usuários com maior número de seguidores tendem a produzir e-WoMs mais significativos, ou seja, com maior influência em comparação a usuários com menos seguidores.

Além disso, Rui, Liu e Whinston (2013) destacam que e-WoMs de caráter positivo estão diretamente associados a um aumento nas vendas de filmes, enquanto e-WoMs negativos correlacionam-se a uma diminuição dessas vendas. Curiosamente, o estudo também revela que o efeito mais significativo nas vendas provém de tweets nos quais os autores expressam sua intenção de assistir a um determinado filme. Portanto, a identificação de usuários influentes no Twitter e o direcionamento do e-WoM favorável para promover produtos pode ser uma estratégia relevante para produtoras e distribuidoras de filmes.

2.3 Sequências de Filmes

Uma sequência pode ser definida como uma obra de literatura, filme, teatro, televisão, música ou videogame que dá continuidade à história ou expande o universo

ficcional de um trabalho anterior. No contexto de uma obra narrativa de ficção, uma sequência geralmente retrata eventos situados no mesmo universo ficcional da obra original, seguindo, na maioria dos casos, a cronologia dos eventos anteriores (Wall Street Journal, March 12, 1991). Em muitos casos, a sequência continua a explorar elementos centrais da história, utilizando os mesmos personagens e cenários. Quando várias sequências são produzidas, isso pode levar à formação de uma série, na qual certos elementos-chave reaparecem repetidamente. Embora a distinção entre múltiplas sequências e uma série seja por vezes arbitrária, algumas franquias de mídia acumulam sequências suficientes para serem reconhecidas como uma série, independentemente de essa estrutura ter sido planejada desde o início ou não.

2.3.1 Classificação das Sequências

A forma mais comum de sequência cinematográfica é a *sequência direta*, que tem como propósito continuar a trama apresentada no filme anterior ou introduzir elementos que foram deixados em aberto ou pouco explorados no primeiro filme.

Uma *sequência legada* dá continuidade à obra original, mas se passa significativamente mais adiante na linha do tempo, frequentemente focando em novos personagens, embora os personagens originais ainda estejam presentes na narrativa. Em certos casos, sequências legadas podem também ser consideradas sequências diretas, ignorando por completo eventos de obras intermediárias e recontando de forma implícita os eventos anteriores. Um exemplo clássico é o filme *Halloween (2018)*, que é uma sequência direta do *Halloween (1978)*.

A *sequência autônoma* ou *spin-off* é uma obra situada no mesmo universo ficcional, mas que possui pouca ou nenhuma conexão narrativa com seu antecessor, podendo ser compreendida independentemente sem a necessidade de conhecimento prévio da série. Exemplos desse tipo de sequência incluem *Annabelle*, *Velozes e Furiosos: Desafio em Tóquio*, e *Animais Fantásticos e Onde Habitam*, derivados respectivamente dos filmes ou série de filmes: *Invocação do Mal*, *Velozes e Furiosos*, *Harry Potter*

Por fim, uma *prequela* é uma sequência que, embora produzida posteriormente à obra original, retrata eventos anteriores aos do trabalho inicial. Um exemplo é o filme

Annabelle 2: A Criação do Mal, que narra a história de um dos antagonistas da saga *Invocação do Mal* antes dos eventos principais da série ocorrerem (SILVERBLATT, 2015).

2.3.2 Sequências como Extensão de Marca

A indústria cinematográfica de Hollywood passou a tratar os filmes como se fossem marcas de produtos de consumo diário. A literatura aponta que a avaliação das sequências de filmes é influenciada pela semelhança percebida entre a marca-mãe e sua extensão (GURHAN-CANLI; MAHESWARAN, 1998; KELLER; AAKER, 1992). Quando essa semelhança é elevada, as sequências tendem a ser associadas à marca-mãe, e o afeto dos consumidores pela obra original é transferido para a sequência. A semelhança percebida, na maioria dos estudos, é definida como uma relação física entre as obras, em termos de sobreposição de recursos ou elementos narrativos. Quando essa sobreposição é significativa, a extensão da marca é percebida como similar e as avaliações tendem a ser mais positivas, comparadas a quando a sobreposição é baixa e a sequência difere substancialmente do filme original (SOOD; DRÈZE, 2006).

Por exemplo, quando uma sequência apresenta um nome semelhante ao do filme original (como *Toy Story 2*, cuja obra original é *Toy Story*) ou mantém um roteiro que se assemelha ao enredo do filme da marca-mãe que obteve sucesso, a probabilidade de a sequência herdar as percepções positivas do público em relação ao original é considerável.

A relevância dos estudos sobre sequências de filmes é amplamente reconhecida na literatura, tanto do ponto de vista acadêmico quanto da gestão cinematográfica. Os gestores dos estúdios de cinema compreendem as sequências como uma estratégia eficaz de gestão de marcas, tratando-as como um "conceito de múltiplos usos que visa amortizar os riscos associados a uma linha específica de produtos" (Wall Street Journal, June 6, 1989; Page 1).

2.4 Sequências e Filmes de Terror

O aumento significativo no número de filmes categorizados como sequência, conforme indicado por (HENDERSON, 2017), teve início na década de 1970, como evidenciado na

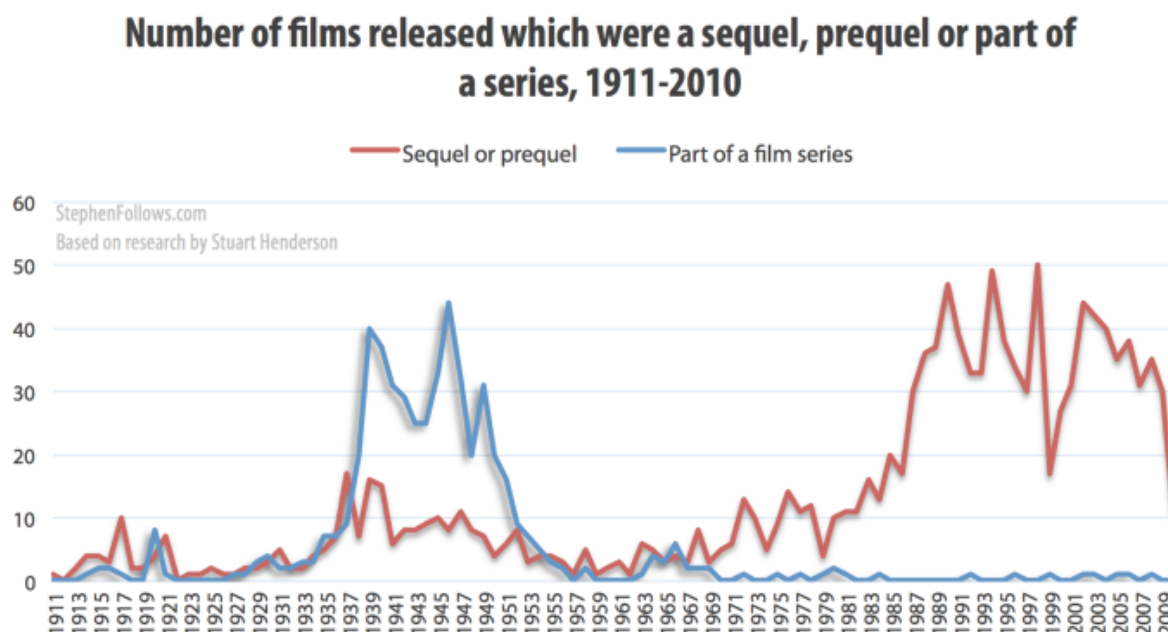


Figura 3 – Filmes lançados que são sequência em comparação com séries de filmes.

Fonte – <https://stephenfollows.com/hollywood-sequels-by-the-numbers/>

Figura 3, que mostra que filmes classificados como sequência ultrapassaram as tradicionais séries de filmes.

Esse crescimento pode ser atribuído à transformação nas categorias e abordagens adotadas pelos estúdios para explorar de forma mais eficaz o potencial de cada produto cinematográfico. Os três gêneros que mais apresentam filmes em formato de sequência são, em ordem decrescente, Aventura, Ação e Terror, conforme apresentado na Figura 4.

Neste estudo, foi decidido focar exclusivamente no gênero Terror, com o intuito de analisar um número considerável de filmes tanto sequenciais quanto não sequenciais, além de restringir a quantidade de tweets coletados dos filmes selecionados.

2.5 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (NLP) é um campo interdisciplinar que combina técnicas de inteligência artificial, linguística e ciência da computação para capacitar máquinas a compreender, interpretar e gerar linguagem humana.

Quando falamos sobre NLP, não estamos apenas discutindo métodos para decifrar palavras em um texto. O verdadeiro objetivo do NLP é muito mais complexo e fascinante.

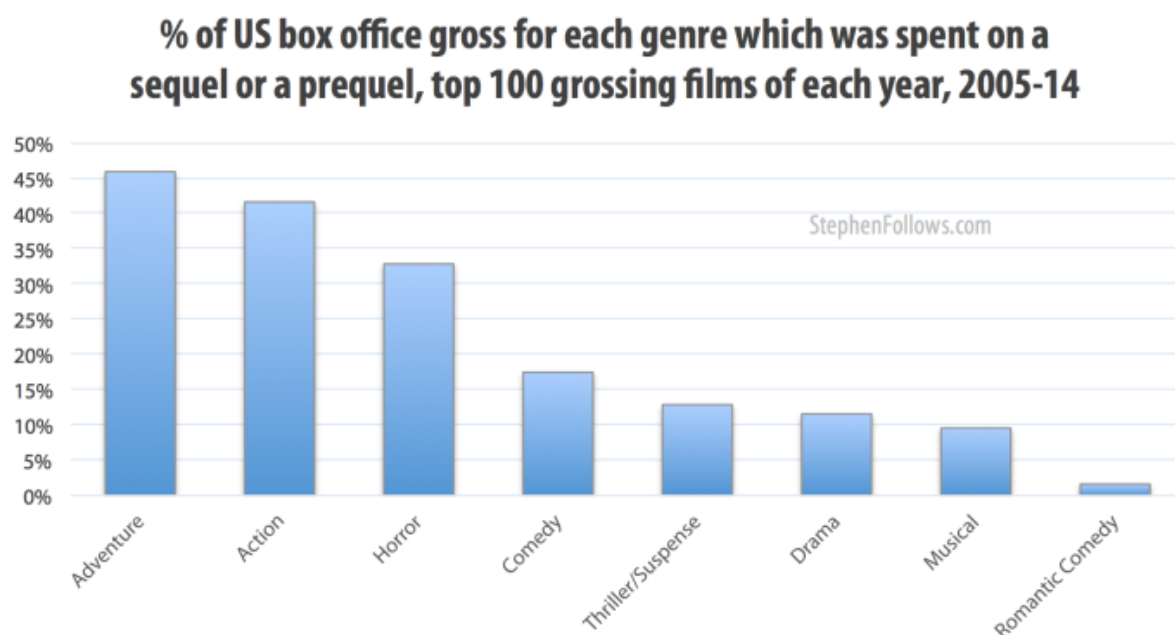


Figura 4 – % da bilheteria bruta dos EUA por gênero, considerando o investimento em sequências ou prequelas. Dados referentes aos 100 filmes de maior bilheteria de cada ano, entre 2005 e 2014.

Fonte – <https://stephenfollows.com/hollywood-sequels-by-the-numbers/>

Em vez de se limitar a traduzir ou reconhecer palavras de forma isolada, o NLP busca entender o significado mais profundo do que está sendo comunicado.

Pense em uma conversa do dia a dia. Quando alguém fala com você, você não apenas ouve as palavras, mas também percebe a intenção por trás delas, as emoções que a pessoa está expressando e o contexto da conversa. O NLP tenta replicar essa habilidade humana, permitindo que os computadores compreendam e interpretem essas nuances.

Para alcançar esse objetivo, o NLP utiliza uma variedade de técnicas avançadas que analisam o texto de forma abrangente. Isso inclui a identificação se uma frase é uma pergunta ou uma afirmação, a detecção de um sentimento positivo ou negativo, e a compreensão da mensagem geral que está sendo transmitida (JURAFSKY; MARTIN, 2008).

2.5.1 BERT

BERT, ou Bidirectional Encoder Representations from Transformers, é um modelo de linguagem inovador criado pela Google AI Language. Desde seu lançamento em 2018,

ele trouxe uma mudança significativa para a forma como os computadores entendem a linguagem humana (DEVLIN *et al.*, 2019).

Antes do BERT, muitos modelos de linguagem baseavam-se em técnicas unidimensionais para entender o texto. Modelos como word2vec e GloVe criavam representações estáticas de palavras, onde cada palavra tinha um único vetor de características, independentemente do contexto em que aparecia (MIKOLOV *et al.*, 2013) e (PENNINGTON; SOCHER; MANNING, 2014). Isso limitava a capacidade desses modelos de capturar o significado real das palavras em diferentes situações.

O BERT mudou essa abordagem ao introduzir um método bidirecional de interpretação do texto, permitindo que o modelo compreenda palavras com base em seu contexto completo, levando em conta tanto as palavras que vêm antes quanto as que vêm depois (PETERS *et al.*, 2018). Esse avanço não apenas melhorou a performance em tarefas de compreensão de leitura e análise de sentimentos, mas também estabeleceu um novo padrão para modelos de linguagem em várias aplicações de NLP.

2.5.2 RoBERTa

RoBERTa, que significa "Robustly optimized BERT approach", é uma variação do modelo BERT desenvolvida pela equipe da Facebook AI Research. Lançado em 2019, o RoBERTa foi projetado para melhorar o desempenho do BERT em várias tarefas de Processamento de Linguagem Natural (NLP) através de otimizações e ajustes no processo de pré-treinamento (LIU *et al.*, 2019). As principais diferenças entre RoBERTa e BERT incluem:

1. Tamanho da Base de dados: RoBERTa utilizou um volume de dados mais extenso e diversificado;
2. RoBERTa utiliza uma abordagem de máscara de linguagem mais dinâmica;
3. RoBERTa removeu a tarefa de previsão da próxima sentença, simplificando o pré-treinamento, se comparado ao BERT.

Nesse trabalho vemos a aplicação de modelos RoBERTa para classificação das emoções e sentimentos dos dados de tweets coletados, com base em estudos realizados pela equipe da universidade de Cardiff (CAMACHO-COLLADOS *et al.*, 2022).

2.6 *Machine Learning* ou *Aprendizado de Máquina*

Machine Learning, ou Aprendizado de Máquina, é uma parte da inteligência artificial (IA) que se foca em criar sistemas que aprendem e se aprimoram com os dados, sem a necessidade de serem programados com regras fixas para cada tarefa específica. Em termos simples, Machine Learning é sobre ensinar computadores a aprender e tomar decisões com base em exemplos, ao invés de seguir um conjunto de instruções pré-determinadas (ALPAYDIN, 2020).

Para entender melhor, pense em como você ensina um amigo a diferenciar fotos de gatos e cães. Você mostraria a ele várias fotos, dizendo se cada uma é de um gato ou de um cachorro. Com o tempo, ele começaria a notar padrões, como o formato das orelhas ou o comprimento do pelo, e assim aprenderia a distinguir entre gatos e cães. De forma semelhante, um modelo de Machine Learning aprende a partir de dados para fazer previsões ou classificações com base em características que ele identificou durante o treinamento (MITCHELL, 1997).

2.6.1 Tipos de Modelos de Machine Learning

Os modelos de Machine Learning podem ser classificados de várias maneiras, e cada tipo é adequado para diferentes tarefas:

1. Modelos Supervisionados: Os modelos supervisionados são treinados com dados que já têm respostas conhecidas. Por exemplo, se você tem uma lista de e-mails e já sabe quais são "spam" e quais são "não spam", você pode usar um modelo supervisionado para ensinar o sistema a reconhecer novos e-mails como spam ou não spam com base nas características dos e-mails anteriores (BISHOP, 2016);
2. Modelos Não Supervisionados: Modelos não supervisionados trabalham com dados sem respostas conhecidas. O objetivo aqui é explorar os dados para descobrir padrões ou organizar informações em grupos. Um exemplo seria analisar as compras dos clientes em uma loja para descobrir quais são os diferentes tipos de clientes sem saber de antemão quantos grupos existem (JAIN; MURTY; FLYNN, 1999);
3. Modelos Semi-Supervisionados e de Aprendizado por Reforço: Além desses, há os modelos semi-supervisionados, que combinam dados rotulados e não rotulados para

melhorar o aprendizado, e os modelos de aprendizado por reforço, onde um agente aprende a tomar decisões através de tentativas e erros, recebendo recompensas ou penalidades por suas ações (SUTTON; BARTO, 2018).

2.6.2 Floresta Aleatória (Random Forest)

O Random Forest é um método de aprendizado de máquina que combina múltiplas árvores de decisão para melhorar a precisão e reduzir o risco de overfitting, utilizando técnicas como bagging e seleção aleatória de atributos. Ele é amplamente utilizado devido à sua robustez e eficácia em lidar com grandes conjuntos de dados, apesar de ser computacionalmente intensivo e menos interpretável que modelos individuais (BREIMAN, 2001), (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.6.3 XGBoost

O XGBoost, uma implementação otimizada da técnica de boosting, tem se destacado como uma ferramenta poderosa em aprendizado de máquina, especialmente em problemas complexos de classificação e regressão. Desenvolvido por Tianqi Chen e lançado em 2016, o XGBoost oferece alta precisão e eficiência, tornando-se uma escolha popular em competições de ciência de dados e aplicações práticas.

Para entender melhor o funcionamento desse modelo, precisamos entender o que seria o Boosting. É uma técnica de ensemble learning onde múltiplos modelos fracos, como árvores de decisão, são treinados sequencialmente. Cada modelo subsequente é ajustado para corrigir os erros cometidos pelos modelos anteriores. O Gradient Boosting, é uma forma de boosting que otimiza a função objetivo utilizando gradientes, permitindo que o modelo melhore progressivamente (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O XGBoost introduz várias inovações que o diferenciam de outras implementações de boosting (CHEN; GUESTRIN, 2016):

1. Regularização: Inclui regularização L1 (Lasso) e L2 (Ridge) para controlar a complexidade do modelo, prevenindo o overfitting;
2. Paralelismo: O XGBoost aproveita o paralelismo das CPUs modernas, acelerando significativamente o processo de treinamento;

3. Manuseio de Dados Faltantes: Implementa uma técnica eficiente para lidar com dados faltantes, otimizando o caminho que os dados ausentes seguem nas árvores de decisão;
4. Importância das Variáveis: Oferece ferramentas para avaliar a importância das variáveis, facilitando a interpretação dos modelos.

O processo de construção das árvores no XGBoost começa com um modelo simples que prevê a média dos valores de saída. A partir daí, novas árvores são sequencialmente adicionadas para ajustar os erros residuais do modelo anterior. Cada árvore nova é construída para minimizar a função objetivo, composta por um termo de perda e um termo de regularização ([HASTIE; TIBSHIRANI; FRIEDMAN, 2009](#)).

A função objetivo do XGBoost combina a perda do modelo com termos de regularização para penalizar a complexidade excessiva. O modelo é ajustado utilizando o gradiente descendente, o que permite que cada árvore melhore incrementalmente a precisão do modelo geral ([CHEN; GUESTRIN, 2016](#)).

Para evitar que o modelo se ajuste excessivamente aos dados de treinamento, o XGBoost aplica técnicas de poda, removendo ramos das árvores que não contribuem significativamente para a melhoria do modelo. Além disso, ele usa um critério chamado "redução máxima de ganho" para decidir onde dividir os nós da árvore, assegurando que apenas divisões que resultem em uma melhoria substancial sejam mantidas ([CHEN; GUESTRIN, 2016](#)).

3 Revisão Sistemática

Esta seção aborda o processo de levantamento bibliográfico, bem como os passos necessários para compor a revisão da literatura. Para fundamentar o estudo, é essencial compreender como as plataformas online podem impactar, de forma positiva ou negativa, um produto ou empresa por meio de avaliações e comentários disponibilizados voluntariamente na internet. Cabe destacar que foi necessária uma atualização dos trabalhos investigados, em função da ampliação do prazo previsto para a conclusão da dissertação.

Inicialmente, este estudo utilizou trabalhos disponíveis nas plataformas IEEE, ACM e Scopus, considerando artigos em inglês e português na seleção.

A revisão sistemática teve como objetivo responder à seguinte pergunta principal:

- "Os Tweets influenciam a possível criação de sequências de filmes?"

Com base nessa pergunta principal, foram elaboradas três perguntas adicionais para direcionar e auxiliar o estudo na busca por referências bibliográficas, a saber:

1. "O Twitter pode ser caracterizado como um meio de e-WOM?"
2. "Quais técnicas e métodos já existentes permitem mensurar e qualificar a influência de tweets?"
3. "Existe uma relação entre os tweets e os critérios de definição para a criação de sequências de filmes?"

3.1 Protocolo de Pesquisa

O protocolo de pesquisa foi desenvolvido com o auxílio da ferramenta START, que tem como objetivo apoiar a construção e execução dos passos do processo de revisão bibliográfica.

3.1.1 Palavras-chaves e sinônimos

Embora o estudo considere trabalhos escritos nos idiomas inglês e português, a maior parte das palavras-chaves utilizadas estão em inglês. Essa escolha se justifica pelo fato de que a plataforma-alvo de estudo é o Twitter e o público-alvo da pesquisa são tweets

em inglês. O foco da pesquisa é o mercado cinematográfico dos Estados Unidos, por ser pioneiro e o de maior faturamento mundial. Além disso, foram consideradas palavras-chaves que pudessem auxiliar na resposta à pergunta principal da pesquisa, cruzando grandes áreas de estudo, como o Marketing.

As palavras-chaves selecionadas foram:

Tabela 1 – Palavras-chaves selecionadas para o estudo

Palavra-chave	Alternativa
Hollywood movies	-
box office revenues	-
brand extensions	-
e-WOM	-
motion pictures	-
movie sequel	movies sequels
opinion analysis	-
sentiment analysis	-
tweet	tweets
twitter	-
twitter social media behaviour	-

Fonte – Brenno Ruschioni de Oliveira, 2024

3.1.2 Critérios de seleção

Para limitar o grande volume de artigos selecionados, foram definidos critérios específicos de seleção para compor o levantamento bibliográfico. Apenas artigos redigidos nos idiomas inglês e português, publicados em revistas de renome e contendo dados oriundos de fontes confiáveis foram incluídos, com o objetivo de garantir a alta qualidade e respaldo científico dos resultados deste trabalho.

O método de pesquisa adotado assegura, desde o início, a qualidade dos trabalhos selecionados, uma vez que foram consideradas apenas ferramentas de busca em bases de dados indexadas dos principais periódicos. As bases de dados utilizadas foram IEEE (Institute of Electrical and Electronic Engineers), a Biblioteca Digital da ACM (Association for Computing Machinery) e o Scopus.

Inclusão/Exclusão	ID	Critério
Inclusão	I1	Analise a influência de tweets
	I2	Trabalhe com técnicas de mineração de dados do Twitter
	I3	Estabeleça relação entre o Twitter e o mercado cinematográfico
	I4	Aborde premissas utilizadas pelas produtoras cinematográficas para a criação de sequências de filmes
	I5	Artigos publicados e disponíveis integralmente em bases de dados científicas ou em versões impressas
	I6	Focalize em sequências de filmes
Exclusão	E1	Que não atendam às especificações deste protocolo
	E2	Que não apresentem resultados estatísticos sobre a influência dos tweets
	E3	Trabalhos que apresentem avaliações sem descrever claramente o método utilizado

Tabela 2 – Critérios de Inclusão e Exclusão

Fonte – Brenno Ruschioni de Oliveira, 2024

3.1.3 Critérios de Inclusão e Exclusão de Artigos

A Tabela 2 apresenta os critérios de inclusão e exclusão adotados na análise quantitativa dos artigos. Todos os critérios de inclusão foram formulados com o intuito de contribuir para a elaboração de respostas às três questões auxiliares, seja para responder uma ou mais delas.

3.1.4 Strings de Busca

Foram elaboradas strings de busca específicas para cada ferramenta de pesquisa mencionada na seção anterior, conforme apresentado na Tabela 3. A formulação dessas strings priorizou a abrangência, utilizando palavras-chave limitadas, deixando a filtragem de resultados mais detalhados para os critérios de seleção subsequentes.

Ferramenta de pesquisa	String de Busca	Artigos
ACM	"[[Abstract: twitter] OR [Abstract: tweet*]] AND [[Abstract: movie*] OR [Abstract: "movie* sequel*"]]"	42
IEEE	"((((Abstract:twitter) OR Abstract:tweet*) AND Abstract:movie*) OR Abstract:movie* sequel*)"	105
Scopus	"TITLE-ABS-KEY((twitter OR tweet OR tweets) AND (movie* OR "movie* sequel*))"	419

Tabela 3 – Strings de Busca

Fonte – Brenno Ruschioni de Oliveira, 2024

3.1.5 Atualização da Revisão Bibliográfica

Devido ao afastamento deste projeto por dois anos, desde a última revisão bibliográfica, houve a necessidade de atualizá-la para verificar potenciais mudanças no estado da arte. Novos estudos foram incorporados à dissertação utilizando o mesmo processo descrito anteriormente, com a única exceção sendo a aplicação de um filtro de data. Nesta nova fase da Revisão Sistemática (RS), foram considerados apenas trabalhos publicados entre 2020 e 2023, visando evitar duplicidades e reduzir o esforço desnecessário de filtragem.

Adicionalmente, estudos foram inseridos manualmente com base em recomendações de meu grupo de estudo e de minha orientadora. Esses artigos foram lidos e incorporados à lista de estudos aprovados, conforme ilustrado na Tabela 2.

Em relação aos resultados das consultas, após a atualização realizada em 2023, a base de dados Scopus foi a que mais retornou estudos. No entanto, é importante notar que as bases IEEE e ACM contribuíram mais significativamente com trabalhos aceitos, conforme demonstrado na Figura 5.

Após a execução das buscas, utilizando as strings de consulta apropriadas, foi realizada uma filtragem inicial. Essa etapa consistiu nas seguintes fases: primeiramente, a identificação e remoção de duplicidades; em segundo lugar, a seleção dos artigos a serem aceitos para a fase qualitativa; e, por fim, a priorização para leitura.

Todas as etapas foram conduzidas com o auxílio do aplicativo START, que atribui automaticamente uma nota de relevância aos artigos com base na interação das palavras-

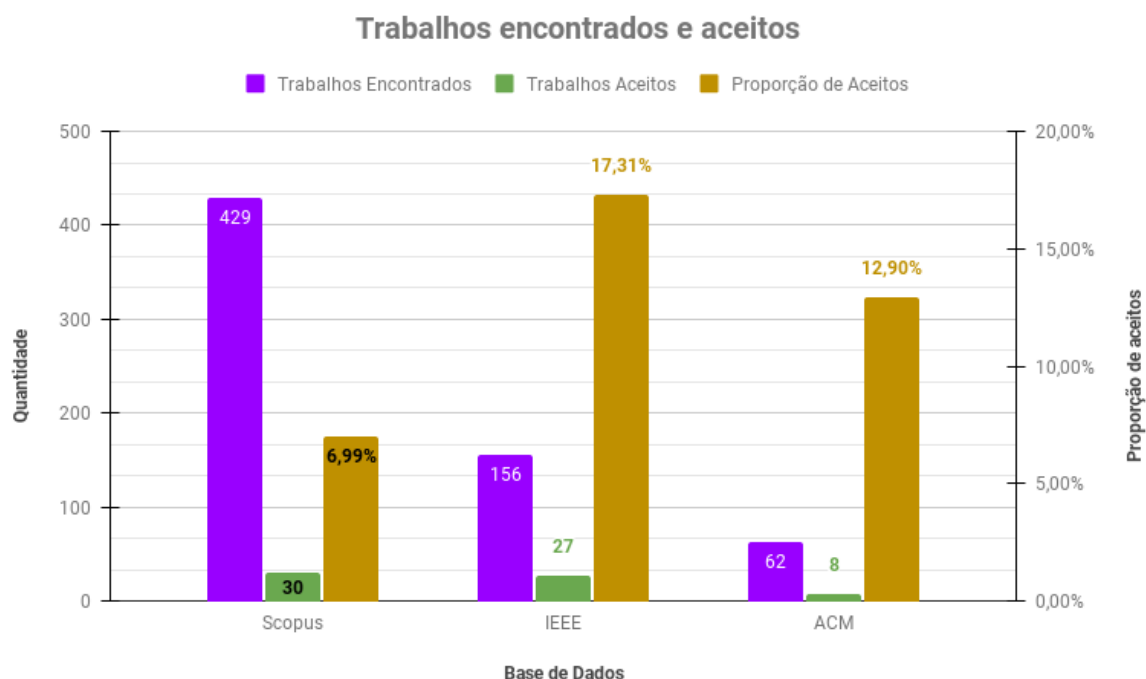



Figura 5 – (a) Quantidade de Trabalhos Encontrados e Aceitos

chave com os textos, utilizando os pesos indicados na Figura 6. Essa nota foi utilizada para a ordenação dos artigos e na classificação final.

Figura 6 – Pesos para Revisão Sistemática

 **Adjust quantitative criteria** ✕

Method for calculating the Score Value

Keywords in title: points per occurrence

Keywords in abstract: points per occurrence

Keywords in keywords: points per occurrence

Fonte – Brenno Ruschioni de Oliveira

Os critérios de inclusão dos trabalhos podem ser verificados na Figura 7. É relevante observar que o volume total de motivos de inclusão excede o número de trabalhos aceitos, pois um artigo aceito pode conter múltiplos motivos de inclusão. Após uma verificação

adicional na fase de extração, alguns trabalhos foram rejeitados por não estarem diretamente relacionados com os objetivos desta dissertação, restando 66 trabalhos aprovados.

Quantidade de Trabalhos por Critérios de Inclusão

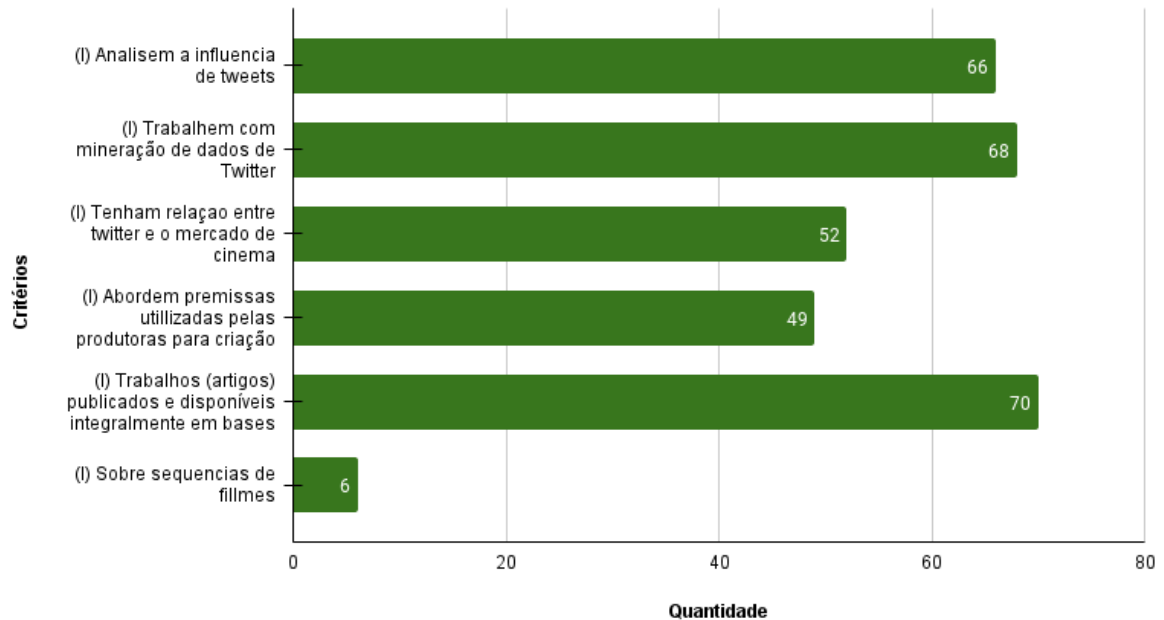


Figura 7 – Quantidade de Trabalhos por Critérios de Inclusão

Para detalhar a distribuição dos trabalhos selecionados, a Figura 8 apresenta a distribuição dos estudos por ano. Observa-se uma concentração de artigos recentes, o que é coerente com a análise de uma ferramenta de microblog relativamente nova. A distribuição dos estudos reflete o crescimento e a expansão do Twitter globalmente. Muitos dos trabalhos encontrados na atualização desta revisão referem-se ao período de 2020 a 2023, com a maioria dos artigos aceitos concentrados no ano de 2019.

Além disso, a Figura 9 exhibe a distribuição dos trabalhos selecionados por tipo de publicação, sejam eles artigos de revistas ou trabalhos de conferências.

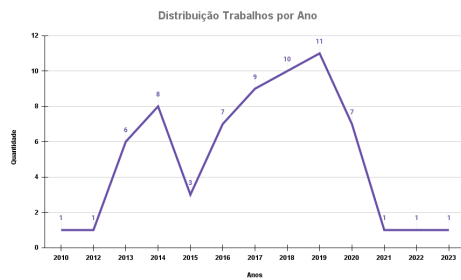


Figura 8 – (a) Distribuição de Trabalhos por Ano

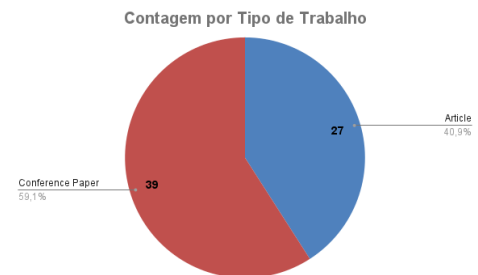


Figura 9 – (b) Tipos de Trabalhos Aceitos

3.1.6 Tabela de Artigos Seleccionados

O quadro 1 mostra as informações principais dos artigos que foram seleccionados para a etapa de extração da Revisão Sistemática.

Quadro 1 – Lista de trabalhos aceitos na revisão sistemática (*Continuação*)

Título original	Autores	Ano ²
Biased k-NN similarity content based prediction of movie tweets popularity	PeÄjka, L. and VojtÄjÄj, P.	2015
Does Twitter matter? The impact of microblogging word of mouth on consumers adoption of new movies	Hennig-Thurau, T. and Wiertz, C. and Feldhaus, F.	2015
Influence of social media on performance of movies	Shruti and Roy, S.D. and Zeng, W.	2014
Impact of tweets on box office revenue: Focusing on when tweets are written	Baek, H. and Ahn, J. and Oh, S.	2014
Movie sentiment analysis based on public tweets	Blatnik, A. and Jarm, K. and Mea, M.	2014
The twitter effect: Social media usage as a contributor to movie success	Treme, J. and VanDerPloeg, Z.	2014
Chronological analysis of the electronic word-of-mouth effect of four social media channels on movie sales: Comparing Twitter, Yahoo!Movies, YouTube, and blogs	Baek, H. and Oh, S. and Yang, H.-D. and Ahn, J.H.	2014
Whose and what chatter matters? the effect of tweets on movie sales	Rui, H. and Liu, Y. and Whinston, A.	2013
Customer engagement, word-of-mouth and box office: The case of movie tweets	Oh, C.	2013
Chatter matters: How twitter can open the black box of online word-of-mouth	Rui, H. and Liu, Y. and Whinston, A.B.	2010
Social Media Mining: Prediction of Box Office Revenue	Choudhery, Deepankar and Leung, Carson K.	2017
Why Watching Movie Tweets Won't Tell the Whole Story?	Wong, Felix Ming Fai and Sen, Soumya and Ching, Mung	2012
Spatio-Temporal Visualization Model for Movie Success Prediction Based on Tweets	Wijekoon, A. W. M. K. S. A. and Sandanayake, T. C. and Jayawardena, K. D. A. A. and Buddhini, A. L. Y. and Ariyawansa, U. K. D. G. S.	2017
Predict Movie Revenue by Sentimental Analysis of Twitter	Sadadi, Hoda and Aloufi, Doaa and Ye, Zilong	2018
Prediction of Movies Box Office Performance Using Social Media	Apala, Krushikanth R. and Jose, Merin and Motnam, Supreme and Chan, C.-C. and Liszka, Kathy J. and de Gregorio, Federico	2013
Empirical Study on the Relationship between Twitter and Movie Box Office Revenue	She, Rui and Guo, Jingzhi and Yang, Yao	2018
Sentiment Analysis of Hollywood Movies on Twitter	Hodeghatta, Umesh Rao	2013
A Review on Basic Methodology of Twitter Base Prediction System	D. K. Zala and A. Gandhi	2018
A Twitter Based Opinion Mining to Perform Analysis Geographically	D. K. Zala and A. Gandhi	2019
How successful movies affect performance of sequels: Signal theory and brand extension theory in motion picture industry	H. Yong and W. Tie-nan and L. Xiang-yang	2013

Quadro 1 – Lista de trabalhos aceitos na revisão sistemática

Título original	Autores	Ano ¹
Sentiment Classification on Movie Reviews and Twitter: An Experimental Study of Supervised Learning Models	Hourrane, O. and Idrissi, N. and Benlahmar, E.H.	2019
Sentiment analysis of movies on social media using R studio	Jaichandran, R. and Bagath Basha, C. and Shunmuganathan, K.L. and Rajaprakash, S. and Kanagasuba Raja, S.	2019
The effects of eWOM volume and valence on product sales an empirical examination of the movie industry	Kim, K. and Yoon, S. and Choi, Y.K.	2019
Analyzing national film based on social media tweets input using topic modelling and data mining approach	Ramos, C.D. and Suarez, M.T. and Tighe, E.	2019
Comparison of sentiment analysis on various twitter #tags using machine learning and deep learning techniques	Rayala Vinod Kumar, C.H. and Lalitha Bhaskari, D. and Srinivasa Rao, P.	2019
The Empire Tweets Back? #HumanitarianStarWars and Memetic Self-Critique in the Aid Industry	Chonka, P.	2019
Does Twitter chatter matter? Online reviews and box office revenues	Vujić, S. and Zhang, X.	2018
Sentiment Manipulation in Online Platforms: An Analysis of Movie Tweets	Lee, S.-Y. and Qiu, L. and Whinston, A.	2018
Predicting motion picture box office performance using temporal tweet patterns	Hossein, N. and Miller, D.W.	2018
Sentiment analysis of movie review using machine learning techniques	Uma Ramya, V. and Thirupathi Rao, K.	2018
A study on data analysis of movie tweets using machine learning techniques	Reddy, M.Y. and Bhavana, V. and Yeshwanth, P. and Santhi, M.V.B.T.	2017
The impact of word of mouth via twitter on moviegoers decisions and film revenues: Revisiting prospect theory: How WOM about movies drives loss-aversion and reference-dependence behaviors	Yoon, Y. and Polpanumas, C. and Park, Y.J.	2017
Electronic word-of-mouth, box office revenue and social media	Baek, H. and Oh, S. and Yang, H.-D. and Ahn, J.	2017
How people utilise tweets on movie selection? the reverse effects of e-WoM valence on movie sales	Kang, H. and Chai, S. and Kim, H.U.	2017
Beyond likes and tweets: Consumer engagement behavior and movie box office in social media	Oh, C. and Roumani, Y. and Nwankpa, J.K. and Hu, H.-F.	2017
Combining structure, content and meaning in online social networks: The analysis of public's early reaction in social media to newly launched movies	Lipizzi, C. and Iandoli, L. and Marquez, J.E.R.	2016
Understanding Twitter as an e-WOM	Hodeghatta, U.R. and Sahney, S.	2016
Twitter sentiment analysis of movie reviews using machine learning technique	Amolik, A. and Jivane, N. and Bhandari, M. and Venkatesan, M.	2016
Social media information diffusion and economic outcomes: Twitter retweets and box office revenue	Oh, C. and Hu, H.-F. and Yang, W.	2016
The overriding influence of social media as the key driver of cinematic movie sales	Moses, C.L. and Olokundun and Ayodele, M. and Omotayo and Adegbuyi and Augusta, A. and Akinbode and Oluwafunmilayo, M. and Inelo, F.	2016

Fonte – Brenno Ruschioni de Oliveira, 2023

Quadro 1 – Lista de trabalhos aceitos na revisão sistemática (*Continuação*)

Título original	Autores	Ano ³
User tweets based genre prediction and movie recommendation using LSI and SVD	S. Bansal and C. Gupta and A. Arora	2016
Real-Time Sentiment Analysis of Tweets: A Case Study of Punjab Elections	Y. Gupta and P. Kumar	2019
Mining online reviews and tweets for predicting sales performance and success of movies	S. S. Magdum and J. V. Megha	2017
A New Sentiment Analysis based Application for Analyzing Reviews of Web Series and Movies of Different Genres	Aishwarya and P. Wadhwa and P. Singh	2020
Twitter Sentiment Mining: A Multi Domain Analysis	S. Shahheidari and H. Dong and M. N. R. B. Daud	2013
Sentiment Analysis on Twitter Data: A New Approach	R. S. Karan and K. K. Shirsat and P. L. Kasar and R. Chaudhary	2018
Neural networks for sentiment analysis on Twitter	B. Duncan and Y. Zhang	2015
Opinion Mining on Movie Reviews	M. R and S. M.R	2019
Predicting iPhone Sales from iPhone Tweets	N. B. Lassen and R. Madsen and R. Vatrappu	2014
Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naive Bayes Classifier	S. Widya Sihwi and I. Prasetya Jati and R. Anggrainingsih	2018
Combining naive bayes and adjective analysis for sentiment detection on Twitter	M. Mertiya and A. Singh	2016
A Statistical and Evolutionary Approach to Sentiment Analysis	J. Carvalho and A. Prado and A. Plastino	2014
Corpus Usage for Sentiment Analysis of a Hashtag Twitter	Herlawati and R. T. Handayanto and D. Setiyadi and E. Retnoningsih	2019
Lexicon-Based Sentiment Analysis for Movie Review Twe-	A. Azizan and N. N. S. A. Jamal and M. N. Ab-	2019

Quadro 2 – Lista de trabalhos inseridos manualmente na revisão

Título original	Autores	Ano ⁴
An Empirical Investigation of Signaling in the Motion Picture Industry	Basuroy, S., Desai, K. K., Talukdar, D.	2006
Brand Extensions of Experiential Goods: Movie Sequel Evaluations	Sanjay Sood, Xavier Drèze	2006
Copcats vs. Original Mobile Apps: A Machine Learning Copycat-Detection Method and Empirical Analysis	Quan Wang, Beibei Li , Param Vir Singh	2018
Fast and frequent: Investigating box office revenues of motion picture sequels	Suman Basuroy, Subimal Chatterjee	2008
Managerial Objectives, the R-Rating Puzzle and the Production of Violent Films	S. Abraham Ravid and Suman Basuroy	2003
Prevision model and empirical test of box office results for sequels	Belvaux, Bertrand, Mencarelli, Rémi	2021
The long-term box office performance of sequel movies	Tirtha Dhar, Guanghui Sun, Charles Weinberg	2012
The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions	Jehoshua Eliashberg, Anita Elberse and Mark A. A. M. Leenders	2015
TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification	Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, Luis Espinosa-Anke	2010

Fonte – Brenno Ruschioni de Oliveira, 2023

4 Trabalhos Correlatos: Tweets Influenciando Hollywood?

4.1 Principais Trabalhos e Pontos Observados

4.1.1 Sequências de Filmes

Um dos estudos mais relevantes identificados a partir da revisão bibliográfica sistemática foi o trabalho de [Dhar, Sun e Weinberg \(2012\)](#), que utilizou um banco de dados abrangente, contendo filmes de Hollywood ao longo de 26 anos, para estimar a prevalência e a eficácia das sequências de filmes ao longo do tempo. Após considerar os efeitos da oferta e da demanda por meio de equações simultâneas, os autores concluíram que as sequências exercem um efeito indireto positivo, relacionado ao lado da oferta, devido ao número significativamente maior de cinemas exibindo essas produções em comparação com os filmes que não fazem parte de uma franquia. Em termos de efeito direto, ou seja, o impacto da demanda, os resultados indicam que as sequências superam os filmes não sequenciais, tanto na presença de consumidores na primeira semana de exibição quanto no total de semanas em cartaz. Além disso, os filmes originais, conhecidos como filmes pais, dos quais as sequências derivam, tendem a apresentar desempenho superior aos filmes não sequenciais, tanto em termos de audiência total quanto de público na primeira semana de exibição.

Curiosamente, [Dhar, Sun e Weinberg \(2012\)](#) ressaltam que, embora as sequências gerem menos audiência total do que os filmes pais, elas alcançam receitas iniciais maiores. Além disso, o impacto das sequências na audiência da primeira semana aumentou ao longo do tempo, embora o número de lançamentos de sequências não tenha acompanhado esse crescimento. Os autores sugerem que uma possível explicação para essa discrepância é o aumento nos orçamentos de produção das sequências em relação aos filmes originais, o que pode resultar em uma redução da margem bruta dentro de uma franquia.

Outro estudo de grande relevância encontrado na revisão sistemática é o trabalho de [Basuroy e Chatterjee \(2008\)](#), que investiga as sequências de filmes como uma extensão de marca de um produto hedônico, correlacionando as receitas de bilheteria das sequências com as dos filmes originais. O estudo também examina se o intervalo de tempo entre o lançamento da sequência e o filme pai, bem como o número de sequências intermediárias, influencia as receitas de bilheteria. Os resultados indicam que, embora as sequências não

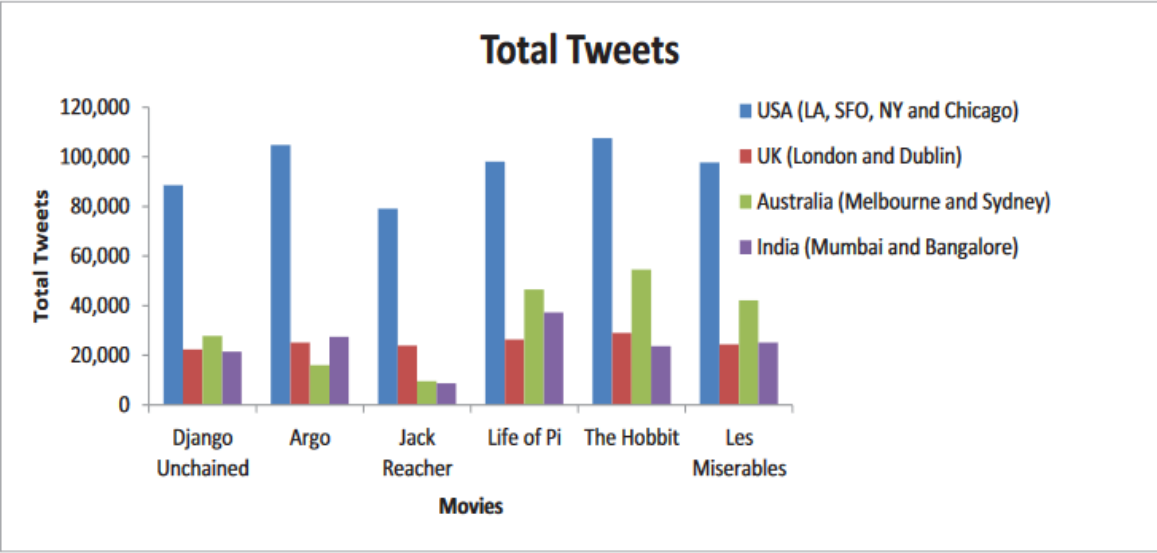
alcancem a mesma receita dos filmes pais, elas superam seus contemporâneos que não são sequências, especialmente quando são lançadas em um intervalo de tempo mais curto em relação ao filme original. Além disso, o estudo revela que as sequências experimentam um declínio mais acentuado nas arrecadações semanais de bilheteria em comparação com os filmes não sequenciais.

O trabalho de [Moon, Bergey e Iacobucci \(2010\)](#) também oferece uma importante contribuição ao investigar como as avaliações de críticos profissionais, comunidades amadoras e do público em geral influenciam as principais métricas de desempenho dos filmes, como receita e avaliações (ratings). Os autores descobriram que uma alta receita do filme original tende a aumentar as avaliações das sequências subsequentes. Além disso, os resultados mostram que um elevado investimento em publicidade, aliado a boas avaliações, maximiza a receita dos filmes. No entanto, o estudo também indica que, embora as sequências gerem receitas maiores, suas avaliações tendem a ser inferiores às dos filmes originais.

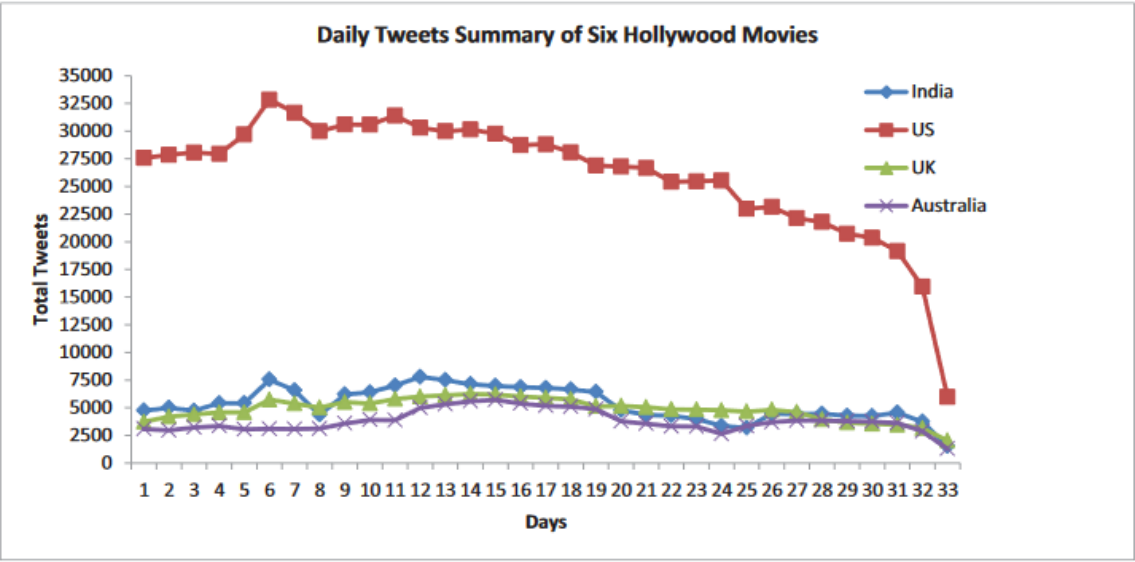
4.1.2 Análise de Dados e Twitter

O estudo de [Hodeghatta \(2013\)](#) revela que a análise de opiniões nas mídias sociais tem crescido significativamente devido ao aumento do volume de informações geradas diariamente. Esse aumento dificulta a obtenção de dados comparado a métodos mais tradicionais, como pesquisas e enquetes. Além disso, o autor destaca que o Twitter é uma das mídias sociais onde as opiniões expressas têm uma influência significativa na comercialização de produtos, podendo gerar uma onda positiva ou negativa no mercado. O foco central de [Hodeghatta \(2013\)](#) é a aplicação de algoritmos de classificação e análise de sentimentos em tweets relacionados a determinados filmes, com o objetivo de compreender o comportamento do mercado em relação aos consumidores e propor melhorias na experiência do cliente.

O trabalho de [Hodeghatta e Sahney \(2016\)](#) é outra referência fundamental para esta dissertação, ao investigar o Twitter como uma ferramenta de transmissão de e-WOM. O estudo demonstra que o Twitter, por sua influência como meio de comunicação de e-WOM, exerce impacto direto no mercado cinematográfico. O experimento envolveu a análise do comportamento do Twitter em sete países, avaliando dois blocos fundamentais do modelo honeycomb: compartilhamento e conversação. Foram estudados 27 filmes, em



Total Tweets Collected across all regions



Summary of Daily Tweets Collected for all movies

Figura 10 – Dados extraídos do Twitter por Hodeghatta, 2013

Fonte – [Hodeghatta \(2013\)](#)

22 cidades diferentes de sete países, abrangendo seis gêneros de filmes e um total de 9,28 milhões de tweets. O comportamento dos usuários do Twitter e as opiniões expressas foram comparados entre países e entre os blocos do modelo honeycomb, e os resultados indicaram diferenças significativas entre os gêneros de filmes e as culturas dos países analisados. Um dos fatores sugeridos para explicar essas variações é a influência cultural, onde algo considerado aceitável em uma cultura pode não ser bem visto em outra. Assim, o consumo de um produto, neste caso, filmes, pode gerar comportamentos e opiniões diferentes quando analisado em diferentes contextos culturais.

O estudo de [Zablocki, Schlegelmilch e Houston \(2019\)](#) reforça a importância da quantidade de tweets como uma variável forte para determinar o consumo de bens, no caso específico deste estudo, filmes. Mais precisamente, o estudo explora como a quantidade de tweets postados pelos consumidores sobre os filmes influencia o sucesso desses produtos no mercado.

4.1.3 Revisões de Filmes

Um dos estudos relevantes encontrados que aborda a importância das avaliações online na percepção da qualidade real dos produtos, com foco em filmes, é o trabalho de [Koh, Hu e Clemons \(2010\)](#). Este estudo investiga em que momento a média das classificações relatadas online corresponde à avaliação média percebida pela população em geral, incluindo tanto aqueles que publicam suas opiniões quanto aqueles que não o fazem, ou seja, avaliadores e não avaliadores. Os autores argumentam que as avaliações feitas pelos consumidores são influenciadas pela cultura na qual estão inseridos, sendo este um fator que afeta de maneira previsível o comportamento de avaliação.

O experimento foi conduzido utilizando dados extraídos de duas plataformas amplamente reconhecidas na área: IMDB.com, uma das maiores bases de dados de filmes do mundo, e o site chinês douban.com. O estudo revelou uma diferença significativa no comportamento de avaliação entre essas duas culturas. Além disso, foram examinados os impactos dos elementos culturais no comportamento de avaliação em uma cultura híbrida — Cingapura. Os resultados indicam que avaliadores com opiniões extremas são mais propensos a publicar suas avaliações, em comparação com aqueles que possuem opiniões moderadas, o que acaba gerando uma influência nas avaliações publicadas online.

O estudo também aponta que esse comportamento de prevalência de avaliações extremas é mais comum entre os reviews analisados nos Estados Unidos do que na China e em Singapura. Como resultado, as avaliações nos mercados chinês e singapurense se mostram mais alinhadas com a realidade da qualidade percebida pelos consumidores, em comparação com as avaliações americanas.

5 Dados e Métodos

Os dados utilizados para as análises deste estudo foram obtidos a partir de diversas fontes, incluindo Twitter, Box Office Mojo, Internet Movie Database (IMDb) e Wikipédia. Para as informações financeiras, elenco, características de produção e dados de gênero dos filmes, utilizamos o conjunto de dados compilado por [SOUZA \(2017\)](#), que engloba aproximadamente 15 mil filmes e suas respectivas informações.

Além desses dados, foram extraídos tweets que complementaram a análise, fornecendo métricas relevantes, tais como contagem de curtidas, compartilhamentos e outras interações. Esses dados de e-WOM (electronic Word-of-Mouth) são fundamentais para examinar a relação entre o comportamento online dos consumidores e o desempenho dos filmes nas bilheterias, assim como o impacto potencial na criação de sequências cinematográficas.

5.1 Extração de Tweets

Para analisar a influência do WOM no Twitter, considerando que os tweets de consumidores são uma forma de e-WOM, foi elaborado um método específico de extração, dada a vasta quantidade de tweets gerados diariamente na plataforma.

Inicialmente, decidiu-se que seriam analisados apenas filmes recentes, a partir de 2013, com o objetivo de coincidir com o lançamento da plataforma Twitter em 2009 e o consequente aumento do volume de dados gerados. Esse critério resultou em uma amostra inicial de 4.662 filmes.

Dada a quantidade significativa de filmes, foi aplicado um filtro adicional: apenas filmes categorizados como "Horror" pelo Box Office Mojo foram selecionados. Incluiu-se qualquer combinação de gênero que contivesse "Horror", resultando em um total de 423 filmes.

Após a seleção dos filmes, foi necessário delimitar o escopo temporal para a extração dos tweets. Devido ao volume elevado de tweets diários, seria inviável realizar a extração sem restrições temporais. Por isso, definiu-se uma janela de 15 dias, abrangendo 7 dias antes e 7 dias após o lançamento oficial nos cinemas de cada filme. Essa escolha visa capturar o volume máximo de interações sociais durante o período de maior relevância do filme nas redes sociais. Além disso, essa delimitação temporal permite a observação

do efeito de causalidade entre tweets e a criação de sequências, dada a natureza de curto prazo do efeito influência [Eliashberg e Shugan \(1997\)](#), [Reinstein e Snyder \(2005\)](#).

Alguns títulos de filmes exigiram ajustes na busca de tweets, principalmente aqueles com nomes simples ou únicos, como substantivos ou nomes próprios, que poderiam gerar ruído nos dados, inflando números ou levando a conclusões errôneas. Um exemplo foi o filme "Annabelle". Para reduzir o risco de coleta de tweets irrelevantes, foi adicionada a palavra "movie" ao final do título, como em "Annabelle movie", limitando a pesquisa a tweets especificamente relacionados ao filme.

Além disso, apenas tweets em inglês foram considerados, uma vez que o idioma abrange a maior parte dos tweets desde o lançamento da plataforma e abrange o principal mercado cinematográfico. A escolha do idioma inglês também facilita a aplicação de modelos de Processamento de Linguagem Natural (NLP), que possuem maior suporte para essa língua.

A coleta de dados foi realizada por meio de um script em Python utilizando a biblioteca *snsrape*. Embora não haja documentação oficial disponível, essa biblioteca é amplamente discutida em fóruns especializados em Python ([Github snsrape](#)). A *snsrape* utiliza um *wrapper* oficial do Python, permitindo a extração de tweets históricos, algo que não é suportado pela API oficial do Twitter, que limita o acesso a tweets recentes.

O script desenvolvido simula a função de busca avançada do Twitter, permitindo pesquisas com base em parâmetros definidos, como *queries* de texto. Esse método permitiu uma coleta eficiente de dados, com o processo de *scraping* e *wrapping* ocorrendo simultaneamente, otimizando a extração e organização dos dados para formar a base utilizada neste projeto.

O código utilizado para a coleta de tweets está disponível no Anexo A desta dissertação. O script foi escrito de forma genérica, mas parametrizado para rodar múltiplos filmes de forma simultânea.

Após a coleta inicial dos tweets, foi constatado um desafio: alguns filmes não retornaram tweets válidos com os parâmetros definidos (como *queries* e *datas*). Isso reduziu o número inicial de filmes de 423 para 386.

5.2 Dados complementares dos filmes

Para complementar os dados de tweets, foram coletadas informações de mercado e desempenho dos filmes selecionados, conforme extraídos e fornecidos por [SOUZA \(2017\)](#), além de dados referentes às características dos filmes. Essas informações visam responder questões essenciais para este estudo, organizando-as em uma base única. Entre as perguntas investigadas estão:

1. O filme é uma sequência de algum outro filme?
2. O filme é um *spin-off* gerado a partir de outro filme?
3. O filme gerou uma sequência?

Para responder a essas perguntas, foram utilizadas as informações disponíveis nas páginas da Wikipédia de cada filme. Nos casos em que as informações necessárias não estavam disponíveis na Wikipédia, seja por ausência de detalhes ou inexistência de uma página específica para o filme, recorreram-se aos dados disponíveis no IMDb ou Rotten Tomatoes.

Observou-se que todos os filmes que possuem uma sequência ou que fazem parte de uma franquia têm essa informação destacada nas respectivas páginas da Wikipédia, facilitando a identificação desses casos.

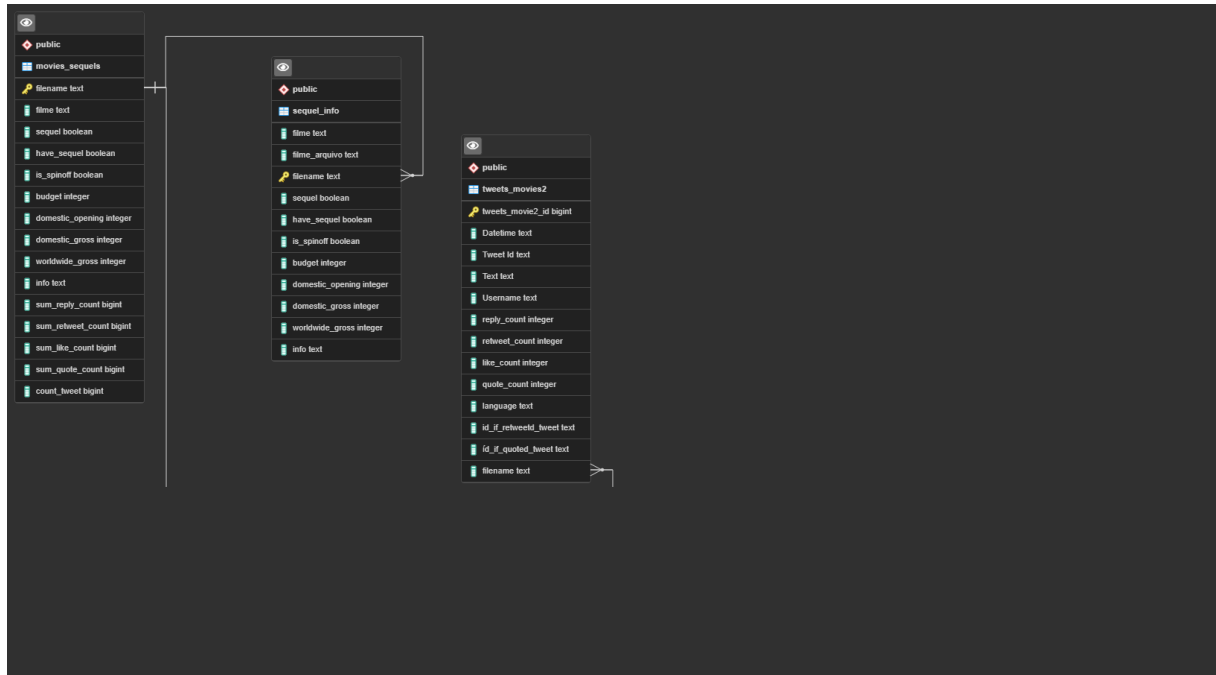
5.3 Banco de Dados de Filmes e Tweets

Após a coleta dos dados mencionados anteriormente, foi construída uma tabela que agrupa todas as informações referentes às características dos filmes, com o intuito de realizar o cruzamento dessas informações com os dados de tweets coletados para cada filme.

Devido ao elevado volume de dados, especialmente relacionado à quantidade significativa de tweets por filme, optou-se por criar um banco de dados simplificado para realizar as operações de tratamento e cruzamento entre as tabelas, garantindo maior eficiência e desempenho no processamento. O modelo relacional das tabelas está representado na Figura 11.

A primeira tabela contém os tweets associados a cada filme (*tweets movies2*), enquanto a segunda tabela contém as informações de características dos filmes (*sequel*

Figura 11 – Modelo Relacional do Banco de Dados de Tweets e Filmes



info). A terceira tabela representa a junção das duas anteriores, sendo esta utilizada nas análises estatísticas subsequentes (*movie sequels*).

5.4 Plataforma Utilizada

Dado o elevado volume de tweets coletados (aproximadamente 4,9 milhões de tweets), foi escolhida a plataforma Google Colab para a execução de todo o processamento dos dados. Optou-se pela aquisição da assinatura paga do Colab, visando obter maior capacidade de memória e velocidade de execução. Um notebook foi desenvolvido para documentar todo o processo de preparação, aplicação e tratamento dos dados pós-resultados, facilitando assim o desenvolvimento contínuo do projeto. Todos os notebooks utilizados estão disponíveis no apêndice deste trabalho.

5.5 Modelos NLP nos Tweets

5.5.1 Seleção dos Modelos

Com o intuito de enriquecer as informações extraídas dos textos dos tweets, foram selecionados métodos de Processamento de Linguagem Natural (Natural Language Pro-

cessing - NLP) para analisar as percepções de sentimentos e emoções contidas em cada tweet coletado. Durante a pesquisa para a aplicação desses métodos, foi identificado um grupo especializado em NLP, vinculado à Universidade de Cardiff no Reino Unido ([Github Cardiff NLP](#)), onde grande parte das aplicações se concentra na análise de redes sociais, com vários estudos publicados utilizando o Twitter como principal fonte para a análise de texto.

Para a análise de sentimentos, foi selecionado um modelo baseado no RoBERTa, treinado em aproximadamente 124 milhões de tweets de janeiro de 2018 a dezembro de 2021, e ajustado para análise de sentimentos utilizando o benchmark TweetEval ([BARBIERI et al., 2020](#)). Este modelo foi amplamente empregado pelo grupo de pesquisa Cardiff NLP no estudo ([LOUREIRO et al., 2022](#)).

No que diz respeito à análise de emoções, foi selecionado outro modelo também baseado no RoBERTa, com abordagem multilabel, treinado em 154 milhões de tweets até o final de dezembro de 2022. Esse modelo foi originalmente utilizado no projeto ([MOHAMMAD et al., 2018](#)). Contudo, visando aprimorar os resultados, a equipe Cardiff NLP realizou ajustes adicionais (tuning) nesse modelo, cujos resultados podem ser observados no trabalho de ([CAMACHO-COLLADOS et al., 2022](#)). O modelo ajustado foi o selecionado para a aplicação neste estudo.

Todos esses modelos foram disponibilizados para a comunidade científica pelo grupo Cardiff NLP em seu repositório no [HuggingFace](#).

5.6 Modelos de Classificação de Texto nos Tweets

5.6.1 Seleção dos Modelos

Dando prosseguimento ao estudo, após a obtenção de informações adicionais sobre os tweets por meio das aplicações dos modelos de processamento de linguagem natural (Análise de Sentimento e Emoção), foi realizado o agrupamento dos tweets por filme, com o intuito de gerar medidas agregadas para cada obra cinematográfica. Esse procedimento visava compreender o comportamento de cada filme em termos de e-WOM, permitindo assim analisar a influência dos tweets na possível criação de sequências de filmes de terror.

Para garantir a robustez dos resultados, foram selecionados dois modelos distintos de classificação para execução das análises, buscando respostas mais precisas: Floresta

Aleatória (Random Forest) e XGBoost. Ambos os modelos de regressão e classificação têm sido amplamente utilizados em estudos que analisam a influência de conteúdo em redes sociais, conforme demonstrado no trabalho de (RUI; LIU; WHINSTON, 2013).

Diversas das variáveis (features) utilizadas neste estudo foram derivadas de outras já coletadas durante o processo de extração de dados dos tweets e das informações financeiras e descritivas dos filmes selecionados. Algumas dessas variáveis foram geradas com base em regras específicas, como por exemplo: DISTRIBUTOR_GROUPED, MPAA_BIN e SEQUEL_SPIN_OFF. Essas regras classificam, respectivamente: as distribuidoras com maior número de filmes (caso estejam entre as 10 maiores, recebem uma marcação); a classificação indicativa do filme (se for adequado para maiores de 16 anos, recebe uma marcação); e a identificação de filmes que possuem sequência ou são derivados de outro (spin-off ou sequência direta).

As variáveis financeiras coletadas, conforme descrito por (SOUZA, 2017), foram ajustadas pela inflação e deflacionadas para o ano de 2019, o último ano considerado para a seleção dos filmes deste estudo. As variáveis mantiveram seus nomes originais, acrescidos do sufixo “DEFLACIONADO” para indicar a padronização dos valores. Esse ajuste foi realizado para garantir maior precisão na aplicação dos modelos, uma vez que todos os valores estavam padronizados em relação ao mesmo ano-base (2019).

Por fim, a variável de data de lançamento foi dividida em duas features: REL_MONTH (mês relativo da data de lançamento) e REL_YEAR (ano relativo da data de lançamento).

As features selecionadas, juntamente com suas descrições resumidas, estão listadas na Tabela 3.

Quadro 3 – Lista de Features

Feature	Descrição
count_valid_tweets	Quantidade de Tweets válidos.
reply_count	Soma de respostas aos Tweets.
retweet_count	Soma de retweets dos Tweets.
like_count	Soma de curtidas obtidas nos Tweets.
quote_count	Soma de Tweets citados (Quoted Retweets).
anger	Valor médio da emoção Raiva.
anticipation	Valor médio da emoção Antecipação.
disgust	Valor médio da emoção Nojo.
fear	Valor médio da emoção Medo.
joy	Valor médio da emoção Alegria.
love	Valor médio da emoção Amor.
optimism	Valor médio da emoção Otimismo.
pessimism	Valor médio da emoção Pessimismo.
sadness	Valor médio da emoção Tristeza.
surprise	Valor médio da emoção Surpresa.
trust	Valor médio da emoção Confiança.
negative	Valor médio do sentimento Negativo.
neutral	Valor médio do sentimento Neutro.
positive	Valor médio do sentimento Positivo.
REL_MONTH	Mês relativo da data de lançamento.
REL_YEAR	Ano relativo da data de lançamento.
DISTRIBUTOR_GROUPED	Flag indicando se o filme é de uma grande distribuidora.
Drama	Flag indicando se o filme é do gênero Drama.
Action	Flag indicando se o filme é do gênero Ação.
Comedy	Flag indicando se o filme é do gênero Comédia.
Horror	Flag indicando se o filme é do gênero Terror.
Adventure	Flag indicando se o filme é do gênero Aventura.
Crime	Flag indicando se o filme é do gênero Crime.
Fantasy	Flag indicando se o filme é do gênero Fantasia.
Biography	Flag indicando se o filme é do gênero Biografia.
Documentary	Flag indicando se o filme é do gênero Documentário.
Animation	Flag indicando se o filme é do gênero Animação.
Thriller	Flag indicando se o filme é do gênero Thriller.
Mystery	Flag indicando se o filme é do gênero Mistério.
History	Flag indicando se o filme é do gênero Histórico.
Sci-Fi	Flag indicando se o filme é do gênero Ficção Científica.
Romance	Flag indicando se o filme é do gênero Romance.
Music	Flag indicando se o filme é do gênero Música.
Family	Flag indicando se o filme é do gênero Família.
Western	Flag indicando se o filme é do gênero Velho Oeste.
Musical	Flag indicando se o filme é do gênero Musical.
War	Flag indicando se o filme é do gênero Guerra.
MPAA_BIN	Flag indicando se o filme é para todos os públicos (classificação indicativa).
RUNTIME	Duração do filme em minutos.
BUDGET_DEFLACIONADO	Orçamento ajustado para o ano de 2019.
DOMESTIC_OPENING_DEFLACIONADO	Receita de Abertura ajustada para o ano de 2019.
DOMESTIC_GROSS_DEFLACIONADO	Receita nacional (EUA) ajustada para o ano de 2019.
INTERNATIONAL_GROSS_DEFLACIONADO	Receita internacional ajustada para o ano de 2019.
WORLDWIDE_GROSS_DEFLACIONADO	Receita mundial ajustada para o ano de 2019.
HAVE_SEQUEL	Flag indicando se o filme possui sequência.
SEQUEL_SPINOFF	Flag indicando se o filme é uma sequência ou spin-off.

Quadro 4 – Tabela de Features dos Modelos de RF e XGBoost.

Fonte – Brenno Ruschioni de Oliveira, 2024

6 Resultados e Discussão

6.0.1 Medidas dos Tweets

Para uma melhor compreensão dos dados de tweets obtidos dos filmes selecionados, os filmes foram divididos em dois grupos: filmes que geraram sequência e filmes que não geraram sequência. A análise seguiu essa divisão em todos os aspectos trabalhados, incluindo análise de dados de tweets, dados financeiros, resultados dos modelos de sentimento e emoção, bem como os resultados dos modelos de aprendizado de máquina para classificação.

Além da categorização dos filmes entre aqueles que geraram sequência ou não, também foi gerada uma amostra dentro dessas categorias, contendo os filmes que obtiveram as maiores receitas mundiais (top 10), visando analisar os dados gerais e dos filmes de maior sucesso com maior precisão. Isso se justifica pelo fato de que filmes que geram mais receita têm maior probabilidade de originar sequências, como apontado por (DHAR; SUN; WEINBERG, 2012).

As variáveis de tweets coletadas e analisadas foram:

1. Quoted Retweets;
2. Likes;
3. Retweets;
4. Replies.

Ao observar essas informações para o grupo de filmes que geraram sequência e aqueles que não geraram, pode-se notar que (Figuras 12 e 13):

1. A média das variáveis de tweets é majoritariamente maior nos filmes que geraram sequência, ou seja, há uma presença mais expressiva (e-WOM) desses filmes nas redes sociais, o que resulta em maior visibilidade (LEE; KIM; KIM, 2011);
2. Considerando o elevado desvio padrão, é interessante verificar os valores do terceiro quartil (top 75%), que confirmam que as diferenças em todas as variáveis permanecem substancialmente maiores nos filmes que tiveram sequência;
3. No caso dos filmes que geraram sequência, nota-se também uma maior presença de *Replies* (respostas aos tweets), em contraste com o grupo de filmes que não geraram sequência, o que indica uma influência adicional nas redes sociais, por meio de maior interação.

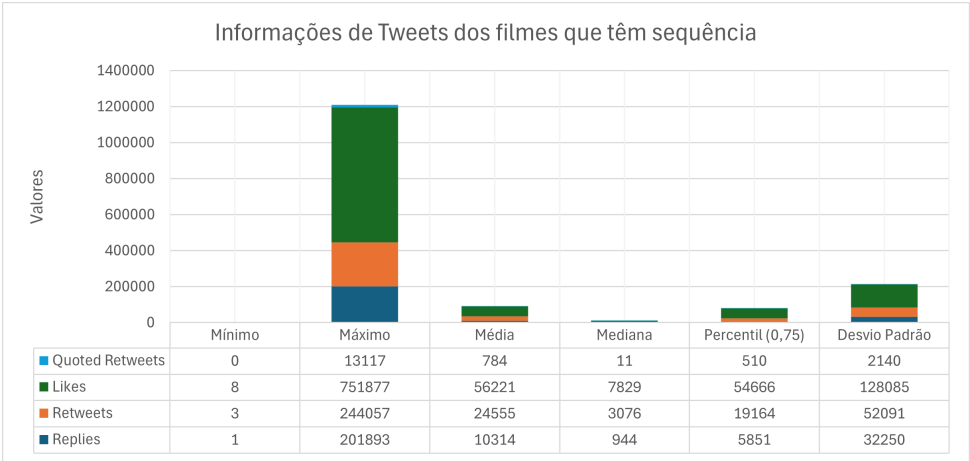


Figura 12 – (a) Informações de Tweets dos filmes que têm sequência

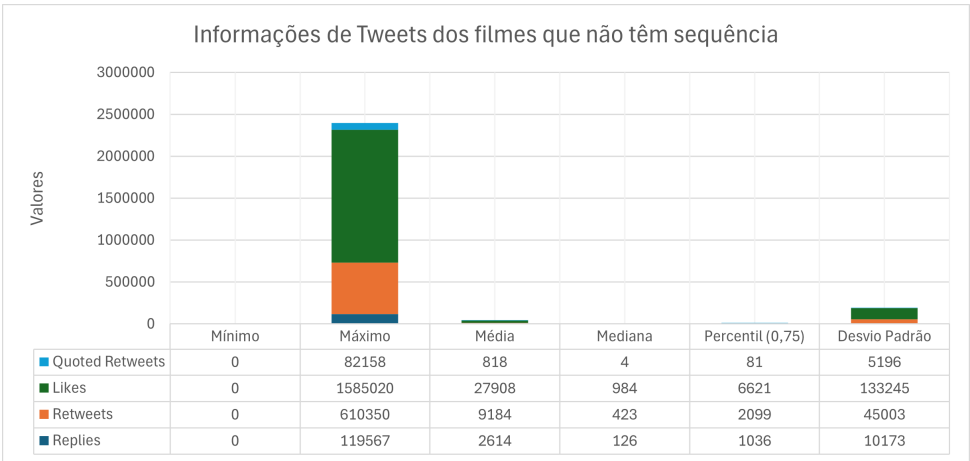


Figura 13 – (b) Informações de Tweets dos filmes que não têm sequência

Fonte – Brenno Ruschioni de Oliveira, 2024

Um aspecto importante a ser considerado na análise é o intervalo de tempo escolhido para a extração dos tweets, que compreendeu uma janela de 15 dias: 7 dias antes e 7 dias após o lançamento oficial dos filmes nos cinemas. Essa escolha foi estratégica para capturar o momento de maior relevância dos filmes nas redes sociais, ao mesmo tempo em que buscou minimizar o efeito de dupla causalidade, conforme discutido anteriormente neste trabalho [Eliashberg e Shugan \(1997\)](#), [Reinstein e Snyder \(2005\)](#). Ao delimitar a coleta de dados a esse período crítico, foi possível reduzir a possibilidade de que a popularidade dos tweets fosse tanto uma causa quanto uma consequência do desempenho dos filmes, permitindo uma análise mais precisa do impacto das interações sociais na decisão de criação de sequências.

Ao examinar o top 10 filmes em termos de receita mundial dentro de ambos os grupos, verifica-se que, em ambos os casos — filmes que geraram ou não sequência —, a presença nas redes sociais foi consideravelmente elevada em comparação com os dados gerais, incluindo as médias. Nos filmes que não geraram sequência, os valores médios foram até dez vezes maiores, enquanto nos filmes que geraram sequência, os valores foram quatro vezes maiores. Isso demonstra uma correlação entre o sucesso financeiro e a influência da marca nas redes sociais, conforme discutido por [Lee, Kim e Kim \(2011\)](#). Os comportamentos observados na análise anterior continuam válidos, inclusive em relação à variável *Replies* (respostas), que apresentou a terceira maior média, atrás apenas de *Likes* e *Retweets*. As *Replies* obtiveram números significativamente mais elevados nos filmes com sequência do que nos filmes sem sequência, com a média das *Replies* chegando a quase 50% do valor médio dos *Retweets* (aproximadamente na proporção de 89 mil para 40 mil, respectivamente). Os detalhes desses números podem ser observados nas Figuras 14 e 15.

6.0.2 Modelos de Análise de Sentimento e de Emoção

Após a coleta dos tweets, foi aplicada uma etapa de limpeza nos dados com o objetivo de excluir tweets com textos inválidos, bem como padronizar o conteúdo textual de alguns tweets. Isso seguiu os procedimentos utilizados em estudos anteriores de processamento de texto, como o de ([BARBIERI et al., 2020](#)), a fim de evitar problemas durante a classificação desses textos na execução do modelo. Além disso, devido ao grande volume de tweets, foi necessário dividir a tabela original em várias partes, facilitando o processo de execução e mitigando o consumo excessivo de memória.

Após a aplicação do modelo de processamento de linguagem natural para análise de sentimentos, foi gerada uma nova coluna contendo as pontuações dos sentimentos, categorizadas como: *negative*, *neutral*, e *positive* (Negativo, Neutro e Positivo). Cada tweet recebeu uma pontuação nessas três categorias, e o sentimento mais proeminente no texto foi representado pela maior pontuação entre as categorias. Para otimizar o uso dos dados coletados, essas pontuações foram divididas em três colunas independentes, permitindo seu uso posterior na aplicação dos modelos de classificação.

A execução do modelo de análise de emoções seguiu um fluxo similar ao modelo de análise de sentimentos, utilizando a mesma base de dados já processada. No entanto, o

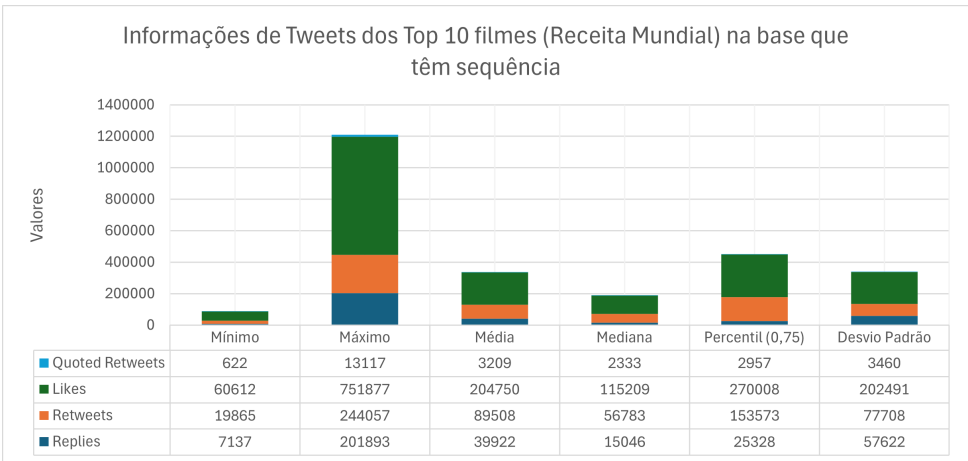


Figura 14 – (a) Informações de Tweets dos Top 10 filmes (Receita Mundial) na base que têm sequência

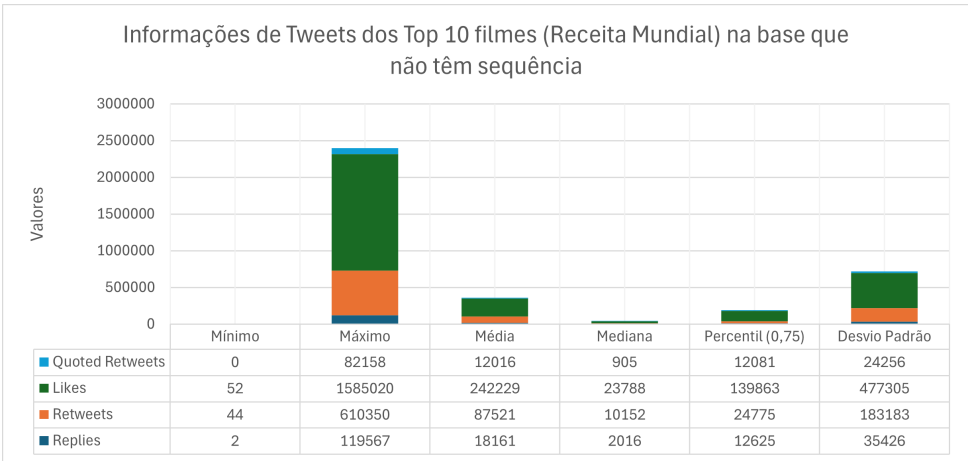


Figura 15 – (b) Informações de Tweets dos Top 10 filmes (Receita Mundial) na base que não têm sequência

Fonte – Brenno Ruschioni de Oliveira, 2024

modelo de análise de emoções demonstrou ser mais custoso em termos de processamento, devido ao número maior de variáveis geradas. Após o processamento, uma nova coluna foi criada para cada emoção, com os resultados numéricos das seguintes emoções: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *love*, *optimism*, *pessimism*, *sadness*, *surprise*, *trust* (raiva, antecipação, nojo, medo, alegria, amor, otimismo, pessimismo, tristeza, surpresa, confiança). Cada tweet recebeu uma pontuação em todas essas emoções, com uma pontuação mais elevada indicando uma representação mais forte da emoção na mensagem e uma pontuação mais baixa indicando uma presença menos significativa dessa emoção. Seguindo o mesmo princípio da análise de sentimentos, essas emoções foram separadas em colunas

independentes, permitindo a utilização de todas as informações dos tweets no futuro modelo de classificação.

Após a aplicação e tratamento dos resultados dos dois modelos, foi criada uma base de dados consolidada, reunindo todos os resultados por meio do ID dos tweets. Dessa forma, foi gerada uma nova base geral contendo tanto os dados originais dos tweets (como número de likes, retweets, etc.) quanto os resultados do processamento de linguagem natural (modelos de análise de sentimento e emoções).

Os códigos utilizados para a execução desses modelos estão disponíveis no Apêndice [A.1.2](#).

A aplicação de ambos os modelos de forma visual pode ser observada no fluxograma apresentado na Figura [16](#).

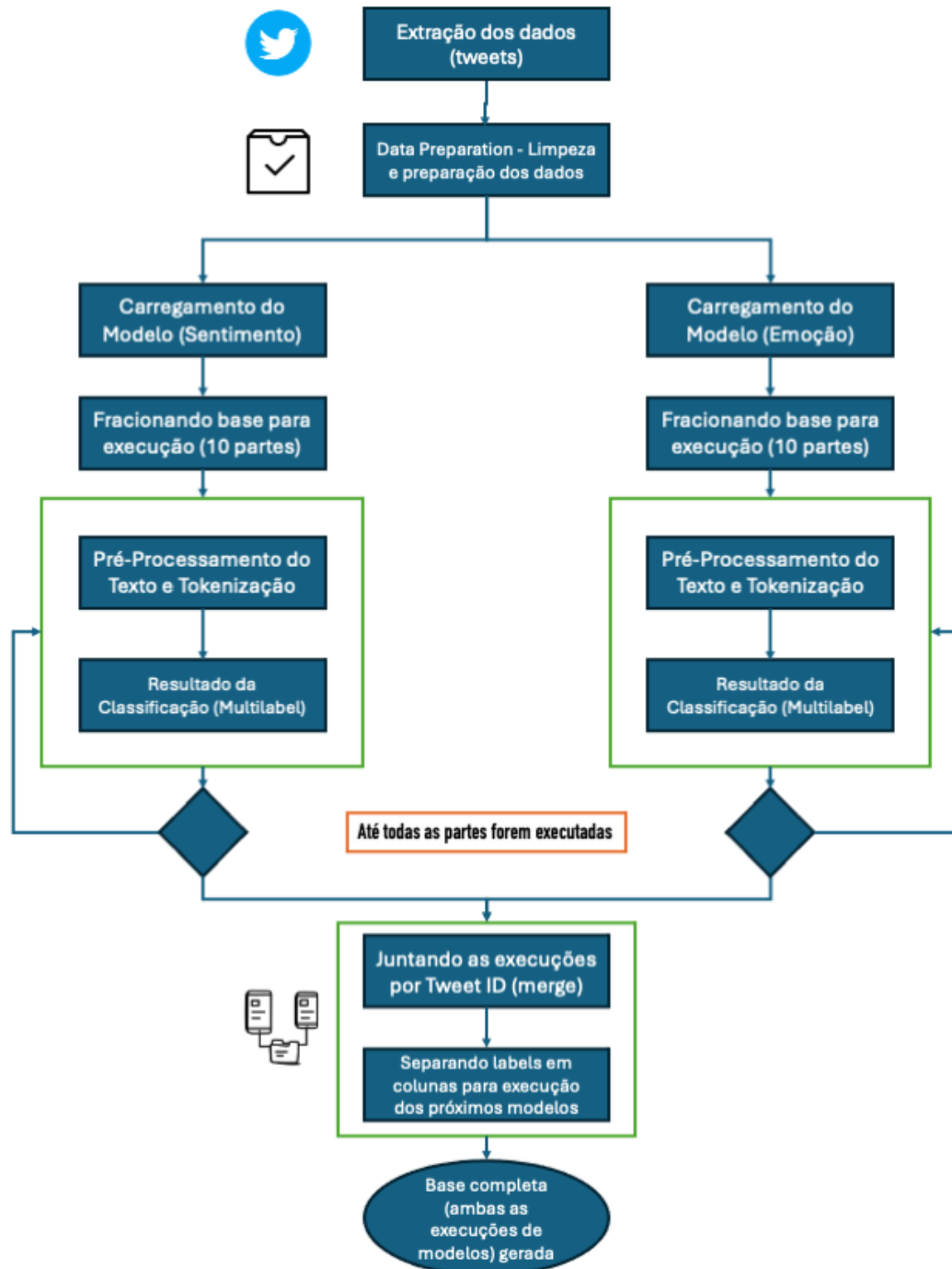
Resultados dos Modelos de Sentimento e de Emoções

Os resultados dos modelos de análise de sentimento e de emoções foram separados para fins comparativos, a fim de proporcionar uma análise mais clara e detalhada de ambos os conjuntos de dados. Com o intuito de aprofundar a análise, também foi selecionada uma amostra composta pelos filmes que geraram ou não geraram sequência, focando especificamente no grupo dos top 10 filmes com maior receita mundial em cada categoria. Esse recorte visa examinar o impacto do sucesso financeiro na criação de sequências, uma vez que, de acordo com [Basuroy e Chatterjee \(2008\)](#), o desempenho financeiro é uma das variáveis mais significativas na decisão de produzir uma sequência cinematográfica.

Análise: Sentimentos

Inicialmente, podemos observar algumas medidas estatísticas dos sentimentos agrupados para todos os filmes analisados, divididos entre as categorias de filmes que geraram sequência e filmes que não geraram sequência. Notamos uma diferença na média dos sentimentos neutros, em que os filmes que geraram sequência apresentam uma maior prevalência de sentimento positivo, aproximando-se bastante do valor neutro (Figuras: [17](#) e [18](#)).

Figura 16 – Fluxograma dos Modelos de Processamento de Linguagem Natural



Fonte – Brenno Ruschioni de Oliveira, 2024

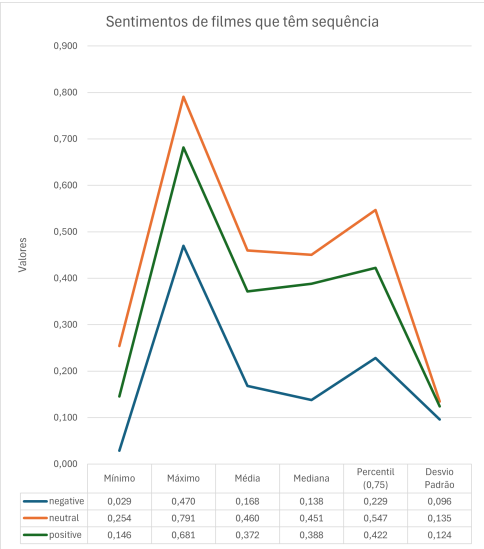


Figura 17 – (a) Medidas de Sentimentos de filmes que têm sequência

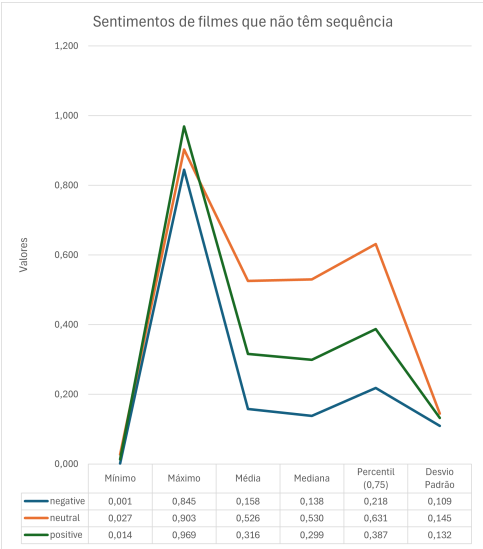


Figura 18 – (b) Medidas de Sentimentos de filmes que não têm sequência

Fonte – Brenno Ruschioni de Oliveira, 2024

Ao extrairmos uma amostra dos filmes que estão entre os dez maiores em termos de receita mundial, podemos notar que o comportamento citado anteriormente se mantém, mas de forma mais acentuada. Os filmes que geraram sequência apresentam uma média de sentimento neutro de aproximadamente 0,653 e positivo de 0,532. Já para os filmes que não geraram sequência, dentro do top 10 de maior receita nessa categoria, os valores de sentimento neutro ficam em 0,574 e de sentimento positivo em 0,293. O valor do sentimento positivo para os filmes com sequência é quase equivalente ao sentimento neutro dos filmes que não geraram sequência, que, nesse caso, representa a maior média entre os três sentimentos (Figuras: 19 e 20).

De maneira geral, ao observarmos apenas as medidas estatísticas dos sentimentos, é evidente uma maior incidência de tweets que indicam sentimentos positivos para os filmes que geraram sequência, entre todos os filmes da base de dados analisada. Esses resultados corroboram o estudo de Moon, Bergey e Iacobucci (2010), no qual o autor demonstra que as avaliações influenciam diretamente no desempenho final de um filme.

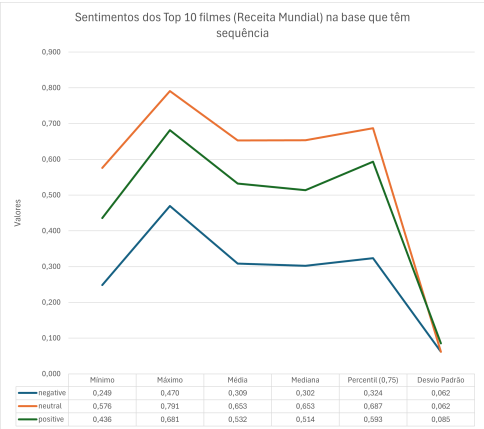


Figura 19 – (a) Sentimentos dos Top 10 filmes (Receita Mundial) na base que têm sequência

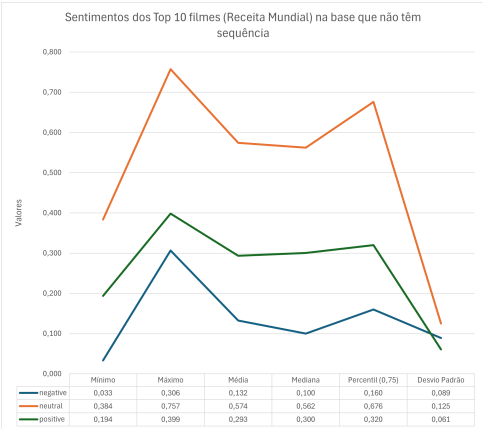


Figura 20 – (b) Sentimentos dos Top 10 filmes (Receita Mundial) na base que não têm sequência

Fonte – Brenno Ruschioni de Oliveira, 2024

Análise: Emoções

Avaliando as emoções obtidas dos tweets dos filmes selecionados, após a aplicação do modelo, e utilizando o mesmo padrão de observação empregado na análise dos sentimentos, é possível destacar alguns pontos relevantes (Figuras: 21 e 22):

1. As três emoções mais prevalentes em ambos os grupos são as mesmas na maioria das medidas observadas (*joy*, *anticipation* e *fear*);
2. No entanto, a medida de valor máximo apresentou algumas variações entre as categorias, com uma presença mais elevada de emoções como *anger* nos filmes com sequência e de *optimism* e *disgust* nos filmes sem sequência.

Ao analisarmos a amostra dos dez filmes com maior receita mundial, podemos identificar comportamentos semelhantes em relação àqueles observados na análise de sentimentos. Além disso, notamos algumas diferenças no comportamento quando comparamos esses resultados com a análise geral, que não separa a amostra dos filmes de maior receita mundial (Figuras: 23 e 24):

1. As três principais emoções observadas (*joy*, *anticipation* e *fear*) se mantêm constantes na maior parte das medidas analisadas;

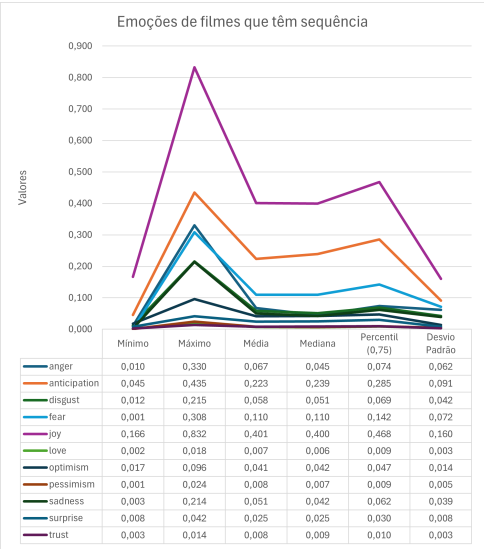


Figura 21 – (a) Emoções de filmes que têm sequência

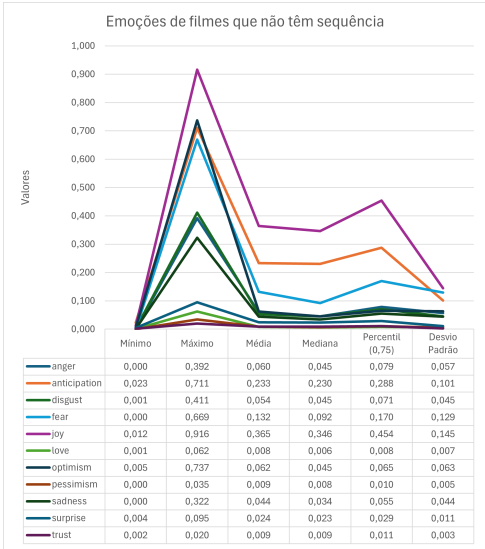


Figura 22 – (b) Emoções de filmes que não têm sequência

Fonte – Brenno Ruschioni de Oliveira, 2024

2. Na categoria de filmes que geraram sequência, as emoções apresentaram, em geral, valores mais elevados quando comparados aos resultados gerais, com destaque para as emoções *anger*, *sadness* e *disgust*, sendo *anger* a emoção que atingiu valores máximos superiores à emoção *fear*;
3. Na categoria de filmes que não geraram sequência, observa-se uma aproximação entre os valores de *anticipation* e *joy*, com algumas medidas mostrando a superação de *anticipation* sobre *joy*, sugerindo altos níveis de expectativa nos tweets extraídos. Isso pode indicar um possível fator que impeça a criação de sequência, conforme abordado por Moon, Bergey e Iacobucci (2010).

6.0.3 Aplicação do RF

Para a aplicação do modelo de Random Forest (Floresta Aleatória), foram selecionadas todas as variáveis do quadro 3 inicialmente, com o objetivo de observar os resultados considerando todas as features disponíveis. Primeiramente, foi definida a variável resposta, HAVE_SEQUEL, pois o intuito era classificar quais variáveis possuem maior influência na possível criação de uma sequência de filme. Após a definição da variável resposta,

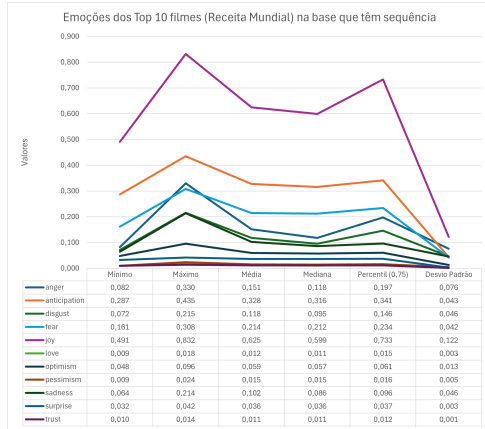


Figura 23 – (a) Emoções dos Top 10 filmes (Receita Mundial) na base que têm sequência

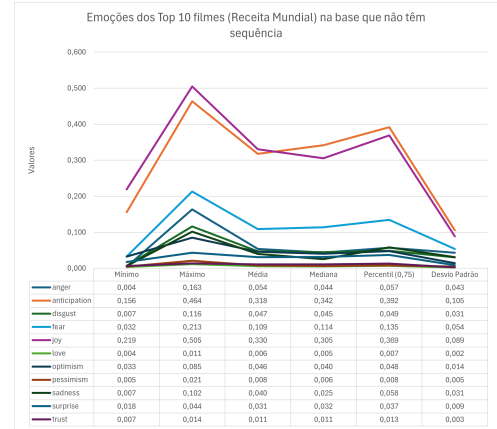


Figura 24 – (b) Emoções dos Top 10 filmes (Receita Mundial) na base que não têm sequência

Fonte – Brenno Ruschioni de Oliveira, 2024

observou-se um desbalanceamento da classe, com um maior número de observações (filmes) classificados como não possuindo sequência (Figura 25).

Para lidar com esse desbalanceamento, foi implementado o método de *over sampling* da biblioteca Python *imblearn*, a fim de complementar os dados de filmes que têm sequência e, conseqüentemente, aplicar o modelo de classificação sem reduzir significativamente a amostra. Dessa forma, evitou-se a aplicação de *under sampling*, que poderia resultar em perda de dados, reduzindo a amostra para apenas 43 filmes e potencialmente introduzindo *overfitting* no processo, dado o número limitado de filmes com sequência. Após o balanceamento, a base de dados foi dividida em treino e teste, e o classificador *Random Forest* foi aplicado à base de treino utilizando os seguintes parâmetros: `criterion = "gini"`, `max_depth = 15`, `min_samples_split = 3`, `n_estimators = 150`, `random_state = 100`.

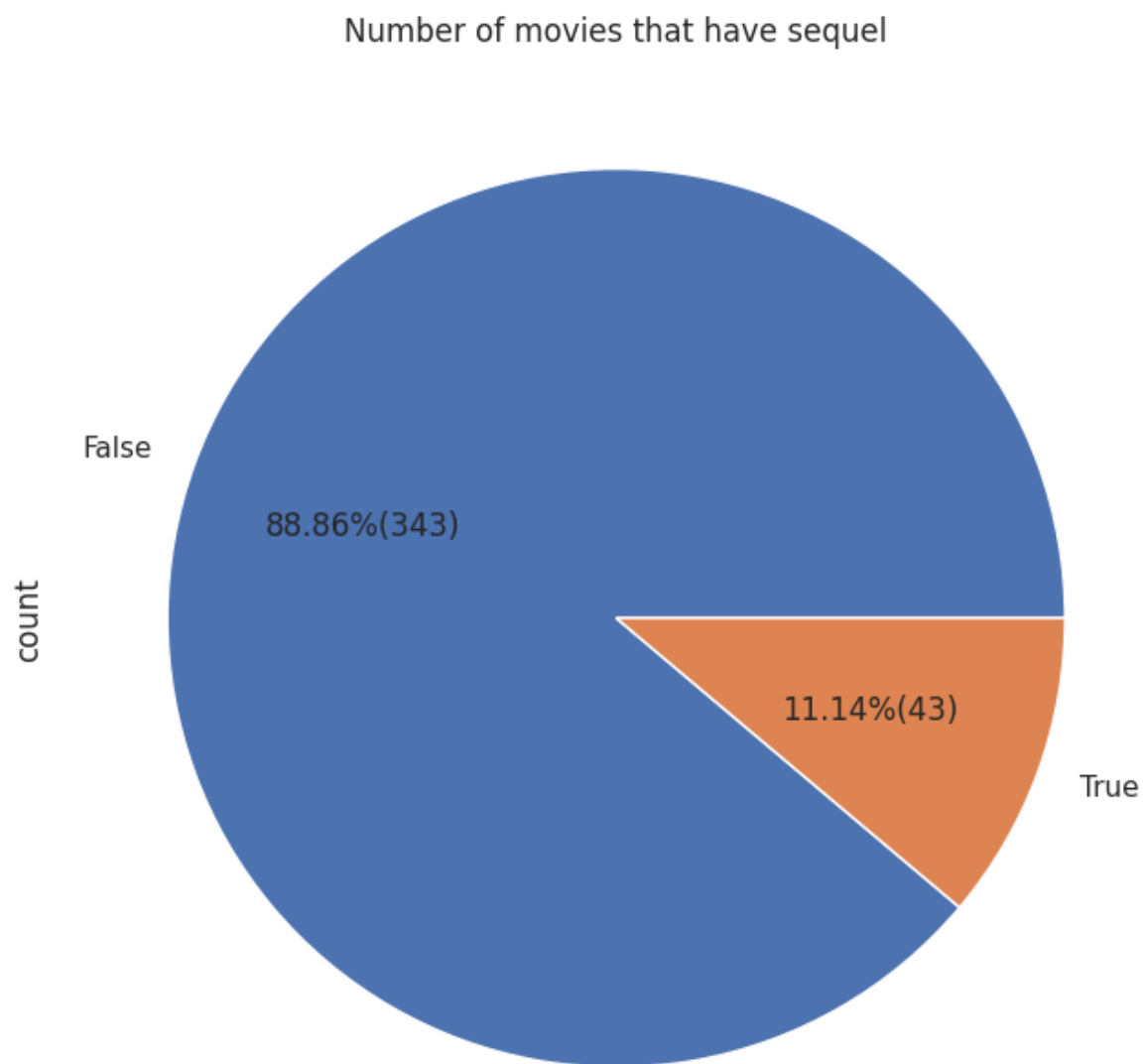
Em seguida, foram identificados os níveis de importância das variáveis (features) e gerada a matriz de confusão, que pode ser observada no quadro 5.

Quadro 5 – Matriz de Confusão RF.

Matriz de Confusão RF			
	Positivo	Negativo	Total
Positivo	89	3	92
Negativo	0	80	80
Total	89	83	N

Fonte – Brenno Ruschioni de Oliveira, 2024

Figura 25 – RF: Filmes que têm ou não sequência



Fonte – Brenno Ruschioni de Oliveira, 2024

A acurácia obtida foi de 0.9825581395348837, e o relatório de classificação pode ser observado no quadro 6.

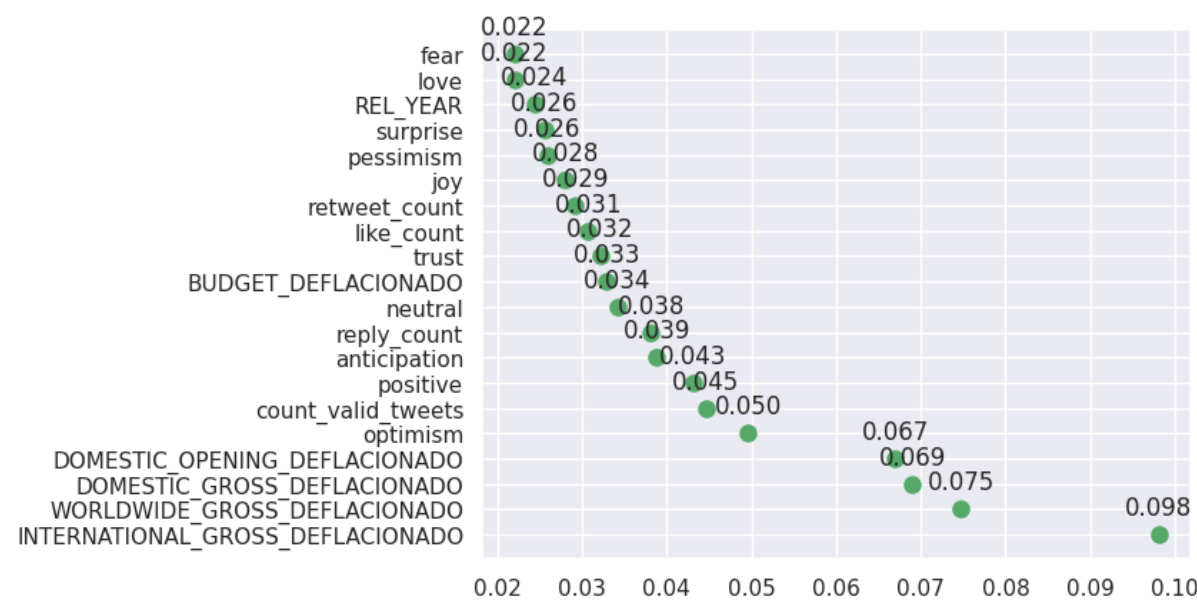
Quadro 6 – Relatório de Classificação RF.

Relatório de Classificação RF				
	Precisão	Recall	f1-score	suporte
Falso	0.97	1.00	0.98	89
Verdadeiro	1.00	0.96	0.98	83
Acurácia			0.98	172
Macro média	0.98	0.98	0.98	172
Média ponderada	0.98	0.98	0.98	172

Fonte – Brenno Ruschioni de Oliveira, 2024

Os resultados da importância das variáveis no modelo *Random Forest* podem ser visualizados na Figura 26.

Figura 26 – Importância das Variáveis no RF



Fonte – Brenno Ruschioni de Oliveira, 2024

O código completo da aplicação está disponível no apêndice A.1.3.

6.0.4 Aplicação do XGBoost

A aplicação do modelo de classificação XGBoost seguiu os mesmos princípios utilizados na aplicação do modelo de Floresta Aleatória (RF). Foram selecionadas todas

as features descritas no quadro de features (3), e a mesma variável resposta foi definida, com o objetivo de comparar os resultados entre os modelos aplicados.

Para tratar o problema de desbalanceamento das classes, foi utilizada a mesma técnica aplicada no modelo de RF, o *over sampling*, utilizando o parâmetro de "não maioria", padrão do método de oversampling (*RandomOverSampler(sampling_strategy="not majority")*). Em seguida, foi criado o objeto com o classificador XGBoost, integrado à biblioteca scikit-learn, para facilitar o processo de classificação.

Na primeira etapa, foi calculada a acurácia inicial do modelo, que apresentou um valor elevado, em torno de 96%. No entanto, como o XGBoost oferece uma ampla gama de parâmetros para ajuste fino (*tuning*), foi criado um objeto para testar diferentes combinações de parâmetros, como o número de folhas, profundidade das árvores, tipo de *booster*, taxa de aprendizado, entre outros. Após a execução dos testes, o classificador vencedor manteve a acurácia de 96%. Um ponto interessante desse processo (detalhado no notebook de execução, disponível no apêndice) é que o ajuste de parâmetros obteve o melhor resultado sem a utilização de alguns *boosters* nativos do XGBoost, como "Dart" e "gblinear", sendo o último baseado em regressão linear, como sugere seu nome.

Com o modelo otimizado, foi aplicada a classificação na base de treino, seguindo o fluxo até a obtenção dos resultados finais. A matriz de confusão pode ser observada no quadro 7.

Quadro 7 – Matriz de Confusão XGBoost.

Matriz de Confusão XGBoost			
	Positivo	Negativo	Total
Positivo	86	6	92
Negativo	0	80	80
Total	86	86	<i>N</i>

Fonte – Brenno Ruschioni de Oliveira, 2024

O relatório de classificação do modelo pode ser observado no quadro 8.

Por fim, foi gerado um gráfico com a importância das features, semelhante ao modelo de Floresta Aleatória. No entanto, alguns resultados se diferem, como pode ser observado na Figura 27.

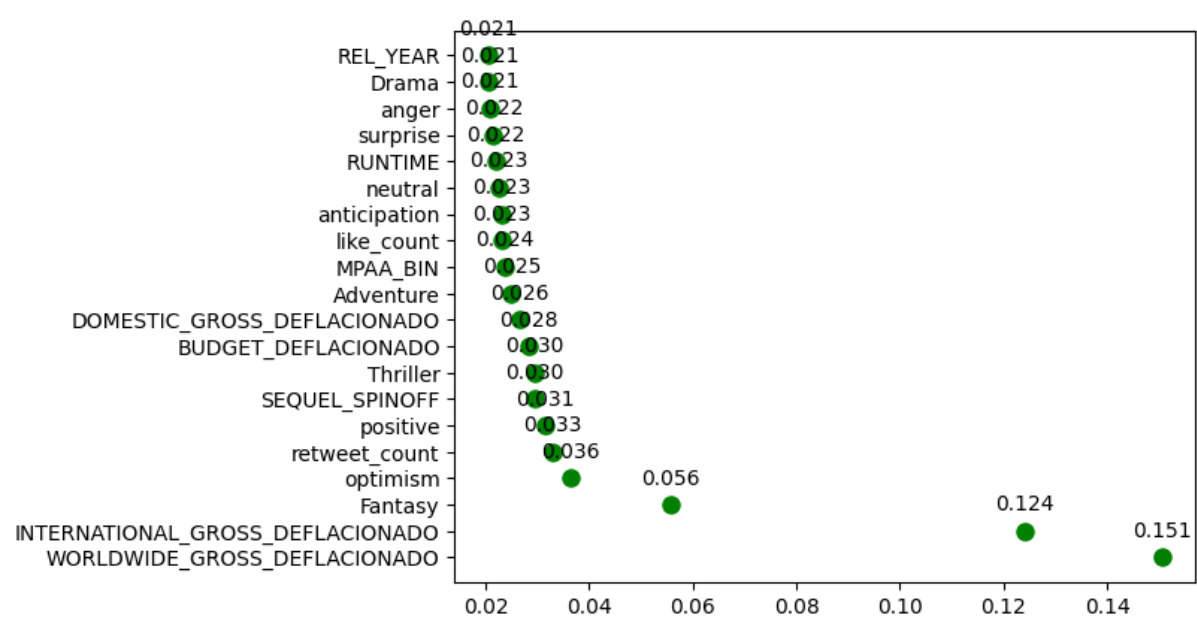
O código completo dessa aplicação pode ser encontrado no apêndice A.1.4.

Quadro 8 – Relatório de Classificação do XGBoost.

Relatório de Classificação XGBoost				
	Precisão	Recall	f1-score	Suporte
Falso	0.93	1.00	0.97	86
Verdadeiro	1.00	0.93	0.96	86
Acurácia			0.97	172
Macro média	0.97	0.97	0.97	172
Média ponderada	0.97	0.97	0.97	172

Fonte – Brenno Ruschioni de Oliveira, 2024

Figura 27 – XGBoost: Importância das Features



Fonte – Brenno Ruschioni de Oliveira, 2024

6.0.5 Resultados sem Balanceamento de Classes

Por fim, ambos os modelos foram executados sem o balanceamento das classes para analisar seu comportamento e verificar a diferença na precisão dos modelos, bem como a possível alteração da influência de cada feature. Primeiramente, observamos os resultados das matrizes de confusão de ambos os modelos (Matrizes 9 e 10), onde foram observados valores menores em todas as variáveis analisadas, incluindo a acurácia. Esta execução sem o balanceamento replicou exatamente o processo anterior, excluindo apenas a etapa de *OverSampling*.

Analisando a importância das features em ambos os resultados, podemos observar, no modelo RF, um comportamento semelhante na ordem de relevância das features,

Quadro 9 – Relatório de Classificação do Random Forest (sem balanceamento de classes).

RF (Sem balanceamento das classes) - Relatório de Classificação				
	Precisão	Recall	f1-score	Suporte
Falso	1.00	0.95	0.97	92
Verdadeiro	0.50	1.00	0.67	5
Acurácia			0.95	97
Macro médio	0.75	0.97	0.82	97
Média ponderada	0.97	0.95	0.96	97

Fonte – Brenno Ruschioni de Oliveira, 2024

Quadro 10 – Relatório de Classificação do XGBoost (sem balanceamento de classes).

XGBoost (Sem balanceamento das classes) - Relatório de Classificação				
	Precisão	Recall	f1-score	Suporte
Falso	0.99	0.91	0.95	94
Verdadeiro	0.20	0.67	0.31	3
Acurácia			0.91	97
Macro médio	0.59	0.79	0.63	97
Média ponderada	0.96	0.91	0.93	97

Fonte – Brenno Ruschioni de Oliveira, 2024

especialmente nas principais (primeiras colocadas), que estão relacionadas a informações financeiras dos filmes. Algumas emoções, como *surprise* e *joy*, ganharam mais relevância, embora permaneçam na metade inferior da tabela de importância.

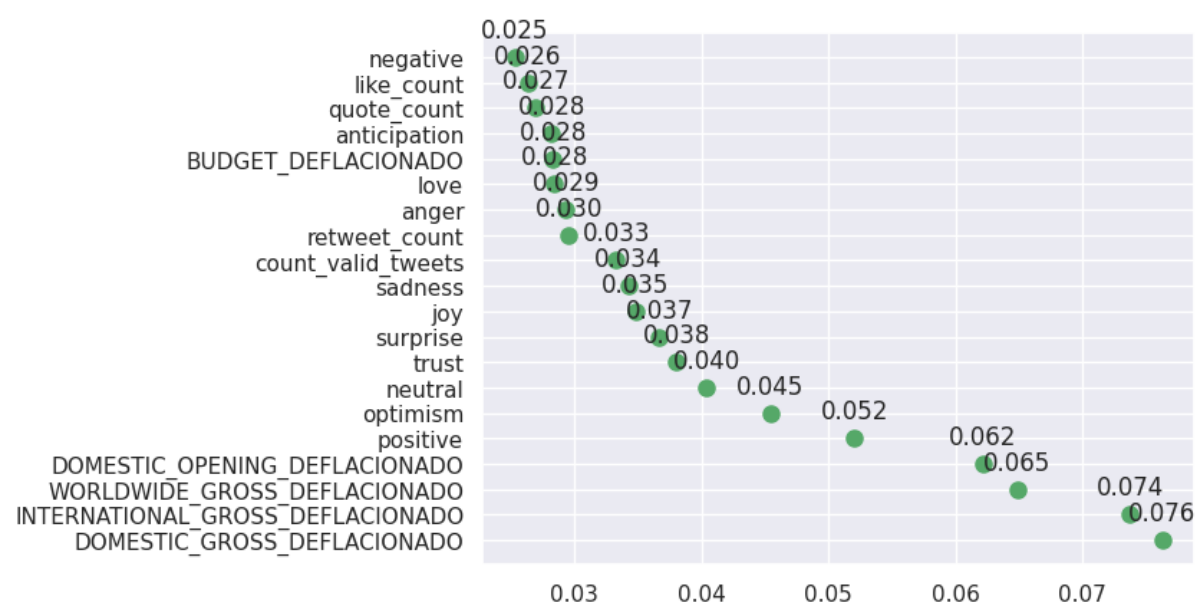
Para o XGBoost, houve um comportamento diferente nas principais variáveis. O *top 5* foi composto predominantemente por features financeiras, sendo a única exceção a feature de sentimento *positive*. Além disso, a distribuição das features apresentou uma menor disparidade entre os valores de importância, resultando em um menor desvio padrão em comparação com os resultados com balanceamento de classes.

Esses detalhes podem ser conferidos nas imagens [28](#) e [29](#).

6.0.6 Comparativo entre os Modelos Random Forest e XGBoost

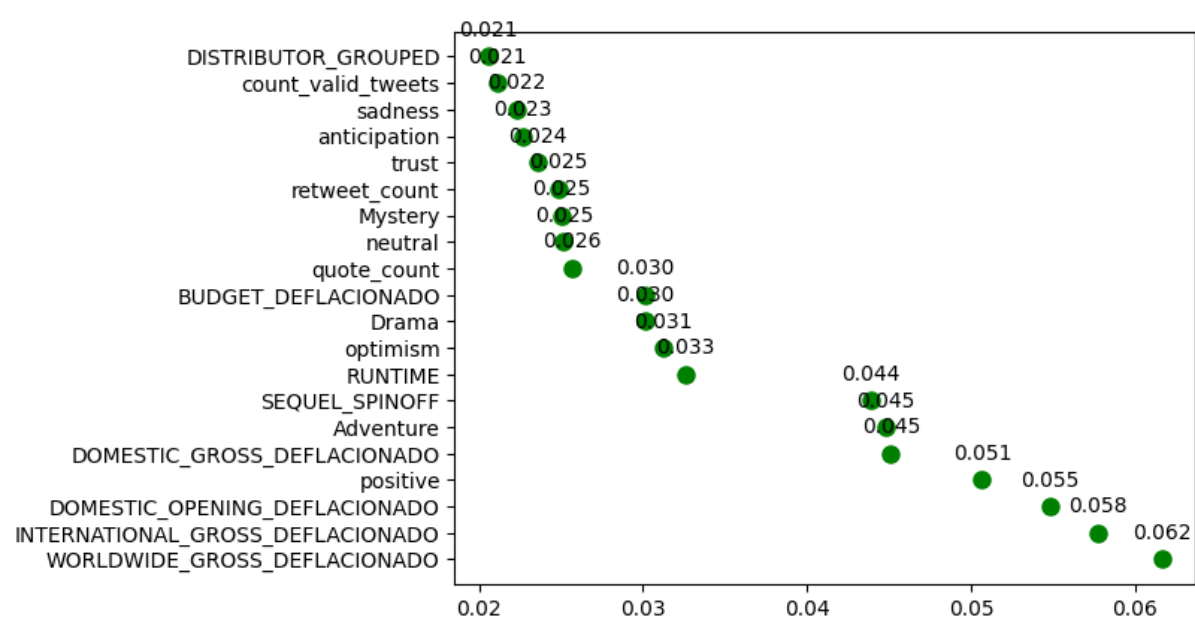
Nesta seção, comparamos o desempenho dos modelos de classificação Random Forest (RF) e XGBoost, avaliando suas performances com e sem o balanceamento das classes.

Figura 28 – RF (Sem Balanceamento): Importância das Features



Fonte – Brenno Ruschioni de Oliveira, 2024

Figura 29 – XGBoost (Sem Balanceamento): Importância das Features



Fonte – Brenno Ruschioni de Oliveira, 2024

Performance com Balanceamento de Classes

Com o balanceamento de classes, ambos os modelos apresentaram resultados robustos. O modelo Random Forest obteve uma acurácia de 0.98, conforme mostrado no quadro 6. O XGBoost, por sua vez, também apresentou uma acurácia elevada de 0.97, como observado no quadro 8.

Apesar da acurácia ligeiramente superior do Random Forest, o XGBoost demonstrou uma maior estabilidade nos parâmetros de recall e f1-score, especialmente na classe positiva (filmes com sequência), o que sugere que o XGBoost pode ser mais confiável na identificação de filmes que têm maior probabilidade de gerar sequências.

Performance sem Balanceamento de Classes

Quando os modelos foram aplicados sem o balanceamento das classes, ambos mostraram uma queda em suas acurácias, mas de maneira mais acentuada no modelo XGBoost. A acurácia do Random Forest caiu para 0.95, enquanto o XGBoost apresentou uma redução mais significativa para 0.91, como ilustrado nos quadros 9 e 10.

O Random Forest manteve uma maior precisão e recall na classe majoritária (filmes sem sequência), mas apresentou uma queda considerável na classe minoritária (filmes com sequência), refletida em um f1-score de 0.67. O XGBoost, por outro lado, teve uma performance mais instável, com uma precisão de apenas 0.20 na classe minoritária, sugerindo uma maior sensibilidade ao desbalanceamento das classes.

Importância das Features

Em ambos os modelos, as features financeiras foram as mais influentes, com destaque para as receitas no geral, como principal determinante na probabilidade de uma sequência. No entanto, as features emocionais, como *joy* e *anticipation*, ganharam mais relevância no modelo XGBoost, principalmente na análise sem balanceamento de classes, conforme mostrado nas figuras 28 e 29.

Conclusão Comparativa

Embora ambos os modelos tenham apresentado alta precisão, o Random Forest mostrou-se mais robusto em termos gerais, especialmente ao lidar com classes balanceadas. No entanto, o XGBoost demonstrou uma capacidade superior de ajuste fino através do tuning de parâmetros, o que pode ser explorado para melhorar ainda mais a performance em cenários específicos. Em situações de desbalanceamento de classes, o Random Forest apresentou uma maior resiliência, enquanto o XGBoost mostrou-se mais suscetível às discrepâncias entre as classes.

7 Conclusão

Este estudo investigou a influência do e-WOM no Twitter sobre a criação de sequências de filmes de terror, com foco em analisar como as interações e sentimentos expressos nas redes sociais podem impactar as decisões das produtoras de cinema. A partir de um conjunto de dados extraídos de fontes como Box Office Mojo, IMDb, Wikipedia e, principalmente, o Twitter, identificamos padrões significativos de comportamento dos consumidores em relação aos filmes analisados.

Os resultados indicam uma correlação significativa entre a quantidade e o tipo de interação gerada no Twitter e a probabilidade de criação de uma sequência. Filmes que geraram sequências apresentaram métricas de engajamento mais elevadas, como retweets, likes e replies, sugerindo que uma maior presença nas redes sociais pode influenciar diretamente as decisões de continuidade de uma franquia cinematográfica.

Além disso, a análise de sentimentos e emoções revelou que tweets com sentimentos positivos e emoções como alegria e antecipação foram predominantes em filmes que geraram sequências. Em contrapartida, filmes que não receberam sequências apresentaram maior volume de tweets com sentimentos neutros ou negativos. Isso destaca que não apenas a quantidade de interações, mas também a qualidade emocional do conteúdo, é crucial para as decisões de produção.

Entre as emoções mais destacadas, alegria, antecipação e medo foram consistentemente observadas nos filmes que geraram sequências, enquanto emoções como raiva e tristeza tiveram uma presença mais significativa em filmes sem sequência, especialmente naqueles com expectativas elevadas, mas que não alcançaram a continuidade. Esses achados corroboram a ideia de que a percepção pública e o estado emocional do público, conforme capturado pelas redes sociais, são elementos críticos na decisão de expandir uma franquia cinematográfica.

Os modelos de aprendizado de máquina aplicados, Random Forest e XGBoost, demonstraram alta acurácia na previsão de quais filmes gerariam sequências, com destaque para variáveis financeiras e métricas de engajamento no Twitter. No entanto, a aplicação dos modelos sem o balanceamento de classes revelou uma queda na precisão, especialmente no XGBoost, ressaltando a importância do tratamento adequado do desbalanceamento de classes para garantir resultados mais confiáveis.

Este estudo contribui para a literatura ao mostrar como o e-WOM no Twitter pode prever o sucesso futuro de filmes e a probabilidade de geração de sequências. As descobertas sugerem que as produtoras de cinema podem se beneficiar significativamente de análises detalhadas das redes sociais para tomar decisões mais assertivas sobre a continuidade de filmes. Futuras pesquisas poderiam expandir este estudo para outros gêneros cinematográficos e considerar variáveis adicionais que possam capturar melhor as nuances culturais e emocionais dos consumidores em diferentes mercados.

Referências

- ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. *Journal of economic perspectives*, v. 31, n. 2, p. 211–36, 2017. Citado na página 18.
- ALPAYDIN, E. *Introduction to Machine Learning, fourth edition*. MIT Press, 2020. (Adaptive Computation and Machine Learning series). ISBN 9780262043793. Disponível em: <<https://books.google.com.br/books?id=tZnSDwAAQBAJ>>. Citado na página 33.
- BARBIERI, F.; CAMACHO-COLLADOS, J.; ESPINOSA-ANKE, L.; NEVES, L. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In: *Proceedings of Findings of EMNLP*. [S.l.: s.n.], 2020. Citado 2 vezes nas páginas 54 e 59.
- BASUROY, S.; CHATTERJEE, S. Fast and frequent: Investigating box office revenues of motion picture sequels. *Journal of Business Research*, Elsevier, v. 61, n. 7, p. 798–803, 2008. Citado 2 vezes nas páginas 45 e 61.
- BISHOP, C. *Pattern Recognition and Machine Learning*. Springer New York, 2016. (Information Science and Statistics). ISBN 9781493938438. Disponível em: <<https://books.google.com.br/books?id=kOXDtAEACAAJ>>. Citado na página 33.
- BREIMAN, L. Random forests. *Mach. Learn.*, Kluwer Academic Publishers, USA, v. 45, n. 1, p. 5–32, oct 2001. ISSN 0885-6125. Disponível em: <<https://doi.org/10.1023/A:1010933404324>>. Citado na página 34.
- BROWN, J.; BRODERICK, A. J.; LEE, N. Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing*, Wiley Online Library, v. 21, n. 3, p. 2–20, 2007. Citado 3 vezes nas páginas 22, 23 e 25.
- CAMACHO-COLLADOS, J.; REZAEI, K.; RIAHI, T.; USHIO, A.; LOUREIRO, D.; ANTYPAS, D.; BOISSON, J.; ESPINOSA-ANKE, L.; LIU, F.; MARTÍNEZ-CÁMARA, E. *et al.* TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Abu Dhabi, U.A.E.: Association for Computational Linguistics, 2022. Citado 2 vezes nas páginas 32 e 54.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. Disponível em: <<http://arxiv.org/abs/1603.02754>>. Citado 2 vezes nas páginas 34 e 35.
- CHEUNG, C. M.; THADANI, D. R. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision support systems*, Elsevier, v. 54, n. 1, p. 461–470, 2012. Citado 3 vezes nas páginas 23, 24 e 25.
- CULNAN, M. J.; MCHUGH, P. J.; ZUBILLAGA, J. I. How large us companies can use twitter and other social media to gain business value. *MIS Quarterly Executive*, v. 9, n. 4, 2010. Citado na página 27.
- DAUGHERTY, T.; HOFFMAN, E. ewom and the importance of capturing consumer attention within social media. *Journal of Marketing Communications*, Taylor & Francis, v. 20, n. 1-2, p. 82–102, 2014. Citado 2 vezes nas páginas 22 e 25.

DELLAROCAS, C. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, INFORMS, v. 49, n. 10, p. 1407–1424, 2003. Citado 2 vezes nas páginas 16 e 22.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <<https://aclanthology.org/N19-1423>>. Citado na página 32.

DHAR, T.; SUN, G.; WEINBERG, C. B. The long-term box office performance of sequel movies. *Marketing Letters*, Springer, v. 23, n. 1, p. 13–29, 2012. Citado 2 vezes nas páginas 45 e 57.

ELIASHBERG, J.; SHUGAN, S. M. Film critics: Influencers or predictors? *Journal of marketing*, SAGE Publications Sage CA: Los Angeles, CA, v. 61, n. 2, p. 68–78, 1997. Citado 3 vezes nas páginas 20, 51 e 58.

GUPTA, P.; HARRIS, J. How e-wom recommendations influence product consideration and quality of choice: A motivation to process information perspective. *Journal of Business Research*, Elsevier, v. 63, n. 9-10, p. 1041–1049, 2010. Citado na página 25.

GURHAN-CANLI, Z.; MAHESWARAN, D. The effects of extensions on brand name dilution and enhancement. *Journal of marketing research*, SAGE Publications Sage CA: Los Angeles, CA, v. 35, n. 4, p. 464–473, 1998. Citado na página 29.

HANNA, R.; ROHM, A.; CRITTENDEN, V. L. We're all connected: The power of the social media ecosystem. *Business horizons*, Elsevier, v. 54, n. 3, p. 265–273, 2011. Citado 2 vezes nas páginas 18 e 26.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer New York, 2009. (Springer Series in Statistics). ISBN 9780387848587. Disponível em: <<https://books.google.com.br/books?id=tVIjmNS3Ob8C>>. Citado 2 vezes nas páginas 34 e 35.

HAUBL, G.; TRIFTS, V. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing science*, INFORMS, v. 19, n. 1, p. 4–21, 2000. Citado na página 26.

HENDERSON, S. *The Hollywood sequel: history & form, 1911-2010*. [S.l.]: Bloomsbury Publishing, 2017. Citado na página 29.

HENNIG-THURAU, T.; GWINNER, K. P.; WALSH, G.; GREMLER, D. D. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of interactive marketing*, Elsevier, v. 18, n. 1, p. 38–52, 2004. Citado na página 25.

HENNIG-THURAU, T.; HOUSTON, M. B.; HEITJANS, T. Conceptualizing and measuring the monetary value of brand extensions: The case of motion pictures. *Journal*

of Marketing, SAGE Publications Sage CA: Los Angeles, CA, v. 73, n. 6, p. 167–183, 2009. Citado 2 vezes nas páginas 16 e 17.

HODEGHATTA, U. R. Sentiment analysis of hollywood movies on twitter. In: IEEE. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. [S.l.], 2013. p. 1401–1404. Citado 2 vezes nas páginas 46 e 47.

HODEGHATTA, U. R.; SAHNEY, S. Understanding twitter as an e-wom. *Journal of Systems and Information Technology*, Emerald Group Publishing Limited, 2016. Citado 3 vezes nas páginas 18, 26 e 46.

HU, N.; KOH, N. S.; REDDY, S. K. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision support systems*, Elsevier, v. 57, p. 42–53, 2014. Citado na página 17.

HUETE-ALCOCER, N. A literature review of word of mouth and electronic word of mouth: Implications for consumer behavior. *Frontiers in psychology*, Frontiers, v. 8, p. 1256, 2017. Citado 4 vezes nas páginas 16, 22, 24 e 25.

HUSSAIN, S.; AHMED, W.; JAFAR, R. M. S.; RABNAWAZ, A.; JIANZHOU, Y. ewom source credibility, perceived risk and food product customer's information adoption. *Computers in Human Behavior*, Elsevier, v. 66, p. 96–102, 2017. Citado 2 vezes nas páginas 23 e 24.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Comput. Surv.*, v. 31, p. 264–323, 1999. Disponível em: <<https://api.semanticscholar.org/CorpusID:12744045>>. Citado na página 33.

JURAFSKY, D.; MARTIN, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. [S.l.]: Prentice Hall, 2008. v. 2. Citado na página 31.

KARNIOUCHINA, E. V. Impact of star and movie buzz on motion picture distribution and box office revenue. *International Journal of Research in Marketing*, Elsevier, v. 28, n. 1, p. 62–74, 2011. Citado na página 17.

KATZ, E.; LAZARSFELD, P. F. *Personal Influence, The part played by people in the flow of mass communications*. [S.l.]: Transaction Publishers, 1966. Citado na página 25.

KELLER, K. L.; AAKER, D. A. The effects of sequential introduction of brand extensions. *Journal of marketing research*, SAGE Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 35–50, 1992. Citado na página 29.

KOH, N. S.; HU, N.; CLEMONS, E. K. Do online reviews reflect a product's true perceived quality? an investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, Elsevier, v. 9, n. 5, p. 374–385, 2010. Citado na página 48.

LAU, G. T.; NG, S. Individual and situational factors influencing negative word-of-mouth behaviour. *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration*, Wiley Online Library, v. 18, n. 3, p. 163–178, 2001. Citado na página 22.

- LEE, D.; KIM, H. S.; KIM, J. K. The impact of online brand community type on consumer's community engagement behaviors: Consumer-created vs. marketer-created online brand community in online social-networking web sites. *Cyberpsychology, Behavior, and Social Networking*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 14, n. 1-2, p. 59–63, 2011. Citado 3 vezes nas páginas 26, 57 e 59.
- LEE, M. K.; SHI, N.; CHEUNG, C. M.; LIM, K. H.; SIA, C. L. Consumer's decision to shop online: The moderating role of positive informational social influence. *Information & management*, Elsevier, v. 48, n. 6, p. 185–191, 2011. Citado na página 23.
- LITVIN, S. W.; GOLDSMITH, R. E.; PAN, B. Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, Elsevier, v. 29, n. 3, p. 458–468, 2008. Citado 2 vezes nas páginas 22 e 23.
- LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. Disponível em: <<https://arxiv.org/abs/1907.11692>>. Citado na página 32.
- LOUREIRO, D.; BARBIERI, F.; NEVES, L.; ANKE, L. E.; CAMACHO-COLLADOS, J. *TimeLMs: Diachronic Language Models from Twitter*. 2022. Citado na página 54.
- LUO, C.; LUO, X. R.; SCHATZBERG, L.; SIA, C. L. Impact of informational factors on online recommendation credibility: The moderating role of source credibility. *Decision Support Systems*, Elsevier, v. 56, p. 92–102, 2013. Citado na página 24.
- MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. Disponível em: <<https://arxiv.org/abs/1310.4546>>. Citado na página 32.
- MITCHELL, T. *Machine Learning*. McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: <<https://books.google.com.br/books?id=EoYBngEACAAJ>>. Citado na página 33.
- MOHAMMAD, S.; BRAVO-MARQUEZ, F.; SALAMEH, M.; KIRITCHENKO, S. SemEval-2018 task 1: Affect in tweets. In: APIDIANAKI, M.; MOHAMMAD, S. M.; MAY, J.; SHUTOVA, E.; BETHARD, S.; CARPUAT, M. (Ed.). *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 1–17. Disponível em: <<https://aclanthology.org/S18-1001>>. Citado na página 54.
- MOON, S.; BERGEY, P. K.; IACOBUCCI, D. Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of marketing*, SAGE Publications Sage CA: Los Angeles, CA, v. 74, n. 1, p. 108–121, 2010. Citado 3 vezes nas páginas 46, 63 e 65.
- NIETO, J.; HERNANDEZ-MAESTRO, R. M.; MUNHOZ-GALLEGO, P. A. Marketing decisions, customer reviews, and business performance: The use of the top rural website by Spanish rural lodging establishments. *Tourism Management*, Elsevier, v. 45, p. 115–123, 2014. Citado 2 vezes nas páginas 22 e 23.

OPITZ, C.; HOFMANN, K. H. The more you know... the more you enjoy? applying 'consumption capital theory' to motion picture franchises. *Journal of Media Economics*, Taylor & Francis, v. 29, n. 4, p. 181–195, 2016. Citado na página 18.

PARK, C.; LEE, T. M. Information direction, website reputation and ewom effect: A moderating role of product type. *Journal of Business research*, Elsevier, v. 62, n. 1, p. 61–67, 2009. Citado na página 23.

PARK, N.; LEE, H. Social implications of smartphone use: Korean college students' smartphone use and psychological well-being. *Cyberpsychology, Behavior, and Social Networking*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 15, n. 9, p. 491–497, 2012. Citado 2 vezes nas páginas 18 e 26.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <<https://aclanthology.org/D14-1162>>. Citado na página 32.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTMELMOYER, L. *Deep contextualized word representations*. 2018. Disponível em: <<https://arxiv.org/abs/1802.05365>>. Citado na página 32.

REINSTEIN, D. A.; SNYDER, C. M. The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The journal of industrial economics*, Wiley Online Library, v. 53, n. 1, p. 27–51, 2005. Citado 2 vezes nas páginas 51 e 58.

ROYO-VELA, M.; CASAMASSIMA, P. The influence of belonging to virtual brand communities on consumers' affective commitment, satisfaction and word-of-mouth advertising. *Online Information Review*, Emerald Group Publishing Limited, 2011. Citado na página 23.

RUI, H.; LIU, Y.; WHINSTON, A. Whose and what chatter matters? the effect of tweets on movie sales. *Decision support systems*, Elsevier, v. 55, n. 4, p. 863–870, 2013. Citado 2 vezes nas páginas 27 e 55.

SHAPIRO, C.; VARIAN, H. *Information Rules: A Strategic Guide to The Network Economy*. [S.l.]: Harvard Business School Press, 2008. v. 30. ISBN 087584863X. Citado na página 18.

SILVERBLATT, A. *Genre studies in mass media: A handbook*. [S.l.]: Routledge, 2015. Citado na página 29.

SOOD, S.; DRÈZE, X. Brand extensions of experiential goods: Movie sequel evaluations. *Journal of Consumer Research*, The University of Chicago Press, v. 33, n. 3, p. 352–360, 2006. Citado na página 29.

SOTIRIADIS, M. D.; ZYL, C. V. Electronic word-of-mouth and online reviews in tourism services: the use of twitter by tourists. *Electronic Commerce Research*, Springer, v. 13, n. 1, p. 103–124, 2013. Citado 2 vezes nas páginas 24 e 25.

SOUZA, T. L. D. E. Os efeitos das revisões críticas online sobre o mercado cinematográfico americano. 2017. Citado 3 vezes nas páginas 50, 52 e 55.

STEUER, J. Defining virtual reality: Dimensions determining telepresence. *Journal of communication*, Wiley Online Library, v. 42, n. 4, p. 73–93, 1992. Citado na página 27.

SUTTON, R.; BARTO, A. *Reinforcement Learning, second edition: An Introduction*. MIT Press, 2018. (Adaptive Computation and Machine Learning series). ISBN 9780262352703. Disponível em: <<https://books.google.com.br/books?id=uWV0DwAAQBAJ>>. Citado na página 34.

VEASNA, S.; WU, W.-Y.; HUANG, C.-H. The impact of destination source credibility on destination satisfaction: The mediating effects of destination attachment and destination image. *Tourism Management*, Elsevier, v. 36, p. 511–526, 2013. Citado na página 24.

VUJIĆ, S.; ZHANG, X. Does twitter chatter matter? online reviews and box office revenues. *Applied Economics*, Taylor & Francis, v. 50, n. 34-35, p. 3702–3717, 2018. Citado na página 19.

WALDFOGEL, J. How digitization has created a golden age of music, movies, books, and television. *Journal of Economic Perspectives*, v. 31, n. 3, p. 195–214, August 2017. Disponível em: <<https://www.aeaweb.org/articles?id=10.1257/jep.31.3.195>>. Citado na página 20.

WANG, Y.; FESENMAIER, D. R. Towards understanding members' general participation in and active contribution to an online travel community. *Tourism management*, Elsevier, v. 25, n. 6, p. 709–722, 2004. Citado 2 vezes nas páginas 17 e 23.

YANG, F. X. Effects of restaurant satisfaction and knowledge sharing motivation on ewom intentions: the moderating role of technology acceptance factors. *Journal of Hospitality & Tourism Research*, SAGE Publications Sage CA: Los Angeles, CA, v. 41, n. 1, p. 93–127, 2017. Citado na página 25.

ZABLOCKI, A.; SCHLEGELMILCH, B.; HOUSTON, M. J. How valence, volume and variance of online reviews influence brand attitudes. *AMS Review*, Springer, v. 9, n. 1, p. 61–77, 2019. Citado na página 48.

Appendices

Todos os códigos que estão marcados como "Notebook" podem ser encontrados no link do meu github: <https://github.com/stuned/Mestrado/tree/main>

[illegible]

```

23 # Export dataframe into a CSV
24 tweets_df1.to_csv('nome_do_filme-sem-retweet.csv', sep=',', index=False)

```

Listing A.1 – Código para extração de tweets

A.1.2 Notebook - Modelos de Emoção e Sentimento (RoBERTa)

```

1 from google.colab import drive
2 drive.mount('/content/drive')
3
4 pip install tweetnlp
5
6 import tweetnlp
7 import tensorflow as tf
8 from transformers import pipeline
9 import timeit
10 from time import sleep
11 from IPython.display import clear_output
12 from tqdm.notebook import tqdm_notebook
13 import pandas as pd
14 import numpy as np
15 import os
16 from datetime import timedelta
17 import json
18
19 # Atribuindo os modelos que serao utilizados
20 model_emotion = tweetnlp.load_model('emotion', model_name='cardiffnlp/
    twitter-roberta-base-emotion-multilabel-latest')
21 sentiment_task = tweetnlp.load_model('sentiment', model_name='cardiffnlp
    /twitter-roberta-base-sentiment-latest')
22
23 ##### MODELO DE EMOCAO #####
24
25 tweets = pd.read_csv ('/content/drive/MyDrive/mestrado - base de dados/
    tweets completos.csv', dtype = {'Unnamed: 0': str, 'Datetime': str, '
    Tweet Id': str, 'Text': str, 'Username': str, 'reply_count': str, '
    retweet_count': str,
26
    'like_count':
    str, 'quote_count': str, 'language': str, 'id_if_retweeted_tweet': str
    , 'id_if_quoted_tweet': str, 'filename': str},

```

```
27         engine='python',encoding='utf-8', on_bad_lines='skip
    ')
28 tweets.head()
29
30 # Retirando valores de texto de tweets nulos
31 tweets = tweets[tweets['Text'].notnull()]
32
33 #checando operacao anterior
34 [tweets['Text'].isnull()]
35
36 # Dividindo arquivo em 10 partes para executar mais rapidamente
37 tweets_split = np.array_split(tweets, 10)
38
39 # atribuindo um dataframe a uma parte do arquivo
40 tweets_test1 = tweets_split[0]
41 tweets_test2 = tweets_split[1]
42 tweets_test3 = tweets_split[2]
43 tweets_test4 = tweets_split[3]
44 tweets_test5 = tweets_split[4]
45 tweets_test6 = tweets_split[5]
46 tweets_test7 = tweets_split[6]
47 tweets_test8 = tweets_split[7]
48 tweets_test9 = tweets_split[8]
49 tweets_test10 = tweets_split[9]
50
51 # Repetir para cada pedaco
52 # Slice 1 tweets_test1
53 result_emotion = []
54 inicio_emotion_time = timeit.default_timer()
55 with tf.device('/device:GPU:0'):
56     for row in tqdm_notebook(tweets_test1.itertuples(index=True, name='
        Pandas'), desc = 'Progress using tqdm_notebook()', total =
        tweets_test1.shape[0]):
57         emotion = model_emotion.predict(getattr(row, "Text"),
        return_probability=True)
58         result_emotion.append(emotion)
59         #print(getattr(row, "Text"))
60         sleep(0.01)
61 tweets_test1["Result_Emotion"] = result_emotion
```

```

62 fim_emotion_time = timeit.default_timer()
63 print ('duracao exec modelo de emocoos: %f' % (fim_emotion_time -
        inicio_emotion_time))
64 print(tweets_test1)
65
66 tweets_test1.to_csv('tweets_test1.csv', encoding='utf-8')
67
68 # replace with your folder's path
69 folder_path = r'/content/drive/MyDrive/mestrado - base de dados/
        tweets_split_emotion'
70
71 all_files = os.listdir(folder_path)
72
73 # Filter out non-CSV files
74 csv_files = [f for f in all_files if f.endswith('.csv')]
75
76 # Create a list to hold the dataframes
77 df_list = []
78
79 for csv in csv_files:
80     file_path = os.path.join(folder_path, csv)
81     try:
82         # Try reading the file using default UTF-8 encoding
83         df = pd.read_csv(file_path, dtype = {'Unnamed: 0': str, '
        Datetime': str, 'Tweet Id': str, 'Text': str, 'Username': str, '
        reply_count': str, 'retweet_count': str,
84                                     'like_count':
        str, 'quote_count': str, 'language': str, 'id_if_retweeted_tweet': str
        , 'id_if_quoted_tweet': str, 'filename': str, 'Result_Emotion': str},
85                             engine='python', encoding='utf-8', on_bad_lines='skip
        ')
86         df_list.append(df)
87     except UnicodeDecodeError:
88         try:
89             # If UTF-8 fails, try reading the file using UTF-16 encoding
            with tab separator
90             df = pd.read_csv(file_path, sep='\t', encoding='utf-16')
91             df_list.append(df)
92         except Exception as e:

```



```

93         print(f"Could not read file {csv} because of error: {e}")
94     except Exception as e:
95         print(f"Could not read file {csv} because of error: {e}")
96
97 # Concatenate all data into one DataFrame
98 big_df = pd.concat(df_list, ignore_index=True)
99
100 # Save the final result to a new CSV file
101 big_df.to_csv(os.path.join(folder_path, 'combined_tweets_emotion.csv'),
102               index=False)
103 ##### MODELO DE SENTIMENTO #####
104
105 tweets_sentiment_csv = pd.read_csv ('/content/drive/MyDrive/mestrado -
106                                     base de dados/tweets_split_emotion/combined_tweets_emotion.csv',
107                                     dtype = {'Unnamed: 0': str, 'Datetime': str, 'Tweet Id': str, 'Text':
108                                               str, 'Username': str, 'reply_count': str, 'retweet_count': str,
109                                               'like_count':
110                                               str, 'quote_count': str, 'language': str, 'id_if_retweeted_tweet': str
111                                               , 'id_if_quoted_tweet': str, 'filename': str, 'Result_emotion': str},
112                                     engine='python', encoding='utf-8', on_bad_lines='skip
113                                     ')
114 tweets_sentiment_csv.info()
115
116 tweets_sentiment_csv = tweets_sentiment_csv[tweets_sentiment_csv['Text'
117                               ].notnull()]
118
119 [tweets_sentiment_csv['Text'].isnull()]
120
121 tweets_sentiment_csv_split = np.array_split(tweets_sentiment_csv, 10)
122
123 tweets_test1_sentiment = tweets_sentiment_csv_split[0]
124 tweets_test2_sentiment = tweets_sentiment_csv_split[1]
125 tweets_test3_sentiment = tweets_sentiment_csv_split[2]
126 tweets_test4_sentiment = tweets_sentiment_csv_split[3]
127 tweets_test5_sentiment = tweets_sentiment_csv_split[4]
128 tweets_test6_sentiment = tweets_sentiment_csv_split[5]
129 tweets_test7_sentiment = tweets_sentiment_csv_split[6]
130 tweets_test8_sentiment = tweets_sentiment_csv_split[7]

```

```
124 tweets_test9_sentiment = tweets_sentiment_csv_split[8]
125 tweets_test10_sentiment = tweets_sentiment_csv_split[9]
126
127 # Refazer esse passo para todas as partes
128 # Slice 1 tweets_test1_sentiment
129 result_sentiment = []
130 inicio_sentiment_time = timeit.default_timer()
131 with tf.device('/device:GPU:0'):
132     for row in tqdm_notebook(tweets_test1_sentiment.itertuples(index=True,
133                             name='Pandas'), desc = 'Progress using tqdm_notebook()', total =
134                             tweets_test1_sentiment.shape[0]):
135         sentiment = sentiment_task.predict(getattr(row, "Text"),
136         return_probability=True)
137         result_sentiment.append(sentiment)
138         #print(getattr(row, "Text"))
139         sleep(0.01)
140 tweets_test1_sentiment["Result_Sentiment"] = result_sentiment
141 fim_sentiment_time = timeit.default_timer()
142 print ('duracao exec modelo de sentimentos: %f' % (fim_sentiment_time -
143         inicio_sentiment_time))
144 print(tweets_test1_sentiment)
145
146 tweets_test1_sentiment.to_csv('tweets_test1_sentiment.csv', encoding='
147         utf-8')
148
149 # replace with your folder's path
150 folder_path = r'/content/drive/MyDrive/mestrado - base de dados/
151         tweets_split_sentiment'
152
153 all_files = os.listdir(folder_path)
154
155 # Filter out non-CSV files
156 csv_files = [f for f in all_files if f.endswith('.csv')]
157
158 # Create a list to hold the dataframes
159 df_list = []
160
161 for csv in csv_files:
162     file_path = os.path.join(folder_path, csv)
```

```

157     try:
158         # Try reading the file using default UTF-8 encoding
159         df = pd.read_csv(file_path, dtype = {'Unnamed: 0': str, '
160         Datetime': str, 'Tweet Id': str, 'Text': str, 'Username': str, '
161         reply_count': str, 'retweet_count': str,
162                                     'like_count':
163         str, 'quote_count': str, 'language': str, 'id_if_retweeted_tweet': str
164         , 'id_if_quoted_tweet': str, 'filename': str, 'Result_Emotion': str,
165         'Result_Sentiment': str},
166         engine='python', encoding='utf-8', on_bad_lines='skip
167     ')
168     df_list.append(df)
169 except UnicodeDecodeError:
170     try:
171         # If UTF-8 fails, try reading the file using UTF-16 encoding
172         with tab separator
173         df = pd.read_csv(file_path, sep='\t', encoding='utf-16')
174         df_list.append(df)
175     except Exception as e:
176         print(f"Could not read file {csv} because of error: {e}")
177 except Exception as e:
178     print(f"Could not read file {csv} because of error: {e}")
179
180 # Concatenate all data into one DataFrame
181 big_df = pd.concat(df_list, ignore_index=True)
182
183 # Save the final result to a new CSV file
184 big_df.to_csv(os.path.join(folder_path, '
185     combined_tweets_sentiment_emotion.csv'), index=False)
186
187 ##### CRIACAO DA BASE COMPLETA #####
188
189 base_completa = pd.read_csv ('/content/drive/MyDrive/mestrado - base de
190     dados/combined_tweets_sentiment_emotion.csv', dtype = {'Unnamed: 0.2'
191     : str, 'Unnamed: 0.1': str, 'Unnamed: 0': str, 'Datetime': str, '
192     Tweet Id': str, 'Text': str, 'Username': str, 'reply_count': str, '
193     retweet_count': str,

```

```

183                                     'like_count':
    str, 'quote_count': str, 'language': str, 'id_if_retweeted_tweet': str
    , 'id_if_quoted_tweet': str, 'filename': str, 'Result_Emotion': str,
    'Result_Sentiment': str},
184         engine='python', encoding='utf-8', on_bad_lines='skip
    ')
185 base_completa.info()
186 base_completa.info()
187 base_completa.head()
188 #base_completa['Result_Emotion_Dict'] = base_completa['Result_Emotion'].
    str.replace('\'', '\"')
189 tweets_test_split = pd.DataFrame()
190 tweets_test_split['Result_Emotion'] = base_completa['Result_Emotion']
191 tweets_test_split['Result_Sentiment'] = base_completa['Result_Sentiment'
    ]
192 tweets_test_split['Tweet Id'] = base_completa['Tweet Id']
193 tweets_test_split.info()
194
195 from numpy import nan
196
197 result_emotion = []
198 result_sentiment = []
199 for row in tqdm_notebook(tweets_test_split.itertuples(index=True, name='
    Pandas'), desc = 'Progress using tqdm_notebook()', total =
    tweets_test_split.shape[0]):
200     dict_emotion = eval(str(getattr(row, "Result_Emotion")).replace("'", "
    \"))
201     dict_sentiment = eval(str(getattr(row, "Result_Sentiment")).replace("'",
    "\"))
202     result_emotion.append(dict_emotion)
203     result_sentiment.append(dict_sentiment)
204 tweets_test_split["Result_Emotion"] = result_emotion
205 tweets_test_split["Result_Sentiment"] = result_sentiment
206 tweets_test_split.info()
207
208 tweets_test_split_teste = pd.concat([tweets_test_split,
    tweets_test_split.pop("Result_Emotion").apply(pd.Series)], axis=1)
209
210 tweets_test_split_teste

```

```

211
212 probability_emotion = tweets_test_split_teste.pop('probability')
213 probability_emotion
214
215 tweets_test_split_emotion_teste = pd.concat([tweets_test_split_teste,
        probability_emotion.apply(pd.Series)], axis=1)
216
217 tweets_test_split_emotion_teste
218 tweets_test_split_sentiment = pd.concat([tweets_test_split_emotion_teste
        , tweets_test_split_emotion_teste.pop("Result_Sentiment").apply(pd.
        Series)], axis=1)
219 tweets_test_split_sentiment
220 probability_sentiment = tweets_test_split_sentiment.pop('probability')
221 probability_sentiment
222 tweets_test_split_sentiment_teste = pd.concat([
        tweets_test_split_sentiment, probability_sentiment.apply(pd.Series)],
        axis=1)
223 base_completa_separada = base_completa
224 base_completa_separada
225 base_completa_separada = base_completa.merge(
        tweets_test_split_sentiment_teste, left_index=True, right_index=True)
226 pd.set_option('display.max_columns', None)
227 base_completa_separada.head()
228 base_completa_separada_rename = base_completa_separada
229 base_completa_separada_rename.columns.map(str)
230 print(base_completa_separada_rename.columns.tolist())
231 base_completa_separada_rename.columns = ['deletar', 'deletar2', 'deletar3',
        ', 'datetime', 'tweet_id', 'text', 'username', 'reply_count', '
        retweet_count', 'like_count', 'quote_count', 'language', '
        id_if_retweeted_tweet', 'id_if_quoted_tweet', 'filename', '
        result_emotion', 'result_sentiment', 'tweet_id2', 'label_emotion', '
        confirmar_emotion', 'anger', 'anticipation', 'disgust', 'fear', 'joy',
        , 'love', 'optimism', 'pessimism', 'sadness', 'surprise', 'trust', '
        confirmar_sentiment', 'label_sentiment', 'negative', 'neutral', '
        positive']
232 print(base_completa_separada_rename.columns.tolist())
233 base_completa_separada_rename['label_emotion'].unique()
234 # Filtrando apenas os resultados com lingua inglesa ("en")

```

```

235 base_completa_separada_filtrada = base_completa_separada_rename[
    base_completa_separada_rename["language"] == 'en']
236 base_completa_separada_filtrada['tweet_id'].equals(
    base_completa_separada_filtrada['tweet_id2'])
237 base_completa_separada_filtrada['label_emotion'].unique()
238 teste_mask = pd.DataFrame()
239 teste_mask = base_completa_separada_filtrada[
    base_completa_separada_filtrada['id_if_quoted_tweet'].isna()]
240 teste_mask.head()
241 #deletando colunas que nao tem uso
242 del base_completa_separada_filtrada["deletar"]
243 del base_completa_separada_filtrada["deletar2"]
244 del base_completa_separada_filtrada["deletar3"]
245 del base_completa_separada_filtrada["confirmar_emotion"]
246 del base_completa_separada_filtrada["confirmar_sentiment"]
247 del base_completa_separada_filtrada["tweet_id2"]
248
249 base_completa_separada_filtrada.head()
250 #limpando a coluna id para nao mostrar os NaN
251 base_completa_separada_filtrada['id_if_retweeted_tweet'] =
    base_completa_separada_filtrada['id_if_retweeted_tweet'].fillna("")
252 base_completa_separada_filtrada['id_if_quoted_tweet'] =
    base_completa_separada_filtrada['id_if_quoted_tweet'].fillna("")
253 base_completa_separada_filtrada.head()
254
255 base_completa_separada_filtrada.to_csv('base_completa_separada_filtrada.
    csv', encoding='utf-8')

```

Listing A.2 – Aplicação dos modelos de sentimento e emoção baseados em BERT

A.1.3 Notebook - Random Forest

```

1 import pandas as pd
2 import numpy as np
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5
6 ! pip install -U imbalanced-learn
7
8 from google.colab import drive

```

```

9 drive.mount('/content/drive')
10
11 base_filmes = pd.read_csv ('/content/drive/MyDrive/Mestrado/mestrado -
    base de dados/movies_full_info_merged.csv', dtype = {'index': str, '
    filename': str, 'count_valid_tweets': int, 'reply_count': int, '
    retweet_count': int, 'like_count': int, 'quote_count': int, 'anger':
    float, 'anticipation': float, 'disgust': float, 'fear': float, 'joy':
    float, 'love': float, 'optimism': float, 'pessimism': float, '
    sadness': float, 'surprise': float, 'trust': float, 'negative': float
    , 'neutral': float, 'positive': float, 'FILME': str, 'ANO': int, '
    REL_MONTH_NAME': str, 'REL_MONTH': int, 'REL_DAY': int, 'REL_YEAR':
    int, 'MARKETS': str, 'DISTRIBUTOR_GROUPED': bool, 'DISTRIBUTOR': str,
    'GENRE': str, 'Drama': bool, 'Action': bool, 'Comedy': bool, 'Horror
    ': bool, 'Adventure': bool, 'Crime': bool, 'Fantasy': bool, '
    Biography': bool, 'Documentary': bool, 'Animation': bool, 'Thriller':
    bool, 'Mystery': bool, 'History': bool, 'Sci-Fi': bool, 'Romance':
    bool, 'Music': bool, 'Family': bool, 'Western': bool, 'Musical': bool
    , 'War': bool, 'MPAA': str, 'MPAA_BIN': bool, 'RUNTIME_HOUR': str, '
    RUNTIME_MINUTES': str, 'RUNTIME': int, 'ACTOR': str, 'CREW': str, '
    SINOPSE': str, 'BUDGET': float, 'DOMESTIC_OPENING': float, '
    DOMESTIC_GROSS': float, 'INTERNATIONAL_GROSS': float, '
    WORLDWIDE_GROSS': float, 'BUDGET_DEFLACIONADO': float, '
    DOMESTIC_OPENING_DEFLACIONADO': float, 'DOMESTIC_GROSS_DEFLACIONADO':
    float, 'INTERNATIONAL_GROSS_DEFLACIONADO': float, '
    WORLDWIDE_GROSS_DEFLACIONADO': float, 'SEQUEL': bool, 'HAVE_SEQUEL':
    bool, 'IS_SPINOFF': bool, 'SEQUEL_SPINOFF': bool, 'INFO': str},
12         engine='python', encoding='utf-8',
    on_bad_lines='skip', index_col=False)
13
14 base_filmes_rf = base_filmes[['count_valid_tweets', 'reply_count', '
    retweet_count', 'like_count', 'quote_count', 'anger', 'anticipation',
    'disgust', 'fear', 'joy', 'love', 'optimism', 'pessimism', 'sadness'
    , 'surprise', 'trust', 'negative', 'neutral', 'positive', 'REL_MONTH'
    , 'REL_YEAR', 'DISTRIBUTOR_GROUPED', 'Drama', 'Action', 'Comedy', '
    Horror', 'Adventure', 'Crime', 'Fantasy', 'Biography', 'Documentary',
    'Animation', 'Thriller', 'Mystery', 'History', 'Sci-Fi', 'Romance',
    'Music', 'Family', 'Western', 'Musical', 'War', 'MPAA_BIN', '
    BUDGET_DEFLACIONADO', 'DOMESTIC_OPENING_DEFLACIONADO', '

```

```
    DOMESTIC_GROSS_DEFLACIONADO', 'INTERNATIONAL_GROSS_DEFLACIONADO', '
    WORLDWIDE_GROSS_DEFLACIONADO', 'HAVE_SEQUEL']]

15
16 base_filmes.shape
17 base_filmes_rf.shape
18
19 base_filmes_rf.head()
20 base_filmes_rf.tail()
21 base_filmes_rf.info()
22 base_filmes_rf.describe()
23 # verificando nulls
24 base_filmes_rf.isnull().sum()
25
26 base_filmes_rf['HAVE_SEQUEL'].value_counts()
27 X = base_filmes_rf.drop(['HAVE_SEQUEL'], axis=1)
28 y = base_filmes_rf['HAVE_SEQUEL']
29
30 y.value_counts()
31 # Show pie plot (Approach 1)
32 y.value_counts().plot.pie(autopct='%.2f')
33
34 # Show pie plot (Approach 2) '%.f.2'
35 sns.set()
36 movies = y.value_counts()
37 movies.plot(kind='pie', title='Number of movies that have sequel',
    figsize=[8,8],
38     autopct=lambda p: '{:.2f}%({:.0f})'.format(p,(p/100)*movies.
    sum()))
39 plt.show()
40
41 #from imblearn.under_sampling import RandomUnderSampler
42
43 # rus = RandomUnderSampler(sampling_strategy=1) # Numerical value
44 #rus = RandomUnderSampler(sampling_strategy="not minority") # String
45 #X_res, y_res = rus.fit_resample(X, y)
46
47 #ax = y_res.value_counts().plot.pie(autopct='%.2f')
48 #_ = ax.set_title("Under-sampling")
49 #y_res.value_counts()
```



```
50
51 ### **Random Oversampling**
52
53 “"not majority"” = resample all classes but the majority class
54
55 from imblearn.over_sampling import RandomOverSampler
56
57 #ros = RandomOverSampler(sampling_strategy=1) # Float
58 ros = RandomOverSampler(sampling_strategy="not majority") # String
59 X_res, y_res = ros.fit_resample(X, y)
60
61 ax = y_res.value_counts().plot.pie(autopct='%0.2f')
62 _ = ax.set_title("Over-sampling")
63
64 y_res.value_counts()
65 ### **Performing RF**
66 X_res.shape
67 y_res.shape
68 from sklearn.model_selection import train_test_split
69
70 X_train, X_test, y_train, y_test = train_test_split(X_res, y_res,
71 random_state=100)
72 X_train.shape
73 y_train.shape
74
75
76 from sklearn.ensemble import RandomForestClassifier
77
78 clf = RandomForestClassifier(criterion="gini",
79 max_depth= 15,
80 min_samples_split= 3,
81 n_estimators = 150,
82 random_state= 100
83 )
84 clf.fit(X_train, y_train)
85 clf.feature_importances_
86 y_pred = clf.predict(X_test)
87 y_pred
88
89 from sklearn.metrics import confusion_matrix
```

```
88
89 confusion_matrix(y_test, y_pred)
90
91 from sklearn.metrics import accuracy_score
92
93 accuracy_score(y_test, y_pred)
94
95 from sklearn.model_selection import cross_val_score
96 cross_val_score(clf, X_train, y_train)
97
98 from sklearn.metrics import classification_report
99
100 print(classification_report(y_pred, y_test))
101
102
103 features = X.columns
104 importances = clf.feature_importances_
105 indices = np.argsort(importances)
106
107 # customized number
108 num_features = 20
109
110 plt.title('Feature Importances')
111 plt.barh(range(num_features), importances[indices[-num_features:]],
112         color='b', align='center')
112 plt.yticks(range(num_features), [features[i] for i in indices[-
113         num_features:]])
113 plt.xlabel('Relative Importance')
114 plt.show()
115
116 #fig, ax = plt.subplots()
117 import itertools
118 from operator import itemgetter
119
120 features = X.columns
121 importances = clf.feature_importances_
122
123 teste = dict(zip(features, importances))
124
```

```
125 s = dict(sorted(teste.items(), key=lambda item: item[1], reverse = True)
    [:20])
126
127 plt.clf()
128
129 # using some dummy data for this example
130 xs = list(s.values())
131 ys = list(s.keys())
132
133 # 'bo-' means blue color, round points, solid lines
134 plt.plot(xs,ys, marker='o', linestyle='', markersize=8, color='g')
135
136 # zip joins x and y coordinates in pairs
137 for x,y in zip(xs,ys):
138     label = "{:.3f}".format(x)
139     plt.annotate(label, # this is the text
140                 (x,y), # these are the coordinates to position the
141                     label
142                     textcoords="offset points", # how to position the text
143                     xytext=(0,10), # distance from text to points (x,y)
144                     ha='center') # horizontal alignment can be left, right
145                                     or center
146
147 plt.show()
```

Listing A.3 – Aplicação do Random Forest

A.1.4 Notebook - XGBoost

```
1 !pip install xgboost
2 !pip install -U imbalanced-learn
3
4 import pandas as pd
5 import numpy as np
6 from sklearn.model_selection import train_test_split
7 from sklearn import metrics
8 import xgboost as xgb
9 import seaborn as sns
10 import matplotlib.pyplot as plt
11
```

```

12 from google.colab import drive
13 drive.mount('/content/drive')
14
15 base_filmes = pd.read_csv ('/content/drive/MyDrive/Mestrado/mestrado -
    base de dados/movies_full_info_merged.csv', dtype = {'index': str, '
    filename': str, 'count_valid_tweets': int, 'reply_count': int, '
    retweet_count': int, 'like_count': int, 'quote_count': int, 'anger':
    float, 'anticipation': float, 'disgust': float, 'fear': float, 'joy':
    float, 'love': float, 'optimism': float, 'pessimism': float, '
    sadness': float, 'surprise': float, 'trust': float, 'negative': float
    , 'neutral': float, 'positive': float, 'FILME': str, 'ANO': int, '
    REL_MONTH_NAME': str, 'REL_MONTH': int, 'REL_DAY': int, 'REL_YEAR':
    int, 'MARKETS': str, 'DISTRIBUTOR_GROUPED': bool, 'DISTRIBUTOR': str,
    'GENRE': str, 'Drama': bool, 'Action': bool, 'Comedy': bool, 'Horror
    ': bool, 'Adventure': bool, 'Crime': bool, 'Fantasy': bool, '
    Biography': bool, 'Documentary': bool, 'Animation': bool, 'Thriller':
    bool, 'Mystery': bool, 'History': bool, 'Sci-Fi': bool, 'Romance':
    bool, 'Music': bool, 'Family': bool, 'Western': bool, 'Musical': bool
    , 'War': bool, 'MPAA': str, 'MPAA_BIN': bool, 'RUNTIME_HOUR': str, '
    RUNTIME_MINUTES': str, 'RUNTIME': int, 'ACTOR': str, 'CREW': str, '
    SINOPSE': str, 'BUDGET': float, 'DOMESTIC_OPENING': float, '
    DOMESTIC_GROSS': float, 'INTERNATIONAL_GROSS': float, '
    WORLDWIDE_GROSS': float, 'BUDGET_DEFLACIONADO': float, '
    DOMESTIC_OPENING_DEFLACIONADO': float, 'DOMESTIC_GROSS_DEFLACIONADO':
    float, 'INTERNATIONAL_GROSS_DEFLACIONADO': float, '
    WORLDWIDE_GROSS_DEFLACIONADO': float, 'SEQUEL': bool, 'HAVE_SEQUEL':
    bool, 'IS_SPINOFF': bool, 'SEQUEL_SPINOFF': bool, 'INFO': str},
16                                     engine='python', encoding='utf-8',
    on_bad_lines='skip', index_col=False)
17
18 base_filmes_rf = base_filmes[['count_valid_tweets', 'reply_count', '
    retweet_count', 'like_count', 'quote_count', 'anger', 'anticipation',
    'disgust', 'fear', 'joy', 'love', 'optimism', 'pessimism', 'sadness'
    , 'surprise', 'trust', 'negative', 'neutral', 'positive', '
    DISTRIBUTOR_GROUPED', 'Drama', 'Action', 'Comedy', 'Horror', '
    Adventure', 'Crime', 'Fantasy', 'Biography', 'Documentary', '
    Animation', 'Thriller', 'Mystery', 'History', 'Sci-Fi', 'Romance', '
    Music', 'Family', 'Western', 'Musical', 'War', 'MPAA_BIN', '
    BUDGET_DEFLACIONADO', 'DOMESTIC_OPENING_DEFLACIONADO', '

```

```
    'DOMESTIC_GROSS_DEFLACIONADO', 'INTERNATIONAL_GROSS_DEFLACIONADO', '
    'WORLDWIDE_GROSS_DEFLACIONADO', 'HAVE_SEQUEL', 'SEQUEL_SPINOFF', '
    'RUNTIME', 'REL_YEAR', 'REL_MONTH']]
```

```
19
20 base_filmes.shape
21 base_filmes_rf.shape
22 base_filmes_rf.head()
23 base_filmes_rf.tail()
24 base_filmes_rf.info()
25 base_filmes_rf.describe()
26 base_filmes_rf.isnull().sum()
27
28 base_filmes_rf['HAVE_SEQUEL'].value_counts()
29
30 X = base_filmes_rf.drop(['HAVE_SEQUEL'], axis=1)
31 y = base_filmes_rf['HAVE_SEQUEL']
32
33 from imblearn.over_sampling import RandomOverSampler
34
35 #ros = RandomOverSampler(sampling_strategy=1) # Float
36 ros = RandomOverSampler(sampling_strategy="not majority") # String
37 X_res, y_res = ros.fit_resample(X, y)
38
39 ax = y_res.value_counts().plot.pie(autopct='%.2f')
40 _ = ax.set_title("Over-sampling")
41 X_train, X_test, y_train, y_test = train_test_split(X_res, y_res,
    random_state=100)
42
43 X_train.shape, X_test.shape
44 len(X_test) / len(X_res)
45
46 ### Criando o objeto com o classificador XGBoost
47
48 classificador_xgb = xgb.XGBClassifier()
49 type(classificador_xgb)
50
51 from sklearn.model_selection import cross_val_score
52
53 #Funciona com scikitlearn
```

[illegible]

```
92         learning_rate=0.01,
93         random_state=100,
94         reg_alpha=0,
95         reg_lambda=1
96     )
97
98 100 * cross_val_score(classificador_xgb_gblinear, X_train, y_train).mean
99     ()
100
101 # Capricho de legibilidade
102
103 classificador_campeao = classificador_xgb_tunado
104
105 # Com o melhor modelo, podemos utilizar a base toda de treino
106 classificador_campeao.fit(X_train, y_train)
107
108 # Podemos realizar a predicao da base de teste!
109 predicoes = classificador_campeao.predict(X_test)
110
111 predicoes[:10]
112
113 y_test
114
115 # Calculando o numero de acertos
116 (predicoes == y_test).sum()
117
118 # Mas qual o tamanho da base de teste?
119 len(y_test)
120
121 acertos = (predicoes == y_test).sum()
122 total = len(y_test)
123
124 acuracia = 100 * acertos / total
125
126 from sklearn.metrics import confusion_matrix
127
128 confusion_matrix(y_test, predicoes)
129
```

```
130 from sklearn.metrics import accuracy_score
131
132 accuracy_score(y_test, predicoes)
133
134 from sklearn.metrics import classification_report
135
136 print(classification_report(predicoes, y_test))
137
138 features = X.columns
139 importances = classificador_campeao.feature_importances_
140 indices = np.argsort(importances)
141
142 # customized number
143 num_features = 30
144
145 plt.title('Feature Importances')
146 plt.barh(range(num_features), importances[indices[-num_features:]],
147         color='b', align='center')
147 plt.yticks(range(num_features), [features[i] for i in indices[-
148         num_features:]])
148 plt.xlabel('Relative Importance')
149 plt.show()
150
151 #fig, ax = plt.subplots()
152 import itertools
153 from operator import itemgetter
154
155 features = X.columns
156 importances = classificador_campeao.feature_importances_
157
158 teste = dict(zip(features, importances))
159
160 s = dict(sorted(teste.items(), key=lambda item: item[1], reverse = True)
161         [:20])
162
163 #plt.clf()
164
165 # using some dummy data for this example
166 xs = list(s.values())
```



```
166 ys = list(s.keys())
167
168 # 'bo-' means blue color, round points, solid lines
169 plt.plot(xs,ys, marker='o', linestyle='', markersize=8, color='g')
170
171 # zip joins x and y coordinates in pairs
172 for x,y in zip(xs,ys):
173     label = "{:.3f}".format(x)
174     plt.annotate(label, # this is the text
175                 (x,y), # these are the coordinates to position the
176                     label
177                     textcoords="offset points", # how to position the text
178                     xytext=(0,10), # distance from text to points (x,y)
179                     ha='center') # horizontal alignment can be left, right
180                                     or center
180 plt.show()
```

Listing A.4 – Aplicação do XGBoost