# Analysis of Twitter Commentary on the Movie Dune

**Jonayed Islam,[1] Faisal Rabbani,[2] Philip Tam [3]**

## Introduction

The analysis was done on the movie Dune. The data collection was done through the Twitter API. 8,000 tweets were collected through the API. 200 tweets were then randomly selected and labelled using an open coding method to determine 8 labels or topics to annotate tweets by. 1,000 randomly selected tweets were then selected to be manually labelled on the topics of Review, Theatre or Cinema mention, Actor/Characters, Awards/Nomination, Cinematography or Videography, Soundtrack, Book, and Other. The labelling was also done with a sentiment annotation indicating if the tweet was positive, neutral, or negative based on the selected topics. The 1,000 tweets were also analyzed with a TF-IDF score to identify the top 10 keywords in each topic. The analysis proved that a large majority of the 1,000 tweets were unusable because of the lack of additional filtering before the open coding section and that the tweets could have been preprocessed more.

## Data

The data was acquired through the Twitter API recent endpoint through with the three keywords: Dune, #Dune, #DuneMovie. The tweets were collected between the dates of 2021-12-03 and 2021-12-06. A total of 8000 tweets matched our initial criterion and were thus collected and saved into a json file. In order to make sure that we only selected tweets that would be meaningful to our analysis, the tweets were filtered to be only English tweets, not retweets, and not replies. To filter tweets that were not in English, we used the langdetect python package. The remaining tweets were then parsed into a single CSV using a python script. In order to reduce bias, a collection of tweets was selected randomly through a random sample method in Pandas. This sampling method was used to choose the 200 tweets for open coding and the 1000 tweets for the final analysis.

## Methods

Once the data is collected, we run an open coding to determine which topics to analyse, then we do a manual annotation of the topic and sentiment for each tweet and, finally, we calculate the top ten words for each topic based on their TF-IDF score.

### Open Coding

The open coding was done with the 200 tweets that were randomly chosen through the sampling method. All three team members individually conducted an open coding of this dataset. This was achieved by analyzing the initial 200 tweets and labelling them based on their potential topics and their level of relatedness to Dune. This allowed us to each establish our view of the top eight topics and then deliberate on which ones best suit the goal of this study. These potential topics were then refined and merged to more specific topics that would each establish a clear inclusion criterion and minimize subjectivity when attributing a topic to a tweet.

### Topic and Sentimental Annotation

After we decided on 8 topics by which to categorize the tweets, a manual annotation was conducted to both assign each tweet a category and indicate whether the sentiment of the text was mostly positive, negative, or neutral. To do this we wrote a simple Python script that printed each tweet, and then prompted the annotator to select a category and a sentiment evaluation. The annotation was then saved into two separate columns created in the pandas dataframe. This was done so as to minimize potential spelling or formatting errors that could compromise our data-set.

A tweet was indicated to have “positive” sentiment if it portrayed positive thoughts, excitement, or anticipation of the movie. Tweets were indicated to be neutral if they either expressed mixed thoughts or did not give an opinion, and if a tweet portrayed negative thoughts or critiques, it was indicated to be of negative sentiment.

After the annotator concluded the manual annotation phase, the two other team members conducted an audit of the annotated dataset to ensure there were no obvious errors or miscategorizations.

### TF-IDF Calculation

In order to calculate the TF-IDF, we first loaded the csv file containing all annotated tweets into a pandas dataframe. We then had to preprocess the text in each of the tweets. For the preprocessing, we removed all links using a regex filter and then we removed all non-alphanumeric characters. The text for each tweet was then tokenized and the frequency of each token for that tweet was stored in a dictionary.

In order to calculate the inverse document frequency (IDF), we first created a dictionary that contains as its key all unique words that appear in all 1000 tweets and as its values the frequency of appearance of those words in the list of tweets. Then, to get the IDF value for each word, we took the base ten logarithm of the division of the number of documents and the frequency of each word. To calculate the total frequency for each topic, we aggregated the word dictionaries of each tweet belonging to a certain topic. We then multiplied the frequency of words in each topic by the corresponding IDF value of that word. With that, we were able to determine the top 10 words with the highest TF-IDF score for each of our 8 topics.

## Results

### Open Coding

The potential topics of the initial 200 tweets were: Review, Experience, Book, Other, Random, Advertisement, Meme, Previous Movie, Sequel Movie, Awards, Facts, Quote, Music, Production, and Sound.

The top 5 topics that were labelled during the open coding were Review with 63 tweets, Experience that was 26, book related that was 20, and Other that was 16. From these potential topics we merged and refined them to 8 topics that can be Review/Commentary, Theatre or Cinema mention, Actor/Characters, Awards/Nomination, Cinematography or Videography, Soundtrack, Book, and Other.

| **Topic** | **Count** |
|---|---|
| Review | 63 |
| Experience | 26 |
| Book | 20 |
| Other | 16 |
| Random | 15 |
| Advertisement | 13 |
| Meme | 8 |
| Previous Movie | 8 |
| Sequel Movie | 7 |
| Awards | 7 |
| Facts | 5 |
| Quote | 5 |
| Music | 4 |
| Sound | 2 |
| Production | 1 |

Open Coding results for each topic

### Topic and Sentimental Annotation

After the manual annotation was conducted, we were able to classify our 1000 tweets based on their sentiment and topic. Our initial results indicate that the majority of tweets presented no sentinel direction. In fact, 54.2% of tweets displayed a neutral sentiment, whereas 6.4% were negative and 39.4% were positive. Overall, 25.1% of tweets were attributed to the “review/commentary” topic and the three next biggest ones with just under 10% of the share were “theatre/streaming service”, “awards/nomination” and “actors/characters”. The remaining three topics had a much smaller share but were more important in our analysis as they had a portion of tweets that presented a sentiment and did not simply offer a neutral perspective.

| **Topic** | **Negative** | **Positive** | **Neutral** | **Total** |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| actors/characters | 12 | 29 | 53 | 94 |
| awards/nominations | 0 | 26 | 5 | 31 |
| book | 8 | 39 | 50 | 97 |
| cinematography or videography | 3 | 28 | 30 | 61 |
| review or commentary | 25 | 185 | 41 | 251 |
| soundtrack or sound quality | 5 | 19 | 9 | 33 |
| theatre or streaming service | 6 | 40 | 51 | 97 |
| other | 5 | 28 | 303 | 336 |
| **Total** | **64** | **394** | **542** | **1000** |

Sentiment and topic allocation of the 1000 tweets

Inevitably, much of the discussion around the film on Twitter was very off-topic and consisted of jokes/memes without any particular sentiment towards the film itself. Therefore, our "other" category accounted for 33.6% of all collected tweets. This is also due to the fact that many tweets had only the #dune or #dunemovie in the tweet along with a picture attached. Also, we did come across many tweets who were using those hashtags or keywords in order to promote unrelated content.

We conducted an initial analysis of word frequency, and wrote a quick script that filtered out emojis, links, non-alphanumerics, and commonly used words such as "the", "and", etc. Our results showed these to be the three most commonly occurring words in each sentiment category by sheer number of usages:

| **Sentiment** | Word 1 | Word 2 | Word 3 |
|---|---|---|---|
| Positive | imax | movie | time |
| Negative | movie | book | boring |
| Neutral | 2021 | movie | time |

Top three words associated with each sentiment

## TF-IDF

We identified the top ten keywords for each topic based on their TF-IDF score. In the topics concerning actors, book, cinematography and soundtrack we observe that tweets contain proper nouns of the person in charge of that section. For example, we see a high TF-IDF score for Denis Villeneuve, the director of Dune, in the cinematography category. Most of the keywords identified in the review topic are generic, except for the word liked. Just like the results in the sentimental analysis section, we can see the impact of the word imax as it has the highest TF-IDF score in theatre and the cinematography topic.

| **Review or commentary** | **Theatre or streaming service** | **Book** |
|---|---|---|
| Commentary | imax | read |
| Movie | seeing | reading |
| Review | watching | book |
| 10 | finally | frank |
| 2021 | screen | herbert |
| 2 | amp | books |
| time | cinema | messiah |
| liked | time | fiction |
| film | watch | series |
| watched | 2 | sci |

| **Actors or characters** | **Cinematography or videography** | **Awards or nomination** |
|---|---|---|

| | | |
|---|---|---|
| chalamet | imax | 1 |
| zendaya | villeneuve | 3 |
| timothee | denis | 7 |
| oscar | shot | 2 |
| paul | africa | zack |
| casting | visual | league |
| heh | digital | justice |
| mother | dunemovie | films |
| timothé | visuals | suicide |
| isaac | scene | squad |

| **Soundtrack or sound quality** | **Other** |
|---|---|
| soundtrack | watching |
| hanszimmer | day |
| zimmer | sand |
| music | 2021 |
| hearing | 1984 |
| composer | david |
| hard | movie |
| hanszimmer | lynch |
| sounds | 2 |
| sound | time |

Top ten keywords by TF-IDF score for each topic

# Discussion

## Open Coding

As the 200 tweets were selected randomly from 8000 there was a lot of variance in the topics of each tweet. Moreover, to determine possible topics each tweet was associated with the general topic that fits it the best. For the 200 tweets that were selected randomly over 15 topics were found.

To describe each topic briefly: Review is a tweet with a score or general review of Dune, Experience is a tweet about going to watch dune, Book is a tweet about book Dune, Other are tweets unrelated to Dune, Random is a tweet related to Dune but not specific, Advertisement is tweet that is a Dune ad, Meme is a tweet about Dune memes, Previous Movie is a tweet about previous Dune movies or shows, Sequel Movie is a tweet about the sequels, Awards is a tweet about the nominations or awards won by Dune, Facts is a tweet about interesting facts in Dune, Quote is a tweet that are quotes from the movies, Music is a tweet about the music in Dune, Production is a tweet about the production in Dune, and Sound is a tweet about the sound in Dune.

These topics turned out to be too extensive and were cut down into 8 which were more descriptive and could be more effective with topic and sentiment analysis. However, if done again a more probable solution would be to clean up the data and filter out for certain tweets that would potentially affect topic creation.

The open coding showed that a lot of tweets were not descriptive of their own content, having to check for their images to identify their contents or even look into the accounts of the tweet to understand what topics they were about.

There was a possibility of filtering tweets by character counts which could have led to more consistent constructive topics and more descriptive topics. These topics that were filtered out would then allow for better results in the topic and sentiment annotation as the text would be more constructive towards longer reviews, experiences, and opinions.

## Topic and Sentiment Annotation

Topic and sentiment analysis for Dune showed that, at least amongst Twitter users, the film was mostly well-received. Outside of our “other - neutral” category (into which spam posts and unrelated tweets were categorized), “review or commentary - positive” was the largest category of tweets. We can also observe that the film has led to discussion over the original Dune book series, as those tweets account for about 10% of the total number of tweets we collected.

Overall, we can get an idea of what viewers liked and disliked about the film from our initial annotation phase. We can see that a lot of the film's positive mentions revolved around mentions of the theatre/cinema experience and the visual elements of cinematography, while a lot of the negative tweets revolved around characters and actors in the film.

In our identification of the top three keywords related to each sentiment, it is interesting to see that most of the keywords are generic such as "movie" or "time". However, for the positive sentiment, the word Imax stood out and thus shows the positive impact on watching the movie in Imax format. We can also see that one of the main gripes viewers had with the film was that they found it 'boring'.

## TF-IDF

In our results, we notice that individual numbers have often appeared as top keywords. This is due to the fact that many tweets do a ranking of different movies and thus use the numbers to denote their rank. With that being said, each of the eight topics presented distinct top keywords.

**Review or commentary**
For this topic, the top three keywords are commentary, movie and review. These keywords are quite descriptive of the category they belong to. During our initial analysis, we saw that tweets in this topic were also followed by hyperlinks to external websites with full reviews. We do see that the keyword in the eight position of this list is "liked", thus showing a sentimental connotation for the movie. It is worth noting that the appearance of the number two in this category is due to the fact that users are wishing for the sequel Dune 2.

**Theatre or streaming service**
For this topic, the top three keywords are imax, seeing and watching. It is interesting to note that many tweets referred to watching the movie in imax format. Watching and seeing were also very popular in this category as many tweets referred to the place where people went to physically watch the movie or stream it.

**Book**
Besides action verbs related to reading or the word book itself, there are also keywords referring to the genre of the book Dune which the movie was based on. Tweets also refer to the author of the novel Frank Herbert.

**Actors or characters**
In this category, we see a lot of tweets praising the main actors Timothee Chalamet as well as Zendaya. We can see the positive sentiment about the choice of actors as twitter users refer to the oscar awards.

**Awards or nomination**
The top four keywords here are numbers which were used to rank the movie to others. We see that movie is heavily compared to justice league by the director Zack Snyder as well as suicide squad.

**Cinematography or Videography**
Just like the theatre topic, the top keyword here is imax, thus referring to the excellence of the movie's imax format. We also note that Twitter users credit the cinematography to the director Denis Villeneuve as his name is the second and third most important keyword. Also, Twitter users frequently mention the scene of the shoot in Africa.

**Soundtrack or sound quality**
Hans Zimmer is frequently mentioned in this topic as he is credited for comping the sounds and music in Dune. Most of the other keywords in this topic are related to the word sound. In the future, it would be interesting to see if considering all words with the same root to be the same would lead to different results.

**Other**
Although this topic contained tweets unrelated to the movie Dune or tweets not fitting in any topic, we did find that the top keyword was watching. This is due to the fact that many users use unrelated trendy hashtags to promote their content for other videos. We also do see that some of the unrelated tweets have to do with an actual dune and thus the reference to sand.

# Group Member Contribution

Overall, we were a very collaborative team. From the onset, we agreed to divide and conquer the tasks and each member accomplished their share of the work.

## Jonayed Islam

Jonayed was in charge of creating the initial task split. He then wrote the code to calculate the TF-IDF score for each topic. He also handled the analysis and discussion of sections concerning the TF-IDF in the report.

## Philip Tam

Philip was in charge of the data collection using the Twitter API, and open coding. He wrote a script to collect the data from Twitter, a script to format the twitter data, and a script to randomly select *n* number of tweets. He is also in charge of the introduction, open coding, and data collection.

## Faisal Rabbani

Faisal was in charge of dataset categorization and sentiment analysis. He wrote a script to help format the data and provide an easy-to-use command line tool to iterate through tweets, and conducted an initial analysis of word frequency per sentiment category. He also is in charge of describing the topic and sentiment analysis throughout the paper.