COMP 370 Final Project - Data Science Project - Movie Release

Written by
Edmund Paquin, Jacob Parent, Kevin Patel

McGill School of Computer Science
3480 Rue University, Montréal, QC H3A 2A7

edmund.paquin@mail.mcgill.ca, jacob.parent@mail.mcgill.ca, kevinkumar.patel@mail.mcgill.ca

Introduction

In this project, we studied the coverage of ten films released in the second half of 2024, primarily in October and November, with a particular focus on the film *Emilia Pérez*. The objective of our work was to determine how much coverage *Emilia Pérez* has in the media in relation to the other films, and to determine what aspects of the film are highlighted in the coverage.

Motivation

To motivate this project, we developed a particular use case in which this exploration would be beneficial. We carried out the project as if we were a media company hired by the stakeholder Netflix (the disseminator of the film *Emilia Pérez*). In particular, we found and analyzed our data as if we were to report to a marketing director at Netflix, responsible for production and improvement of film promotion campaigns. For each new Netflix film, the marketing director's goal is to oversee the creation of a promotional campaign that leads to widespread positive recognition of the film for minimal cost.

The marketing director therefore needs to know three key pieces of information after every film release: a) how much coverage the film receives in the media relative to other films; b) what aspects of the film are emphasized in the coverage; and c) which keywords appear most frequently within articles in those aspect categories. This is exactly the information which we found in the project.

With the information from a), the marketing director can determine how much to invest in the promotional material of the film to affect its relative coverage in the media. The information from b) is used to show what general aspects of the movie the media tends to focus on. It is an intermediary step to arriving at c). The information from c) reveals what exact keywords are being discussed in the media and can show if there is a particular evaluative leaning in the coverage. This data is used by the marketing director to determine whether there needs to be a change in the content of the promotional material to highlight a different area of the film.

With the above use case as a clear, practical motivation for the project, we arrived at the following guiding question: given a sample of news articles regarding movies released between October 11th and Nov 10th, 2024, what percentage of the articles focus on *Emilia Pérez*, which aspects of the film are predominantly mentioned, and what keywords are associated with each aspect?

Data

We began by collecting the articles covering the films of interest. To create our dataset, we went through several iterations to ensure quality and relevance. Initially, we used NewsAPI to retrieve as much information as possible about our specific movies in JSON format. Each entry in the raw dataset includes the following fields: name of the source, author, title, description, URL, URL to image, published date, and content of the specific article. To generate this data, we searched using movie keywords (e.g., movie titles) and applied an English-language filter. This process resulted in 522 articles initially, all related to our ten selected movies. These ten films were chosen due to proximity of release as well as amount of articles available, as is discussed in the following sections.

Data Collection Methods

We created a Python script to retrieve articles for specified movies from NewsAPI.org. There were two main limitations with this collection method which needed to be addressed. First, we were limited in the amount of daily requests allowed. The maximum number of articles we could fetch for a movie was 100. We also found that NewsAPI doesn’t allow fetching all 100 articles in a single request. Instead, it allows retrieving a limited number of articles at a time. This limited set is called a "page." We decided each page would contain up to 20 articles. So, we created a loop that went through pages to collect articles (Page 1: Fetches articles 1–20, Page 2: Fetches articles 21–40, … , Page 5: Fetches articles 81–100).

The second limitation is that we could only fetch articles up to one month old. Because of this limit, when getting articles about other films for comparison to *Emilia Pérez*, we had to focus on articles from within a narrow time frame. We decided to search for *all* the blockbuster films released in 2024 as well as the well-known (meaning popular actors, directors, etc.) films releasing in and around November. This data collection was done on November 10th. Specifically, we searched for articles from the prior month about these movies: *Blitz*, *Conclave*, *Emilia Pérez*, *Juror #2*, *Gladiator II*, *Joker: Folie à Deux*, *Wicked*, *Moana 2*, *The Apprentice*, *Megalopolis*, *Beetlejuice Beetlejuice*, *Deadpool & Wolverine*, *Dune: Part Two*, *Despicable Me 4*, *Inside Out 2*, *Kung Fu Panda 4*, *Twisters*, *Godzilla x Kong: The New Empire*, *Bad Boys: Ride or Die*, *Kingdom of the Planet of the Apes*, *The Piano Lesson*, and *Venom: The Last Dance* (we based our movie selection on the Wikipedia page 'List of American films of 2024'—see References).

Finally, we selected the nine movies with the most related articles, along with Emilia Pérez. We ended up with 10 movies, adding up to 500+ articles. These movies have release dates falling mostly within October to November.

Recognizing Bias in the Data

Naturally, there was some inherent bias in this process of article selection. Importantly, the number of articles available about a particular film is a function of the amount of time it has been released; that is, very recently released films may be prone to having fewer related articles in our data collection period, and the same is true of films released many months ago. This bias was not possible to correct for, and it will be considered during data interpretation. Additionally, the selections are biased towards American blockbuster films, as these films naturally have far more media coverage than lesser-known films.

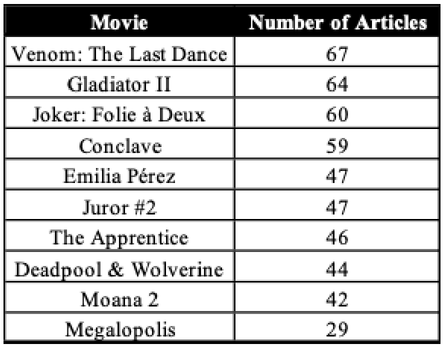
Data Cleaning Steps

The data had to be cleaned and organized before being used for analysis. We found duplicate entries in the dataset (e.g., articles with the same url). These were removed, reducing the dataset to 511 articles. Then, we cleaned the JSON data by selecting only the following fields to create our CSV file: Movie (keyword used for the search), Article Title (the title of the news article), Description: (a brief summary of the article), URL (the link to the full article), and Content (the start of the text content of the article).

Secondly, we had to filter out irrelevant articles. We identified six articles that were unrelated to the movies, despite using appropriate filters: Regarding the film *The Apprentice*, there were three articles discussing Donald Trump’s TV show instead of the movie. Regarding *Megalopolis*, there were two articles about a ship named "Megalopolis" and China’s megalopolis of Guangzhou. Finally, regarding *Venom: The Last Dance*, there was one non-English article despite applying an English-language filter.

Final Dataset Breakdown

The final cleaned and organized dataset consists of 505 articles, distributed across the 10 movies as follows. This structured dataset serves as the foundation for our analysis, ensuring that only relevant data is included.



Movie	Number of Articles
Venom: The Last Dance	67
Gladiator II	64
Joker: Folie à Deux	60
Conclave	59
Emilia Pérez	47
Juror #2	47
The Apprentice	46
Deadpool & Wolverine	44
Moana 2	42
Megalopolis	29

Methods

Open-Coding and Typology/Annotation

After collecting our set of articles, we started with a joint open-coding process on the first 200 articles. We knew we would need two labels for each article, to answer the first two sub-questions as follows:

For the first sub-question, “Given a sample of news articles regarding movies released between October 11th and Nov 10th 2024, what percentage of the articles focus on Emilia Pérez?”, we initially devised a label involving two categories: either Emilia Pérez was mentioned or it was not.

For the second sub-question, “What aspects of the film are most predominantly mentioned?”, we used our general knowledge of movies and the industry to draft an initial typology. This included labels (categories) like the movie’s plot, visual aspects, acting cast, box office performance, media/public reception, cultural impact, and other miscellaneous topics.

Refining the Typologies

While coding the first 200 articles, we quickly realized that the first label needed a more robust typology to catch nuanced categories. Some articles only briefly mentioned Emilia Pérez while focusing on something else entirely. It did not make sense to stick with a black-and-white approach, so we expanded to four categories: *Focus on "Emilia Pérez", Focus on Other Movies, Mentions "Emilia Pérez", and Does Not Mention "Emilia Pérez".*

To handle ambiguous cases, we created a “focus rule” in the taxonomy guide. If the article title didn’t clearly indicate its focus, we reviewed the content to decide: Assign to Category 1 if *Emilia Pérez* was clearly prioritized, or to Category 2 if another movie was clearly prioritized; assign to Category 3 or Category 4 if Emilia Pérez was only mentioned briefly or not at all. For vague titles (e.g., "2024’s Best Netflix Movies"), we checked the opening content to determine if Emilia Pérez was included. If not, the article was classified as Category 4.

The resulting typology allowed us to label every film descriptively regarding to what extent it involved *Emilia Pérez*.

Final Typology for the Second Question

For the second label (regarding movie aspects), we refined the categories based on our coding process and finalized the following typology:

Production and Behind-the-Scenes: This category includes all articles focusing on the creation process of the film. This can involve its soundtrack, filming process, costume creation, budgeting, and so on. Also included are interviews with directors and producers that focus on the creation process or otherwise non-public elements of the film. Finally, articles on post-credit scenes are included here. Positive example: “Hugh Jackman's Alternate Costume In Deadpool And Wolverine Cost $100,000”. Negative Example: “Ryan Reynolds Says Marvel Execs Are “Obsessed” With Channing Tatum’s Gambit”. The latter article focuses more on cast and crew.

Plot and Themes: These articles focus on the plot of the film, as well as any general thematic content related to it. Articles focusing on the script of the film are included in this category. The articles may include an explanation of the script, a question about the script, or an explanation of the script. Positive example: “Venom: The Last Dance Director Explains The Film's Ending”. Negative example: Making the Big Hair Big Enough in “The Apprentice”. The latter belongs in the production category.

Cast and Crew: This category includes articles focusing on the cast and crew. The term “cast and crew” refers to all individuals involved in the film in any way, meaning that those on- and off-screen are included (actors, directors, producers, etc.). This category includes articles discussing a particular cast or crew member's role in the film. It also includes direct interviews with a cast or crew member about subjects not related to the creation of the film itself (such interviews belong in the production and behind-the-scenes category). Positive example: “The ‘Juror #2’ cast still can’t believe they got to work with Clint Eastwood”. Negative example: “Paul Mescal Learns How to Fight in Training Featurette for 'Gladiator II'. The latter belongs in the production category.

Reception and Reviews: Included in this category are articles providing a general consensus of the public’s reception of the film, or editorial articles directly providing a review. Positive example: “‘Gladiator II’ First Reactions Call Ridley Scott’s Sequel “Unhinged”, “Deliciously Cinematic And Machiavellian”: “Can’t Believe Ridley Pulled This Off””. Negative example: “Moana 2: Special Look And New Poster Have Been Released.” Notable edge case: “TAXI DRIVER Writer Paul Schrader Offers His Brutally Honest Thoughts on JOKER: FOLIE À DEUX”. This involves cast and crew, but focuses on a review of the film, so it belongs here.

Marketing and Promotion: This category includes articles focusing on persuading readers to watch the film or buy film related merchandise. Also included are articles focusing on awards received by the film. In general, the category includes articles whose focus is to promote the film either directly or indirectly. Positive example: “Gladiator II: 6 New Character Posters And A Score Featurette”. Negative example: “Venom 3 Features A Tiny Ghostbusters Easter Egg You Probably Missed.” While this article does serve to promote the film, it focuses more on a plot element and belongs in plot and themes.

Cultural Impact: Included in this category are articles discussing events or ideas in society that are inspired, affected, or catalyzed by the film. These events may not directly involve the film or any of its personnel, but can be tied back to the film in some way. Positive example: “Martha Stewart Says Ryan Reynolds is Not so Funny in Real Life” . Negative example: “Moana 2: First Single & Full Tracklist Released by Disney”. This fits more in the marketing and promotion category.

Box Office Performance: Included are articles focusing on the amount of ticket sales at box office, or any focus on the financial performance of the movie. Positive Example: ““Moana 2” Sets Record First Day Ticket Sales”. Negative Example: “Hugh Jackman's Alternate Costume In Deadpool And Wolverine Cost $100,000.” Despite mentioning budget, this article belongs in the production and behind the scenes category.

Broader Industry Trends: This category includes articles that address the effect of the film in the broader context of the film industry. Articles discussing how other movies are affected, what films may come next, releases of a new related series, and so on are all included. Any article focusing on how a film has changed any paradigm within the film industry is to be included. Finally, articles focusing on events that occur within the film industry are included, such as an article about a particular studio acquiring the rights to some film. Positive Example: “2025 Will Officially Be The Year Of Pedro Pascal.” Negative Example: “Are Netflix Normies Ready for Emilia Pérez?”. This example has a scope outside of an individual film, which suggests it belongs in this category; however, it focuses more on cultural impact than the film or film industry.

With the two taxonomy guides finalized, we were able to annotate the entire dataset of articles. For the annotation process, we used a CSV file with the following column names: Movie, Article Title, Description, URL, Content, Movie Topic, and Article Topic. As mentioned earlier, when we encountered ambiguous cases, like when the title or description wasn’t clear enough, we clicked on the URL to read the full article. Interestingly, sometimes, the URLs led to videos, like TV show segments, rather than traditional written articles. This made the annotation process even more interesting, as we weren’t just working with text-based content but also with multimedia elements!

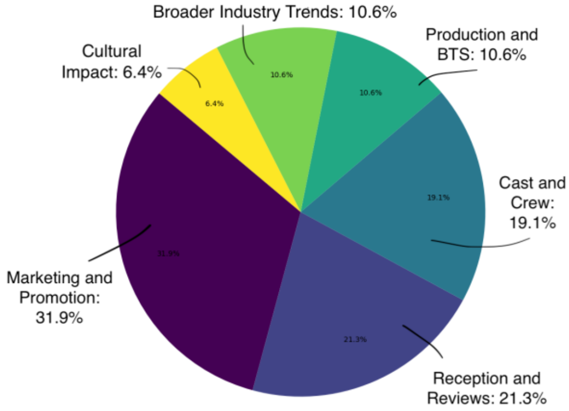
TF-IDF Analysis

Finally, having made annotations to address the first two sub-questions of the guiding question, we were left only with the sub-question: *“What keywords are associated with each aspect [of Emilia Pérez mentioned in the articles]?”* To answer this question, we conducted a TF-IDF analysis of the words present in the title and description of each article that focuses on or mentions Emilia Pérez. We included the default list of English stop words from the Python scikit-learn module to filter to maximally informative results. The results of this analysis, as well as the results pertaining to the other two sub-questions, are in the following section.

Results

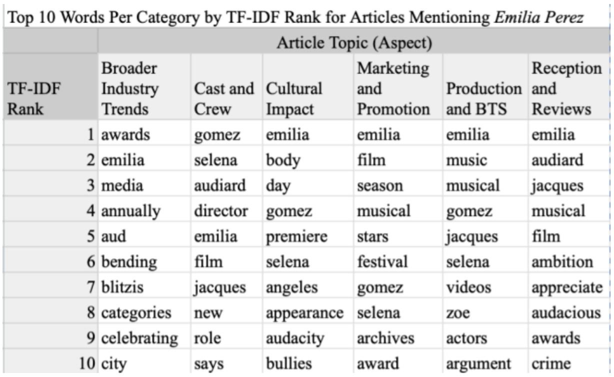
From the first feature annotation (the “movie topic” label), we were able to determine the proportion of the articles that mention *Emilia Pérez.* We counted the number of articles with labels “focus on Emilia Pérez” OR “mentions Emilia Pérez”, and determined that 9.3% of the articles involve Emilia Pérez, compared to the maximally-appearing film of Venom: The Last Dance at 13.6%, and the minimally-appearing film of Megalopolis at 5.7% of articles.

From the second feature annotation (the “article topic” label), we were able to determine the breakdown of aspects discussed in articles pertaining to each film. Using a combination of both labels, we were able to select only the articles relevant to Emilia Pérez and then determine the distribution of aspects discussed within only those articles. The following chart indicates that data.



Above: Distribution of Article Topics for Emilia Pérez.

Finally, the TF-IDF analysis run on the articles relevant to Emilia Pérez returned the following words.



Top 10 Words Per Category by TF-IDF Rank for Articles Mentioning *Emilia Perez*

	Article Topic (Aspect)					
TF-IDF Rank	Broader Industry Trends	Cast and Crew	Cultural Impact	Marketing and Promotion	Production and BTS	Reception and Reviews
1	awards	gomez	emilia	emilia	emilia	emilia
2	emilia	selena	body	film	music	audiard
3	media	audiard	day	season	musical	jacques
4	annually	director	gomez	musical	gomez	musical
5	aud	emilia	premiere	stars	jacques	film
6	bending	film	selena	festival	selena	ambition
7	blitzis	jacques	angeles	gomez	videos	appreciate
8	categories	new	appearance	selena	zoe	audacious
9	celebrating	role	audacity	archives	actors	awards
10	city	says	bullies	award	argument	crime

Discussion

Key Topics for *Emilia Pérez*

The analysis of articles about *Emilia Pérez* revealed that certain aspects of the movie received more attention than others. The most prominent article topic was "Marketing and Promotion," which made up **31.9%** of the articles. This focus on marketing highlights how the film relied heavily on pre-release buzz, especially through Selena Gomez. Keywords like "season," "stars," and "festival" suggest that the promotional strategy relied on leveraging Selena Gomez’s popularity and the premise of the film as a musical to attract attention. This matches common trends in the film industry, where high-profile celebrities are often used to generate early hype.

The "Reception and Reviews" category was the second most discussed, accounting for **21.3%** of the coverage. Words like "ambition," "audacious," and "appreciate" suggest that most articles framed the film positively, focusing on its creativity and bold approach. The inclusion of the keyword "crime," however, hints at mixed reactions, as blending the crime and musical genres may have divided critics and audiences. The word "audacious" really fits this genre mix because it shows how the film took a big risk by blending two very different styles—crime and the lively nature of a musical. While this boldness drew praise from some for its originality, others may have found the combination difficult to connect with. This tension between the film's ambition and what audiences expected most likely played a big role in how people reacted to it.

The "Cast and Crew" topic, which accounted for **19.1%** of the coverage, focused mainly on Selena Gomez. TF-IDF results highlighted her name ("Gomez" and "Selena") as key terms, showing that her involvement played a major role in shaping the media narrative. Jacques Audiard, the director, also received notable attention, with keywords like "Audiard" and "director" emphasizing his contribution to the film’s artistic identity. Together, these factors positioned *Emilia Pérez* as both a celebrity-driven project and one with a strong creative vision. This mix of star power and a well-regarded director likely helped the film stand out in a competitive media landscape. It also reflects how the entertainment industry often relies on well-known personalities to generate interest in smaller or more experimental films.

The "Production and Behind-the-Scenes" topic (10.6%) highlighted the creative work involved in making the film, especially its musical elements. ords like "musical," "videos," and "music" show the emphasis on the artistic processes behind *Emilia Pérez*. Similarly, the "Broader Industry Trends" category (10.6%) explored how the film fit into larger conversations about the movie industry. Keywords like "awards" and "media" suggest these discussions focused on its potential for recognition and its connection to current trends, such as the rise of celebrity-driven projects and experimentation with genres. Lastly, the "Cultural Impact" topic (6.4%) framed the movie as a significant cultural event, particularly during its premiere. Words like "audacity" and "premiere" highlighted the film's boldness and artistic ambition, which played a key role in shaping how it was received.

Comparison to Other Films and Media Perception

When compared to other films in the dataset, *Emilia Pérez* stood out because it focused more on artistic and cultural themes than financial success. Unlike big hits like *Joker: Folie à Deux* and *Venom: The Last Dance*, which were all about “Box Office Performance” and used words like “million,” “weekend,” and “office” in their promotion, *Emilia Pérez* highlighted creativity and a unique approach. Similar to *Gladiator II* and *Moana 2*, it relied on pre-release campaigns to gain attention but leaned more on originality than on meeting commercial expectations.

The media coverage of *Emilia Pérez* was mostly positive, with words like “appreciate,” “awards,” and “ambitious” showing praise for its vision. However, terms like “crime” and “musical” suggest that its mix of genres caused some mixed reactions, as it didn’t connect with every critic or viewer. This highlights a common trend in the industry: independent films like *Emilia Pérez* are often judged by their creativity, while blockbusters are mostly evaluated by how much money they make.

Limitations and Future Directions

There are some limitations to this analysis that need to be noted. First, NewsAPI.org, which primarily gathers articles in English, provided the dataset for the study. This implies that viewpoints from non-English media were left out, particularly in areas where Emilia Pérez might have had a different reception. Due to this restriction, some media outlets might have been overrepresented, which could have skewed the analysis. Secondly, because the dataset only spanned a small time frame, it may not accurately reflect how opinions about Emilia Pérez changed over time, especially after its release or premiere.

To address these issues, future work could take several directions. Exploring how *Emilia Pérez* was discussed in non-English media would offer a more diverse perspective and help understand its global reception. For instance, regions with different cultural contexts might have unique takes on the film’s crime musical genre. Sentiment analysis could also be a helpful way to measure the tone of the coverage, making it easier to see whether reactions were mostly positive, negative, or mixed. Also, using a larger dataset with articles from a longer time period could help us understand how media coverage changed over time, especially after the film reaches a wider audience. These steps could lead to a more comprehensive understanding of the film’s impact and reception in different contexts.

Conclusion

Recall that the project was motivated by the Netflix marketing director example use case. The question we sought to answer was: given a sample of news articles regarding movies released between October 11th and Nov 10th, 2024, what percentage of the articles focus on *Emilia Pérez*, which aspects of the film are predominantly mentioned, and what keywords are associated with each aspect?

We found and analyzed data in a manner capable of answering this question. First, we found that *Emilia Pérez* appears in a middling 9.3% of the articles collected. This percentage would give an idea to a marketing director of the dominance of the film in the media. Secondly, we found that the most commonly discussed aspects of the film were “Marketing and Promotion”, “Reception and Reviews”, and “Cast and Crew.” Thirdly, we conducted analysis within the above categories to find the most-mentioned keywords within each aspect, returning useful information to a marketing director. These keywords included the exact actors of interest, the general sentiments surrounding the film, and the results of promotional strategies. This analysis thoroughly answers the guiding question and satisfies the proposed use case.

Group Member Contributions

The project went smoothly, thanks to the strong organization of our team. From the very beginning, we prioritized collaboration and made sure to divide the workload equally. We relied on tools like GitHub to share our code, plots, and data, allowing us to build on each other’s contributions without missing a beat. Like any successful team project, communication was key, and we kept each other updated through a lively and ongoing Discord chat. This setup ensured we were always on the same page and could quickly address any questions or issues as they came up.

Regarding contributions, everyone participated in defining the research questions, laying the foundation for the entire project. Jacob took charge of the data collection process, making sure we had all the materials needed to move forward. Data annotation was a collective effort, with everyone contributing equally to tag and categorize articles while reviewing each other’s work for consistency. Edmund focused on developing the use case and conducting the TF-IDF analysis. Kevin rounded things off by interpreting the results and leading the discussion, ensuring that the findings were clear and well-explained.

Overall, our teamwork and open communication made this project not only efficient but also enjoyable. Each member brought their unique strengths to the table, and together, we're proud of what we accomplished.

References

List of American films of 2024. (2024, November 29). In *Wikipedia*. https://en.wikipedia.org/wiki/List_of_American_films_of_2024