

COMP 598 Final Projects

Altamirano Mayoral, Teresa

260868349 teresa.altamiranomayoral@mail.mcgill.ca

Liang, Marlene

260928668 marlene.liang@mail.mcgill.ca

Yadollahi, Parsa

260869949 parsa.yadollahi@mail.mcgill.ca

Introduction

Shang-Chi and the Legend of the Ten Rings, released in theaters on September 3, is the first Marvel Studios film to feature an Asian lead and a predominantly Asian cast. Its highly anticipated release has garnered many people's attention who took it to Twitter to comment on the movie. Most of the discussion concerned its scenes, characters, diversity, and visuals. More specifically, the bus fight scene, the character Morris, the Asian cast, and the quality of the visual effects emerged as the primary concerns for each respective topic. Other topics covered in the tweets included advertisements and streaming options, such as Disney+, for the movie. Many users compared the film to other releases from Marvel Studios film such as Black Widow and Black Panther. The majority of the tweets (30.4%) offered positive reviews of the movie and mentioned that they would be interested in rewatching it. Overall, the tweets provided positive reviews of the film, crediting its success to the memorable scenes, the unique characters, and the cast diversity.

Data

Data Collection

We collected 10,000 tweets from Twitter under the query "Shang Chi" from November 16 to November 19, 2021. More specifically, approximately 3333 tweets were collected each day. Instead of the full movie title "Shang-Chi and the Legend of the Ten Rings", the query "Shang Chi" was used to retrieve the data since it had been the primary keyword used by the official Marvel Entertainment Twitter account to announce and promote the film. Therefore, the resulting tweets from this query would be more representative of the actual discussion surrounding the movie. The dataset with the 10,000 tweets was stored under a TSV file in order to clean the data.

Data Cleaning

The dataset was filtered by language, uniqueness, hashtags and keywords in the content of the tweet. We only analysed the English tweets. This filtering step was done by using the language property of the tweets in the data. The uniqueness of the tweets was ensured by using Pandas python library.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

More specifically, we used the `df.drop_duplicates` function. Next, we selected a small sample of around 200 tweets to determine the most used hashtags. The list was composed of different variations of the movie title, characters' names and actors' names. To see the complete list, refer to Appendix A. Notice that the empty string was used as a valid hashtag. This is because many relevant tweets didn't contained a hashtag. Because of this, we also parsed the content of the tweets and compared them to a list of relevant keywords. If the content of the tweet contained at least one of the relevant keywords, the tweet was kept, else it was removed from the dataset. Refer to Appendix B to see the full list of relevant keywords used. Finally, the number of remaining tweets per day was output giving:

- Day 1: 643
- Day 2: 480
- Day3: 467

These number were further reduced in order to fit the 1000 tweets constrain such that around 333 tweets per day were in the dataset. The complete clean data was output as a CSV document in order to be annotated.

Methods

Open Coding & Annotation

We used the following open coding procedure to develop our topology for the Shang-Chi-related tweets. First, we randomly picked 200 tweets from our dataset of a thousand tweets. Next, each team member went through a subset of those 200 tweets, each covering roughly 1/3 of them, and came up with potential topics. Once we had finished forming ideas on the categorization, we met and brainstormed on which topics were most relevant.

Once the topics were well defined, we annotated around 333 tweets, both for sentiment and topic. Throughout the process, we encountered exceptional cases. When this happened, we would consult the team, and we would modify the definitions of the topics to make them more precise.

TF-IDF Calculations

To calculate the tf-idf (Term Frequency-Inverse Document Frequency) we used a similar approach to that used in assignment 8. We cleaned each tweet in our dataset by first

removing all ‘@’ characters, links, emojis and removing all hashtags while keeping the text that follows it, then discarding all stopwords and punctuation using the NLTK library in Python and finally replacing all capital characters with their respective lowercase letter. We then calculate the tf-idf with respect to the topic characterization defined by the product of the following:

$$tf - idf(w, c) = tf(w, c) \times idf(w)$$

Where the left-hand side of the product is defined as:

$$tf - idf(w, c) = \text{the count of the word } w \text{ in the category } C$$

And the right-hand side of the product is defined as:

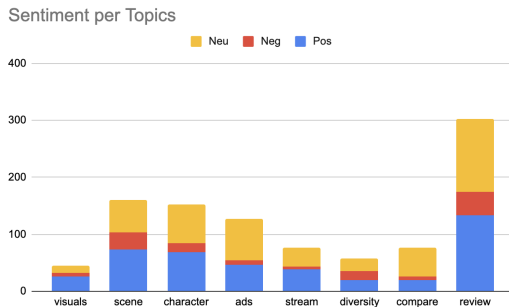
$$idf(w) = \log \frac{\text{the total number of categories}}{\text{the number of categories using word } w}$$

Results

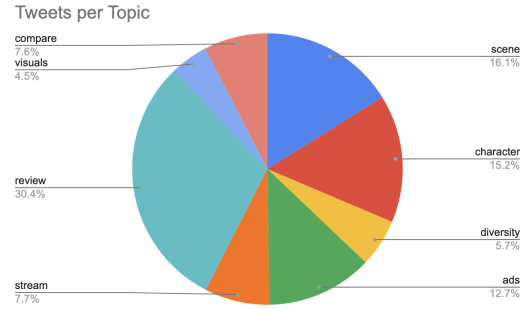
Topic Selection and Characterization

The 8 topics, their definitions and characterizations are listed in Appendix C. The characterization of each topic is defined as the top 10 words with highest tf-idf scores.

Sentiment per Topic The stacked bar chart in figure 4.2 reveals the different sentiments of each topic. Given the fact that the movie’s official reviews were relatively high on many reviewing platforms, scoring a 91% on Rotten Tomatoes and a 7.5/10 on IMDb, it is unsurprising that most of the user’s sentiments were either neutral or positive. For each topic, there are more tweets expressing a positive sentiment than negative sentiment. Additionally, there is roughly an equal amount of tweets that display a positive sentiment and a neutral sentiment.



Tweets per Topic The pie chart bellow reveals the distribution of tweets based on the topic. We notice at first glance that the two main topics discussed in most tweets are either review, about a character or a specific scene in the movie with 30.4%, 15.2% and 16.1% respectively. We also notice a relatively high contribution being 12.7% from tweets which focus heavily on advertisement. The other topics such as diversity, compare, visuals and stream were all around the 6% or 7%.



Outliers

Certain keywords appearing in Appendix C such as “halal” in visuals, “eat” in diversity, as well as ”calling”, ”day”, ”origin”, and ”tier” in review appear to be outliers and does not seem to tie in with the discussion surrounding the movie. However, tweets containing these words often discussed the movie in the rest of the tweet.

To learn more about the different outliers per topic please refer to the [discussion section](#).

Discussion

Sentiment

Overall, the tweets related to the movie had a *positive* sentiment. The second most prominent sentiment was *neutral* followed by *negative* sentiment. Most positive sentiment talked about the scenes, visuals, characters, diversity, and reviews. Ads and stream also had a majority of positive sentiment since it was an advertisement strategy to promote the movie. As for compare, most of the sentiment was neutral. However, a surprising result is that reviews has almost as many *neutral* tweets as it has *positive*. Indeed, while annotating the tweets, we noticed that many review liked and gave positive feedback about certain aspects of the movie and other negative aspects. Thus, we annotated such tweets as *neutral*. This could also explain why after computing the tf-idf for reviews, we only got *positive* words or *neutral* words.

Scene

As suggested by the keywords “bus”, “scene”, and “fight”, viewers were primarily interested in the fighting sequences in the movie and were especially interested in the bus fight scene. Many tweets were commenting on the realism and feasibility of the bus scene after a real-life driver revealed the technicalities involved. Since Shang-Chi and the Legend of the Ten Rings is marketed as a superhero film by Marvel Studios, it is unsurprising that viewers are mainly concerned with the quality of the action sequences in the movie. However, many users were still impressed by the complexities in the scene and the fact that the actor performed his stunts.

Character

The primary keywords for this topic are “morris”, “wenwu”, “ben”, “kingsley”, “michelle”, and “yeoh”, suggesting that viewers are interested by the character Morris, Wenwu, and

the actress Michelle Yeoh, as well as the actor Ben Kingsley. While the movie centers its plot around Shang-Chi for most of the storyline, viewers are also concerned by other side characters that stand out. Many tweets expressed their admiration for Morris, a fantastical character in the movie, and mentioned how adorable it was. Other tweets also highlighted Wenwu, Shang-Chi's father in the movie, indicating that viewers were interested by the family dynamic displayed throughout the movie. The fans also displayed their love for the actress Michelle Yeoh and were enthusiastic for her performance in the movie as Shang-Chi's aunt. Finally, Ben Kingsley was also mentioned a certain number of times. He is the actor that played "The Mandarin" during Iron Man 3. In Shang-Chi he appears as a prisoner of the real leader of The Ten Rings and helps the heroes in their adventure. Tweets mentioning the actor offered mixed reviews of his performance in the movie. Some users praised his performance while others said that his performance distracted from the plot.

Diversity

The topic of *diversity* was also largely discussed by many of the tweets, most of them praising the film for its accurate representation of Asian culture. The specific keywords for this topic were "simuliu", "chinese", "matters", "eat", "asian" and "representation". The keywords "asian", "representation" and "matters" were mentioned in tweets where users were complimenting the diversity of the cast. Since Shang-Chi and the Legend of the Ten Rings is the first Marvel movie to feature a predominantly Asian cast, it was important for the viewers that the movie portrayed Asian culture in a sincere tone that avoided typical clichés and stereotypes which are often present in other movies. Most viewers agree that this intention was well portrayed for this movie. However, the keywords "chinese" and "eat" were mentioned in tweets that carried racist connotations, suggesting that movies featuring *diversity* can also receive racist backlash.

Visuals

The keywords for the topic *visuals* are "fanart", "lighting", and "cgi". The word "fanart" was employed by users who tweeted fanart that they had made of the movie, suggesting that the movie had memorable visuals that viewers wanted to recreate. The word "lighting" and "cgi" was brought up by users who had mixed feelings about the visual effects of the movie. Since the movie featured many fantastical elements that had to be digitally rendered, it was important for the audience that these elements were naturally incorporated into the movie so as to not diminish the quality of a scene. However, certain tweets expressed their disappointments in the rendering of these visual effects while others praised the qualities of the visual effects.

Ads

Another recurrent topic of the tweets was of advertising nature. Approximately, 18% of the tweets sampled qualified as "ads". Most of the tweets aimed to advertise the movie itself, but many other aimed to advertise shows or products in link

to the movie's franchise. This can be explained since Disney and streaming services use Twitter as a main platform to advertise their content. On the other hand, the movie had mostly positive reactions from the public (see [Sentiment](#)). Because of this, many people released media content related to the movie and they used Twitter to advertise this material. Moreover, the fact that the words found through the tf-idf computation are not fully related make sense since the *ads* topic was for any sort of advertisement excluding streaming. Finally, since ads are only the 12% of the tweets, we can imagine that the words found in appendix C are probably from the same product or seller promoting their services or products in similar ways.

Stream

Streaming services are among the most prominent ways the general audience tends to watch content. Since Disney+ owns a Twitter account and hosts Shang-Chi and the Ten Rings on their streaming platform, they used the opportunity to promote the newly released film. Other streaming services also used the movie to promote themselves on Twitter. Movie distribution services also took to Twitter to promote the film and their service, such as its availability on Blu Ray.

Review

The *review* topic of discussion that had the most engagement is the review topic, representing 30.4% of the collected tweets. Since many people go on social media to share their opinions of newly released movies, it is natural to expect that most tweets will be reviews of the movie. Additionally, reviews posted on Twitter are mostly from online film critics or from the general audience and is generally less academic. Therefore, the sentiment observed from these tweets will reflect the general audience's sentiment towards the movie. Indeed, the keyword for this topic was "rewatch" in the context that they were interested in rewatching the movie another time. Most of the tweets in this category had a positive or neutral sentiment, suggesting that the film was well received by the general audience. Moreover, it was interesting to see that there is no negative word that was salient for tweets about reviews. This is probably because the main sentiment around review was positive, followed by neutral. On the other hand, most comments with negative connotation about the movie did so "implicitly". In other words, they didn't use "negative" words. Finally, there are some outlier words such as "day", "finished", "origin", "calling", "tier" and "yet". It is unclear why "day", "origin", "calling" and "tier" showed up in our results. This could be because some of the reviews seemed to wonder before reaching a conclusion. As for "finished" and "yet", they both made allusion to the action of watching the movie and in the usual template of a review, they are often accompanied by an opinion.

Compare

As the Marvel Cinematic Universe gets more complex over each new release for the past ten years, it is almost impossible for Twitter user fans to avoid comparing the new movie with old productions. The main comparison is with Spider-Man. Indeed, as the new Spider-Man movie is about to be

released at the end of the year, fans have been speculating about theories surrounding how Shang-Chi will fit in the cinematic universe. Additionally, fans couldn't resist comparing the movie to the previous MCU production, Black Widow, released earlier in 2021. Furthermore, since there were a lot of scenes that featured martial arts, viewers would also compare the martial arts featured in Shang-Chi with the ones featured in Daredevil. Finally, since the movie featured cultural representation, notably Asian representation, it has often been compared to Black Panther, another MCU movie that was heavily praised for its accurate depiction of African culture.

Group Member Contribution

Marlene Liang collected the tweets, designed the topics, annotated the data, and contributed to the introduction and discussion portion of the written report.

Teresa Altamirano Mayoral filtered and cleaned the tweets, helped design the topics, annotated the data, and contributed to the data and discussion portion of the written report, and to the formatting of the LaTeX file.

Parsa Yadollahi designed the topics, annotated the data, characterized the topics, wrote the script to compute the tf-idf and created the graphs showing the results. He also contributed to the methods, results, and discussion portion of the written report.

References

Creton, Destin D. (Director). (2021). Shang-Chi and the Legend of the Ten Rings

IMDb.com. (2021, September 3). Shang-Chi and the legend of the Ten rings. IMDb. Retrieved December 14, 2021, from <https://www.imdb.com/title/tt9376612/>.

Shang-Chi and the legend of the Ten rings. Rotten Tomatoes. (n.d.). Retrieved December 14, 2021, from https://www.rottentomatoes.com/m/shang_chi_and_the_legend_of_the_ten_rings.

Appendix

A: Hashtags Filter

- shangchi
- katychen
- xialing
- wenwu
- mengerzhang
- tenrings
- awkwafina
- meng
- benkingsley
- shangchiandthelegendofthetenrings
- tonyleung
- thetenrings
- simuliu
- tonyleungchiuwai
- michelleyeoh
- falachen
- yingli
- (The empty string was also considered)

B: Relevant Keywords

This list was formed after parsing a sample of a 100 tweets and collecting the most relevant keywords in the content of the tweets. It was found that the keywords and the hashtags used were similar with the exception that in the content of the tweets there were mentions marked by the character '@'. Moreover, the empty string was excluded from the relevant keywords.

- shangchi
- katychen
- xialing
- wenwu
- mengerzhang
- tenrings
- awkwafina
- meng
- benkingsley
- shangchiandthelegendofthetenrings
- tonyleung
- thetenrings
- simuliu
- tonyleungchiuwai
- michelleyeoh
- falachen
- yingli
- @shangchi
- @simuliu
- shang chi
- shang-chi

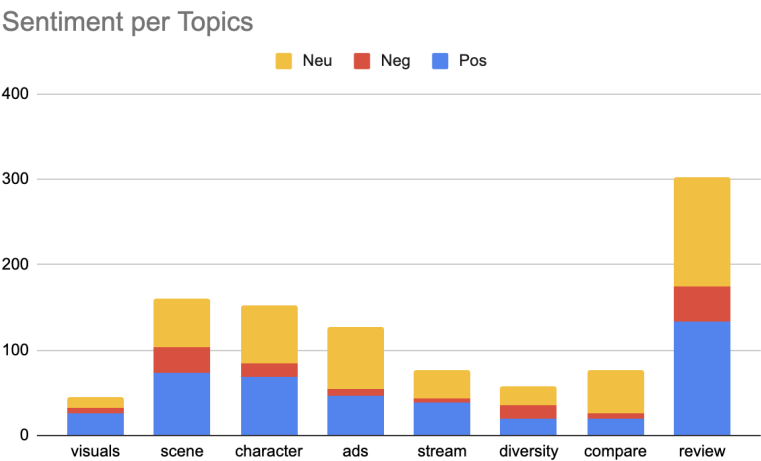
C: Topics Construction

All the topics are placed in order of precedence. In other words, if a tweet belongs to more than one topic, the tie breaker is the precedence.

Topics	Definitions	Keywords
ads (17 posts)	Tweets having the purpose of advertising the movie, products or media content related to the movie. If the tweet refers specifically to a streaming service, <i>stream</i> takes precedence.	vote: 12.477 pcas: 12.477 theactionmovie: 11.09 medusa: 8.318 available: 8.318 est: 8.318 nobis: 8.318 digest: 6.238 stay: 6.238 bonus: 6.238
stream (16 posts)	Tweets mentioning different streaming services the movie is hosted on. If the tweet advertises non-streaming viewing services, it also belongs to <i>stream</i> .	streaming: 13.863 stream: 11.09 bluray: 9.704 disneyplusday: 8.318 plus: 6.931 households: 6.238 disneyplus: 5.64 available: 5.545 rednotice: 4.159 revisit: 4.159
diversity (25 posts)	Tweets focusing on diversity displayed in the movie. The notion of diversity relates to external diversity (nationality, appearance, religion etc.) and internal diversity (race, gender, age, ethnicity, sexual orientation, etc.).	simuliu: 8.318 chinese: 6.238 matters: 6.238 eat: 6.238 asian: 5.64 representation: 5.545 success: 4.159 cheung: 4.159 country: 4.159 countries: 4.15
Scene (53 posts)	Tweets focusing on specific scenes. This can be done implicitly or explicitly. Tweets focusing on a scene should only mention ideas that are related to the plot of the movie as opposed to the visual effects of a scene. A tweet is also considered a scene if it relates to story lines within the MCU Universe.	scene: 36.737 bus: 28.419 fight: 15.98 driver: 15.249 nothing: 14.712 father: 12.477 shorts: 10.397 kill: 10.397 aim: 9.704 deleted: 9.704

character (47 posts)	Tweets focusing on a specific character or a set of characters. This topic is also relevant when the tweet discusses an actor or actress that portrays a character in the movie.	loves: 8.318 morris: 6.931 wenwu: 6.11 ben: 5.545 kingsley: 5.545 michelle: 5.545 yeoh: 5.545 katy: 5.545 guy: 4.904 leung: 4.23
compare (11 posts)	Tweets comparing Shang-Chi with other movies (usually Marvel movies).	spiderman: 16.636 blackwidow: 8.827 panther: 8.318 daredevil: 8.318 spiderman: 6.238 lord: 6.238 nowayhome: 6.238 homecoming: 6.238 home: 5.545 worldwide: 5.545
visuals (15 posts)	Tweets focusing on the beauty and cinematography displayed in the movie. This topic also includes any discussion surrounding the visual effects in the movie. Visual recreations such as fanart of certain movie scenes are also considered to be in this topic	fanart: 5.545 contrast: 4.159 lighting: 4.159 halal: 4.159 cgi: 4.159 design: 4.159 digitalart: 4.159 characterdesign: 4.159 effects: 2.773 charge: 2.773
reviews (16 posts)	Tweets reviewing the movie and giving a personal opinion. This can be done implicitly or explicitly. If there is more information that concerns another category, the other topic has precedence. Also applies if it mentions re-watching or watching since it shows interest or not in the movie.	rewatch: 10.397 day: 8.827 dope: 8.318 finished: 8.318 origin: 8.318 calling: 8.318 masterpiece: 8.318 tier: 8.318 holy: 8.318 yet: 7.847

D: Sentiment per Topic



E: Tweets per Topic

