COMP370 Final Project - Movie Release

Written by:
Jacob Lerner, Shuaishuai Jiang
Jaewon Moon, Rohomutally Zafeerah

Abstract

Within the media, the announcement or expectation of a new film initiates a wave of news stories. It is crucial for a media firm to comprehend the main ideas of these pieces and to compare the quantity of attention a movie receives to that of its peers. Our team looked into how the movie "Oppenheimer" was covered in the news with this objective in mind. This study aims to identify the primary themes found in news reports about the latter and a few other carefully chosen films that were released at the same time. At the same time, we wanted to determine how visible "Oppenheimer" was in relation to the other films.

1 Introduction

Media and movies share a close relationship, where the announcement or release of a movie starts a conversation on all media platforms, whether it be through debates, reviews, or critiques. Understanding news article themes and emphasizing a movie's relevance is crucial for media organizations at the intersection of media and film in order to engage consumers and capitalize on changing entertainment trends.

In this context, our team delved into the exploration of news coverage surrounding the movie 'Oppenheimer'. Our analysis, based on a dataset of over 500 articles from NewsAPI.org in English, focused on movies released concurrently. To ensure relevance, articles were filtered using a selected set of keywords. Employing an open coding methodology, we annotated each article into one of seven distinct topics, offering a nuanced exploration of Oppenheimer's narrative threads. Leveraging tf-idf scores, we identified the ten most relevant words in each category. Ultimately, our analysis aims to understand the distinctive features and characteristics of each identified theme, offering key insights into how the media represents this movie.

In our findings, the top three article topics concerning 'Oppenheimer' emerged as 'Actor/Director,' 'Release and Promotion,' and 'Economical Aspects.' Notably, 9.50% of the total articles were dedicated to the movie, acknowledging and addressing potential time and geographical bias.

2 Datasets

For our analysis, we conducted an API call to NewsAPI.org, obtaining 504 English-language articles related to 'Oppenheimer' and 21 other movies released in proximity to the latter's debut, including notable titles like 'Barbie', 'Blue Beetle', and 'The Nun II'. To ensure the relevance of the gathered articles to our focus on cinematic narratives, we applied filters on the API call, using movie-related keywords such as 'Film', ='Actor', 'Plot', 'Scene', 'Review', 'Genre', 'Character', and 'Audience'. In structuring our dataset for analysis, we identified the following key columns and information as pertinent:

- Movie: the film discussed in the article
- Title: the title of the article
- Description: overview of the article's content
- Article Date: the release date of the article
- URL: the link to the article

The data, initially retrieved in JSON format, was processed and organized into an Excel format to facilitate efficient data annotation and enable swift filtering or processing for an initial exploration of the dataset.

It is important to note that our use of NewsAPI.org came with a limitation, restricting our access to news articles from no earlier than a month prior. In Figure 1, we have higher peaks in media coverage for movies such as "Five Nights at Freddy's" and "The Holdovers," both of which were released in November, aligning with the timeframe of our data collection. This indicates a time bias in our dataset, with higher article counts for movies released around the period when we retrieved our data.

As part of our data quality control, we removed duplicates and examined the dataset for any missing or invalid entries. This ensures the reliability of our dataset, providing a solid foundation for our subsequent analysis.

In the final stage, we performed data annotation, categorizing each article into 7 distinct topics to fulfill the requirements of our analysis. Further details on the annotation process can be found in Section 3.

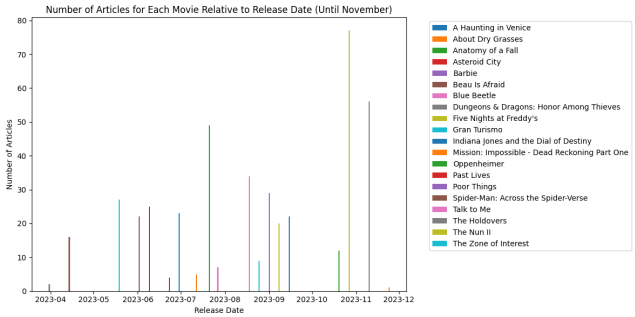


Figure 1: Proportion of articles in dataset based on Movie Release Date

3 Methods

3.1 Movie Selection

To ensure a fair representation of contemporaries with ”Oppenheimer”, we curated a list of 22 movies released around the same period as its debut. These selections were made based on their comparable revenue to ”Oppenheimer”, presuming that the extent of media coverage often aligns with a movie's popularity, reflected in its earnings.

With the constraints of data retrieval, we faced limitations in obtaining information for all 22 movies. Consequently, we strategically selected two specific movies, namely ”The Color Purple” and ”Poor Things” that only premiered within our selected time frame, but not playing in theatres yet. These movies were added to our dataset due to the abundance of articles associated with them, all of which were highly relevant to our chosen topics.

For clarity and uniformity in our dataset, we opted to filter content based on the English language parameter during data retrieval. This decision ensures a consistent language throughout, enhancing data comprehensibility for our analysis.

3.2 Dataset Refinement

After retrieving the data and removing duplicates, we discovered a considerable number of articles unrelated to the movies, especially for movies such as “Oppenheimer”, “Talk to Me”, and “Poor Things”, as these titles are frequently used outside the context of movies. To address this issue, we refined the movie list to include only those containing specific film-industry-related keywords, such as ”Film”, ”Plot”, ”Scene”, etc. This step aimed not only to refine the dataset but also to eliminate any non-movie-related content and enhance the precision of our subsequent analyses.

3.3 Open Coding

Next, we selected a random subset of 200 articles from the initial 504 retrieved articles for the initial open coding process. During this stage, we identified 10 distinct categories based on the article titles and descriptions. We then underwent multiple iterations to further refine these categories, consolidating them into 7 overarching categories: Production, Actor/Director, Storytelling, Release and Promotion, Economic Aspects, Comparison, and Review. Subsequently, we systematically annotated the remaining articles based on these finalized 7 topics. This meticulous refinement process aimed at constructing an ontology ensuring comprehensive coverage for our analysis.

4 Results

4.1 Topic Selection

When selecting topics we had three main thoughts in mind

(1) Each article can fit into at least one topic
(2) Topics are objective and clear
(3) Not to have a garbage collector topic such as other

With that in mind decided on the topics

- Production
- Actor/Director
- Economic Aspects
- Story Telling
- Release and Promotion
- Review
- Comparison

Production - This is relevant to any part of the article that talks about choices on why certain design aspects were chosen, moving forward with certain spin-offs and new projects related to the first movie.

Actor/Director - Relating to any articles that mention Actor/Director statements, opinions, comparisons, and next projects that will include said Actor/Director.

Economic Aspects - This is defined as any article that speaks about the money aspects of the movie. For example, Box-office, investments, companies' earnings being impacted because of the movie, cumulative sales vs expectations.

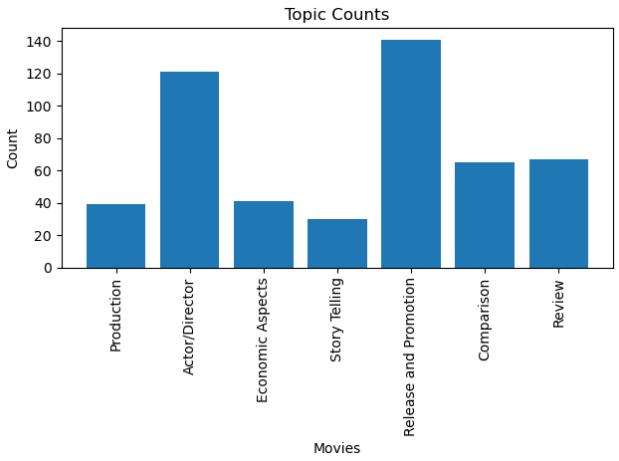
Story Telling - A simpler topic which is defined as articles mentioning plots or spoilers.

Release and Promotion - Which theatres we can watch the movie, any dates regarding streaming and when out of theatres, and relevant facts about promotion leading up to the movie.

Review - Articles that review either aspect of the movie or the movie itself.

Comparison - This topic includes comparing two or movies in any aspect, Examples include awards, plots, top 10 lists.

4.2 Total Engagement



The top 3 topics were Release and Promotion, Actor/Director and Review. The most popular topic was Release and Promotion which accounted for 27.9% of total articles (141 articles). Actor/Director accounted for 24% and Review totaled 13.3%

4.3 Oppenheimer Engagement

higher TF-IDF value indicates how important the term is within a document relative to a collection of documents. We found the top 10 words with the highest TF-IDF scores for each topic.

- Production
  The top 10 words with the highest TF-IDF scores for the ”production” category are: ”causeway,” ”japan,” ”viral,” ”fireworks,” ”negret,” ”defends,” ”bombings,” ”daniel,” ”gear,” and ”site.”
- Actor/Director
  For the ”actor/director” category, the highest TF-IDF scores are associated with words such as ”willem,” ”dafoe,” ”oprah,” ”steven,” ”winfrey,” ”sagaftra,” ”aster,” ”favorite,” ”payne,” and ”van.”
- Economic Aspects
  In the ”economic aspects” category, the significant terms include ”revenue,” ”quarter,” ”amc,” ”blockbusters,” ”cinemark,” ”prepandemic,” ”levels,” ”imax,” ”record,” and ”specialty.”
- Story Telling
  Words with the highest TF-IDF scores in the ”storytelling” category comprise ”priest,” ”spreading,” ”irene,” ”facetoface,” ”valak,” ”thrillerimdb,” ”imdb,” ”games,” ”genre,” and ”unbelievable.”
- Release and Promotion
  The ”release and promotion” category is characterized by terms like ”disney,” ”netflix,” ”date,” ”digital,” ”haunting,” ”jones,” ”dial,” ”streaming,” ”venice,” and ”indiana.”
- Review
  The ”review” category is represented by terms like ”law,” ”haunting,” ”friend,” ”performances,” ”zone,” ”reviews,” ”joe,” ”biden,” ”tone,” and ”disturbing” with the highest TF-IDF scores.
- Comparison
  Finally, in the ”comparison” category, words such as ”european,” ”nominations,” ”categories,” ”gotham,” ”zone,” ”awards,” ”led,” ”nominees,” ”lead,” and ”von” have the highest TF-IDF values.

4.4 Oppenheimer Focus (Q1)

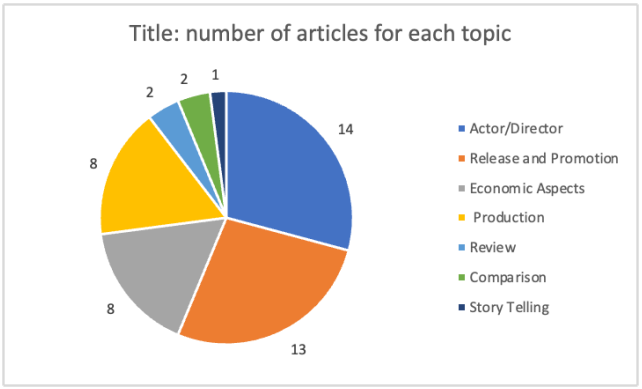
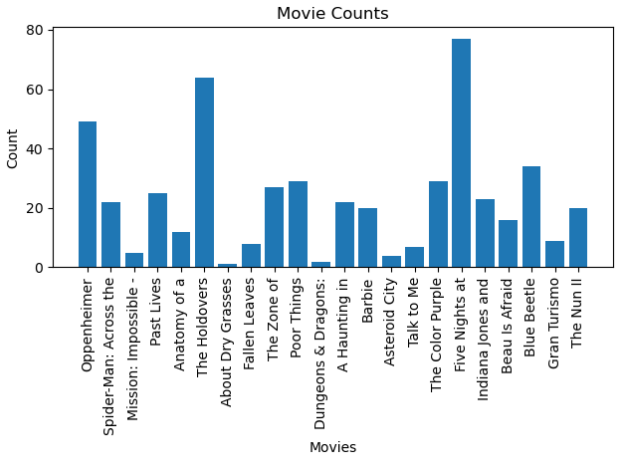


Figure 2: Proportion of articles in dataset based on Movie Release Date

Based on the distribution of articles, it is evident that the majority of articles about Oppenheimer primarily focus on the aspects related to the Actor/Director, Release and Promotion, Economic Aspects, and Production of the movie. Specifically:

4.5 Coverage (Q2)



As a percentage, our movie, Oppenheimer represented 9.50% of the total coverage (49 articles) while the only two movies that saw more articles were The Holdovers (12.67%: 64 articles) and Five Nights at Freddy’s (15.24%: 77 articles). The average number of articles per movie was 22.95, the median was 21, and the standard deviation was 19.02.

As a whole, it is clear Oppenheimer received a lot of coverage relative to other big box office releases in 2023.

5 Discussion

5.1 Evidence of Bias

As mentioned in Section 2 we knew there would be a possibility of time bias and it is evident based on figure 1 that it has occurred. The top 2 movies with the most articles are the newest released movies. However, we still felt as if picking the biggest box office movies in 2023 with time bias would give a more accurate comparison then all movies released in the same month as Oppenheimer. The reason for this is because so many movies in that time got close to no press surrounding them. Netflix releases several movies a month which doesn't give a real representation of how much coverage Oppenheimer receives vs other movies. We would know the results before doing any data collection. In conclusion, while we experienced time bias, we still believe the way we selected our movies gives a better comparison than other explored options.

From the use of English language filters, a step integral to our data preprossessing, we encountered geographical bias. This filtering mechanism, while essential for language consistency, inadvertently introduces a potential bias towards Western perspectives, thereby excluding valuable insights originating from non-English sources. Notably, China ranks as the country with the third-highest box office revenue for Oppenheimer, yet our analysis does not encompass articles from Chinese sources. This limitation could lead to a lack of diverse perspectives in our research findings.

5.2 Analysis of tf-idf

For the "Actor/Director" category, the names of actors and directors have higher TF-IDF scores. This could be because in this category, the focus is on individual personalities, and names are crucial in the context of movies. For the "Economic Aspects" category, words related to economic aspects in the theatre industry have high TF-IDF scores. This suggests that specific economic terms or jargon related to the theatre are significant in this category. For the "Storytelling" category, high TF-IDF scores for words related to Christianity might be attributed to the specific content of movie articles about "The Nun II," where the storytelling prominently features themes related to Christianity. For the "Release and Promotion" category, names of streaming services have high TF-IDF scores, indicating that these names are closely associated with discussions related to the release and promotion of movies. For the "Review" category, words with high TF-IDF scores vary, suggesting that the choice of words in movie reviews depends on the individual authors. Different reviewers may use diverse vocabulary and emphasize different aspects of a movie in their reviews. For the "Comparison" category, words like "nominations," "lead," and "nominees", "awards" imply that comparisons often involve discussions about the success or recognition of different movies, in the context of awards or achievements.

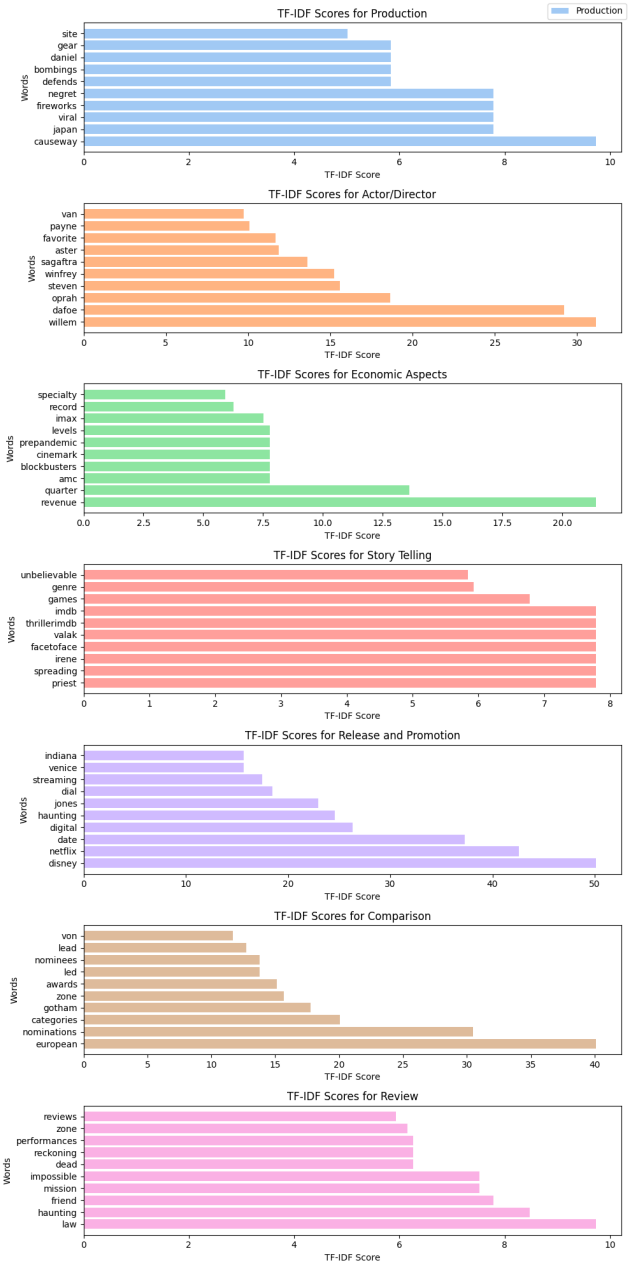


Figure 3: TF-IDF Scores Visualization

5.3 Topics of the movie(Q1)

As per our result in Section 4.4, the main article topic were Actor/Director-related content about Oppenheimer, which can be attributed to the significant influence and reputation of the director, Christopher Nolan. Nolan's involvement in the movie likely draws attention to aspects related to his role in the creative process. This extends to the second most common topic, 'Release and Promotion', mirroring the anticipation surrounding the film's launch on streaming platforms. Additionally, there is a notable abundance of articles dissecting the 'Economic Aspects' of the movie is driven by

audience curiosity surrounding Oppenheimer's post-release success and financial impact. Precisely, professionals in the industry are interested in the film's box office performance, revenue, and budget specifics as crucial metrics for gauging its success and industry influence.

The topic 'Story Telling' has the least amount of articles, potentially influenced by the time bias in our data. Discussions and analyses regarding a movie's plot usually surface in the immediate aftermath of a movie's release. While 'Oppenheimer' was released in July 2023, data spans articles published between October and November 2023, contributing to the reduced coverage of storytelling aspects. In fact, considering the other prominent article topics surrounding the movie presently, this could potentially suggest a natural progression in media focus from one aspect to another over the course of time following the release of a movie.

5.4 Coverage Comparison (Q2)

Several factors contribute to Oppenheimer receiving significant coverage compared to other major box office releases in 2023. The articles on Oppenheimer primarily focus on actor/director aspects and release and promotion. As discussed above, one of the major contributors to this attention could be the director's reputation. Compared to that, the movie 'Five Nights at Freddy's ', amassed 77 articles, the most out of all the movies. The majority of articles delved into the themes of "Economic Aspects" and "Story Telling". These articles, collected within the time-frame of the movie's release in November, highlight a focus on discussions about the film's financial success and narrative elements. This suggests that articles for movies just released are primarily centered around the box office performance and plot details of the movie. Extending our remark in the section 5.3, we can reiterate that recent movies tend to elicit a greater number of articles centered around "Story Telling", while films released earlier, such as Oppenheimer, exhibit fewer articles on this topic but have an augmented emphasis on Actor/Director-related or other content.

5.5 Typology

Among the seven categories, Release and Promotion accounted for the highest number of occurrences, totaling 140 articles. This was closely followed by articles related to the actors and directors involved in the movie. This dominance in content might be attributed to the time bias mentioned in 5.1. Since our data collection focused on newly published articles following the movie release, it's reasonable to observe a higher count in promotional content.

On the other hand, Storytelling emerged as the least prevalent category, comprising approximately 30 articles. This lower count may suggest that in the immediate aftermath of the movie's release, there might be a scarcity of in-depth analyses or detailed discussions on the movie's narrative or storytelling aspects. Articles within this category often delve deeper into discussing the movie's elements in detail, possibly indicating that such comprehensive analyses might surface later in the coverage cycle.

6 Conclusion

As a result, our thorough examination of the media coverage of "Oppenheimer" and its predecessors offers important new understandings into the nature of news pieces and the variables affecting the volume of coverage. According to our research, Oppenheimer received a significant amount of coverage—9.50% of all coverage—with an emphasis on content relating to actors and directors, release and promotion, economic aspects, and production.

The preponderance of content pertaining to actors and directors highlights the important role played by the film's acclaimed director, Christopher Nolan. This added to the film's increased media exposure. The next most frequently discussed topic was "release and promotion," which was apt given the buzz and conversations around the movie's release on several platforms.

When comparing the articles of Oppenheimer to another, we observed a shift in media focus over time. For instance, recent movies tend to steer attention toward narrative elements, while older releases, like Oppenheimer, had a heightened emphasis on Actor/Director-related or or economical-focused content. This shift reflects the dynamic evolution of media coverage patterns surrounding film releases.

The dependability of our results was strengthened by our methodological approach, which included open coding, tf-idf analysis, and dataset refinement, despite potential biases like time and geographic limitations in our dataset. The use of tf-idf scores brought each topic's important terms to light, giving readers a more complex grasp of the news coverage's unique aspects.

Our approach, which acknowledges its limits (time bias, geographic filtering, etc.), made it possible to compare Oppenheimer more meaningfully with other 2023 big box office films.

7 Group Member contributions

- Jacob proactively led the team as the group leader. He reached out to each group member by email, scheduled the first meeting, distributed work, assigned a fair amount of tasks to each member, and created deadlines for us. Additionally, he worked on the results section of the report, which included Q2 and topics selected/definition
- Zafeerah worked on the abstract, introduction, dataset and discussion part covering Q1.
- Shuaishuai worked on the methods and discussion section covering the Typology.
- Jaewon worked on the results part, covering Q1, TF-IDF, the discussion part covering Q2, and the contributions of each group member section.
- Throughout the project, everyone collaborated on open-coding and coding parts together. Fortunately, no one faced difficulties attending either in-person or online meetings.