Insper
Economia

Marcos Felipe Dullius Silva

Predição de bilheterias de filmes brasileiros em estágio de pré-produção

São Paulo
2022

Marcos Felipe Silva

Predição de bilheterias de filmes brasileiros em estágio de pré-produção

Trabalho de Conclusão de Curso apresentado ao programa de graduação em economia como requisito parcial para obtenção do bacharelado em Ciências Econômicas

Orientador: Prof. André Filipe de Moraes Batista

São Paulo

2022

Silva, Marcos Felipe
Predição de bilheterias de filmes brasileiros em estágio de pré-produção
Marcos Felipe Dullius Silva. - São Paulo, 2022
X f.

Monografia: Faculdade de Economia
Insper Instituto de Ensino e Pesquisa.
Orientador: Prof. André Filipe de Moraes Batista
1. Cinema brasileiro. 2. Análise emocional. 3. Predição de bilheteria.

São Paulo

2022

Marcos Felipe Silva

Predição de bilheterias de filmes brasileiros em estágio de pré-produção

Trabalho de Conclusão de Curso apresentado ao programa de graduação em economia como requisito parcial para obtenção do bacharelado em Ciências Econômicas

Orientador: Prof. André Filipe de Moraes Batista

Banca examinadora

_____

André Filipe de Moraes Batista
Orientador

_____

Michel Fornaciali
Avaliador

São Paulo
2022
Resumo

Esse trabalho consiste numa análise de filmes brasileiros e seus respectivos arcos narrativos, de forma a poder compreender quais estilos de história possuem maior apelo ao público, indicado pela bilheteria gerada, e quais estilos possuem maior apelo à crítica, conforme notas aferidas por especialistas do campo cinematográfico. Com essas noções e, munido de outras informações sobre o filme que podem ser obtidas antes do início da produção - como o conteúdo do roteiro - cria-se um modelo de predição de performance para os filmes.

Palavras-chave: Cinema brasileiro. Análise emocional. Predição de bilheteria.

Abstract

This work consists of an analysis of Brazilian films and their respective narrative arcs, in order to understand which story styles have greater appeal to the public, indicated by the generated box office, and which styles have greater appeal to critics, according to notes given by experts in the cinematographic field. With these notions and, armed with other information about the film that can be obtained before the start of production - such as the content of the script - a performance prediction model for the films is created.

Keywords: Brazilian cinema. Emotional analysis. Box-office forecast.

Sumário

1 Introdução **8**

2 Apresentação preliminar da literatura relacionada ao tema **10**

3 Discussão preliminar da metodologia **12**

4 Resultados esperados **13**

5 Base de dados 16**6 Metodologia** 20**5**

7 Resultados 20**6**

8 Conclusão **21**

9 Referências bibliográficas **22**

1 Introdução

A indústria criativa, como a cinematográfica, por muito tempo se mostrou como algo imprevisível, que apenas aqueles que tivessem muita experiência no ramo poderiam buscar prever a performance de certas obras ou indicar o que deveria ou não receber atenção. Contudo, trabalhos recentes como Lash et al. (2019), Lee et al. (2016) e Lu et al. (2019) indicam que, por meio de dados, a tomada de decisão e o desenvolvimento da estratégia em meio a essa indústria pode ser aprimorado. Dessa forma, o ambiente de negócios torna-se menos incerto, uma vez que as decisões são baseadas em dados. Uma menor incerteza permite maiores investimentos, o que por sua vez abre espaço para novas criações artísticas, que anteriormente poderiam ser preteridas devido a decisões viesadas.

No Brasil, a indústria de cinema nacional possui dificuldades em competir com filmes estrangeiros, principalmente com as grandes produções de Hollywood que possuem orçamentos muito maiores, Michel et al. (2014). Já existem mecanismos que procuram trazer mais atenção aos filmes brasileiros, como as cotas em salas de cinema e, muitas vezes, as produções recebem, por meio da Lei Rouanet, dinheiro para a produção. Sob o olhar de um modelo de predição de sucesso, o governo poderia embasar as suas decisões de investimentos, focando ou em um maior apelo popular ou em um maior apelo da crítica, gerando, assim, uma maior eficiência do dinheiro público.

Desde a Grécia clássica, as histórias são classificadas em grandes grupos conforme a forma de seus arcos narrativos. Baseado no estudo de Reagan et al. (2016) pode-se dividir as histórias em seis principais grupos, fundamentados na flutuação de sentimentos ao longo do filme. Por exemplo, se a história começa, de forma geral, triste e, ao longo do filme, há um aumento da felicidade, instaura-se um arco denominado ‘dos trapos à riqueza’ (do inglês ‘rags to riches’). Com esse tipo de divisão, se torna possível lidar com diferentes tipos de narrativas, sendo que cada qual terá diferentes performances conforme seus gênero (ação, comédia, drama, etc) e seus orçamentos.

Com diferentes combinações de estruturas narrativas, gêneros e orçamentos forma-se a segmentação que será o alicerce do modelo de predição.

2 Apresentação preliminar da literatura relacionada ao tema

A literatura sobre o qual esse trabalho se baseia pode ser subdividida em quatro esferas de estudo: Indústria criativa, cinema no Brasil, análise de emoções e métodos de predição.

Sobre a indústria criativa, Goff et al. (2016) observam a indústria de cinema como possuindo dois lados, o da bilheteria e o da crítica, e conclui que esses lados são pouco correlacionados e respondem de formas distintas ao orçamento, gênero e conteúdo dos filmes. Carillatt et al (2017) realizam uma meta análise sobre o valor das grandes estrelas de cinema e concluem que a popularidade de um artista impacta mais a bilheteria do que seu reconhecimento artístico (medido pelas premiações concedidas). Addis et al. (2018) ao estudarem a nota da crítica especializada, a bilheteria inicial e a quantidade de nomeações ao Oscar no elenco; puderam concluir que apenas o primeiro elemento se mostrou relevante para explicar a avaliação do público geral.

Sobre o cinema no Brasil, Morelli-Mendes et al. (2016) traçam um panorama sobre a indústria criativa brasileira, especialmente a cinematográfica, averiguando um crescimento da mesma e um aumento, nas nações desenvolvidas, do interesse dos governos sobre a economia criativa. Michel et al. (2014) ao analisarem a indústria de cinema brasileira de 1995 até 2012 conseguiram notar a importância das políticas de fomento à cultura e como essas devem ser mantidas e aplicadas a diferentes elos da cadeia de valor do cinema para promover uma maior competitividade e diminuir os níveis de concentração de mercado. Gimenez (2021) analisou o impacto de políticas públicas realizadas pela ANCINE que buscavam desenvolver o cinema em diferentes regiões do país – principalmente fora de São Paulo e Rio de Janeiro. O trabalho concluiu que houve um aumento no número de filmes produzidos por diferentes unidades da federação, contudo, a performance desses filmes não acompanhou o aumento no volume.

Dentro da esfera de análise de emoções temos Reagan et al. (2016) que determinaram os seis principais arcos narrativos a partir de textos de ficção:

‘Dos trapos à riqueza’ - Ascensão

‘Da riqueza aos trapos’ - Queda

‘Homem dentro do buraco’ - Queda e ascensão

‘Ícaro’ - Ascensão e queda

‘Cinderela’ - Ascensão, queda e ascensão

‘Édipo’ - Queda, ascensão e queda

Chu et al. (2017) replicaram o experimento de Reagan para vídeos curtos da internet, obtendo resultados similares. Hur et al. (2016) e Kim et al. (2018) elaboraram métodos de predição de bilheteria baseando-se na análise emocional das críticas elaboradas pelos espectadores.

Outros métodos de predição são vistos em Ahmed et al. (2019) que formam a projeção de bilheteria a partir de dados históricos do elenco (atores, diretores, roteiristas, etc). Lash et al (2019) desenvolvendo um modelo de predição baseado não só no elenco, mas também no conteúdo da história e na data de lançamento. Lee et al. (2016) adicionam um elemento baseado na teoria de narrativa transmídia, o qual permite melhor interpretação e acurácia de predição. Os autores também propõem um método ensemble – uso de vários métodos de predição diferentes para melhorar a performance dessa predição – cuja performance se mostrou superior a estudos anteriores que usavam algoritmos de aprendizagem de máquina. Lu et al. (2019) trazem o método de análise conjunta para identificar quais os principais fatores que influenciam a bilheteria de filmes.

3 Discussão preliminar da metodologia

Para a análise proposta neste trabalho será usada uma base de dados contendo características de filmes e seus roteiros. Para a análise textual, na ausência dos roteiros

originais como documentos públicos de fácil acesso, a análise será feita através das legendas. Juntando essas informações, cria-se uma base de dados cujos textos serão divididos palavra por palavra (tokens), que serão classificadas conforme o sentimento que elas indicam; as palavras são então recompostas ao texto completo que será sub-amostrado em cem partes, as quais permitirão criar as curvaturas que indicarão, por meio de uma análise de cluster, dentro de qual arco narrativo aquele filme mais se assemelha. A metodologia usada é baseada em Del Vecchio et Al. (2021), sendo o enfoque no cenário brasileiro a principal diferença entre os estudos.

Após essa alocação em diferentes grupos, análises podem ser feitas para identificar como cada um desses grupos narrativos se comporta perante diferentes cenários de orçamento e gênero narrativo. Consequentemente, as produções subsequentes são classificadas dentro desses grupos e a bilheteria e a avaliação da crítica são calculados.

4 Resultados esperados

É esperado que filmes de diferentes arcos narrativos performem de forma diferente. Outras análises do cinema brasileiro sugerem que filmes de comédia tenham melhores resultados na bilheteria. Uma hipótese que se tem é de que a crítica e o público geral possuam percepções bastante diferentes, havendo, portanto, filmes aclamados pela crítica que não trazem resultados em bilheteria, sendo ineficientes financeiramente.

Também é objetivo verificar se o modelo de predição formulado seja bastante acurado a ponto de ser passível o seu uso por aqueles que determinam o orçamento de filmes, de forma que seus investimentos sejam mais seguros e com maior retorno.

5 Base de dados

Para esse estudo, a principal base de dados são as legendas dos filmes, obtidas via web scraping do site “ytssub.com”. Outros dados sobre os filmes foram obtidos nos sites “www.imdb.com” e “www.the-numbers.com”.

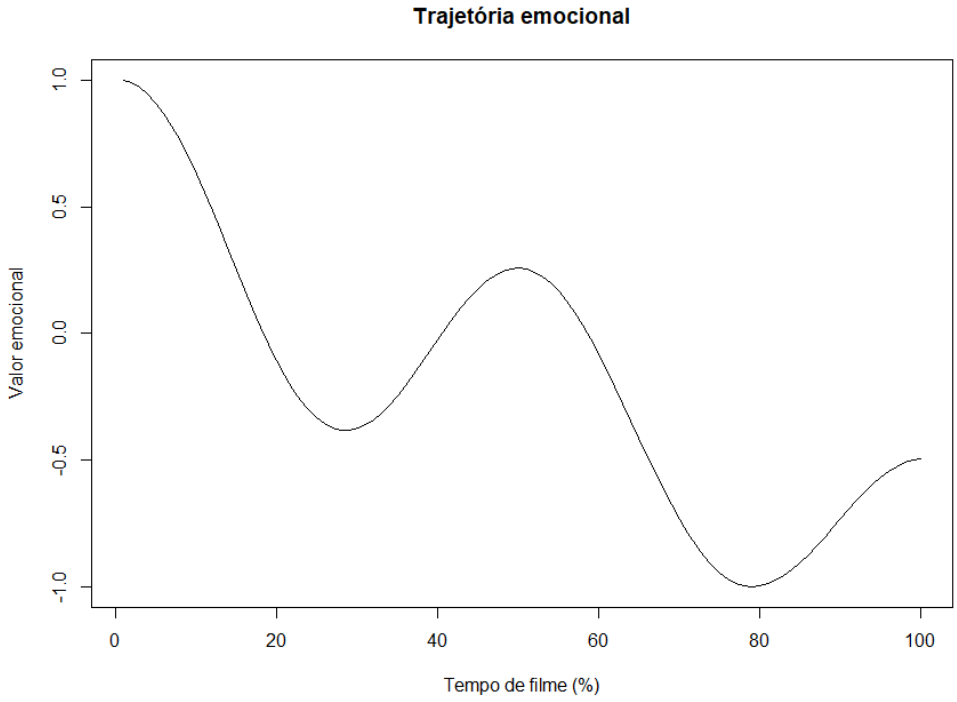
Na base do IMDB temos mais de 9000 filmes, havendo filmes desde 1902, como “Viagem à Lua” de Georges Méliès. Entretanto, ao combinar essa base com as legendas obtidas e os dados de receita, esse número de filmes reduz para um patamar de 3000.

Apesar de, inicialmente, a ideia fosse trabalhar exclusivamente com filmes brasileiros, a dificuldade de encontrar dados, seja via scraping seja via órgãos responsáveis, fez necessário o uso de uma base de dados que contenha filmes nacionais e internacionais.

6 Metodologia

Para analisarmos os dados dos roteiros, foi usada a biblioteca “syuzhet” na linguagem R. Para tal, o roteiro foi transformado em blocos de sentenças, as palavras comuns que não agregam muita informação (denominadas stopwords) foram eliminadas e foram aferidos valores conforme a emoção atribuída à determinada palavra, sendo palavras positivas, como “feliz”, identificadas com valores positivos e negativas, como “medo”, identificadas com valores negativos. O roteiro foi então subdividido em cem partes, cada qual tendo como valor de emoção, a média dos valores das palavras que fazem parte daquele subgrupo. Num roteiro de mil palavras, a média das dez primeiras representaria o primeiro ponto.

Gráfico 1 - Exemplo de trajetória emocional



Fonte: Elaborado pelo autor

Definidas as trajetórias narrativas de cada filme, o próximo passo foi clusterizá-los, de forma a deixar juntos aqueles filmes cujos arcos narrativos fossem mais similares.

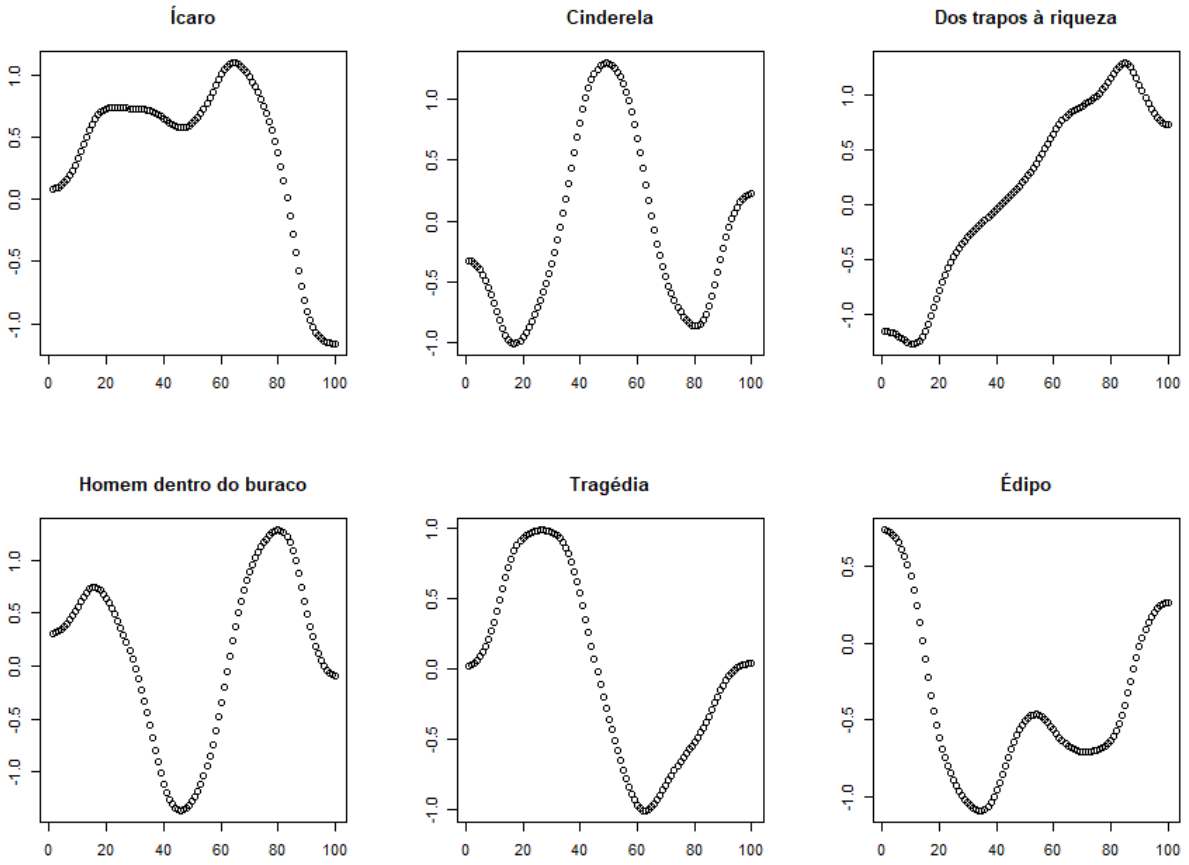
Com a definição dos clusters, foi possível analisar o impacto dessa divisão na acurácia de modelos preditivos.

7 Resultados

Após a clusterização em grupos de diferentes tamanhos (4, 6 e 8), a melhor divisão foi, como esperava a literatura, de seis clusters. Para esse número de clusters o modelo de predição de bilheteria tinha a melhor performance e estava mais em linha com a literatura apresentada anteriormente.

No gráfico 2, é possível ver seis linhas de arcos de narrativa, cada qual representa um dos seis clusters. Esses arcos foram nomeados conforme a literatura. É notável que, exceto para movimentações logo no começo ou bem para o final do filme, esses arcos se comportam de forma bastante condizente com a literatura.

Gráfico 2 - Arcos de narrativa para cada cluster



Fonte: Elaborado pelo autor

Com a divisão dos clusters, é possível analisar, para cada grupo, sua performance média em bilheteria, para identificarmos se há diferenças entre as médias de cada grupo. Para essa análise, foi feita uma ANOVA, seguida de um teste de Bonferroni, cujos resultados estão na tabela 1.

Tabela 1: P-Valores obtidos do teste de Bonferroni, para verificar quais clusters são diferentes entre si.

Cluster	Ícaro	Cinderela	Dos trapos à riqueza	Homem dentro do buraco	Tragédia	Édipo
Ícaro						
Cinderela	6,04%					
Dos trapos à riqueza	9,28%	7,02%				
Homem dentro do buraco	5,26%	5,67%	3,61%			
Tragédia	4,88%	6,68%	6,22%	5,18%		
Édipo	4,97%	4,42%	3,51%	6,54%	6,05%	

Podemos então afirmar que, a um nível de significância de 10%, todas as médias entre os grupos são estatisticamente diferentes. Mesmo para um nível mais rigoroso de 5%, ainda temos algumas diferenças de médias entre os grupos.

Sabendo que as médias são diferentes, usaremos o cluster a que o filme pertence como variável preditiva (dummy) para regressões lineares.

Tabela 2: Coeficientes da regressão

Cluster	Ícaro	Cinderela	Dos trapos à riqueza	Homem dentro do buraco	Tragédia	Édipo
Receita	-2,01	0,89	-3,16	5,8***	-3,38*	-2,27
Avaliação	-0,16	0,02	0,82	-1,4*	1,66*	-0,18

Ao observar a tabela 2, podemos perceber que temos poucos valores estatisticamente relevantes. Mas podemos inferir que filmes com arco narrativo Homem no buraco ganham, na média, 5,8 milhões de dólares a mais de receita. Já as tragédias ganham 3,38 milhões a menos. Ainda mais interessante, em linha com o conhecimento popular, filmes aclamados pela crítica não são tão bem vistos pelo público geral, impactando a bilheteria e vice-versa. Isso é notado quando vemos os resultados da regressão para esses dois arcos no quesito avaliação, os sinais dos coeficientes são invertidos. Tragédias ganham menos na bilheteria, mas tem melhores avaliações da crítica e Homem no buraco ganha bastante na bilheteria, mas nas avaliações sai perdendo.

Esses resultados preliminares já permitem traçar estratégias dentro da indústria cinematográfica. Seja com o objetivo de arrecadar muito dinheiro ou ganhar prêmios, a escolha do arco narrativo já se mostrou importante decisão.

8 Conclusão

A ideia principal que pode-se abstrair desse trabalho é que, mesmo na indústria criativa, dados e seu uso de forma inteligente permite ganhos antes impossíveis. Decisões baseadas em dados dentro da indústria do cinema são possíveis, é preciso desmistificar esse ramo, como algo incerto e sem lógica. Foi evidenciado que os milhares de filmes estudados, podem ser simplificados em seis grupos conforme seus arcos narrativos.

Dessa forma, como fechamento, temos as limitações, as conclusões e possíveis melhorias possíveis.

Como limitações, temos a falta de dados que pudessem traçar um cenário mais específico sobre a indústria cinematográfica brasileira. É provável que economias e culturas diferentes consumam diferentes tipos de conteúdos em intensidades diferentes, dessa forma, não ter muitos dados sobre o cenário nacional dificulta o processo de inferência dos resultados obtidos para a indústria nacional.

Mas mesmo assim, foi possível, de forma mais generalizada, estabelecer que é possível melhorar a eficiência da indústria cinematográfica com dados da pré-produção, no caso o roteiro, aumentando a expectativa de bilheteria ou de avaliação. Dessa forma, o risco de cada projeto diminui, abrindo portas para projetos mais arrojados que possam quebrar mais paradigmas e trazer novos ares à sétima arte.

Como melhorias e desenvolvimentos futuros, seria interessante buscar outros dados da pré-produção, como estúdio, atores, diretor, equipe técnica, etc. Essas variáveis podem ajudar a mensurar o sucesso esperado do filme e servir como base na determinação do orçamento do projeto.

9 Referências bibliográficas

Marco Del Vecchio, Alexander Kharlamov, Glenn Parry & Ganna Pogrebna (2021). Improving productivity in Hollywood with data science: Using emotional arcs of movies to drive product and service innovation in entertainment industries, Journal of the Operational Research Society, 72:5, 1110-1137, DOI: 10.1080/01605682.2019.1705194

Ahmed, U., Waqas, H., & Afzal, M. T. (2019). Pre-production box-office success quotient forecasting. Soft Computing. doi:10.1007/s00500-019-04303-w.

Allahbakhsh, M., & Ignjatovic, A. (2015). An iterative method for calculating robust rating scores. IEEE Transactions on Parallel and Distributed Systems, 26(2), 340–350. doi:10.1109/TPDS.2013.215

Hur, M., Kang, P., & Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and Independent subspace method. Information Sciences, 372, 608–624. doi:10.1016/j.ins.2016.08.027

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. Journal of Management Information Systems, 33(3), 874–903. doi:10.1080/07421222.2016.1243969

Lee, K., Park, J., Kim, I., & Choi, Y. (2018). Predicting movie success with machine learning techniques: Ways to improve accuracy. Information Systems Frontiers, 20(3), 577–588. doi:10.1007/s10796-016-9689-z

Lu, W., & Xing, R. (2019). Research on movie box office prediction model with conjoint analysis. International Journal of Information Systems and Supply Chain Management, 12(3), 72–84. doi:10.4018/IJISSCM.2019070104

Addis, M., & Holbrook, M. B. (2018). Is movie success a judgment device? When more is not better. Psychology & Marketing, 35(12), 881–890. doi:10.1002/mar.21141

Carrillat, F. A., Legoux, R., & Hadida, A. L. (2018). Debates and assumptions about motion picture performance: A meta-analysis. Journal of the Academy of Marketing Science, 46(2), 273–299. doi:10.1007/s11747-017-0561-6

Goff, B., Wilson, D., & Zimmer, D. (2017). Movies, mass consumers, and critics: Economics and politics of a two-sided market. Contemporary Economic Policy, 35(2), 269–277. doi:10.1111/coep.12180

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. Journal of Management Information Systems, 33(3), 874–903. doi:10.1080/07421222.2016.1243969

Chu, E., & Roy, D. (2017). Audio-visual sentiment analysis for learning emotional arcs in movies. 2017 IEEE International Conference on Data Mining (ICDM) Conference Paper, pp. 829–834. doi:10.1109/ICDM.2017.100

Das, B., & Chakraborty, S. (2018). An improved text sentiment classification model using TF-IDF and next word negation. arXiv preprint arXiv:1806.06407.

Hur, M., Kang, P., & Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and Independent subspace method. Information Sciences, 372, 608–624. doi:10.1016/j.ins.2016.08.027

Kim, Y., Kang, M., & Jeong, S. R. (2018). Text mining and sentiment analysis for predicting box office success. KSII Transactions on Internet and Information Systems, 12(8), 4090–4102.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science, 5(1), 31–43. doi:10.1140/epjds/s13688-016-0093-1

Gimenez, F. (2021). Política de fomento ao cinema: a questão do estímulo à regionalização da produção de filmes no Brasil. (Portuguese). Revista Eptic Online, [s. l.], v. 23, n. 2, p. 62–75.

Morelli-mendes, C., De almeida C. D. (2016) O desenvolvimento da economia criativa no Brasil: uma perspectiva através da indústria cinematográfica brasileira. (Portuguese). Verso e Reverso, [s. l.], v. 30, n. 75, p. 196–207.

Michel, R. C., Avellar, A. P. (2014) Indústria cinematográfica brasileira de 1995 a 2012: estrutura de mercado e políticas públicas. Nova Economia, [s. l.], v. 24, n. 3, p. 491–514.

Ruy, K. dos S. (2009) Estratégias e apostas da indústria cinematográfica brasileira contemporânea. (Portuguese). Sessões do Imaginário, [s. l.], v. 1, n. 22, p. 69–76.

Luiz de Figueiredo, J. (2019) O sistema produtivo da indústria do cinema brasileiro e sua dispersão concentrada. (Portuguese). Gestão e Desenvolvimento (18075436), [s. l.], v. 16, n. 2, p. 62–94.