An Analysis of the 2021 Dune Film Release and Reception

Kate Kim (260839625), Patrick Janulewicz (260929866), Navneet Kaur (260779742)

Abstract

This project aimed to investigate various discussions taking place on Twitter around the movie Dune. Given the boundless talk happening around the movie, we chose to explore and review the buzz on the 2021 film adaptation of Dune, a well-known science fiction novel by Frank Herbert. First, data (tweets) was collected from December 4 - December 7, 2021. Then, upon reading through the first 200 tweets 3-8 topics were chosen to categorize the entire dataset. Although there were many potential topics that could have been chosen, the final six categories were generic reactions and reviews, specific commentary on the film, reference or comparison to the book, expectations and desires of the film, anecdotes/stories involving the film, and promotions/merchandise/ads. After data wrangling and processing, the dataset was manually annotated with one of these categories as well as a sentiment label of positive, negative, or neutral. The final results revealed that the movie was very well received by the masses. As seen in Figure 2, of the 1000 tweets that were analyzed, 482 of them expressed positive opinions on the movie, 356 of them had neutral judgements while only 162 of them had a negative experience. Categorical analysis of the tweets also displayed that the generic review category had the most amount of tweets followed by anecdotal stories and in-depth movie critiques. The TF-IDF score of the data was then calculated in two different ways in order to obtain the most frequently used words associated with each category and sentiment. The first approach used a manually written script while the other one used a TF-IDF vectorizer provided by the python library, scikit-learn. While the words outputted by the manually written TF-IDF script were more distinct and suited to each category, the words outputted by the scikit-learn tfidf tool overlapped with each other and were not as felicitous to the defined category.

Introduction

Data analysis lies at the heart of all marketing industries. The movie industry is no exception to this rule. In today's age of technology, as social media assimilates more and more into our daily lives, the success of a movie is measured by the buzz it creates on social media platforms. Twitter is the most widely used social media platform for celebrity engagement and promotions. Hence, due to a striking celebrity presence on Twitter, it makes for a well acclaimed platform for sharing opinions on latest music, movies, books, TV series and more. In order to make this feature more utilizable, Twitter provides students and researchers with a tool called the Twitter Developer Platform. Twitter Developer platform provides budding data scientists and analysts with the tools required to analyze the events unfolding on twitter. This project utilizes this feature of Twitter in order to gather data and derive meaningful conclusions from it by performing data analysis on it.

The process of exploratory data analysis requires three primary skills: web scraping or data collection, data wrangling or data processing, and data presentation. The process of web scraping or data collection involves gathering data pertinent to a project by utilizing Application Program Interfaces (APIs). An application program interface (API) is a connection between two computers or pieces of software. Python offers a library called 'requests' used for making API calls. For this project, a GET request is made to the Twitter API to gather data revolving around the Dune Movie. To execute this, a Twitter Developer Platform account was utilized.

The tweets collected after web scraping can then be analyzed via Natural Language Processing (NLP) analysis. This project uses a metric called TF-IDF (Term Frequency - Inverse Document Frequency). TF-IDF is used to calculate the importance of a word or n-gram of words for a particular set of documents. [For details on TF-IDF vectorizer, see the following link: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html]. Term frequency refers to how many times a word appears throughout the document. Inverse document frequency ensures that common words (such as "the", "is", and "a") that may not be specific to the document are not given too much importance thereby giving a larger weight to unique words. Therefore, the TF-IDF metric accounts for the uniqueness of the word in the document along with its frequency of existence. The calculation of the TF-IDF scores allow for better understanding of the topics and words that appear throughout the dataset.

These tweets can also undergo sentiment analysis to identify the thought and emotional states of Twitter users on a specific movie. While different natural language processing

(NLP) algorithms can be implemented to extract sentiment automatically, manual annotation can be used to get better and more accurate labels. Using these annotations gives insight on the audience that expressed their feelings on Twitter.

Data

Early on in the data collection process, it became apparent that tweets related to the movie were scarcer than anticipated. As a result, the window was extended from 72 hours to 96 hours and covered December 4-7, both inclusively. The window was separated into twelve equally spaced periods; each period had a duration of eight hours. For each period, up to 100 tweets were collected. If there were not 100 tweets in a given eight hour block, the script collected as many tweets as possible. If there were over 100 tweets in a given eight hour block, only the first 100 were chosen. This process was carefully designed to garner a global response to the movie all while mitigating biases. For instance, 8 hour blocks ensured that multiple time zones were included in the data collection, and that each block contained a non-negligible portion of the world's English speakers. Furthermore, by allowing fewer than 100 tweets to be chosen from any given block, regions with fewer English speakers simply had fewer collected tweets, thus avoiding disproportionately weighing their voices. While it is impossible to entirely eliminate bias, this method was deemed to be the most optimal.

As previously stated, data was scarce, and the filtering parameters were designed with that in mind. The data was collected using simple yet efficient filters. The only words required in a given tweet were "Dune" and "movie" (both case insensitive). This removed unrelated posts about sand dunes and the desert, but also did not impose any further restrictions on the content of the tweet. Originally, 1072 tweets were collected over the 96 hour window. This data was then saved, with the extra 72 tweets serving as replacements in case any inconsistencies were found during annotation. More details on this replacement process can be found in the "Topics" subsection of "Methods".

Methods

Topic Selection

After the collection of the tweets, the first 200 of them were read through to determine potential topics. Each person came up with their own list of categories and these were then compiled into one big list. These included critical reception, soundtrack, cinematography, actors and actresses, related media and promotions, book comparisons, expectations, random stories, satire, production of the movie, and more.

After much debate on the potential topics, they were combined and narrowed down to six overarching topics: generic reactions/reviews, specific comments on cinematography/scenes/soundtrack/acting, comparisons and/or reference to the book, expectations and desires before and after watching the film, personal stories/anecdotes/troll tweets, and Dune merchandise/promotion/media/ads. For a more in-depth overview of the topics, their selection, and examples, see Figure 1. These six categories were chosen specifically because they were general enough to encompass many of the tweets but still specific enough that there was not a lot of overlap with each other.

Then to ensure that these six categories were actually sufficient to label the majority of tweets without overlap, we read over the first 100 tweets together and labelled them. This approach helped to clarify what type of tweet belonged to which topic and gave allowed for a general consensus that the correct topics were chosen.

Additionally, one extra label was given to tweets that were not found to be relevant to the 2021 film adaptation of Dune. This included posts referencing only David Lynch's 1984 adaptation of Dune, and some related to actors, actresses, and characters with the surname "Dune". The 72 most irrelevant tweets were consequently removed, bringing the size of the dataset to exactly 1000.

Annotation

The finalized dataset was manually annotated. It was first split into three groups - one person annotated 500 tweets, a second person annotated 400 tweets, and a third did 300 tweets. Considering that there were 1072 tweets at the time of annotation, some of the tweets were annotated more than once. During the manual annotation, each tweet was given a label for one of the six chosen topics and another label for its sentiment (one of positive, neutral, or negative). After each person finished their annotations, the labels were merged into one file. Tweets with more than one annotation were given the label of the majority vote. If all annotations were different, they were looked at again before deciding the best label.

Calculation of TF-IDF

Before computing the TF-IDF, a few design decisions were made. All words were converted to lowercase before analysis, making the analysis case-insensitive. Stopwords were also removed from that start; the list can be found at [https://gist.githubusercontent.com/larsyencken/1440509/raw/53273c6c202b35ef00194d06751d8ef630e53df2/stopwords.txt].

All words containing the character "@" were removed entirely. This was to eliminate potential usernames or Twitter handles from appearing in the TF-IDF calculation. Usernames would be quite unique and could potentially have high TF-IDF score. However, they often refer to arbitrary Twitter accounts and were therefore deemed uninteresting. URLs were also removed in a similar fashion, as their contents were not particularly meaningful. Finally, non-ascii characters such as emojis were also removed, and words that appeared fewer than 5 times across all valid categories were ignored.

The TF-IDF values for categories, sentiment, and the combined section were each calculated using the formula

$$\text{TF-IDF} = \text{TF}(w, g) \times \log\left(\frac{N}{\text{DF}(w)}\right)$$

where $\text{TF}(w, g)$ is the number of times a word w is mentioned in a group g, N is the total number of groups, and

Topic	Description and reason for selection	Examples
Generic reactions/reviews	Tweets that refer to the movie without a specific critique. Majority of the first 200 tweets were simple reactions to the watching the movie.	1) Dune movie good 2) Neeed Dune to hurry up and be over. I'm so over this movie. Let me out!! 😭
Specific comments on cinematography/scenes/soundtrack/acting	Tweets that mentioned more specific details regarding the movie whether it be on the cinematography, music, acting, or directing. These posts tend to have more details than generic reactions and specific reasons as to why they liked/disliked the film.	1) I LOVED #Dune. Never expected not to. Villeneuve is a genius. I loved the world, cast, brilliant pacing and that score. It feels a bit incomplete (but mostly because part 2 is coming 2023). Chalamet is a bona fide movie star, as is Ferguson. An awe-inspiring movie. 4.25/5 🌟 2) Just watched Dune. Fantastic Sets and Costumes Brilliant Cinematography and Visuals FX Great Cast but Confusing as fuck. Understood the whole chosen one stuff, but the voice, imperium, betrayal plot twist and a bunch other stuff made ZERO sense to me #Dune #movie #Review
Comparisons or reference to the book	Tweets that compared the film to the book or referred to it in general. Many of these tweets also expressed some reactions and reviews but was made into its own category since the film was directly adapted from the novel.	1) Just saw Dune, on the same week that I finished the book for the first time. Loved the movie and the visuals. The story is of course better in the books but I loved the how they adapted the movie! 2) I remember reading Dune, and not really getting the acclaim. Same for this latest movie adaptation.
Expectations and desires before and after watching the film	Tweets that mentioned either how people felt before going into the movie or what they would have liked changed after watching. While some tweets contained reception, these were differentiated by an expression for change rather than a review of what they saw. It also contained tweets that compared this film to other movies and franchises as an expectation of Dune being better or worse.	1) Sting, who played the Harkonnen scion Feyd-Rautha, was great as Mack the Knife in Bertolt Brecht's Threepenny Opera (1989). They should have been cast as an older Lady Fenring in the new Dune movie. 2) Gonna go see Dune. I have extremely low expectations, but it will be interesting to see a movie in the theater for the first-time in... a couple years? 3) Even then when the movie was free in @hbomax and with the brand name of marvel what #Dune has done is phenomenal. It will outgross so many first movie biggies in
		franchises including Batman Begins, Into the spider verse, captain America etc. that released Pre pandemic. Amazing
Personal stories/anecdotes/troll tweets	Tweets that include a story or anecdote regarding the film but not a discussion or review of the film itself. This category also included conspiracy theories and unfounded claims regarding the movie.	1) Back home icing a very swollen and sore forearm for the evening. Damn lucky I still have a right hand that's attached to my body. Gonna blunt the embarrassment with some bourbon and the new Dune movie. 2) dune is just a movie for Big Water to sell us more of those water bottles that make us drink half a gallon a day
Dune merchandise/promotion/media/ads	Tweets that included Dune merchandise, advertisements, and promotions. It did not matter if they were made by accounts officially affiliated with the creators of Dune or if they were independent artists.	1) Amazon will tell you the #Dune movie art book is sold out, but you can order it direct from Simon and Schuster and get it in five days. 2) My Fan Art for Dune Movie #DuneMovie #shotoniphone 6s
Irrelevant	Irrelevant tweets are those that do not have any relevance to the 2021 version of the film Dune. This included tweets that refer to the 1984 Dune	1) with the one & only exquisite @ginacarano as the sexy badass Cara Dune. I know it's not a movie a but IDGAF 2) #NowWatching @9Gem DUNE (1984). It is a pug movie, I just realised.

Figure 1: Further details on topic selection

DF(w) is the document frequency of w (i.e. the number of groups that contain w).

Additionally, as a comparison, the Sci-Kit Learn TfidfVectorizer was also applied to each topic and sentiment. The text was converted in a way consistent with the first technique. Words were cast to lowercase, and invalid characters, stopwords, and punctuation were all removed. In both techniques, there was no other filtering or normalization done to the text such as lemmatization or stemming.

Results

Topics and Sentiment

Each tweet with the same topic and/or sentiment were grouped together from the dataset. The groups from largest to smallest for the six topics were (see Figure 2 for precise numbers):

1. Generic reactions and reviews
2. Stories, anecdotes, and trolls
3. Comments on cinematography, soundtrack, and acting
4. Media, merchandise, promotions, and ads
5. Expectations and desires regarding the film
6. Comparison to the book

The number of tweets in each group were roughly the same with the exception of generic reactions and reviews. This category had almost 200 more tweets than the next biggest. The smallest group was book comparison with only 85 total tweets from the 1000 in the dataset.

Next, for the sentiment, the vast majority of the tweets were labelled as positive or neutral (see Figure 3), with positive having 482 tweets and neutral having 356 tweets. There were only 162 negatively labelled tweets, less than one-fifth of the total dataset.

Lastly, for the combined sentiment and topic groups (see Figure 4), the largest group by far is positive reactions with 231 tweets. None of the other groups had over 100 tweets, and the smallest group (negative media/ads) only had 4 total tweets.

TF-IDF

The TF-IDF scores were calculated both manually and automatically (see Figure 5 for the full list of top words). The manual calculation of the TF-IDF yielded more unique words for each label such as 'fetus', 'shit', 'blunt', 'fan', and 'reporter' and had little overlap with each other. The top associated words for the automatic calculation were more common and could be generalized to almost every label such as watch, story, amazing, movies, theater, and bad.

Discussion and Conclusion

Overall, the results were very similar to the expected outcomes. The majority of the tweets in regards to Dune were simple reactions or more extensive commentaries on the film. There were also a fair amount of comparisons or references to the book, as well as bizarre stories and anecdotes involving the film. It was also not surprising to see that majority of the tweets were positive (almost half) and a very small minority were negative.

It was however interesting to see that most of the tweets labelled with expectations or desires were actually considered neutral rather than positive or negative. Considering that this category was for tweets expressing things they would like changed and what they were expecting of the film, it was expected that they would carry more emotion. Also, since the amount of positive tweets outnumbered the neutral and negative tweets, it was surprising to see that several of the categories had more neutral than positive tweets. It was not surprising to see that the negative categories were always the smallest.

However, these results somewhat make sense. Many of the stories were neutral since the stories themselves did not have much to do with the actual movie. Similarly, many of the promotions and media were neutral because they did not reflect on the movie; they were rather trying to sell the film or film-related objects to the general population. Thus, if majority of the positive tweets were categorized in generic reaction/reviews and critiques of cinematography, then it is understandable that some of the categories have more neutral than positive tweets.

For the TF-IDF analysis, the manually calculated top ten words for both the topics and sentiment show a clear shift in vocabulary when talking about different subjects and expressing different feelings. Each of the top ten words are quite distinct and specific to the label itself. For instance, some of the top ten words associated with the book category are "adapted", "book", "written", "frank" (author's first name), and "herbert" (author's last name). These words are all related to the book or writing. This is also seen with the top words for the negative sentiment. A majority of the top ten words such as "sorry", "weird", "waste", and "worst" have negative connotations.

There is a noticeable difference with the automatically calculated top ten words using Sci-Kit Learn's TfidfVectorizer. Although the same pre-processing was applied to the tweets and the only difference was the actual TF-IDF calculation, the top ten words from the TfidfVectorizer for each label overlapped more and had more words with the same root. Specifically for the topics, the word "time" showed up in nearly every category (5/6) and "book" showed up in 4 out of 6 categories. Furthermore, one of "watch", "watched", or "watching" showed up in every topic. This difference may be due to a slight variation in how the TfidfVectorizer calculates the scores as well as the fact that it implements smoothing of the IDF. Additionally, lemmatization or stemming of the word may have prevented the issue of "watch", "watched", and "watching" showing up many times.

In all, the results were not very surprising. With the manual calculation of the TF-IDF, many of the top words associated with each label were fairly appropriate and specific to the topic or sentiment. The categories most relevant to the film's reception were review, cinematography, and film adaptation. These sections focus mainly on people's attitude towards the movie. They contain genuine reactions, comments, and emotions from the film's audience. Moreover,

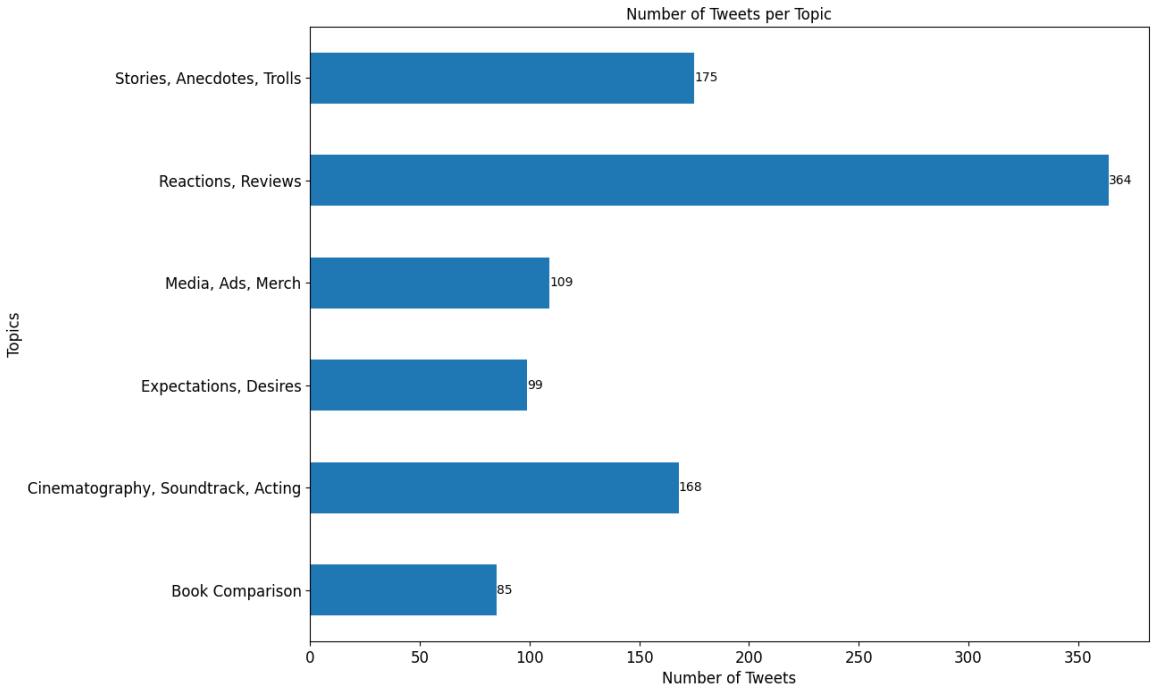

Figure 2: Number of tweets per topic

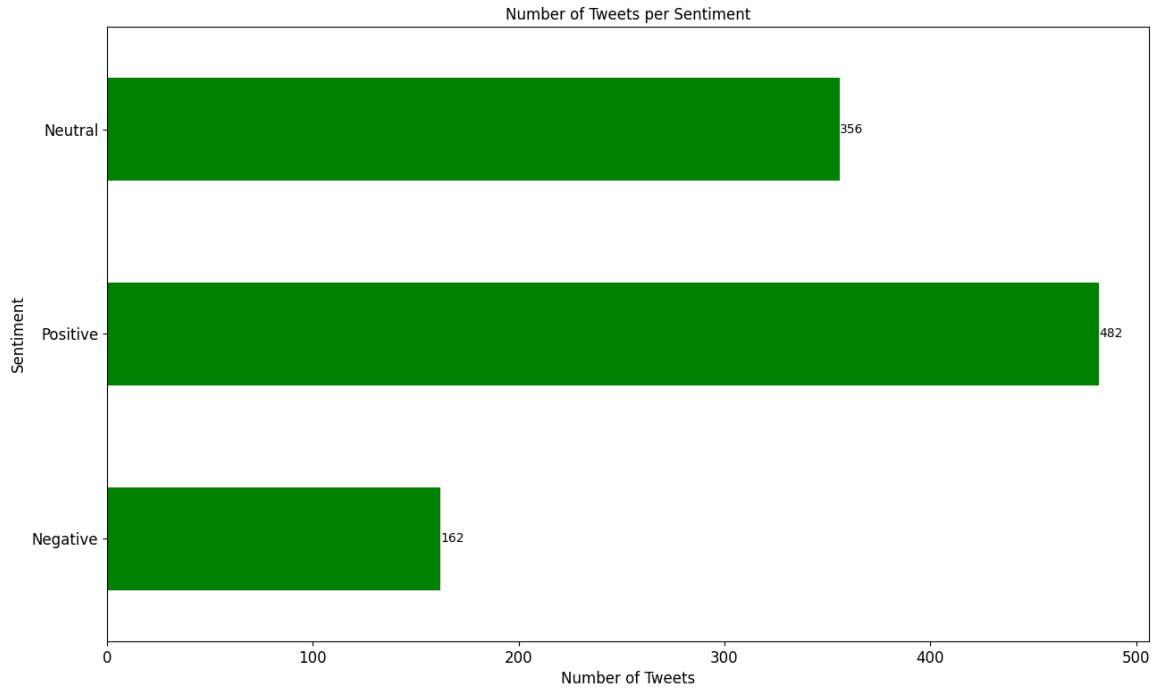

Figure 3: Number of tweets per sentiment

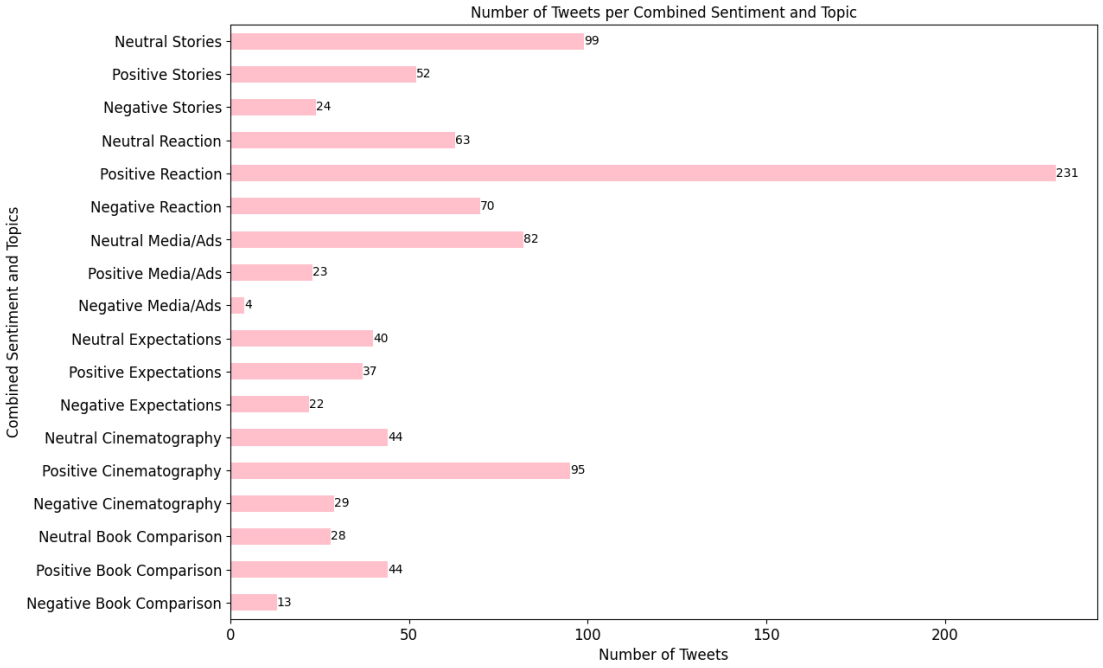

Figure 4: Tweets per combined sentiment and topic group

Topics	**Top Words (manual TF-IDF)**	**Top Words (sci-kit learn TfidfVectorizer)**
Reaction	probably, fetus, times, sound, amazing, liked, imax, third, ones, favorite	time, watch, watched, loved, book, seen, love, bad, watching, ive
Cinematography/soundtrack/acting	director, sound, casting, stunning, cinematography, chalamet, visuals, atreides, score, amazing,	time, watching, story, villeneuve, amazing, book, world, chalamet, love, film]
Book comparison	adapted, written, books, try, frank, reading, started, blunt, herbert, course	book, read, books, reading, love, written, half, watching, story, ve
Expectations	Harkonnen, sequel, win, spider, idk, mcu, written, ones, spiderverse, franchise	book, watch, lynch, excited, series, seen, time, amp, don, movies
Personal anecdotes	imax, stopped, theater, home, poster, theatre, premiere, started, eternals, waiting	time, review, movies, via, frank, series, entertainment, dunemovie, watch, arrakis
Media/merch/ads	ridley, scott, via, villeneuves, entertainment, arrakis, herberts, reporter, white	time, watching, imax, watch, day, theater, watched, love, re, im

Sentiment	**Top Words (manual TF-IDF)**	**Top Words (sci-kit learn TfidfVectorizer)**
Neutral	ridley, video, fetus, based, via, home, scott, excited, win, theres	book, time, watching, watch, read, lynch, movies, watched, star, bad
Positive	times, listening, score, forward, shit, fan, stunning, weekend, hour, released	book, time, watch, love, loved, watched, amazing, seen, imax, watching
Negative	sorry, cause, sting, weird, waste, comes, happened, serious, worst	time, boring, book, seen, story, lynch, watch, watching, watched, bad

Figure 5: Top ten words based on TF-IDF scores for topics and sentiment

they contain minimal spam or unhelpful information. As a whole, it is quite clear that these three sections were overwhelmingly positive. In contrast, the anecdotes, media, and expectations categories are far less useful to understanding the film's reception. As previously stated, many of the tweets from these categories were not particularly helpful to understand the audience's opinion. The amount of valid commentary in these sections, whether positive or negative, was dwarfed by content that did not provide adequate insight into the question at hand.

Looking at Figure 3, it may appear that a large portion of the audience was simply indifferent about the movie. For instance, the number of tweets that were neutral and negative added up to be greater than the number of positive tweets. However, it is important to keep in mind what was discussed above. The neutral tweets were largely contained inside categories that did not provide great insight into the audience's appreciation of the film. Looking at Figure 4, it is apparent that the more relevant sections had a stronger net positive sentiment. Furthermore, it is worth noting in particular the "book comparison" and "cinematography" section of Figure 4. It is incredibly hard for a movie adaptation to please loyal fans of a novel. However, the negative comments surrounding the adaptation were rather sparse. The film community can also be rather difficult to please, as enjoyers of cinema have higher standards than the average person watching a film. Despite this, the negative commentary on acting, directing, music, and overall cinematography was quite limited. It can therefore be concluded that fans of the book, adamant moviegoers, and the general population agree on one thing: Dune was a quality film. The media company can be confident that its reception was one of contentment.

Statement of Contributions

Kate Kim, Patrick Janulewicz, and Navneet Kaur collaborated for this project fairly. Navneet Kaur applied for the Twitter Developer Platform account to experiment with filters, query parameters, and Twitter API calls. Then she wrote the very first version of the script that was employed to make a GET request to Python. Patrick Janulewicz then improved this script in addition to writing a script to clean the data. The team then worked together to remove all unnecessary tweets and irrelevant data that was gathered in the process. After the data was collected, the data was manually annotated with Navneet annotating 300 tweets, Patrick annotating 400 tweets, and Kate annotating 500 tweets. The team then together came up with the categories relevant to the tweets. Navneet Kaur suggested the idea of implementing TF-IDF using the inbuilt Python library in order to have a baseline to compare the results of the manual script. Then, Patrick Janulewicz implemented the manual TF-IDF calculations for the topics and sentiment labels. Afterwards, Kate Kim implemented the automatic calculation of the TF-IDF using Sci-Kit Learn and additionally created all of the figures seen in the report. The report was written by all three members; specifically, Navneet wrote the Abstract, Introduction, Discussion and Conclusion along with formatting the figures and the table of categories in the document. Patrick wrote about the Data, TF-IDF calculation, and Discussion, and Kate wrote the Abstract, Methods, Results, and Discussion and Conclusion.