COMP370 Final Project- Movie Release

Written by Hyeonmin Soh, Aamir Shivji, Jonathan Shiang

Abstract

The goal of the project is to use news articles to understand the visibility and reception happening around the film "*The Substance*" by comparing it to other films that have been released around the same time. The important aspects of this question that we aim to probe are: What aspect of The Substance is the main focus of the article, and how much news coverage has The Substance received relative to other films released in the same month.

Introduction

Why did we choose "The Substance"?

The Substance has quickly emerged as a significant topic across various internet platforms, particularly on Twitter. Its recent debut, along with its cultural impact and potential to become a cult classic, makes it a compelling subject for analysis. This choice aligns well with the goals of this project, as we aim to explore the visibility and reception of the film compared to other movies released during the same period. Specifically, we want to determine what aspects of *The Substance* are emphasized in news articles and how much coverage the film has received relative to its peers.

Another factor that influenced our choice is the film's production journey. Initially set to be released by Universal Pictures, the movie was ultimately passed over, with rights later acquired by Mubi, a smaller streaming service known for curating independent films. This history makes *The Substance* particularly interesting to study, as it provides an opportunity to examine the media coverage of a film that does not have the backing of a major studio or an established fanbase.

By selecting *The Substance*, we hope to shed light on how an independent film navigates the competitive media landscape, especially when compared to blockbuster releases with pre-existing popularity. This contrast between an underdog film and more established titles makes the analysis all the more intriguing.

Why did we choose our other movies?

To effectively compare the coverage of *The Substance*, we chose several other films released in October 2023. Most of these films are part of popular franchises with large, devoted fanbases. These selections were intentional, as they provide a baseline for understanding how established movies dominate news coverage.

The movies we chose for Comparison are: *Alien: Romulus, Speak No Evil, Smile 2* and *Venom 3*. An obvious bias presented itself here as almost all of these films are part of the Horror genre. We believe this bias is justified, as this overlap happened naturally due to the significant number of horror films released recently. This coincidence also allows us to focus our analysis within a single genre, making the comparisons more cohesive and meaningful.

By contrasting *The Substance* with these highly visible films, we can examine not only how much coverage it received but also the nature of the articles written about it. This comparative approach allows us to analyze whether *The Substance* managed to carve out its own space in the media and how its reception differs from that of blockbuster movies.

Data

To answer our main questions: "What aspect of the movie was the focus of the article" and "How much coverage did the movie receive relative to other movies that came out at a similar time", we built a dataset containing articles that covered the selected movies- however as we will later discuss, the dataset was not a perfect one. For each article in our final dataset, we maintained the following information:

- Title
- Author
- Publication date
- Source name
- URL

Our dataset was collected using NewsAPI. Given the free plan's limitation to 100 articles per query, we designed multiple queries to broaden the scope of collected content. We had two different approaches to collecting our data:

Approach 1 involved a snowball sampling method, where we began with a set of keywords relevant to our chosen movies, followed by iteratively expanding our keyword set by extracting relevant terms from the content of the collected articles. This iterative process allowed us to refine the scope of the dataset over several iterations while ensuring diversity in coverage.

Approach 2 used a predefined set of focused queries to collect articles via the NewsAPI.

Both approaches aimed to capture the most relevant articles by using a ***lookback period of 30 days***, a timeframe that aligns with the film's release and early reception, a critical period for analyzing visibility and reception. To ensure relevance, we applied filters for English-language articles using the ***language='en'*** parameter.

In the end, we stuck with the ***snowball sampling*** approach and collected a total of 616 articles related to *The Substance* and its comparison films. This method began with a small set of words (e.g., "The Substance," "Alien: Romulus", "Horror Movies") and dynamically expanded to include additional relevant terms discovered in the collected articles. This adaptive approach helped reduce bias by broadening the scope of keywords and ensuring that the dataset better reflected the overall media conversation surrounding these films.

The collected articles were saved in a JSON file, and then TSV format with the fields listed above. The dataset was parsed to take out any duplicate articles and articles labelled as [removed], which reduced our total article number to 502. Even with 500 articles, and careful collection using relevant queries, we still ended up with articles that were either movie related, but unrelated to our selected movies, or articles that were completely unrelated to movies.

Movie	Number of articles
The Substance	87
Smile 2	30
Alien: Romulus	75
Speak No Evil	8
Venom 3: Last Dance	13
Total Unrelated	289

Figure 1shows the article distribution:

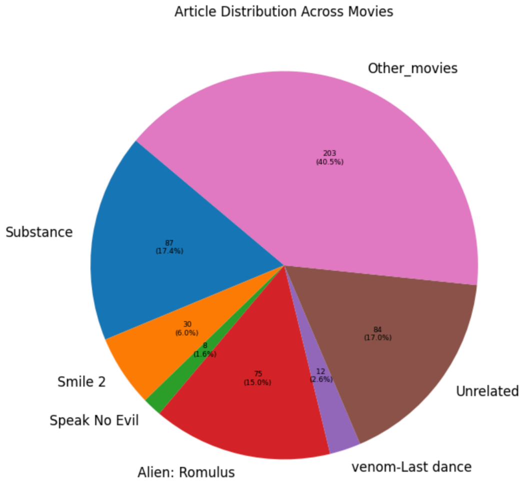


Figure 1: Relative coverage per movie in our dataset

Acknowledging Biases and "other_movies"

A significant limitation of our dataset is the presence of unrelated movie articles, particularly "upcoming movie announcements" for films not yet released during our collection window. This issue stems from the broad scope of keywords and the general nature of media coverage. While it would have been ideal for us to have more articles that are relevant to our selected films, we view these high numbers (203 for other movies and 84 for unrelated content) in a somewhat positive manner. It demonstrates the general issues that easily arise when collecting data automatically using NewsAPI, and also does provide more context to the current media landscape - there is always a heightened interest for new films which only emphasizes the competitive media environment for movies like *The Substance.*

In the case of ***bias****,* we implemented several measures to ensure fairness and relevance. Using the snowball sampling approach, we dynamically expanded our keyword set based on the content of collected articles, reducing selection bias by reflecting organic media discussions. We filtered for English articles only using the language = 'en' parameter to maintain consistency and limited our collection to a 30-day window around *The Substance*'s release to ensure temporal alignment with comparison films. Duplicate and placeholder articles (e.g., "[removed]") were excluded to eliminate redundancy and low-quality content.

Despite these efforts, some biases remain, such as the presence of irrelevant articles and a potential genre bias due to the focus on horror films. However, these limitations are acknowledged and, where possible, mitigated through filtering and cleaning processes. Overall, the dataset feels very representative of the mainstream media landscape. We do believe that such a dataset would look completely different had we used twitter as our source for data.

Methods

The **Data** section of this report provides more detail on our data collection methodology. Therefore, this section will focus more on how we developed the typology and the tf-idf analysis.

Developing our Typology and Annotation Phase

We performed an ***open coding*** on 200 articles at random to begin hypothesizing possible categories. We did three rounds of open coding, generating a category name based solely off of the title of the article in the first round, followed by refining and redefining those names in the second round, and finally making each label more comprehensive in the third round. The first round was performed by each group member whereas the second and third rounds were done together. This was a great way to not only discuss our own thought processes and ways of labeling an article, but to try and tackle a lot of disagreements that could possibly arise when annotating all 500 articles.

We initially started with 14 categories before narrowing it down to the final 8 that will be discussed in ***Results***. Categories like *awards and industry recognition, plot, reviews, rankings, Production, interviews, box office returns, streaming sites, sequel, franchise etc.* were all viable categories, but adjustments could be made. We realized that a lot of the categories that we had initially come up with were very nuanced and could be grouped up with other categories. Some categories that we initially came up with, such as "Awards and Industry recognition" did not appear very frequently or could easily be defined within a different category. Others like "plot" felt too vague and ended up under the umbrella term "review" as a lot of plot based articles often involved some sort of a critique on the film itself , or a recap of the film, which served as an appetizer for an "upcoming movie announcement". "Sequel" and "franchise" could be combined to "upcoming movie announcement" which was more comprehensive and accurately defined that subset of articles. The third round essentially helped take care of overlaps to narrow down the number of categories.

We also had to take care of unrelated content, which was done by creating two distinct categories- both of which are unrelated, one unrelated to movies and the other to ***selected*** movies, a very important distinction that needed to be made.

We then used the final 8 categories to annotate all 502 articles.

TF_IDF analysis

In order to conduct a proper TF-IDF analysis, and retrieve the important, unique, and relevant terms in the title of our articles, we coded two different algorithms to calculate word frequencies. Both algorithms had the following keywords: alien, Romulus, substance, smile, venom, speak, evil. This was done since the title for our movies would top all of the TF-IDF scores if not removed.

The first algorithm we created extracted the TF_IDF scores for all of our articles cumulatively (all 502). We pulled the top 20 most frequent words, and used the formula learned in class to calculate a TF_IDF score and ranking for all 20 words. Once we had extracted our criteria, we turned to matplotlib to help build a bar graph, sorted from highest to lowest, to show our most significant words. The reason for the sorting method was to help the viewer easily determine importance in a consistent order, and make terms comparison straightforward.

The second algorithm, was a more specialized, targeting each individual category created from our typology and building TF_IDF scores for the top 20 words under each one. We separated our data by having the algorithm only pull articles under specific annotation categories. Otherwise the process was very similar to our first algorithm with the same excluded words and formula to determine our TF_IDF scores. Once insights from each individual category (8 total) were pulled, matplotlib was used to create bar graphs, sorted highest to for the unique terms. Although with a significantly smaller data set the TF_IDF scores were much lower than our first algorithm.

Results

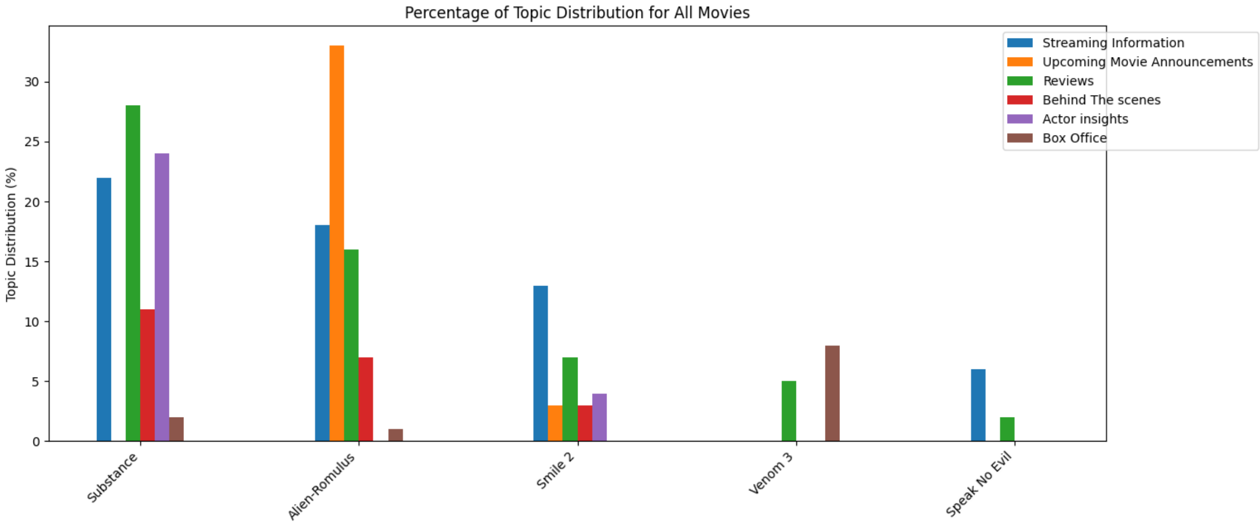
Final Typology Definitions

After performing an open coding, we ended up with the following topics for our typology:

- **Streaming Information:** This includes articles that highlight details on where and how a movie can be watched, including streaming platforms, release dates and subscription requirements. A *positive example* is "New Horror Movies Streaming This Week for Halloween 2024" by Cameron Bonomolo which provides streaming sites for each movie on the list, such as MUBI for The Substance. A *negative example* is "3 great Hulu movies you need to stream this weekend (October 25-27)" by Blair Marnell which may seem like an article providing streaming information, but the article itself is much more opinionated and critiques three movies streaming on Hulu. An *edge case* that is Aamir's personal favorite are the numerous articles highlighting how paramount+ users can stream the first 7 minutes of Smile 2 for free if they remain grinning on their webcam for all 7 minutes.

- **Review:** These articles serve as a review for the movie, and often include the plot of the film, a critique, and those that explore the theme, cultural impact, narrative and style of the movie. A review of an actor's performance is also part of this category. An *edge case* for this is articles that rank movies, as these opinionated articles usually provide some form of a critique and synopsis of the movie's plot. A *Positive example* is "Demi Moore's The Substance Earned All my Adoration By Giving Me Something I Rarely Experience As A Horror Fan" by Nick Venable which concretely critiques the movie and addresses the plot.

- **Upcoming Movie Announcements:** Articles announcing or detailing the development of new movie projects. Most articles that fit into this category are related to existing franchises, they could be sequel or reboot announcements. A *positive example* is "After the success of Alien: Romulus, Ridley Scott is working on a brand new Alien movie" by Molly Edwards announcing production of Alien: Earth, a sequel to Alien: Romulus is underway. There is no obvious negative example, given the simplicity of this category, but an interesting *edge case* is articles that speculate the release of a film or a sequel to it, such as "Smile 3 Could Completely Flip Smile 2's Gemma Twist With Brilliant Skye Replacement Option", which does not directly say that Smile 3 will be coming to cinemas, but rather is fan and director speculation at what a potential sequel to Smile 2 could be about.
- **Behind the Scenes:** Articles that explore behind-the-scenes elements of a movie, often through the perspective of actors or filmmakers. They might include photos, anecdotes, or insights into character portrayals and production details. This includes choreography, costuming, makeup, set design, relationships between the actors and director, the director's style of filmmaking etc. A *positive example* is "Alien: Romulus Used A Certain Seasonal Delicacy To Simulate Chestburster Sounds" by BJ Colangelo covers how the sound team created the sound of characters being annihilated by the antagonist (the alien) in the film. A *negative example* is "Ridley Scott's Biggest Regret Involves Two Legendary Sci-Fi Franchises" by Joe Roberts which does indeed give an insight about production of Ridley Scott's films (director of Alien: Romulus), but none of which are relevant to our selected films.
- **Actor Insights:** These articles focus more on the Actor from a recently released, and popular movie. Articles in this category detail the actor's feelings towards one of the selected movies, red carpet coverage, the actor's personal life & beliefs, and their acting journey as a whole. A *positive example* is "Smile 2 Finally Gives Naomi Scott The Role She Deserved 5 Years After $1 Billion Breakout" by Megan Hemenway which highlights Naomi Scott's casting career pre-Smile 2. A *negative example* is "For Naomi Scott, Playing a Troubled Pop Star in 'Smile 2' Meant Alternating Between Joyful Dance Rehearsals and Bloody Hallucinations" by Christian Zelko which may be perceived as part of this category but actually ends up being a review on Naomi Scott's performance in Smile 2 and would fall in the "review" category instead.
- **Box Office:** Articles focusing on how much a selected film earned at the box office. "Venom: The Last Dance,' 'The Wild Robot,' and Specialized Hits Thrive" by Tom Breuggemann where the main focus is Venom 3's box office earnings. Any article that does not mention box office returns is an automatic negative example. Articles that do mention box office returns are usually exclusively talking about a film's earnings and not much else.
- **Other Movies:** These articles may fall into one of the above categories but the focus is on a film other than the ones we have selected. A *positive example* is The First Reactions to *Nosferatu* Call It 'Pure Evil,' 'Erotic,' and 'Devilishly Bloody' by Isaiah Colbert which is indeed a review to a movie, but a movie we did not



select, as the film itself (Nosferatu), has not been released yet. Articles about The substance, Venom 3, Smile 2, Speak No Evil and Alien: Romulus are automatic negative examples.

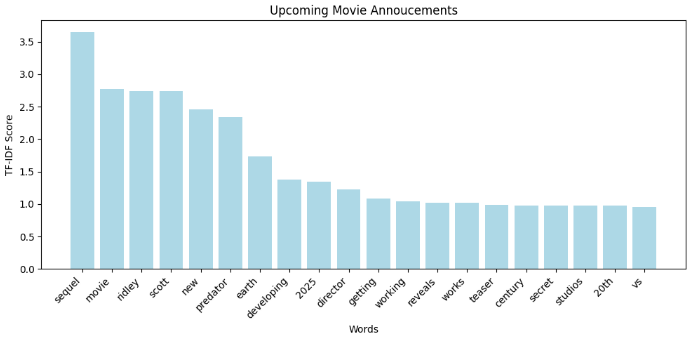
- **Unrelated Content:** These articles do not cover movies at all. A *positive example* is “Restore Our Joy in You” by Marshall Legal, which is an article about the Gospel.

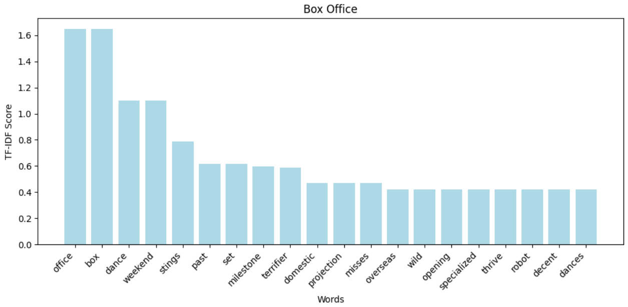
Annotation Results

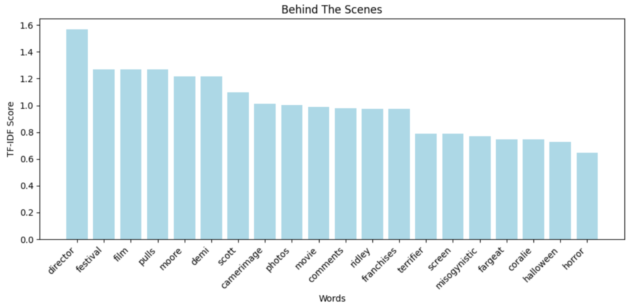
Using the above typology, we were able to manually annotate the entire dataset, and generate a valid topic distribution for our movies to accurately compare them. The ***Bar chart on page 4*** accurately summarizes these results

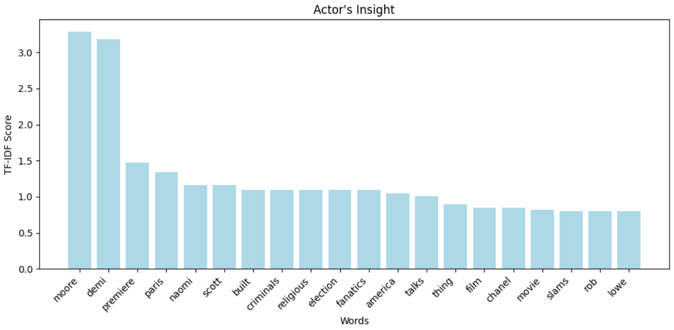
TF-IDF Results

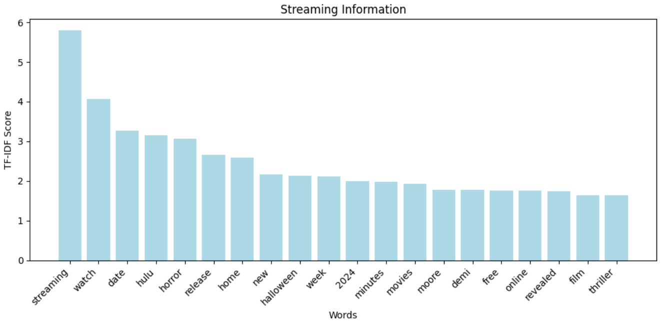
Below are the results of the tf-idf analysis for each category, as well as a cumulative result.

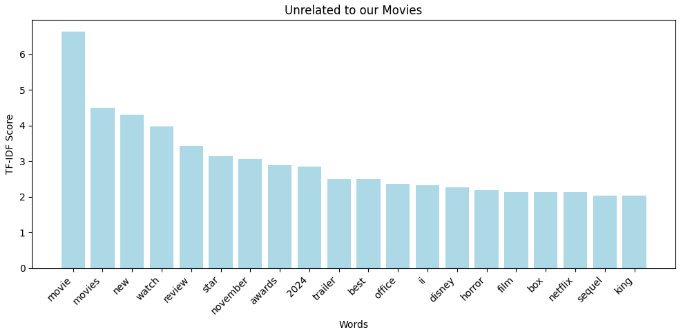


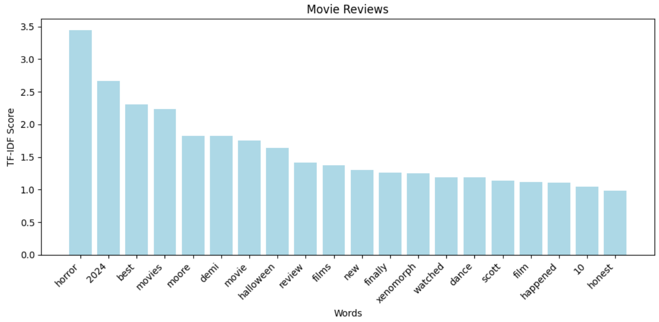


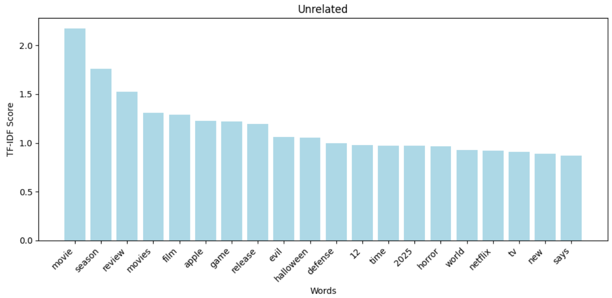


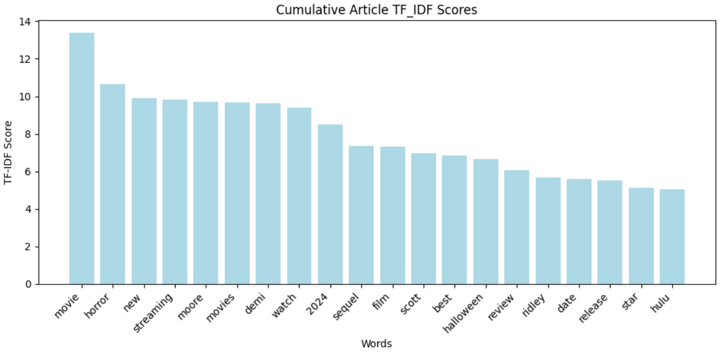












Discussion

To answer the question “*What aspect of The Substance is the main focus of the article?*”, we have to look at the topic distribution across all our films, obtained from our annotation phase.

Our analysis of this article distribution reveals several interesting insights into the media coverage received: Of the 213 articles related to our 5 chosen movies, there was a variation between the number of articles that fell under each categories from our typology. Certain categories had significantly more articles than others and led to the TF_IDF score axes being different between graphs. The most common category was Streaming Information with 59 articles or 27.7% being under this classification. Following this was Reviews with 58 articles or 27.2%, Upcoming Movie Announcements with 36 or 16.9%, Actor Insights with 28 or 13.1%, Behind the Scenes with 21 or 9.9%, and Box Office with the least at only 11 articles or 5.2% coverage.

What aspect of The Substance is the main focus?

Of the 87 articles covering our key movie The Substance, there is a pretty wide distribution of articles among all the categories. 28 (or 32.2%) of all articles are about reviewing the plot and narrative of the film, followed by 24 articles (or 27.6%) focusing on Actor’s insights, 22 articles (or 25.3%) focusing on streaming information, 11 articles (or 12.6%) on behind the scenes, 2 articles (or 0.3%) on Box office earnings, and 0 articles (0%) relating to upcoming movie announcements.

As a result of its vast category coverage with sufficient articles under most categories, The Substance plays quite a big role in ***TF_IDF scores***. Firstly, in terms of Reviews, it again accounts for 28 of 58 review articles or 48.2% and has the largest impact on this category. Consequently, terms like "Demi", "Moore", and "Horror" are among the top 6 words with the highest TF_IDF scores. Secondly, it covers 24 of 28 Actor Insight Articles or (85.7%) meaning The Substance articles have a huge impact on the unique and relevancy of words, with "Demi", "Moore", "Premiere", "Paris", being the most common words as it refers to the actress's appearance in a red carpet event. Thirdly, it accounts for 22 of the 59 (37.2%) articles regarding Streaming Information resulting the prevalence of words like "Demi" and "Moore" among the top 20 words. Fourthly, with 11 of the 21 Behind the Scenes articles being about The Substance its not surprising it has an impact, resulting in "Demi" and "Moore" once again but also "director", "festival", and "pulls" these being about the recent withdrawal of the movie from a film festival over controversies. Although it has minimal to no impact in the Box Office and Upcoming Movie Announcement scores.

Reviews were the primary focus of many articles, but there was also significant coverage of ***actor insights***. This focus can be attributed to the protagonist Demi Moore’s career comeback and the film’s narrative, which critiques Hollywood’s mistreatment of women and its unrealistic beauty standards—issues that align closely with Moore’s own experiences. The heavy coverage of actor insights reflects this compelling overlap between the film’s story and Moore’s personal journey.

The lack of articles on upcoming movie announcements was unsurprising. As a standalone film with a self-contained message, *The Substance* is not part of a franchise and does not lend itself to sequel speculation, unlike the comparison films in this study.

How much news coverage has The Substance received relative to other films released in the same month?

T*he Substance* garnered notable media attention, especially considering its status as a standalone film. Our dataset comprises 502 articles, with 213 directly related to our five selected movies. Among these, *The Substance* accounted for 87 articles, representing 40.8% of the coverage. This is particularly significant given the competition from franchise films released in the same period.

For context, *Alien: Romulus*, part of an established franchise, secured 75 articles (35.3%). *Smile 2* had 30 articles (14%), *Venom 3* 13 articles (6.1%), and *Speak No Evil* 8 articles (3.8%).

TF_IDF discussion for the other movies

Speak no Evil was the movie with the least coverage out of our 5 chosen movies, with only 8 articles out of our dataset addressing the film. Out of these 8 articles, 6 of them were related to Streaming Information (75%), while the other 2 were about reviews for the film (25%), with 0 article coverage in any other category. Additionally, due to its very low article frequency, Speak No Evil features had minimal impact on the TF_IDF scores in our graphs. However, this is not to say that there is none, as since 6 of the articles related to the movie speak about how to stream it, it boosted the score for the words Streaming, Horror, and Thriller since those words directly relate to the film.

Although *Venom 3* had the second least coverage amongst our 5 chosen movies, it also accounts for the majority of Box Office articles in our dataset. Only covering by 13 articles total, 5 of which are Reviews of the movie (38.5%) and the other 8 being Box Office information (61.5%), it can be easy to assume Venom 3 has little impact on our TF_IDF scores, but, this is very much not the case. This is because, out of the 11 articles under the Box Office category, Venom 3 accounts for 8 of them. As a result, key words like "Dance", "Weekend", "Stings", and "Milestone" all address Venom 3's successes or failures at the box offices during different time periods. While on the other hand, with only 5 of the 58 review articles being about this film, there is little impact it has on the TF_IDF score for that category.

Of the 30 articles related to "*Smile 2*", there was a notable emphasis on streaming information, which accounted for the majority of articles (13/30 or 43.3%). Not just "Smile 2", but articles from other selected movies also fall under the Streaming Information category. This implies a strong interest from the audience in watching the movie on digital platforms. Among the Streaming Information category overall, it accounts for 13 of 59 articles and has somewhat of an impact, resulting in the commonality of key words like: "minutes" which refers to the fact that you can watch the first 7 minutes of the Smile Movie trailer online. Its next highest category Reviews covers 7 articles or 23.4% of total Smile 2 articles. Though compared to the total number of Review category articles of 58 it had minimal impact. This can also be said about the remaining categories with only 3 articles in Behind the Scenes and 3 articles in Upcoming Movie Announcements. However, Smile 2 does have a lot of relevance in the Actor Insights category as even though it only accounts for 4 of the 28 articles, the words "Naomi" and "Scott" can be seen in the top 6 words, this being the name of the lead actress.

"*Alien Romulus*" stood out as a significant presence in our dataset with 75 articles about the film. 33 articles (44%) were about Upcoming Movie Announcements, 18 articles (24%) were about Streaming Information, 16 articles (21.3%) addressed Reviews on the movie and its plot, 7 articles (9.3%) were about Behind the Scenes, only 1 article (1.4%) about Box Office information, and none (0%) on Actor Insights.To keep our dataset organized and inclusive, all articles that fall under Alien Romulus, not only refer to the one specific movie but also anything related to the larger franchise. These include future sequels or instalments like Alien Romulus 2, Alien vs Predator 2, Ridley Scott's new Alien Sequel, Alien Covenant 2, and the new Alien: Earth TV series. As a result out of the 75 Alien Romulus articles, only 26 (or 34.7%) of them are directly addressing the movie itself, while the other 49 (65.3%) are about the larger franchise. As for the movie/franchise's impact on TF_IDF scores, it has minimal or no impact on Box Office and Actor Insights with only 1 (1.4%) and 0 (0%) articles respectively. However, it does have a large impact on the other categories. Firstly for Upcoming Movie Announcements, Alien Romulus covers 33 of the 36 articles in this category (91.7%) leading key words like "sequel", "ridley", "scott", "predator", "earth" popping up as a result of future projects that are coming within the franchise. Secondly, for Streaming Information covering 18 of the 59 total articles (30.5%), words like "revealed" and "hulu" are directly related to the Alien, referring to Romulus's presence on the Hulu streaming site, and the reveal of upcoming Alien projects. Thirdly, with 16 of the 58 (27.6%) review articles being here, words like "xenomorph" and "scott" can be seen relating to the name of the alien and the name of director respectively. Finally, for Behind the Scene with 7 of 21 articles in this category (33.3%), words like "scott", "ridley", and "franchises" can be seen in the top 20 words, referring actions of the Alien's creator.

Group Member Contributions

- *Hyeonmin Soh* - Data collection scripts, open coding, typology formation, data annotation, TF_IDF analysis, graph creation, report writing (“abstract”, “introduction”, “data”, “methods”, “results”), report editing

- *Aamir Shivji* - Data collection scripts, open coding, typology formation, data annotation, report writing (“abstract”, “introduction”, “data”, “methods”, “results”)

- *Jonathan Shiang* - Data collection scripts, open coding, data annotation, report writing (“methods”, “discussion”), TF_IDF analysis, graph creation