Análise Comparativa dos Artigos de Filmes Indicados ao Oscar em Três Versões da Wikipédia

Cecília Junqueira V. M. Pereira
ceciliajunq@ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais

Marisa Affonso Vasconcelos
marisa.vasconcelos@gmail.com
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais

Ana Paula Couto da Silva
ana.coutosilva@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais

ABSTRACT

This work analyzes the differences in how Oscar-nominated films in the 'Best Picture' category are addressed in the English, Spanish and Portuguese versions on Wikipedia. Using text and graph analysis techniques, the connectivity and semantics of the articles in these languages were investigated, revealing that each language highlights different aspects of the films, reflecting the cultural and linguistic priorities of their respective communities.

KEYWORDS

Wikipédia, Oscar, grafo, diferença semântica, idiomas

1 INTRODUÇÃO

A Wikipédia é uma das principais fontes de informações atualmente, com artigos em 303 idiomas[1], o que a torna um repositório valioso de diferenças culturais entre diversas línguas. Diante dessa diversidade, há um crescente interesse em compreender como diferentes idiomas abordam temas específicos e em que elementos esses idiomas dão mais ênfase, a fim de captar possíveis indicações sobre as culturas das sociedades e os assuntos de maior interesse [5–7].

Tal análise é valiosa, pois revela a diversidade linguística e as tendências globais, mostrando o que é mais predominante em cada idioma. Além disso, permite identificar os vieses presentes nas diferentes versões da Wikipédia e como a disseminação da informação varia entre os idiomas, resultando em lacunas na representação do conhecimento. Em razão das diferentes ênfases culturais, certos temas podem ser mais explorados e publicados em algumas línguas do que outras. Dessa forma, entender quais Wikipédias são mais completas (abordam mais temas) e quais têm uma gama menor de assuntos abordados pode levar colaborações entre as diferentes versões da plataforma, visando homogenizar a cobertura dos tópicos e enriquecer a disseminação de informação nos vários idiomas.

Esse estudo visa entender quais tópicos são mais destacados pelos editores da Wikipédia em diferentes idiomas e como essas escolhas refletem variações culturais e linguísticas. Foram analisados artigos sobre filmes indicados ao Oscar de "Melhor Filme" em inglês, espanhol e português. A escolha desses idiomas se deve à escassez de estudos em espanhol e em português, em contraste com a abundância de análises em inglês. Os resultados mostram que, semanticamente, o espanhol enfatiza interações e comportamentos sociais, enquanto o português foca em detalhes de localização, e o inglês se concentra em aspectos temporais e de movimento. A versão em inglês tem uma rede de referências mais integrada, enquanto as versões em espanhol e português possuem comunidades mais definidas e menor densidade de conexões. Essas diferenças refletem diferenças culturais nas descrições e organização dos filmes.

[1] https://en.Wikipedia.org/wiki/Wikipedia

In: IV Concurso de Trabalhos de Iniciação Científica (CTIC 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTIC'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2596-1683

2 TRABALHOS RELACIONADOS

Diversos estudos têm investigado como diferentes idiomas abordam temas na Wikipédia. Um deles analisou os artigos em inglês, francês e russo, explorando quais tópicos se destacam em cada idioma e identificando padrões no comportamento dos leitores em relação à profundidade do conhecimento buscado [5]. Os resultados indicam que os tópicos em tendência estão fortemente relacionados ao entretenimento, com a cobertura midiática influenciando o aumento da busca por artigos sobre esses temas.

Outro estudo focou na descrição das Guerras Mundiais nas Wikipédias em inglês, alemão, francês e italiano [6]. Esse trabalho revelou que as edições tendem a enfatizar mais os aspectos das guerras que envolvem os países associados a cada idioma. Além disso, um estudo comparou as versões em inglês e português sobre as línguas Indígenas brasileiras [7], identificando diferenças significativas nas práticas editoriais, que embora distintas, alcançam níveis semelhantes de qualidade através de dinâmicas temporais variadas. As análises em [1] mostraram como a Wikipédia reflete o mercado musical global, focando nas diferenças na representação de gêneros musicais e artistas masculinos e femininos.

Diferente dos estudos anteriores, este trabalho explora como diferentes temas são abordados nas versões da Wikipédia em português e espanhol, uma análise não realizada anteriormente. A análise dos artigos coletados nos três idiomas de interesse fornece uma visão geral sobre quais assuntos são melhor representados em cada idioma, revelando lacunas de conhecimento e possíveis vieses.

3 METODOLOGIA

Para conduzir a análise proposta, foi adotada a seguinte abordagem:

Passo (1): Coleta dos títulos dos filmes. Os títulos dos filmes indicados à categoria "Melhor Filme" foram extraídos da página oficial da premiação[2], totalizando 587 filmes. As informações coletadas incluíram o título do filme, o ano da edição do Oscar, e a indicação de vitória. Como a página lista apenas os títulos em inglês, foi realizado o mapeamento para suas versões em português e espanhol para a coleta dos artigos correspondentes na Wikipédia.

Passo (2): Coleta dos artigos da Wikipédia. Utilizou-se a API da Wikipédia, especificamente, a função *Wikipedia.search* do pacote

[2] https://www.oscars.org/oscars/ceremonies

"Wikipedia" em Python[3], para buscar artigos de filmes. A função *Wikipedia.search* retorna uma lista de títulos correspondentes à consulta. Para identificar os títulos relevantes, utilizou-se o módulo "SequenceMatcher" do pacote Python "difflib", com um valor de similaridade superior a 0, 6. Caso a similaridade fosse menor, foi feita uma busca manual para cerca de 20% dos títulos em inglês. Após identificar os artigos em inglês, a API da Wikipédia foi novamente usada para obter as versões em espanhol e português[4], coletando os respectivos artigos.

Passo (3): Coleta dos artigos e metadados dos títulos. Após a normalização dos filmes e de suas páginas correspondentes no passo anterior, utilizou-se a função *Wikipedia.page* para coletar o conteúdo textual de cada artigo, além de links e referências a outras páginas. Após a coleta dos dados, o número de artigos de filmes encontrados para cada idioma foi: 587 para inglês e português, e 572 para espanhol. Assim, a cobertura da descrição desses filmes é de 100% para inglês e português e de 97, 4%. para o espanhol.

4 ANÁLISE DE CONECTIVIDADE

Nessa seção é analisada a conectividade entre os artigos de cada versão da Wikipédia, oferecendo uma visão de como as informações são referenciadas e revelando possíveis diferenças na priorização e organização das informações em diferentes culturas.

A análise considerou 572 filmes com artigos nas três versões da Wikipédia. Para cada artigo, foram extraídos links para outros artigos da Wikipédia na mesma versão, selecionando apenas aqueles dentro do dataset. As relações de referência entre os artigos dos filmes foram usadas para criar um grafo, onde cada filme é um vértice e as arestas direcionadas representam essas referências. Apenas os filmes que referenciam ou são referenciados pelo menos uma vez (ou seja, que têm grau diferente de zero) foram incluídos no grafo, correspondendo a 84, 15%, 44, 46% e 44, 97% do artigos em inglês, espanhol e português, respectivamente. Os grafos gerados podem ser visualizados na Figura 1. A construção dos grafos foi feita com a ferramenta *NetworkX*[5]. Para analisar a composição das comunidades, foi executado o algoritmo Louvain para identificar as comunidades de cada grafo e visualizamo-nas com a ferramenta *Gephi*[6]. A análise das comunidades não revelou relações semânticas ou de gênero entre os filmes. As relações de referências predominantes eram entre filmes com diretor e/ou atores em comum.

A Tabela 1 sumariza as principais métricas de cada grafo. A análise revela diferenças significativas na estrutura e conectividade entre os idiomas. O grafo em inglês tem maior densidade de conexões, refletida em métricas, como o número de vértices, arestas e grau médio, superando os grafos em espanhol e português. O inglês também tem o maior valor de *closeness* médio, indicando que os vértices estão mais próximos, e o maior *betweenness* médio, sugerindo que os artigos em inglês desempenham um papel mais central nas conexões. O grafo em espanhol mostra maior modularidade, com comunidades mais bem definidas, enquanto o inglês tem maior transitividade e coeficiente de agrupamento médio, sugerindo uma maior coesão local e tendência de agrupamento. Esses resultados mostram que a versão em inglês possui uma rede de referências mais integrada, enquanto as versões em espanhol e português têm estruturas menos conectadas e com comunidades mais bem definidas.

[3] https://pypi.org/project/Wikipedia [4] https://www.mediawiki.org/wiki/API:Langlinks
[5] https://networkx.org/ [6] https://gephi.org/

Tabela 1: Principais métricas dos grafos

Métricas	Inglês	Espanhol	Português
Número de vértices	494	261	264
Número de arestas	1.381	343	325
Diâmetro	12	11	12
Grau máximo de entrada	29	9	16
Grau máximo de saída	17	7	13
Grau médio	2,7955	1,3141	1,2310
Betweenness médio	0,0045	0,0003	0,0008
Closeness médio	0,1117	0,0099	0,0158
Densidade	0,0056	0,0050	0,0046
Transitividade	0,0906	0,0513	0,0317
Modularidade	0,5140	0,7742	0,7320
Coefic. de agrupamento médio	0,0671	0,0180	0,0270

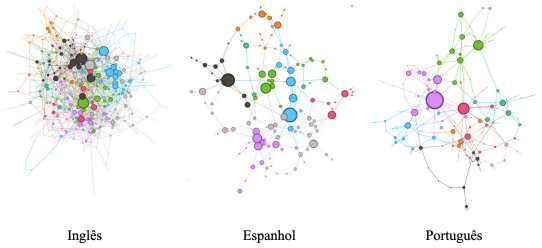


Figura 1: Grafo dos artigos em cada idioma, com nós coloridos segundo as comunidades definidas pelo algoritmo Louvain.

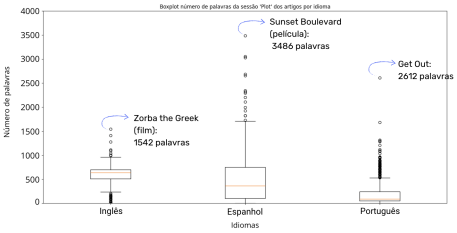


Figura 2: Variação nas descrições do enredo dos filmes.

5 ANÁLISE TEXTUAL ENTRE VERSÕES

Essa seção analisa como os enredos dos filmes são descritos em cada versão da Wikipédia, com o foco na seção "Enredo" (ou "Plot" em inglês e "Argumento" em espanhol). A comparação abrange 471 filmes que contêm essa seção nos três idiomas, permitindo uma análise consistente das variações culturais.

Primeiro, verificou-se se os textos em português e espanhol eram traduções literais do inglês. Não foi observada semelhança significativa, sugerindo que não são traduções diretas. Em seguida, foram comparados os tamanhos das descrições dos enredos em cada versão da Wikipédia. A Figura 2 apresenta boxplots que ilustram o

Figura 3: Nuvens de palavras dos enredos.

número de palavras em cada idioma. A mediana dos boxplots indica que metade dos textos em inglês é significativamente maior do que em português e espanhol. Isso pode ser explicado porque 76% dos editores da Wikipédia contribuem primeiro para a versão em inglês[7]. As descrições em português são, em média, mais curtas, enquanto os textos em espanhol mostram uma maior variação, incluindo *outliers*, como o filme "Sunset Boulevard".

A seguir, investigou-se a frequência das palavras mais usadas em cada idioma, gerando nuvens de palavras com o pacote *wordcloud*[8], conforme mostrado na Figura 3. Os textos dos 471 enredos foram concatenados, transformados em tokens e as *stopwords* foram removidas com a ferramenta NLTK[9]. As nuvens de palavras revelaram uma predominância de termos do âmbito familiar e palavras masculinas, como "família", "filho", "father" e "padre", indicando que muitos filmes indicados ao Oscar abordam temas da esfera familiar. Embora essa conclusão vá além do escopo do estudo, sugere-se que as relações entre os personagens são um foco importante para os editores da Wikipédia. Além disso, os textos em inglês contêm mais verbos, sugerindo maior ênfase no desenvolvimento e na dinâmica da história em comparação com os textos nos outros idiomas.

Também foram analisadas as frequências das funções sintáticas nos enredos em diferentes idiomas revelou que, embora a frequência é diferente nos três idiomas, os verbos são mais frequentes em inglês, enquanto substantivos e adjetivos predominam em português. Além disso, observou-se que os adjetivos são geralmente pouco utilizados, o que sugere que os editores da Wikipédia não priorizam a caracterização dos personagens.

5.1 Análise de Categorias Semânticas

A terceira análise utilizou a ferramenta *PyMUSAS*[10] para classificar termos em categorias semânticas e identificar padrões em tópicos e sentimentos na seção "Enredo" de cada idioma, excluindo categorias sintáticas e gramaticais. Foram identificadas 107 categorias nos textos analisados cujas frequências foram normalizadas e comparadas entre idiomas usando o teste de Kruskal-Wallis [4]. As diferenças significativas foram analisadas entre os pares de idiomas. A Figura 4 mostra essas diferenças nas frequências das categorias entre os idiomas comparados. Valores positivos indicam maior frequência no primeiro idioma mencionado na figura, enquanto que negativos indicam maior frequência no segundo idioma.

A análise comparativa das categorias semânticas entre português, inglês e espanhol revela diferenças significativas na ênfase de cada idioma. O português apresenta uma maior frequência em categorias relacionadas a localização e direção e atributos físicos, enquanto o inglês destaca-se em nomes de indivíduos e locais específicos, aspectos temporais e movimentos. Em comparação com o espanhol, o português também mostra mais frequência em localização e avaliação, enquanto o espanhol se foca em termos gerais e abstratos, além de ações sociais, refletindo uma abordagem mais abstrata e um interesse maior em contextos sociais. Entre o espanhol e inglês, o espanhol destaca-se em ações sociais, estados e processos, e termos gerais e abstratos, enquanto o inglês se concentra mais em nomes de pessoas, avaliação e processos de posse, indicando uma maior atenção a referências específicas e julgamentos de valor. Essas diferenças evidenciam como cada idioma evidencia a descrição dos filmes e refletindo características culturais distintas na comunicação e interpretação dos conteúdos.

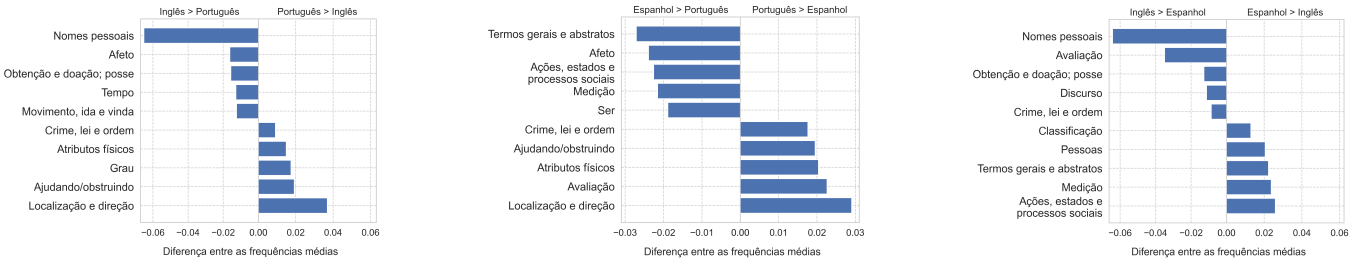
5.2 Análise dos Tópicos dos Filmes

Nessa seção, analisa-se a similaridade semântica entre grupos de filmes em cada idioma, utilizando *embeddings* que são representações vetoriais contínuas capturando o contexto e o significado das palavras. Com a ferramenta *BERT*[11], obtiveram-se os *embeddings* que refletem com precisão o contexto dos textos. A análise por gênero foi descartada devido à predominância do gênero Drama, que representa cerca de 40% dos filmes no Oscar, resultando em diferenças semânticas pouco claras entre os gêneros.

A partir dos *embeddings*, aplicou-se o algoritmo *K-Means* para identificar clusters de filmes semanticamente semelhantes em cada idioma. O número ideal de clusters foi determinado como 4 para inglês e espanhol, e 5 para português, com base no *Coeficiente Silhouette*. Para visualizar esses clusters, utilizou-se a técnica *t-SNE*, que reduz a dimensionalidade dos dados. Testaram-se diferentes valores para o hiperparâmetro *Perplexity*, encontrando-se que o valor ideal era 355, garantindo uma redução eficaz. A Figura 5 ilustra os clusters formados para cada idioma. Após obter os clusters, analisaram-se os filmes que compõem cada cluster e para atribuir uma descrição que resuma a abordagem comum das histórias dentro de cada conjunto (Tabela 3).

Foi feita também uma análise comparativa dos clusters de filmes entre os três idiomas, utilizando a métrica de Normalized Mutual Information Score (NMI), que avalia a similaridade entre duas clusterizações, variando de 0 a 1, onde valores mais altos indicam maior correspondência entre os clusters. A similaridade entre os clusters em inglês e português apresentou um NMI de 0.045, e entre espanhol e português, o NMI foi de 0.033, ambos indicando uma correspondência bastante baixa entre os agrupamentos desses idiomas. Em contraste, a similaridade entre os clusters em espanhol e inglês foi um pouco mais alta, com um NMI de 0.185, sugerindo uma correspondência moderada. Esses resultados sugerem que, enquanto há alguma sobreposição entre os padrões de clustering dos filmes em inglês e espanhol, os clusters nos diferentes idiomas tendem a ser bastante distintos, refletindo diferenças nas características culturais e contextuais dos filmes em cada língua.

[7] https://en.wikipedia.org/wiki/Wikipedia:Wikipedians
[8] https://pypi.org/project/wordcloud
[9] https://www.nltk.org/
[10] https://ucrel.github.io/pymusas/
[11] https://pypi.org/project/bertopic/0.3.1/



(a) Português vs. Inglês. **(b) Português vs. Espanhol.** **(c) Espanhol vs. Inglês.**

Figura 4: Diferenças semânticas nas categorias do PyMUSAS.

Tabela 3: Clusters por idioma.

Idioma	Cluster	# Filmes	Descrição	Cluster (EN)	Cluster (ES)	Cluster (PT)
Inglês	0	143	Decisões cruciais na vida dos personagens	-	3 (4, 76%)	3 (4, 76%)
	1	142	Mudanças drásticas na vida dos personagens	-	0 (9, 82%)	0 (6, 20%)
	2	89	Histórias de guerra e conflitos	-	1 (6, 79%)	0 (4, 97%)
	3	79	Dramas familiares e românticos	-	2 (2, 42%)	2 (3, 84%)
Espanhol	0	161	Enfrentamento de conflitos pessoais	1 (9, 82%)	-	0 (5, 82%)
	1	81	Relações de poder e influência	2 (6, 79%)	-	0 (4, 31%)
	2	127	Relacionamentos dos personagens	0 (7, 30%)	-	0 (4, 31%)
	3	93	Mudanças significativas no estilo de vida	1 (2, 10%)	-	3 (2, 45%)
Português	0	154	Papel social dos personagens na sociedade	1 (6, 20%)	0 (5, 82%)	-
	1	56	Mudanças na visão do personagem principal	0 (4, 56%)	2 (3, 35%)	-
	2	77	Conflitos e suas descrições detalhada	3 (3, 84%)	2 (3, 46%)	-
	3	111	Relações e interações entre personagens	1 (6, 09%)	0 (4, 86%)	-
	4	59	Biografias reais ou fictícias	0 (3, 96%)	0 (2, 72%)	-

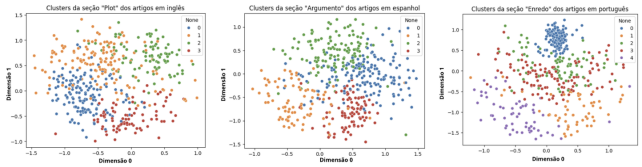


Figura 5: Clusters dos artigos de cada idioma visualizados com o método t-SNE.

Em seguida, foi analisada a interseção entre clusters de dois idiomas diferentes. O resultado dessas interseções é mostrado na Tabela 3 nas colunas "Cluster (EN)", "Cluster (ES)" e "Cluster (PT)", que indicam qual cluster de cada idioma é mais semelhante ao cluster da linha e a taxa de semelhança. A baixa porcentagem de semelhança entre os clusters dos idiomas reforça a diferença semântica na forma como as histórias dos filmes são abordadas nos três idiomas. Exemplos que ilustram essa diferença incluem o filme "Uma Mente Brilhante" que, na Wikipedia em português, foca nas relações dos personagens, enquanto, em inglês e espanhol destaca os conflitos enfrentados pelo personagem. Outro exemplo é "E o Vento Levou", que no inglês enfatiza às decisões cruciais dos personagens, no espanhol foca nos relacionamentos dos personagens, e no português no papel social dos personagens.

6 CONCLUSÃO E TRABALHOS FUTUROS

Este estudo analisou os artigos das versões em inglês, espanhol e português da Wikipédia sobre os indicados à categoria "Melhor Filme" do Oscar. Os resultados mostram diferenças significativas na forma como os filmes são retratados: a versão em inglês tem uma rede de referências mais integrada, enquanto as versões em espanhol e português mostram comunidades mais definidas e menor densidade de conexões. Essas variações semânticas e de conectividade refletem as distintas abordagens culturais e prioridades informacionais de cada comunidade linguística.

Durante o estudo, foram enfrentadas limitações como o BERTopic, que não forneceu resultados satisfatórios sobre os tópicos predominantes em cada artigo. Em trabalhos futuros, serão exploradas outras ferramentas de análise de tópicos, como o LDA[2] e o LaBSE [3], e avaliado diferentes públicos para aprofundar a compreensão das diferenças na percepção dos filmes entre falantes desses idiomas.

Agradecimentos: Trabalho financiado pela FAPEMIG e CNPq.

REFERÊNCIAS

[1] A. Pappu A. Wang and H. Cramer. 2021. Representation of Music Creators on Wikipedia, Differences in Gender and Genre. In *ICWSM*.
[2] David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.
[3] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. *ACL* 1 (2022), 878–891.
[4] W. Kruskal and W. Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621.
[5] V. Miz, J. Hanna, N. Aspert, B. Ricaud, and P. Vandergheynst. 2020. What is Trending on Wikipedia? Capturing Trends and Language Biases Across Wikpedia Editions. In *WebConf.*
[6] A. Smith and L. Lee. 2022. War and Pieces: Comparing Perspectives About World War I ans II Across Wikipedia Language Communities. *COLING* (2022).
[7] M. Vasconcelos, P. Mizukami, and C. Pinhanez. 2024. Disappearing without a Trace: Coverage, Community, Quality, and Temporal Dynamics of Wikipedia Articles on Endangered Brazilian Indigenous Languages. In *Proc. of ICWSM.*