COMP370 Group Project

[Repo](https://github.com/tgompper1/movie-release-analysis)

<https://docs.google.com/spreadsheets/d/1uO0WqME5e33dZQQGo2TsNU4x87QoXfo-W7PcsuqtN8E/edit?usp=sharing>

11/24

* ~~Open coding 200 articles~~ 
  + ~~3-8 topics~~
* ~~Annotate full set (500)~~
* ~~compute the 10 words in each category with the highest tf-idf scores~~ 
  + ~~(to compute inverse document frequency, use all 500 posts that you originally collected.)~~
* ~~% of each category per film~~
* ~~% of 500 per topic~~
* ~~% of barbie articles per topic~~
* ~~Number of barbie articles mentioning other films, split by percent per other film~~
* **What aspect of the movie was the focus (topic) of the articles**
  + **Film in culture => strong cultural impact**
* **How much coverage the movie received relative to other movies that came out at a similar time** 
  + **Vs oppenheimer:**
  + **Vs new movies**
  + **Eg how much is it still being talked about compared to movies coming out now and movie that came out at the same time**
* **Introduction** (0.5 page) – General overview and key findings laura
* **Data** (0.5 page) – describe your dataset. This should include statistics relevant to the project – the number of articles you collected, filtering you did, and any design decisions you had to make around the collection of this content.
  + How did we avoid bias
* **Methods** (0.5 page) - explanation and justification for what you did. Focus on the design decisions you made NOT listed in this document that impacted your results. Laura
* **Results** (1 pages) - share all your findings including the topics selected (and their definitions), topic characterization, and topic engagement. Mill
* **Discussion** (1 pages) - interpret your results in terms of what they reveal about the way each candidate was being discussed and perceived. Make extensive use of your results to justify your interpretations. Tess
* **Group Member contributions** (0.25 page) - a description of the contributions each group member made to this project.

# Project Description

Movie Release

### Overview

Your team has been hired by a media company that wants to understand the news reporting currently happening around the film “(insert a recently-released movie that your team selected here)”. They have indicated that they are especially concerned with visibility and reception relative to other movies that have come out at a similar time. Specifically, they want to know:

1. **What aspect of the movie was the focus (topic) of the article**
2. **How much coverage the movie received relative to other movies that came out at a similar time**

You will conduct this analysis and submit a report discussing your findings.

### Analysis Details

Your analysis will draw on news articles drawn from **NewsAPI.org**. To inform your analysis, you should collect **500+ articles** (total) on the movies you consider in your analysis. **Ensure to collect articles in a way that does not bias towards or against coverage volume of your selected movies**. You should set filters such that all 500 posts have a very high likelihood of being related to one of the movies AND all are in English.

To develop your topics, conduct an **open coding on** **200 articles** (approach the exercise requiring each article to belong to exactly one topic). For your open and later codings, just use the title and opening of the article (i.e., you don’t need to read the entire article). You should aim for between **3-8 topics in total**.

Once your topics have been designed, manually **annotate the entire set of the 500 article**s in your dataset.

Characterize your topics by **computing the 10 words in each category with the highest tf-idf scores** (to compute inverse document frequency, use all 500 posts that you originally collected.

# Final Report

Your report should be written using the **Camera Ready AAAI format** (you may use either Word or Latex): https://aaai.org/authorkit24-2/.

The template formatting (e.g., font, font size, spacing, citation style) should be followed strictly. The report structure should consist of the following sections (the lengths are suggestions):

1. **Introduction** (0.5 page) – General overview and key findings
2. **Data** (0.5 page) – describe your dataset. This should include statistics relevant to the project – the number of articles you collected, filtering you did, and any design decisions you had to make around the collection of this content.
3. **Methods** (0.5 page) - explanation and justification for what you did. Focus on the design decisions you made NOT listed in this document that impacted your results.
4. **Results** (1 pages) - share all your findings including the topics selected (and their definitions), topic characterization, and topic engagement.
5. **Discussion** (1 pages) - interpret your results in terms of what they reveal about the way each candidate was being discussed and perceived. Make extensive use of your results to justify your interpretations.
6. **Group Member contributions** (0.25 page) - a description of the contributions each group member made to this project.
7. **References** (< 1 page) - this is an optional section should you reference other works in your report.

The report must be between 5 and 7 pages in length, not including references. Figures are encouraged – but should be used to maximum effect (fluffy or otherwise unnecessary images that do not make strong contributions to the report will lead to point deductions)

# Grading

Each **group will submit one report** which will receive one grade that all members of the group will share. The one exception to this is in the case of strong evidence of delinquent group members (including but not limited to details in the Group Member contributions section). In this case, each member's grades may be adjusted up or down as appropriate.

### Rubric

| **Criteria** | **Points (100 in total)** | **Details** |
| --- | --- | --- |
| Style | 10 | Is the text written in a clear, concise way? Is good grammar and spelling employed throughout? |
| Data collection correctness | 10 | Was the dataset prepared correctly? Was sampling done to avoid problematic biases? Did it have baseline characteristics that would allow this study to deliver meaningful insights? |
| Topic design validity | 15 | Was a process followed that would produce valid topics? Insufficient details should be treated the same as if something was not done. |
| Topic validity | 15 | Are the topics appropriate to the task? Are they well -defined? Are they defined to minimize subjectivity? |
| Annotation quality | 10 | Does the annotation process give us confidence in the quality of the annotations? |
| Results | 20 | Are all results requested present? Do the results make sense? Are outliers or unusual trends appropriately explained? |
| Findings | 20 | Are insightful interpretations provided? Are these grounded in results? Do the findings integrate results and prior knowledge in a sound, well -reasoned way? |

Recent movies:

* **Oppenheimer**
* **barbie**
* Mission impossible
* Ant man and the wasp
* **Asteroid city**
* **Transformers**
* The Flash

Tasks:

-get articles (150-200 articles) by Friday Nov 24th

- tess: asteroid city

- laura: oppenheimer

- Mingchen: barbie

- mill: transformers

!all 500 posts have a very high likelihood of being related to one of the movies AND all are in English. -> give priority to articles that talk about your movie AND barbie

Oppenheimer articles:

* There is a business called Oppenheimer Holdings that is unrelated to the movie, many articles talk about the business and are not relevant to the project

Filter out spiderman

Filtered out trolls

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Potential title: The Barbie Movies Persistence in Mainstream Culture and Film Post-Release

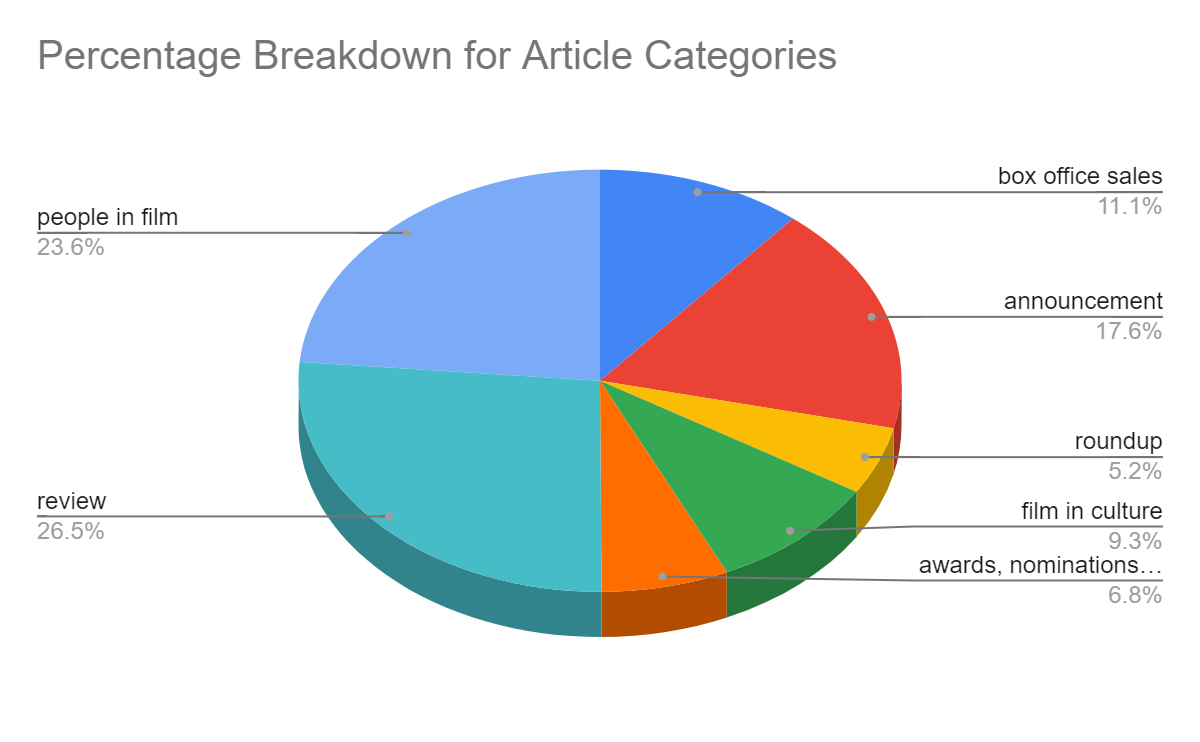
# INTRODUCTION - Laura

-attempted to understand the audience’s attitude and impressions of the Barbie movie in itself and as opposed to other newly-released movies, in particular Barbie’s main competitor Oppenheimer due to them being released on the same date. (July 21 2023)

-found that Barbie is: the most talked about in subject of Film in culture: mainstream media and pop culture.

-competing with Oppenheimer, while Oppenheimer had a more positive impact in reviews, box office sales and announcements which showed that Oppenheimer was more anticipated than Barbie and more delved into by critics, Barbie remained a major subject to talk about in terms of pop culture and awards, which shows that while it was not as anticipated as its rival before and after release, it made its mark amongst the common population and fans of dolls alike.

# DATA - Mingchen



# METHODS - Laura

* Used Google Sheets to annotate our data and collect relevant statistics. A useful tool in retrieving counts of all categories per a certain movie and counts of all movies per a certain category was *Data > Columns* stats. This has helped us get counts in order to generate pie charts for a more readable, visual representation of the analysis.

# RESULTS - Mill

To understand different aspects of the movie, we created and categorized each movie article into 7 topics. First, we will introduce the topics and their definitions.

1. **Box office sales**: articles that refer to the amount of sales the box office is expected to make, has made, or failed to make regarding a movie.
2. **Roundup**: articles that group movies based on release dates or genre
3. **Announcement**: articles that provide an announcement, making public some information related to a movie or event such as announcing the release date of a movie/event, the platforms where the movie is shown, or information regarding previews and trailers ahead of the cinematic release.
4. **Film in culture**: articles that cover pop culture and mainstream trends related to recent movies that have been released.
5. **Awards, nominations and film festivals**: articles related to the announcement of awards and/or nominations of movies or actors and film festivals.
6. **Review**: articles that provide a review or general impressions and feedback regarding a movie.
7. **People in film**: articles that specifically refer to staff in the film industry, involved in the making of a film such as actors, directors etc.

The topics “Box office sales” and “Awards, nominations and film festivals” help us gain more insight into the success of the movies in terms of achievements such as profits and awards. “Film in culture” conveys a similar idea, showing us the success of the movies in terms of popularity and how they impact the audience outside of the theater. However, compared to the two topics mentioned previously, “Film in Culture is less official and anyone can contribute. The “Roundup” category consists of articles that mention multiple movies, whereas other topics primarily focus on a single movie. The “Announcement” and “Review” categories both focus on general aspects of the movie. Articles related to general information before the movie’s release are categorized as an “Announcement,” while after the movie’s release, they are considered a “Review.” To categorize any article, we can first ask ourselves whether the article can fit into any topic other than “Announcement” and “Review.” If the answer is no then it has to be either “Announcement” or “Review,” depending on the time period and general topic of the article. Thus, our typology is comprehensive.

Next, we will discuss the topic engagement and characterization.

*Five Nights at Freddy’s* constitutes 25.3% of the articles in “Box office sales,” making it the movie with the highest engagement within this topic. This is also reflected in the tf-idf score, where the words “freddy” and “nights” both receive a score of 0.22. *Five Nights at Freddy’s* is the second-highest in “Roundup” at 15.6%, while *The Marvels* hold the highest percentage of 18.8% in this category.

Although *The Marvels* constitutes the most of the“Roundup” category, it only makes up 4.64% of the “Announcement” category and 0% of “Film in culture” and “Awards, nominations and film festivals.”

*Saltburn,* the newest movie in our focus*,* has the highest percentage of 18.2% in“Announcement.” The words related to *Saltburn* include “keoghan” and “saltburn” and have tf-idf scores of 0.14 and 0.13, respectively. Conversely, the oldest movie in our focus, *Transformers: Rise of the Beasts*, makes up a very low percentage of the “Announcement” category at 4.6%.

Articles about *Barbie* make up 45.6% of “Film in culture” which is significantly higher than other movies. The top 10 words in this topic contain multiple Barbie-related words, including “barbie,” “doll,” and “mattel.” The lowest topic that *Barbie* contributes to is “Review,” with a percentage of 3.1%. It is the second-lowest in this topic, with *Transformers: Rise of the Beasts* being the lowest at 2.5%.

In “Awards, nominations and film festivals,” *Anatomy of a Fall* makes the highest contribution at 32.6%. However, there is no word related to this movie present in the tf-idf score of this topic. The only movie-related word that appears is “barbie,” which is the second-highest movie at 23.3%. Many movies have “Awards, nominations and film festivals” as their lowest topic engagement. This includes *Oppenheimer, Five Nights at Freddy’s, Priscilla,* and *The Marvels. Transformers: Rise of the Beasts* makes up 7% of this topic. Although the number does not seem impressive, this topic is actually the one in which *Transformers* has the highest engagement.

*Five Nights at Freddy’s* makes up the highest percentage of the “Review” category with 20.2% And *Priscilla* and *The Marvels* are the second-highest contributors after *Five Nights at Freddy’s with* 19.6%.Also, *Priscilla* and *The Marvels* have the highest engagement with “Review” compared to other topics*.*

From the tf-idf score of “Review,” the words related to *Five Nights at Freddy’s*, “freddy” and “nights”, have a combined score of 0.47. *The Marvels*-related word, “marvels” and “marvel”, have a combined score of 0.41. Lastly, *Priscilla*-related words, including “priscillia,” “presley” and “coppola,” have a total score of 0.53, which is the highest.

In “People in film,” 20% of the articles are *The Hunger Games: The Ballad of Songbirds and Snakes.* Half of the tf-idf word list in this topic is comprised of *The Hunger Games-*related words including “hunger,” “games,” “rachel,” “zegler” and “ballad.” This movie has the highest engagement in “People in film,” but it has 0% engagement in “Roundup.”

To summarize, *Five Nights at Freddy’s* has the highest engagement in 2 topics. *Saltburn*, *Barbie*, *Anatomy of a Fall* and *The Hunger Games* each have the highest engagement in 1 topic. The movie with the least engagement overall is *Transformers,* where its highest contribution is 7% *in* “Awards, nominations and film festivals.”

—----------------------------------------------------------------------------------------------------------------------------

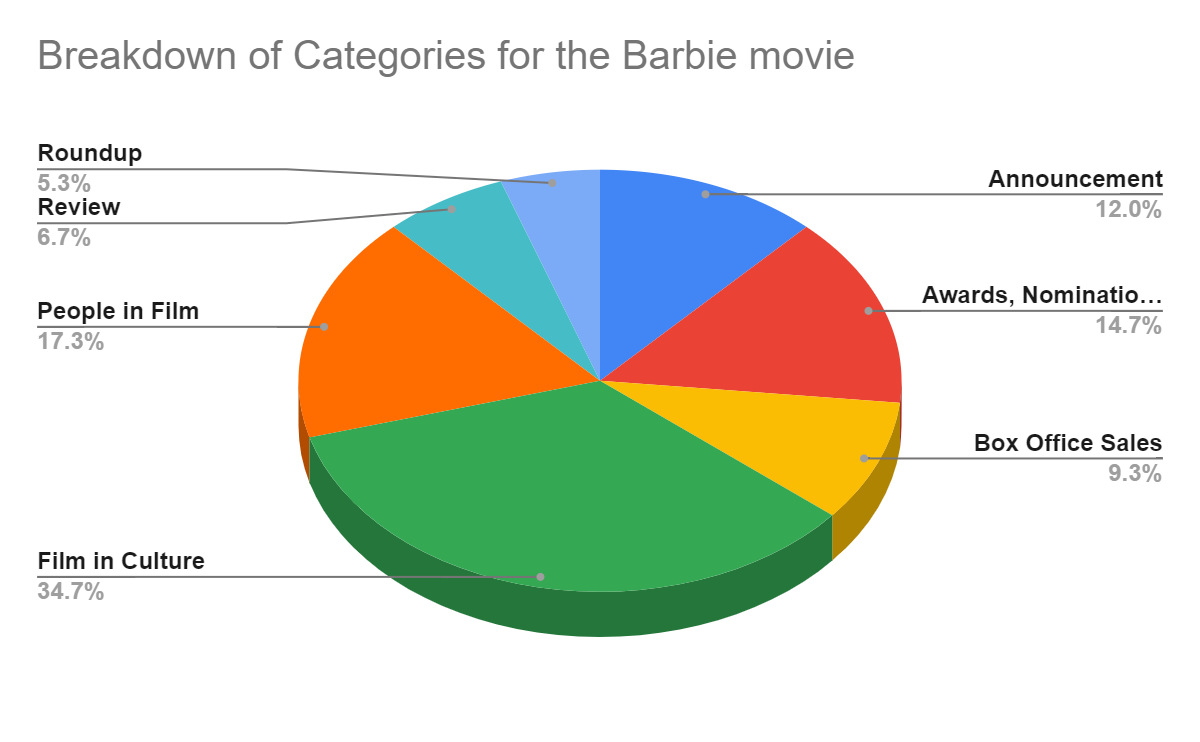
Definitions:

1. **Box office sales**: articles that refer to the amount of sales the box office is expected to make, has made, or failed to make regarding a movie.
2. **Roundup**: articles that group movies based on release dates or genre
3. **Announcement**: articles that provide an announcement, making public some information related to a movie or event such as announcing the release date of a movie/event, the platforms where the movie is shown, or information regarding previews and trailers ahead of the cinematic release.
4. **Film in culture**: articles that cover pop culture and mainstream trends related to recent movies that have been released.
5. **Awards, nominations and film festivals**: articles related to the announcement of awards and/or nominations of movies or actors and film festivals.
6. **Review**: articles that provide a review or general impressions and feedback regarding a movie.
7. **People in film**: articles that specifically refer to staff in the film industry, involved in the making of a film such as actors, directors etc.
8. **Announcement:**
9. **- trailer: 0.26**
10. **- new: 0.18**
11. **- movie: 0.14**
12. **- film: 0.14**
13. **- keoghan: 0.14**
14. **- hunger: 0.13**
15. **- digital: 0.13**
16. **- saltburn: 0.13**
17. **- games: 0.13**
18. **- release: 0.13**
19. **Awards, nominations and film festivals:**
20. **- awards: 0.45**
21. **- nominations: 0.36**
22. **- film: 0.23**
23. **- gotham: 0.22**
24. **- festival: 0.21**
25. **- year: 0.19**
26. **- grammy: 0.19**
27. **- barbie: 0.17**
28. **- best: 0.13**
29. **- strangers: 0.13**
30. **Box office sales:**
31. **- box: 0.43**
32. **- office: 0.43**
33. **- million: 0.30**
34. **- weekend: 0.23**
35. **- freddy: 0.22**
36. **- nights: 0.22**
37. **- marvels: 0.16**
38. **- opening: 0.15**
39. **- 78: 0.11**
40. **- mcu: 0.10**
41. **Film in culture:**
42. **- barbie: 0.30**
43. **- buy: 0.18**
44. **- halloween: 0.17**
45. **- movie: 0.15**
46. **- mattel: 0.12**
47. **- year: 0.12**
48. **- games: 0.12**
49. **- hunger: 0.11**
50. **- owlkitty: 0.11**
51. **- doll: 0.11**
52. **People in film:**
53. **- elordi: 0.26**
54. **- jacob: 0.23**
55. **- nolan: 0.21**
56. **- hunger: 0.19**
57. **- games: 0.19**
58. **- christopher: 0.17**
59. **- rachel: 0.15**
60. **- zegler: 0.15**
61. **- ballad: 0.14**
62. **- movie: 0.14**
63. **Review:**
64. **- freddy: 0.24**
65. **- marvels: 0.24**
66. **- nights: 0.23**
67. **- priscilla: 0.23**
68. **- marvel: 0.17**
69. **- movie: 0.16**
70. **- games: 0.16**
71. **- film: 0.16**
72. **- presley: 0.16**
73. **- coppola: 0.14**
74. **Roundup:**
75. **- movies: 0.44**
76. **- watch: 0.24**
77. **- like: 0.17**
78. **- marvels: 0.17**
79. **- november: 0.15**
80. **- best: 0.13**
81. **- mcu: 0.13**
82. **- check: 0.12**
83. **- freddy: 0.12**
84. **- new: 0.12**

# DISCUSSION - Tess

In collecting and analyzing data on seven newly released movies: *Transformers: Rise of the Beasts*, *The Marvels*, *Saltburn*, *Priscilla*, *A Ballad of Songbirds and Snakes*, *Five Nights at Freddy’s*, and *Anatomy of a FallI*, and one movie released at the same time as *Barbie*: *Oppenheimer* we discovered much about how *Barbie* has persisted in both mainstream culture and recent film discussion five months post-release.

Of the 65 collected and annotated articles regarding only *Barbie* and 10 articles on *Barbie* and one other movie we discovered that 34.7% of articles on *Barbie* were categorized as “Film in Culture”, this was followed by 17.3% “People in Film”, and 14.7% “Awards, Nominations and Film Festivals.



These statistics tell us that *Barbie*, even five months after its cinematic release, can still be seen as having a major impact on mainstream culture. This is particularly notable when we make a comparison to *Oppenheimer*, another highly anticipated movie released on the same date as *Barbie*. When we look within the “Film in Culture” category of articles alone, we see that *Barbie* made up 45.6% of articles in this category and *Oppenheimer* made up just 5.3%. It is also worth noting that *Barbie* is seen in “Film in Culture” articles almost the same amount as all seven recently released movies combined (these make up 49.3% of the articles in this category). This indicates the extreme relevance and impact of *Barbe* on mainstream culture.

We also see *Barbie*’s large cultural impact when we look at the 10 words in the “Film in Culture” category with the highest tf-idf scores:

• barbie: 0.30

• buy: 0.18

• halloween: 0.17

• movie: 0.15

• mattel: 0.12

• year: 0.12

• games: 0.12

• hunger: 0.11

• doll: 0.11

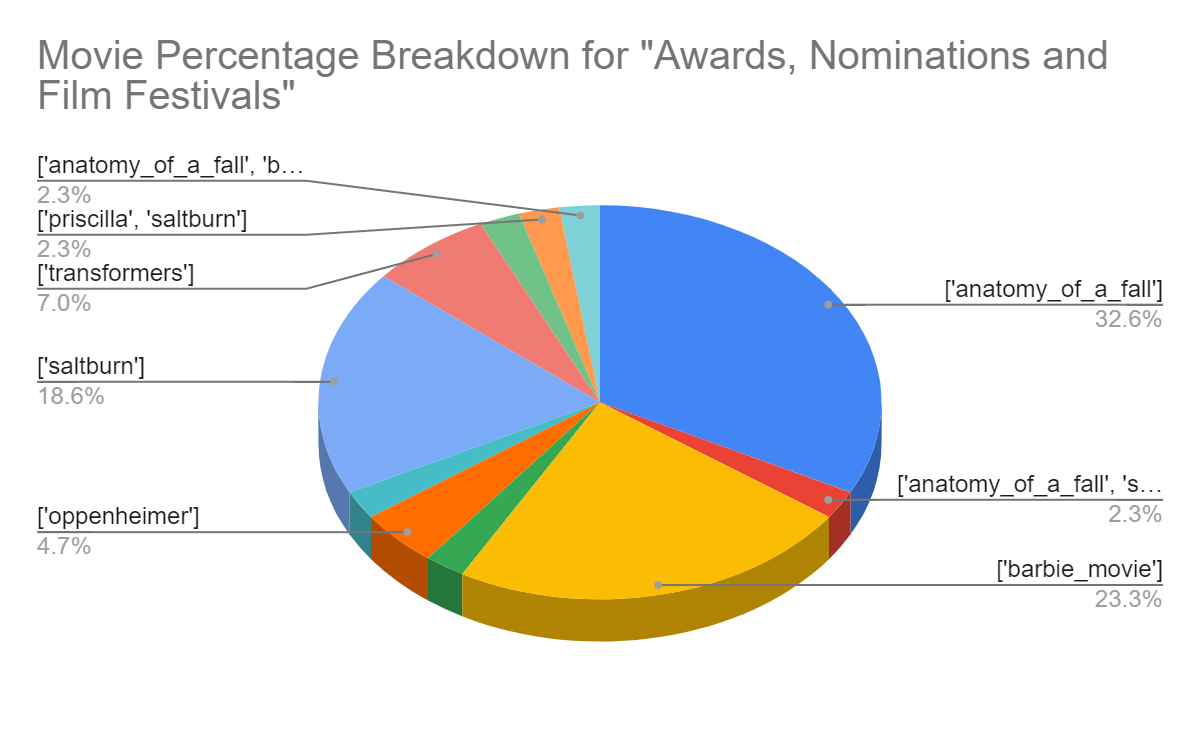
• owlkitty: 0.11

The word “barbie” receives a score of 0.30 and *Barbie-related* words (“barbie”, “mattel” and “doll”) sum to an impressive combined score of 0.53. It is perhaps most interesting that "Barbie” receives a score of exactly double the word “movie”(which ranks fourth in the list with a score of 0.15), a word which we would expect to score quite highly given our dataset of movie-related articles. And we do see this in the high tf-idf scoring words in other categories, with “movie” coming in third for the “announcement” category, “film” in third for “awards, nominations and film festivals”, and “movies” coming in first for “reviews”. This once again indicates *Barbie*’s continuing relevance in mainstream culture.

As it is now established that *Barbie* is very present in mainstream culture despite being five months post-release, I will discuss how *Barbie* has also persisted in film discussion and maintains its relevance in the film industry despite the many new, highly anticipated movies coming out in November. When we look at articles that discuss *Barbie* and another movie, we see that *Oppenheimer* is discussed seven times, and *Saltburn*, *Anatomy of a Fall*, and *The Marvels* are each discussed in one article along with *Barbie*.

Due to the cinematic hype around “Barbenheimer” it makes sense that when discussed along with another movie, *Barbie* is discussed the most with *Oppenheimer*. But it is interesting that they are both still being discussed five months later.

Another indicator of *Barbie*’s post-release relevance in film discussion is that it is being discussed in relation to *Saltburn* and *Anatomy of a Fall* which make up 18.6% and 32.6% of the “Awards, Nominations, and Film Festivals” category. The prevalence of these two films in the “Awards, Nominations, and Film Festivals” category indicates their success and appreciation as high-quality films and *Barbie* is right up there with them despite being five months older than the other two films.



Let us now dive deeper into the “Barbenheimer” situation. As previously mentioned, we know that *Barbie* and *Oppenheimer* were both highly anticipated films that were released on the same day creating a pop-cultural moment coined as “Barbenheimer”. Both *Barbie* and *Oppenheimer* received large amounts of press at the time of their release and we still see articles being produced about them today, as we observe in our data collection. However, the two films are being talked about in largely different ways five months post-release.

Of our collected articles, we have 10.6% on *Barbie*, 11.5% on *Oppenheimer,* and 1% that discusses both. Although we have a similar number of articles for both movies, they deal with vastly different subjects. Here is a breakdown of how much each movie made up each of our categories:

• Roundup: *Barbie*- 12.5%, *Oppenheimer*- 12.5%

• Review: *Barbie*- 3.1%, *Oppenheimer*- 4.9%

• People in Film: *Barbie*- 8.3%, *Oppenheimer*- 17.9%

• Film in Culture: *Barbie*- 45.6%, *Oppenheimer*- 5.3%

• Box Office Sales: *Barbie*- 10.3%, *Oppenheimer*- 22.6%

• Awards, Nominations and Film Festivals: *Barbie*- 23.3%, *Oppenheimer*- 4.7%

• Announcements: *Barbie*- 7.4%, *Oppenheimer*- 15.7%

*Barbie* articles make up much more of the “Film in Culture” and “Awards, Nominations and Film Festivals” than *Oppenheimer* which strengthens our conclusion of *Barbie*’s cultural relevance when we consider how much it outperforms a movie released at the exact same time. The two movies perform the same and similarly in the “Roundup” and “Review” categories respectively. This makes sense as five month old movies are unlikely to receive many new reviews and as many people have already seen the movie due to the extreme hype around their releases, it would not be of much benefit to each movie to mention them in roundups.

What is worth discussing is how *Oppenheimer* outperforms *Barbie* in the “Box Office Sales” and “People in Film” categories. I argue that *Barbie*’s lack of performance in these categories in relation to the similarly released movie does not imply a detriment to *Barbie*’s long term performance. *Oppenheimer*’s outperformance in the “Box Office Sales” category makes complete sense when we consider that *Oppenheimer* was shot and released in IMAX and *Barbie* was only re-released in IMAX in September. As we can see in the data, a large number of the *Oppenheimer* articles refer to IMAX so this statistic does not imply a larger success of *Oppenheimer* but a larger focus on the financials of IMAX movies.

Similarly, *Oppenheimer*’s outperformance in the “People in Film” category does not imply a lack of longevity for *Barbie*. In the data, we see that 1/3 of the *Oppenheimer* articles in the “People in Film” category focus on the movie’s director, Christopher Nolan. Similarly, in or tf-idf scores for the category, “Nolan” receives a 0.21 and “Christopher” receives a 0.17 making them the third and sixth highest scoring words of the category. Thus, this statistic does not imply a larger success of *Oppenheimer* over *Barbie*, but a larger press focus on Christopher Nolan as a director.

Overall, our data collection, annotation and analysis revealed that *Barbie* remains a successful and culturally relevant film, even five months post release, and also continues to be relevant in current film discussions.

# Group Member contributions

• Tess Gompper

• Data collection, open coding, and annotation

• Discussion section

• Laura Ladaru

• Data collection, open coding, and annotation

• Mill Gai

• Data collection, open coding, and annotation

• Mingchen Ju

• Data collection, open coding, and annotation