# Text Analysis and Spatial Data for Economists

Joy Albertini

June 2024

# 1 Project

This project contains several analysis of films, several Jupiter notebooks And working UI, please read the **README.MD**

# 2 Research question

My research question investigates whether a film's rating is influenced by political reviews. Nowadays, it seems politics significantly impacts film ratings. For instance, platforms like Rotten Tomatoes have adopted methods to mitigate this by displaying ratings from only verified users (this approach also helps reduce the influence of bot-generated reviews). I aim to develop a similar tool for IMDb. This tool would scan all text reviews of a film and classify them as neutral, left-leaning, or right-leaning. After classifying the reviews, it would compute a new rating for the film based on user settings (for example, considering only neutral reviews). The goal of this tool is to provide a rating that more accurately reflects the film's content, free from bias that could significantly alter the film's overall rating. If we excluded the more politically-oriented reviews, the rating might differ, which could provide an answer to my research question.

# 3 Scientific Contribution

My project, which examines the impact of politically-oriented reviews on film ratings, could contributes to understanding how subjective biases influence different public platforms. The subject of my project parallels the research discussed in "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards" by Werner Antweiler and Murray Z. Frank and "Measuring Economic Policy Uncertainty" by Scott R. Baker, Nicholas Bloom, and Steven J. Davis. Both studies delve into how biases can shape perceptions in public forums and economic indicators. By developing a tool to classify and filter reviews based on their political orientation, I aim to provide a clearer, more unbiased view of film ratings, similar to how these studies address biases in their respective fields.

# 4 Data scraping

I was unable to find a dataset suitable for my project, as all available datasets containing IMDb user reviews lacked links to the corresponding films. Furthermore, I needed the original IMDb ratings for my analysis. Consequently, I decided to scrape the necessary data myself. I extended the PyMovieDb GitHub Repository to enable scraping of user reviews from IMDb. This task proved challenging, particularly because the IMDb page for user reviews (see example reviews) loads them dynamically at the click of a 'load more' button. Therefore, I employed Selenium, which uses the Chrome browser to automate the process of repeatedly clicking this button until all reviews are displayed. Once fully loaded, I scraped all the data in one go. The collected data is stored in a folder within the project named `Data/{Film ID}`.

# 5 Review Classification

## 5.1 Train datasets

Finding a dataset suitable for training spaCy was challenging, as most available datasets contained politicized texts labeled simply as `left` or `right`. Moreover, these texts were often significantly different from product reviews. The best dataset I found that maintained consistent labeling in the training set is the JyotiNayak political ideologies dataset. However, this too was limited to classifications of just left or right. To enhance my model, I created my own datasets using ChatGPT. I requested ChatGPT to generate a dataset of politicized film reviews for both left and right ideologies, as well as a neutral dataset focusing mainly on the film's qualities, whether positive or negative. The datasets created by ChatGPT were decent, but they contained many very similar sentences, which could potentially degrade the model's performance. I have generated more 3000 labelled entries using ChatGPT. I combined these two datasets to train spaCy more effectively.

# 6 Model

## 6.1 Model training

To train the model, I used spaCy 3.7, employing the `textcat` pipeline with exclusive classes, as each text must be classified into only one label. Specifically, I increased the `ngram_size` to 3 to consider a wider range of n-grams, which significantly improved the results. Additionally, I incorporated `en_core_web_lg` into the pipeline as a pre-training step to represent the text as vectors. This enhancement enables the model to better understand the relationships between words, which is later utilized by `textcat`. It was also essential to ensure rapid predictions due to the large volume of documents I have.

## 6.2 Review classification accuracy

Let's consider three film examples. Figures 3, 4, and 5 present some statistics. The first bar plot, `political counts`, represents the number of reviews per political affiliation. The second bar plot displays the quantity of extreme votes; films that are down-voted or promoted by user campaigns are likely to receive ratings of 1s and 10s. The box plot titled `rating by politics` shows the distribution of votes according to political affiliation. The last box plot illustrates the distribution of confidence in the categorization of the document.

**Considered good Non-Political** How to Train Your Dragon is a family-oriented film featured in IMDb's top 250 list of all time, and is probably free political content. As shown in Figure 3, most reviews are classified as neutral. The box plot `rating by politics` illustrates that the distribution of ratings across the three political categories is fairly consistent, though ratings from right-leaning viewers are slightly lower. According to the bar chart on political outliers, there are very few politically motivated ratings of 1.

| Type | Review rating | IMDB rating | Weighted average Rating | Mean Rating |
|---|---|---|---|---|
| N | 8.66 | | | 8.38 |
| N + L + R | 8.53 | 8.1 | 8.1 | 8.31 |
| All | 8.56 | | | 8.33 |
| Docs: | 547 | 804267 | | |

Symbols explained in section **Rating Comparison**. **Review Rating** is derived from the written reviews I analyzed, while **IMDB Rating** reflects the aggregate rating of all reviews, including those without textual content from IMDb. The **Weighted Average Rating** is calculated using the formula: Weighted Average Rating $= \frac{(\text{Review Rating} \times |\text{reviews}|) + (\text{IMDB Rating} \times |\text{IMDB reviews}|)}{|\text{reviews}| + |\text{IMDB reviews}|}$

**Considered bad Non-Political** Jaws: The Revenge is widely considered one of the worst movies and is not associated with any political discourse. As illustrated in Figure 4, the majority of reviews are neutral; the distribution of votes between neutral and right-leaning reviews is identical, while left-leaning reviews tend to be slightly more biased towards lower ratings. However, all ratings tend towards the lower end of the scale. Additionally, the analysis of outliers reveals no significant skew in the data.

| Type | Review rating | IMDB rating | Weighted average Rating | Mean Rating |
|---|---|---|---|---|
| N | 3.30 | | | 3.13 |
| N + L + R | 3.39 | 3 | 3 | 3.19 |
| All | 3.22 | | | 3.11 |
| Docs: | 258 | 49796 | | |

From the two examples above, we can conclude that the model does not associate negativity or positivity with left or right-leaning biases, as both have distributions of ratings similar to those of neutral reviews, this will change when we consider the next example.

**Considered Political** The first purge is described with tags such as social injustice, abuse of power, political oppression, class conflict, and social commentary TAG the First Purge clearly exhibits its political nature. As shown in Figure 5, the number of left-leaning and right-leaning reviews is significantly higher compared to the previous two examples. The distributions of left and right-leaning ratings differ notably from the neutral ones, with their medians being substantially lower. Additionally, the bar plot for outliers

shows a larger presence of politicized ratings at 1.0. From the latest analyses, we can conclude that the

| Type | Review rating | IMDB rating | Weighted average Rating | Mean Rating |
|---|---|---|---|---|
| N | 5.22 | | | 5.21 |
| N + L + R | 5.18 | 5.2 | 5.2 | 5.19 |
| All | 4.70 | | | 4.95 |
| Docs: | 458 | 71410 | | |

model is capable of discerning political reviews from non-political ones in general. However, it struggles to accurately distinguish between left-leaning and right-leaning reviews. This is further supported by the box plot regarding confidence levels, as shown in all three examples; we observe significantly higher confidence in neutral reviews compared to those identified as left or right.

# 7 Rating comparison

Now that we have developed a model capable of distinguishing between non-political and political reviews, we can utilize it to compute different ratings based on the data sets we select. We have primarily chosen three sets of data:

1. $N$ **(Neutral Reviews)**: This set contains only neutral reviews.
2. $N + L + R$ **(Neutral + Left + Right)**: This set includes neutral reviews as well as those identified as left-leaning or right-leaning, excluding extreme ratings of 1 and 10.
3. $ALL = N + L + R + L1 + L10 + R1 + R10$ **(Complete Set with Extremes)**: This set comprises all reviews, including those with extreme ratings of 1 and 10.

The tables shwon in the examples displays various ratings for each film, along with the IMDB rating, and a combination of the two: the Weighted Average Rating and the Mean Rating.

## 7.1 Considered non political

The list of film to compute such distribution is written in the `rating_analysis.ipynb`.
To evaluate whether film ratings are influenced by political reviews, I will compute the differences between the review ratings and the actual IMDB ratings, and plot the distributions. Figure 1 demonstrates that, generally, the differences in ratings are not as significant as one might expect compared to the actual IMDB ratings composed of thousand of ratings, which is surprising. The red dashed line is the mean of all the sets ($N$, $N + L + R$, $All$)

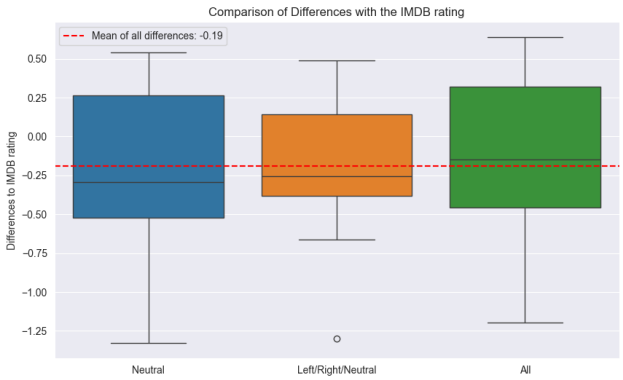


Figure 1: Distribution difference review rating and IMDB rating

## 7.2 Considered political

The list of films used to compute this distribution is specified in `rating_analysis.ipynb`. Comparing Figure 2 with Figure 1, it is evident that the political distribution of differences is more sparse than the non-political one. Interestingly, we observe larger outlier values; in Figure 1, the range of differences extends from [0.50 to -1.25], whereas in Figure 2, the outlier values are more contained within [0.8, -0.8]. Based on this distribution, we can affirm that the answer to my initial question holds true, albeit with minor significance.

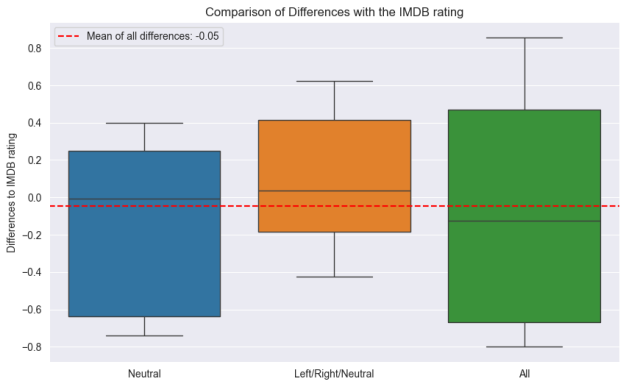


Figure 2: Distribution difference review rating and IMDB rating

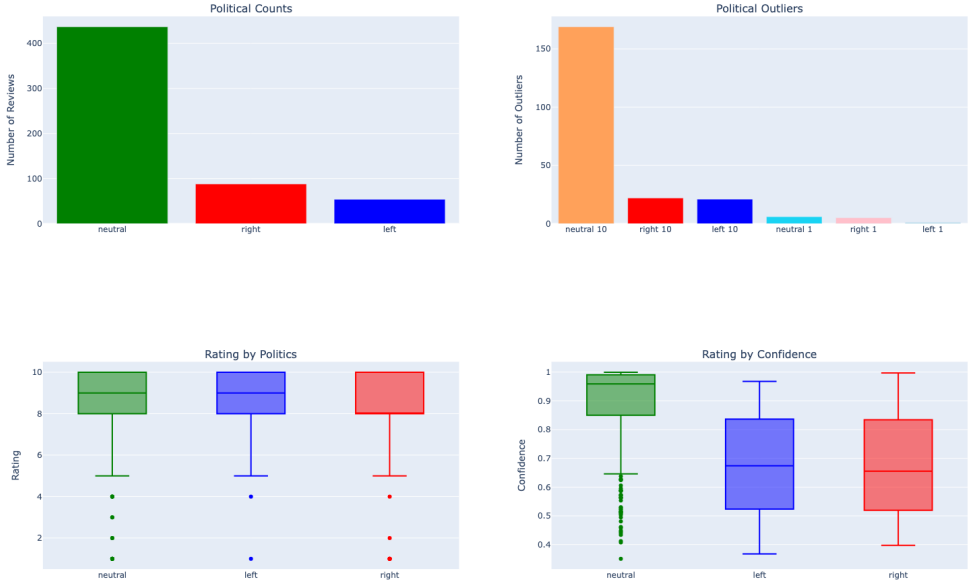


Figure 3: How to train your dragon stats

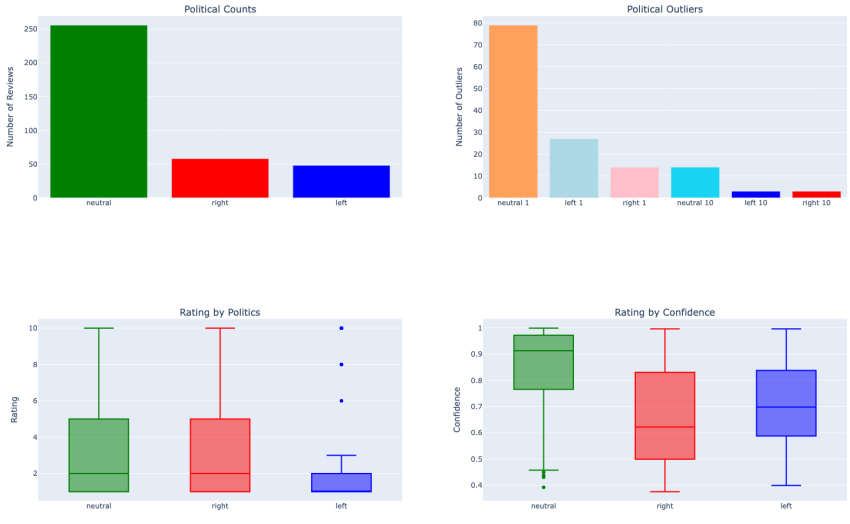


Figure 4: Jaws the revenge stats

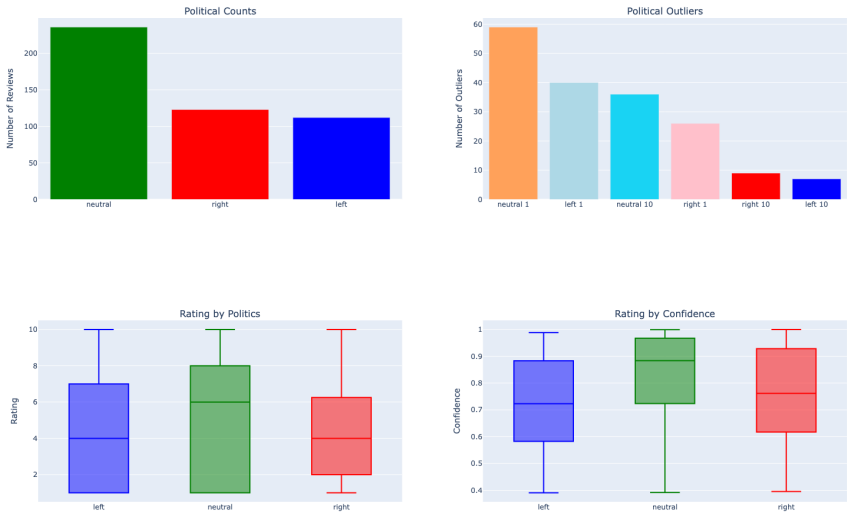


Figure 5: The first purge stats

# 8 Answer to the research question

As previously discussed, it appears that the average rating does not significantly deviate from the IMDb rating, despite the latter being composed of thousands more reviews. Notably, films with a political focus show a wider distribution of rating differences compared to non-political ones, suggesting that politically charged reviews slightly alter ratings more than other reviews. Surprisingly, the ratings show minimal differences from the IMDb ratings, with the maximum discrepancy being only -1.25. This leads me to speculate that IMDb might already employ methods to ensure more neutral reviews on their site, although this is purely speculative. The answer to my initial question—whether a film's rating is influenced by political reviews—would be affirmative. However, the dataset is too small for statistical significance; a more robust analysis would require examining at least 100 films undergoing the same evaluative procedures.