COMP 598 Final
Project 2: Movie Release

December 13, 2021

Asu Simla Ayduran 260822715
Zara Horlacher 260785813
Ariela Guigui 260768760

Introduction

For film media companies, quantitative statistics on the public perception of a movie is invaluable information. Generating appropriate and relevant data, however, presents a significant challenge. In this paper, we use data science tools to uncover discourse and favorability surrounding the film Dune, a Sci-fi movie released in 2021. First, we perform data collection by retrieving 1,000 tweets discussing Dune. Second, data annotation is conducted through open coding, where we obtain 7 most commonly discussed topics in the data-set and manually assign one such topic to each tweet. In this step, we also attribute a sentiment to each tweet, reflecting how positive, negative, or neutral the tweet considers the film. Finally, we perform an analysis of the annotated data-set by computing the 10 words in each topic with the highest TF-IDF score as well as each topic's most frequent sentiment. Our findings show that the most salient topics discussed around the film are Cast, Screenplay, Cinematography, Production, Director, Soundtrack, and Fans, with Cast being the most frequently mentioned. Further, the response to Dune has been generally positive, whereby the aspect consisting of the most positive ratio is the soundtrack.

Data

Our dataset consisted originally of 1,000 tweets related to the movie Dune. To collect these tweets, we first created a twitter developer account and started making lists of key words and hashtags which were curated to target tweets with a high likelihood of being related to our movie[1]. At first we thought of keywords such as "Dune" and "Dune Movie", as well as "Frank Herbert" and "Denis Villeneuve", the author and director of Dune. We also looked at a combination of the word "movie" and "Zendaya" or "Timothee Chalamet" - Dune's principle actors to increase the likelihood that the tweets about the actors were relevant to our movie choice.

To perform tweet collection, our code consisted of pandas library for data manipulation, datetime library for the three day window, requests library as well as json library to output our results to a json file. We also decided to save results to a csv file, as a json file was better for the annotation section and a csv file would be better for the computation section. One issue was that the maximum tweets per request was 100, so we had to do a few requests to get to the required 1000 tweets.

When scraping, we could only accumulate 623 unique tweets in the span of 3 days, short of 377. We therefore needed to diversify the list of keywords and find other areas within the film to touch upon, which led us to search for additional popular twitter hashtags related to the movie. Ultimately, we added some extra main characters, such as Paul Atreides, Lady Jessica, Chani, Vladimir Harkonnen, Duncan Idaho and Stilgar to raise our number of tweets to 1117. This served as the finalized data set after tweet collection and was the baseline for which we performed data annotation.

Methods

Data Annotation

To conduct annotation of the tweet dataset, we started by open coding. This involved looking through a subsection of the already collected data, 200 tweets, and deriving 7 most commonly discussed topics. These topics are: Cast, Screenplay, Cinematography, Production, Director, Soundtrack, and Fans.

After our open coding, we revisited the entire dataset and performed manual annotation, attributing one of the mentioned topics to each tweet. We also assigned a sentiment to each tweet, taking up a value of 1, -1, or 0, representing positive, negative, and neutral respectively. This attribute reflects how favourable the writer of the tweet sees the film. Since the goal of this project is to explore Dune's public perception, each tweet's sentiment is therefore crucial to what insights we can draw from data analysis. Thus, after performing data annotation, our dataset consisted of three columns: tweet (tweet content), topic, and sentiment.

Data Analysis

The methodology of data analysis was two-fold: first, we computed 10 highest TF-IDFs per topic, then we generated each topic's most frequent sentiment.

To obtain greater insight into the favorability of a film through tweets, it is applicable to explore the most commonly used words. In doing so, we use term frequency-inverse document frequency (TF-IDF), which assesses how

relevant a particular word is to a collection of texts. The equation of TF-IDF is as follows:

$$tfidf_{i,j} = tf(i,j) \times log(\frac{N}{df_i})$$

Whereby, *tf* is the number of times a word *i* occurs in a document *j*, *N* is the total number of topics, and *df* is the number of topics that use word *i*. Moreover, we computed the 10 words in each topic with the highest TF-IDF score. This was achieved by first collecting all words belonging to tweets of a particular topic, removing any stop words[4], and then generating the TF-IDF scores of words in each topic. The reason for removing stop words was to prevent common filler words from being considered; for instance, the word "the" is very commonly used, however its frequency does not necessarily provide insight into a writer's sentiment.

We also computed the sentiment most commonly associated with each topic. This was performed by counting the number of positive, negative, and neutral sentiments per topic, and finding which was most frequent. Such results would allow us to see the favorability of different aspects about Dune. Further, we also calculated the ratio of positive tweets per category by dividing the number of positive tweets in a topic by the total number of tweets (positive, negative, and neutral) of that topic.

All discussed computations in data analysis were performed by using Python and other common technologies, like Pandas library.

Results

Topics: Definitions and Examples

Before delving into our computed results, we will unpack the definitions and examples of each topic resulting from open coding, as mentioned in the *Methodology section*. This subsection of results showcase the 7 most salient topics discussed around Dune.

1. Cast : Any tweet related to the movies cast members and characters that they play.

"Timothée Chalamet spent that whole movie simping for Zendaya"

2. Screenplay: Any tweet related to the movies script and its relation with the book.

"@JamesCWarne I loved reading Dune, especially the first book and really enjoyed watching the movie. Can't wait to see it again. I am glad I read the book first though."

3. Cinematography: Any tweet related to the movies scenes and visuals.

"A Tribute To The Wide Shots of Dune Wideshots Cinematography DenisVilleneuve Johann"

4. Production: Any tweet related to more than one aspect of the movie and any tweet that just simply state the viewer's opinion about the whole movie.

"@moonchildinmono i wonder if they'll see Dune at some point which recently been in SK movie theaters. that movie was so good"

5. Director: Any tweet related to the movie's director, Denis Villeneuve.

"@guardianfilm If there was one great thing to come out of this film, it was this Denis Villeneuve's direction."

6. Soundtrack: Any tweet related to the movie's score or music.

"I just saw Dune in theaters and I am blown away by just absolutely everything, especially the music and I don't think I will ever forget what I just saw and I cannot recommend it more DuneMovie"

7. Fans: Any tweet related to the people who saw the movie more than once and artwork or merchandise created by the fans or for the fans.

"Got I want to go see dune again Im a bit over halfway through the original book and ive seen the movie 4 times... a 5th wouldnt hurt"

Computations: TF-IDF and Sentiments

As for our generated results, we first computed the number of tweets collected per topic. This provides insight into which aspects of Dune are most frequently discussed relative to others, which is useful for the media company to know what areas of the film are generating buzz. As seen in Figure 1, out of our 513 tweets, 176 were related mentioned Dune's cast, making "Cast" the most frequently discussed topic. By contrast, "Soundtrack" is the least mentioned topic amongst the dataset.

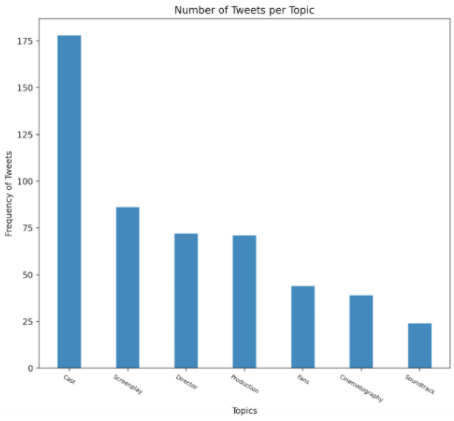

Figure 6: Graph of tweets by topic

Next, we calculated the 10 highest TF-IDF scores per topic. Looking at Figure 2, we see that within the top 10 words of tweets pertaining to topic "Cast", half of the words are names of celebrities part of the cast members of Dune.

Tweets related to "Screenplay", on the other hand, mostly mention the names of Dune's author, Frank Herbert, or words describing literature (Figure 3).

For the topic "Cinematography", the word with the highest TF-IDF score is "imax" (13.9) while the lowest is "amazing" (6) (Figure 4).

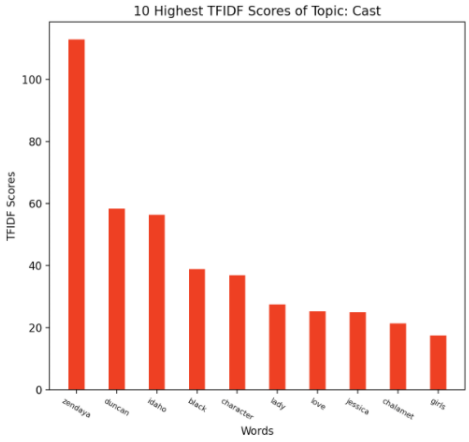

Figure 2: Tweets pertaining to the ”Cast” topic

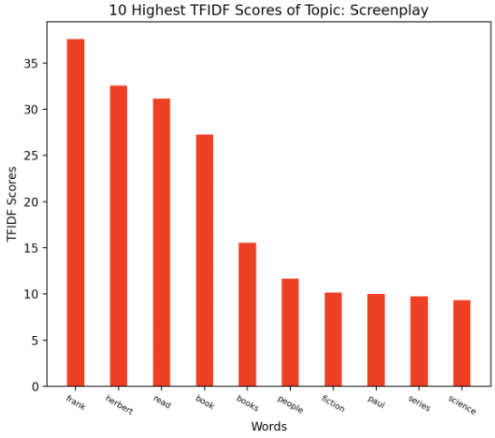

Figure 3: Tweets pertaining to ”Screenplay” topic

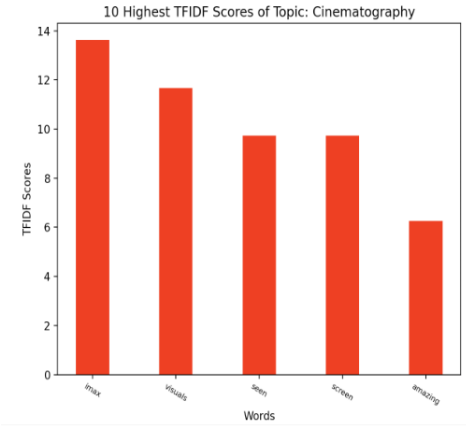

Figure 4: Tweets pertaining to ”Cinematography” topic

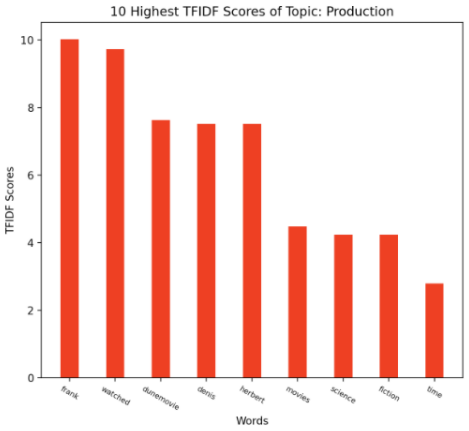

Figure 5: Tweets pertaining to ”Production” topic

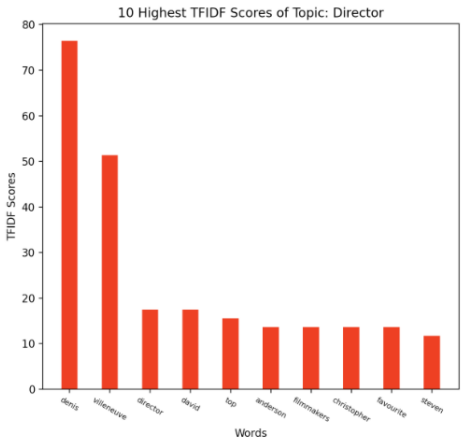

Figure 6: Tweets pertaining to ”Director” topic

Furthermore, Figure 5 reflects most frequently used words in tweets of the category “Production”, which includes the names of both author and director of Dune.

For those belonging to topic “Director”, we can observe various director names, with “Denis” (78.2) and “Villeneuve” (52.6) having the highest TF-IDF scores (Figure 6).

Tweets of the next category, “Soundtrack”, mostly contain words related to music (Figure 7), such as “music” itself which has a score of 20.

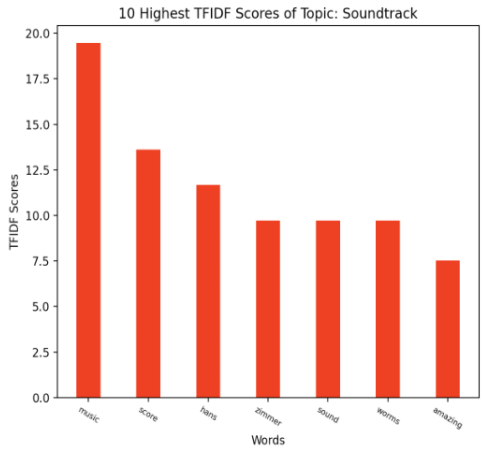

Figure 7: Tweets pertaining to "Soundtrack" topic

Finally, topic "Fans"' greatest TF-IDF word is "timotheechalamet", as per Figure 8.

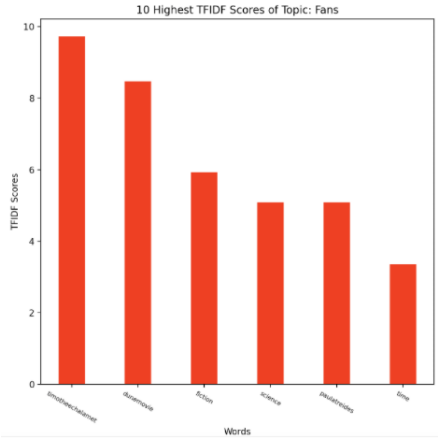

Figure 8: Tweets pertaining to "Fans" topic

Moving on, we extracted each topic's most frequent sentiment, meaning the sentiment of tweets repeated most frequently per topic. Upon performing computation, we found the mode of each topic to be 1.0, reflecting positive reactions for every category. Furthermore, as seen in Figure 9, the soundtrack achieved the highest ratio of positive feedback (73 percent), versus screenplay, which has the least positive ratio (57 percent). That being said, all topics were still discussed in a more positive than negative manner.

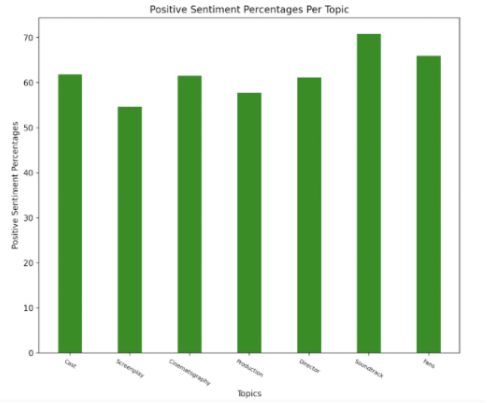

Figure 9: Positive and Negative Feedback Chart

Discussion

Our findings show that the most salient topics discussed around the film Dune were Cast, Screenplay, Cinematography, Production, Director, Soundtrack, and Fans. After computing the relative engagement with these topics, we observed that "Cast" was the most frequently discussed amongst our dataset of tweets. This may be due to the various famous and critically acclaimed actors and actresses part of the cast, which accelerated discourse surrounding the film. For instance, Zendaya and Timothee Chalamet play two of the main roles in Dune and have both won numerous awards. As such, they already have an established fan base, resulting in more online interaction once their engagement with the film is announced. This further explains why the words with the highest TF-IDF score in the topic of "Cast" are the names of actors (Figure 2), mentioning both names Zendaya and Timothee Chalamet precisely.

By contrast, the category with the least number of tweets is "Soundtrack". This is a surprising discovery, since the soundtrack of Dune was curated and scored by Hans Zimmer, a notable composer responsible for the soundtrack of other critically acclaimed movies such as Inception. What this indicates is that there is the least amount of buzz surrounding the music of Dune, relative to its other aspects. We can theorize this is because other elements of the movie were even more inspiring and worth discussion, such as the cast or director, leaving the soundtrack less noticeable.

Next, we discuss the top 10 TF-IDF scoring words in each tweet category.

1. Cast : The first three words that have the highest TF-IDF are "zendaya", "Duncan" and "Idaho". Zendaya is an actress that plays one of the main characters in the movie while Duncan Idaho is a character played by the actor Jason Mamoa. From this result, we can conclude that Zendaya and Jason Mamoa's performances were most appealing to the viewers compared to other members of the cast. We can also infer that their presence in the movie generated the most publicity, which is useful for media companies to further strategize PR. For instance, to maximize engagement,

they can continue to use these two particular actors to market the film, since they know Zendaya and Mamoa have a strong online presence.

2. Screenplay: The first three words that have the highest TF-IDF are "frank", "herbert" and "read". The Dune movie is adapted from a book with the same name and Frank Herbert is the writer. Therefore, we can conclude that viewers of the movie talked about the original book as well and frequently compared it to the screenplay or movie script.

3. Cinematography: The first four words that have the highest TF-IDF are "Imax"," visuals" ,"seen" and "screen". IMAX is a type of screen that shows movies with detailed visuals. Therefore, for the tweets related to this, we can conclude that most viewers watched the movie on an IMAX screen for detailed visuals. This provides significant insight into how Dune's media team can further market the movie's cinematography, by advocating for IMAX viewings.

4. Production: The first three words that have the highest TF-IDF are "frank", "watched" and "dunemovie". From these words, we conclude that viewers who stated their opinion were also discussing the book since Frank Herbert is the author of the book with the same name.

5. Director: The first three words that have the highest TF-IDF are "denis", "villeneuve" and "director". Canadian director Denis Villeneuve directed the Dune Movie. Consequently, viewers who tweeted about this topic mostly spoke about their opinion about Villeneuve's adaptation.

6. Soundtrack: The first four words that have the highest TF-IDF are "music", "score", "hans" and "zimmer". Hans Zimmer Composed the score of the movie which is the soundtrack. Therefore, we can conclude that viewers wanted to discuss the composer and his work in the context of Dune. This shows us that having a well-known composer, Hans Zimmer, generates online discussion and increases publicity of a film. Some viewers tweeted that they are streaming the soundtrack from streaming services.

7. Fans: The first three words that have the highest TF-IDF are "timotheechalamet", "dunemovie" and "fiction". From these words, we can conclude that the fan base of the movie were mostly attracted to the actor Timothee Chalamet and the fiction element. This is an unsurprising discovery, since Chalamet has generated a large amount of online traffic over the past year and was trending on twitter prior to Dune, creating a great amount of publicity.[5] We argue that his fan base is therefore very active on twitter, resulting in his name being the most mentioned under topic Fans.

Finally, we analyze how positive/negative the response to the movie is. As mentioned in the results section, the most frequent sentiment of each topic is positive. This further implies that the overarching reception of the film is positive, which can be confirmed by the ratings of 83 percent on Rotten Tomatoes and 8.2/10 on IMDB.[2,3] What's more, the topic of "Soundtrack" received the most positive sentiment ratio of 73 percent, indicating that the music score of Dune had the least negative response, relative to the amount of tweets under its category. We posit that its high ratio is due to Soundtrack's small number of tweets, in fact the least out of all topics (See Figure 1), which also imply that it was the least discussed. Nevertheless, the ratio still reflects that the music was not very controversial, meaning it did not result in massive discussion, but the discussion that was had was generally positive.

Challenges and Limitations

The primary challenge this project faced pertains to the dataset size. While manually annotating our data, we realized that some collected tweets were not related to the movie. One reason for this was due to the keyword selection. For example, "Zendaya" was one of the filters used to collect our tweets, as she is one of the actresses of Dune. However, she is also simultaneously playing in another movie, thus many of the collected tweets were in fact discussing that movie instead of Dune. For instance, while the tweet "I would pay good money to see Tom Holland and Zendaya in a Fred Astaire/Ginger Rogers typa movie" contains the keyword "Zendaya", it has no relationship to the movie Dune itself. As a result, even though we collected and filtered a total of 1117 tweets, almost half of the tweets were removed due to lack of applicability. Consequently, our dataset was much smaller and may have resulted in less accurate results.

To produce more precise statistics, we could explore other complex mechanisms to collect only relevant tweets. One such method could be classifying and filtering for tweets we want using Machine Learning algorithms such as Natural Language Processing. Nevertheless, such classification techniques are beyond the scope of this project and can be deferred for future works.

Group Member Contributions

Ariela started the project off by completing the data collection with filtering and formatting the document in the LaTeX Template. Zara performed computations for the results section, including the TF-IDF and sentiment calculations. Simla developed topics and annotated the collected data. We all worked on our respective parts of the document and all helped each other edit the whole document and make sure it meets all specifications.

References

1. 'Display Purposes - Best dunemovie Hashtags for Instagram, TikTok, YouTube in 2021'. Accessed 13 December 2021. https://displaypurposes.com/hashtags/hashtag/dunemovie.
2. Dune. Accessed 13 December 2021. https://www.rottentomatoes.com/m/dune_2021.
3. 'Dune (2021) - IMDb'. Accessed 13 December 2021. https://www.imdb.com/title/tt1160419/.
4. Gist. 'Larsyencken's Gists'. Accessed 13 December 2021. https://gist.github.com/larsyencken.
5. The Diamondback. 'Why Is the Internet so Obsessed with Timothée Chalamet?', 29 April 2020. https://dbknews.com/2020/04/29/timothee-chalamet-call-me-by-your-name-gen-z/.