Uma análise sistemática dos métodos avaliativos para classificação de filmes

Gabriel D'Luca Souza Viana[1], Vitor Antônio de Lima Silva[2]

Universidade Federal de Pernambuco – Centro de Informática
Caixa Postal 7.851 – 50.732-970 – Recife, PE – Brasil
gdsv@cin.ufpe.br, vals@cin.ufpe.br

Abstract

Este documento foi elaborado com o objetivo de agregar uma série de informações referentes à nossa pesquisa, mais especificamente sobre os mecanismos que sumarizam os principais escores avaliativos do Metacritic. Com o auxílio de diversas ferramentas disponíveis para ciência dos dados, buscamos compreender as principais variáveis que regem e exercem influência nos principais critérios de avaliação do site – a exemplo do Metascore e do escore global dos usuários –, assim como suas inter-relações, como uma forma de melhor compreender e delinear uma visualização do cenário contemporâneo de resenhas de filmes.

1. Introdução

Um dos mais tradicionais websites para agregação de críticas, o Metacritic sintetiza avaliações de publicações reputadas sobre um determinado filme, álbum, jogo eletrônico ou série de televisão. Desde 1999, o site é responsável por realizar uma agregação de notas e converter para uma escala de 0 à 100 pontos, reunindo a nota atribuída por cada crítico em uma única avaliação ponderada global, denominada Metascore. Este relatório tem como objetivo reunir as análises e estudos que conduzimos no Metacritic, a fim de verificar o cenário atual das resenhas dos filmes baseando-se em avaliações do The New York Times.

[1]Estudante de graduação do curso bacharelado em Sistemas de Informação
[2]Estudante de graduação do curso bacharelado em Sistemas de Informação

2. Método

Objetivamos conduzir análises no popular website para agregação de críticas, o Metacritic. Para a concretização de nosso estudo, utilizamos a linguagem de programação Python 3 ao decorrer de diversas etapas — desde a coleta até a análise. A partir desta, desenvolvemos *scripts* para realizar a coleta dos filmes avaliados pelo The New York Times conforme o site, integrando os dados obtidos com os dados de detalhes de cada filme. Em seguida, realizamos etapas de pré-processamento para que pudéssemos transformar os dados de forma que estes se tornassem compátiveis com as análises que pretendíamos realizar.

Isso feito, o próximo passo fora utilizar de diversas bibliotecas da linguagem de programação que possibilitassem a realização de nossas análises. Através da elaboração de visualizações, buscamos compreender os critérios que influenciam as diferentes avaliações do site — a exemplo do escore global dos críticos (denominado Metascore) e dos usuários — em um nível macro, mas sem perder de vista as interrelações e correspondências existentes a nível micro.

3. Coleta

Inicialmente, realizamos a análise do código-fonte das páginas que desejávamos extrair — mais especificamente, da página da publicação do The New York Times (`http://www.metacritic.com/publication/the-new-york-times`) e da página de detalhes de cada filmes (`http://www.metacritic.com/movie/call-me-by-your-name/details`). Isso feito, desenvolvemos *scripts* que automatizassem estas coletas. Para isso, utilizamos as bibliotecas *Requests* e *BeautifulSoup*, de Python, que foram essenciais para este processo. Modelamos o dados coletados no formato de um DataFrame, conforme utilizado pela biblioteca *Pandas*, e salvamos os conteúdos extraídos em arquivos no formato CSV.

4. Pré-processamento

Adiante, realizamos a criação de um *notebook* a partir da ferramenta *Jupyter Notebook*, utilizada para a criação de um documento que possibilitasse o acom-

panhamento de cada etapa realizada pela equipe. Inicialmente, realizamos a leitura dos arquivos CSVs gerados durante a coleta e partimos para a etapa de pré-processamento dos dados. Removemos dados que foram perdidos durante a extração das páginas, convertemos cada dado para seu tipo de atributo correspondente e, por conseguinte, discretizamos alguns dados em intervalos que corroborassem com a execução da análise. Como extraímos um grande conjunto de dados, a alternativa que escolhemos para o tratamento dos dados ausentes fora simplesmente ignorá-los durante a análise, visto que isso não impactaria tanto nossas observações. Por fim, realizamos a integração dos dados em um único arquivo CSV a ser utilizado para a análise.

5. Análise exploratória

Finalmente, partimos para a análise explorátoria. Iniciamos nossa pesquisa buscando responder as seguintes perguntas: "quão forte é a influência no Metascore por parte da nota atribuída pelos críticos do The New York Times?"; "dentre os críticos do The New York Times, quais atribuem um maior número de notas positivas, mistas e negativas em suas avaliações?".

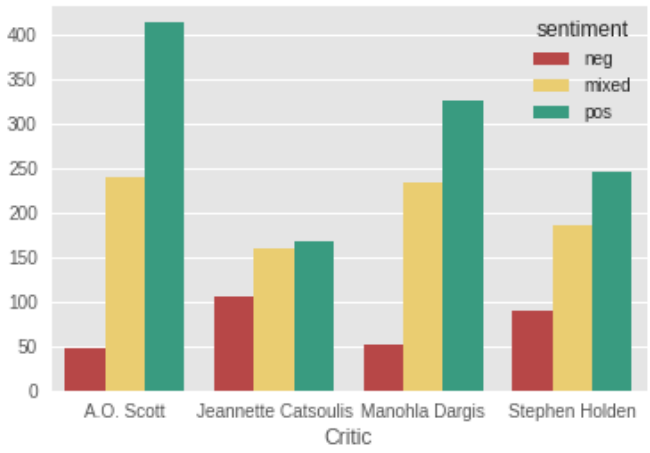

Figure 1: Distribuição das avaliações dos quatro principais críticos do The New York Times.

No geral, é possível identificar que nem todos os críticos possuem a mesma distribuição de notas internamente. Tratando-se do crítico A.O Scott, por exemplo, podemos verificar que há a atribução de uma quantidade substancial de avaliações positivas em relação aos demais. Similarmente, também é possível verificar que críticos como a Jeannette Catsoulis apresentam um maior balanceamento de notas, com menos da metade de avaliações positivas em relação ao A.O Scott e pouco mais do dobro de avaliações negativas em relação ao mesmo, conforme podemos verificar na Figura 1.

Ainda assim, é preciso levar em consideração, também, que uma influência negativa do Metascore não significa que esta é uma relação desfavorável — pelo contrário, se um crítico atribui mais notas negativas em suas avaliações, uma redução concomitante do Metascore significa que este ainda é ativamente influente.

	coef	SE	t	p-value
Intercept	56.250000	0.717279	78.421364	0.000000e+00
C(sentiment)[T.neg]	-18.122340	1.772475	-10.224313	5.120661e-24
C(sentiment)[T.pos]	18.175121	0.901523	20.160461	3.117044e-83
C(reviewer)[T.Jeannette Catsoulis]:C(sentiment)[mixed]	-4.797170	1.136256	-4.221911	2.518615e-05
C(reviewer)[T.Manohla Dargis]:C(sentiment)[mixed]	-1.750000	1.020868	-1.714228	8.662440e-02
C(reviewer)[T.Stephen Holden]:C(sentiment)[mixed]	-4.925676	1.087168	-4.530739	6.185020e-06
C(reviewer)[T.Jeannette Catsoulis]:C(sentiment)[neg]	-5.740867	1.947321	-2.948084	3.230393e-03
C(reviewer)[T.Manohla Dargis]:C(sentiment)[neg]	2.930033	2.236458	1.310122	1.902882e-01
C(reviewer)[T.Stephen Holden]:C(sentiment)[neg]	-2.949882	1.999788	-1.475098	1.403261e-01
C(reviewer)[T.Jeannette Catsoulis]:C(sentiment)[pos]	-5.107756	1.018646	-5.014258	5.739222e-07
C(reviewer)[T.Manohla Dargis]:C(sentiment)[pos]	-2.089736	0.823520	-2.537566	1.122955e-02
C(reviewer)[T.Stephen Holden]:C(sentiment)[pos]	-4.641447	0.895681	-5.182033	2.390149e-07

Table 1: Distribuição das avaliações dos quatro principais críticos do The New York Times por sentimento.

Em uma análise mais detalhada, é possível verificar a quantificação desta influência varia conforme o sentimento por crítico, conforme pode-se conferir na Tabela 1. Interpretam-se os resultados como a seguir: considerando as resenhas mistas do crítico A.O Scott como base, é possível ver que este costuma atribuir

notas mistas maiores em suas resenhas do que em comparação à outros críticos, enquanto Stephen Holden costuma atribuir notas mistas menores que os demais. Quanto às avaliações negativas, estima-se que Jeannette Catsoulis atribua as menores notas. Já entre o polo de reviews positivas, Jeannette atribui as mais baixas, enquanto A.O Scott introduz, novamente, notas superiores em relação aos demais.

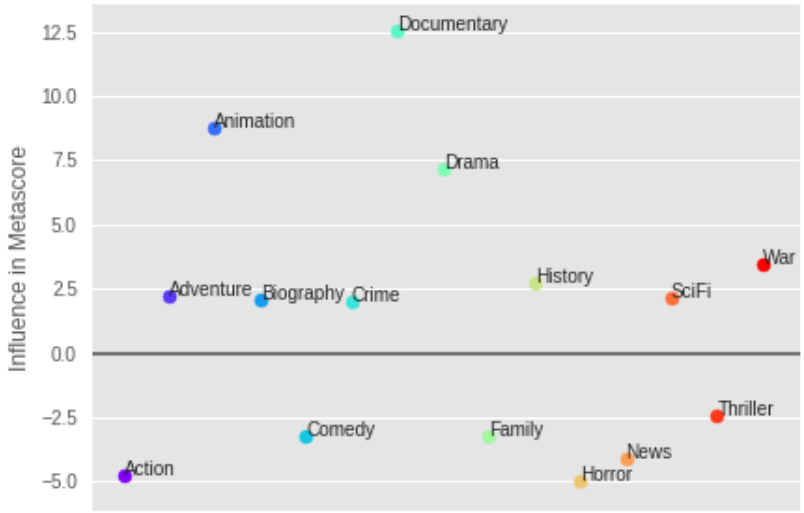

Figure 2: Influência da categorização por gênero no Metascore.

Seguimos para a próxima pergunta: "quais gêneros constumam ser atribuidos a maiores e menores índices de aclamação [dos críticos]?". Conforme a Figura 2, é possível verificar que, em uma escala de primeiro nível e estimando-se um valor padrão para o Metascore, os documentários, filmes de animação e de drama estão associados a Metascores maiores (e consequentemente, a maiores índices de aclamação no site), enquanto filmes de ação e filmes de terror se encontram associados a Metascores abaixo do que o habitual.

Por conseguinte, para possibilitar uma resposta mais concreta, elaboramos, ainda, *boxplots* para verificar a distribuição interna do Metascore e dos escores

atribuídos ao The New York Times nos cinco principais gêneros, conforme pode ser verificado nas Figuras 3 e 4, que se encontram dispostas na página logo abaixo.

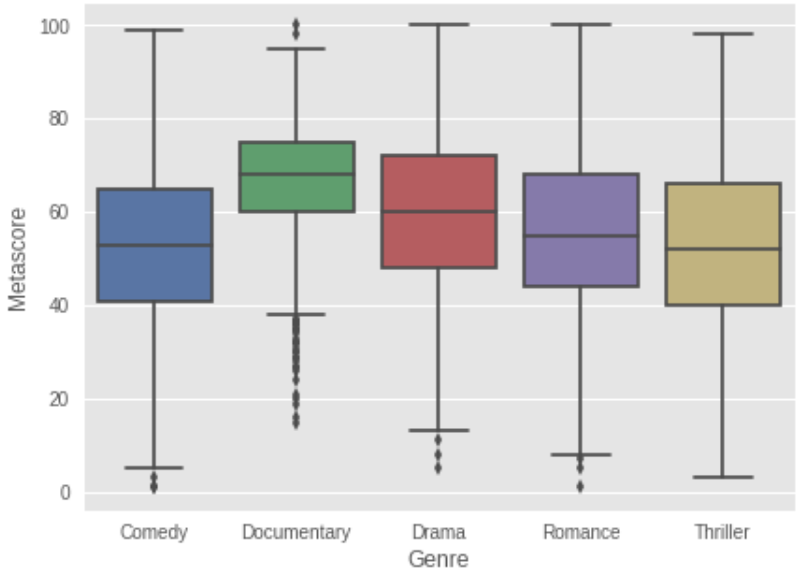

Figure 3: Distribuição do Metascore por gênero.

Partiremos, agora, para outra pergunta que buscamos responder: "o quanto o Metascore influencia no escore global dos usuários?". Para esta pergunta, também buscamos levar em consideração o fator das décadas de lançamento do filme, conforme podemos verificar na Figura 5, onde a imagem da esquerda corresponde aos anos 2000, e a da direita aos anos 2010.

É possível ver que a influência do Metascore sobre o escore global dos usuários aumentou um pouco em relação à última década, apesar de ambas se demonstrarem igualmente pertinentes de que esta influência realmente existe e é significativa, onde atualmente há um aumento de aproximadamente 7 pontos no escore dos usuários a cada 20 pontos adicionados no Metascore.

Para concluírmos, a última pergunta que buscamos responder foi: "existe alguma associação entre as demais características do filme e o Metascore?". Sumarizando o que encontramos em nossas análises, conforme podemos verificar

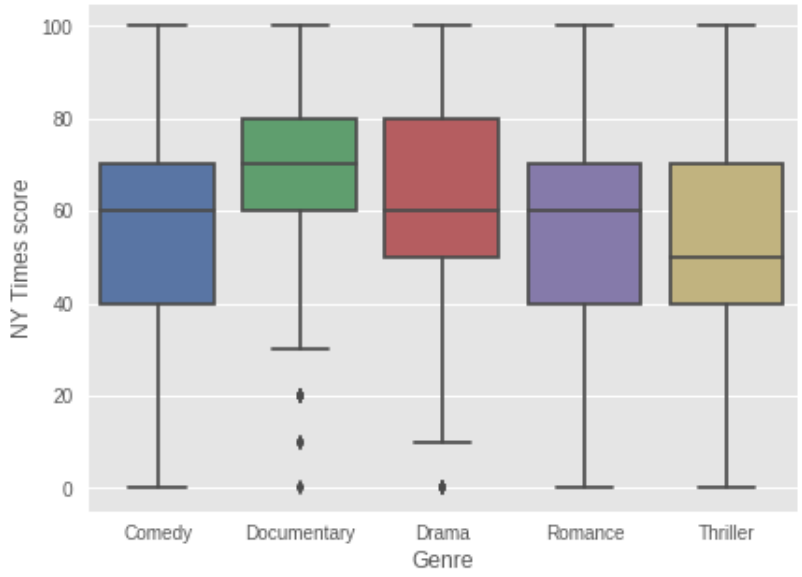

Figure 4: Distribuição das notas do The New York Times por gênero.

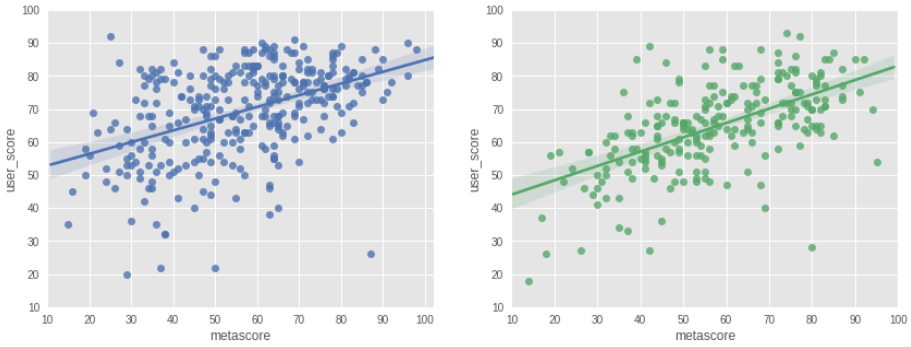

Figure 5: Influência do Metascore no escore global dos usuários.

na Figura 6 e considerando os filmes em inglês como padrão, é possível ver que os filmes em outros idiomas tendem a possuir escores superiores. Os filmes japoneses possuem a maior influência no Metascore com sua categorização, seguidos pelos filmes alemães, franceses, logo após, espanhóis. Assim sendo, é possível verificar que os idiomas também se apresentam fatores consideráveis e de influência.

Figure 6: Influência da categorização por idioma no Metascore.

Ao contrário do que esperávamos, conforme pode ser visto na Tabela 2, ao conduzirmos uma análise mais cuidadosa com os principais diretores, é possível ver que estes não possuem uma influência propriamente dita, visto que o p-value apresentado por estes é alto — o que indica que não podemos rejeitar a hipótese de que estes não são influentes. Além disso, é possível verificar que o alto p-value da estatística F, mostrado logo após a tabela, permite-nos concluir que, com a inclusão dos diretores como variavéis, não é possível rejeitar a hipótese nula de que não estamos melhorando a qualidade de nosso modelo.

Finalizaremos aqui a seção de análises como uma forma de manter o balaceamento de conteúdo neste relatório. Não obstante, foram conduzidas diversas análises adicionais visando analisar as inter-relações presentes nos métodos avaliativos do site e estas podem ser conferidas na íntegra nos *notebooks* disponibilizados no repositório do GitHub da nossa pesquisa (`https://github.com/if1015-datascience-ufpe/2017-2-projeto-metascience`).

	coef	SE	t	p-value
Intercept	63.00	4.437449	14.197347	3.092062e-12
C(director)[T.Spike Lee]	-4.80	6.275500	-0.764879	4.528539e-01
C(director)[T.Steven Soderbergh]	10.00	6.008334	1.664355	1.108931e-01
C(director)[T.Steven Spielberg]	13.75	6.656173	2.065752	5.141077e-02
C(director)[T.Woody Allen]	1.00	6.008334	0.166435	8.694063e-01

```
0.050416371540606957
```

Table 2: Influência dos diretores no Metascore.

6. Classificação de texto

Por fim, fizemos modelos de classificação de texto para tentar predizer o sentimento do autor de uma resenha a respeito de um filme, dado o seu texto. Esses classificadores são treinados associando uma considerável quantidade de palavras aos seus respectivos sentimentos. Ao fazer essa associação, os algoritmos podem analisar cada palavra de uma resenha e, usando o que aprendeu com os dados previamente passados para ele, atribuí-la a uma classe positiva, negativa e neutra.

A fim de otimizar os modelos, as resenhas passaram por uma série de tratamentos para maximizar a quantidade de informação extraída de cada uma das palavras no seu texto. Para isso, utilizamos um tokenizador inteligente da biblioteca NLTK (do inglês, *Natural Language Toolkit*). Após a tokenização, cada uma das palavras é convertida para letras minúsculas e passa pelo processo de lematização, que consiste em deflexionar as palavras, buscando seu lema. Em seguida, como forma de reforço, as palavras são tratadas por *stemming* para minimizar a quantidade de palavras não lematizadas corretamente. Por fim, as palavras recebem suas respectivas classes e estão prontas para serem passadas para os classificadores.

Utilizamos 70% (cerca de 8400) dos dados como conjunto de treino, enquanto os outros 30% (cerca de 4000) foram utilizados para testar a acurácia de cada

um dos classificadores. A partir disso, foram treinados 7 modelos de classificação de texto, entre eles:

- Naive Bayes Classifier
- Multinomial Naive Bayes Classifier
- Linear Regression Classifier
- Stochastic Gradient Descent Classifier
- Support Vector Classifier
- Linear Support Vector Classifier
- Nu Support Vector Classifier

Ao verificar a acurácia média dos modelos apresentados acima, obtivemos uma média de apenas 55.30%. Verificando a divergência entre os resultados dos classificadores, decidimos criar um novo classificador: o classificador por voto. Este funcionaria como um detector de moda entre os resultados dos outros classificadores, classificando a resenha com a classe mais comum entre eles. A acurácia desse classificador por voto também ficou próximo da média, sendo, geralmente, o segundo melhor classificador, perdendo apenas para o Multinomial Naive Bayes, que mostrou ser o melhor em todos os casos e acertando cerca de 60% das predições. Em seguida, foram feitas investigações com três diferentes conjuntos de testes a fim de encontrar *bias* nos classificadores, o primeiro conjunto foi composto por resenhas exclusivamente negativas, enquanto o segundo por resenhas exclusivamente positivas e o terceiro, neutras.

Descobrimos, a partir disso, que o Support Vector Classifier classificava todas as resenhas como positivas, pois apresentou acurácia de 100% para o conjunto de teste exclusivamente positivos e 0% para os conjuntos neutros e negativos, o que nos levou a descartá-lo. Não obstante, ao fazermos algumas predições com frases contendo palavras significativas, todos os modelos se mostraram bastante razoáveis, classificando como positiva a frase “I love this movie!”, enquanto a

frase “Even though I think it’s average, my wife loves it!” fora classificada como positiva ou neutra e, por fim, a frase “What a terrible movie, I was bored the whole time!”, como negativa.

7. Considerações finais

A possibilidade da realização de nossas análises nos permitiu uma melhor compreensão dos critérios de influência entre os diversos classificadores avaliativos do Metacritic. Por sua vez, a elaboração de *notebooks* e utilização de diversas bibliotecas que conduzissem nosso estudo demonstraram a utilidade e a importância do domínio destas ferramentas no campo de ciência dos dados, bem como a do conhecimento da linguagem de programação Python.

Por conseguinte, as diversas visualizações mostradas ao decorrer deste relatório corroboram com o processo de delineamento da influência de diversos fatores nos escores. A partir das análises realizadas, fora possível identificar que essas avaliações, apesar de se apresentarem correlatas em um primeiro momento, divergem internamente conforme a consideração de atributos diversos. Tivemos a oportunidade de ter hipóteses atestadas, contestadas, e visualizar o surgimento de novas ao decorrer de nossas análises.

Uma vez que conseguimos realizar um mapeamento geral do panorama dos critérios avaliativos do site, há a possibilidade de condução de estudos posteriores que busquem priorizar a estimação destes, utilizando os diversos critérios de influência que encontramos para o cálculo de um peso único para a publicação — de forma similar ao site (`http://www.metacritic.com/about-metascores`). Considerando uma equipe maior e um período de médio a longo prazo, poderiam ser consideradas, ainda, as demais publicações neste processo, buscando estimar e realizar normalizações para cada escore de modo a tornar as predições mais assertivas, mas sem perder de vista a ótica inter-relacional e sistemática necessária para sua concretização.