Comp 598 Final Project

Project 2: Movie Release

Angele Park Collin (260898256), Calla Lu (260895843), Vitoria Soria (260894388)
angele.parkcollin@mail.mcgill.ca, calla.lu@mail.mcgill.ca, vitoria.larasoriacastelolima@mail.mcgill.ca

Introduction

The task we chose for our final project was Project: 2 Movie Release. The task was to help a media company understand the discussions currently happening around a recently released film. To get more specific answers regarding the audience's response the company wanted us to analyze based on certain factors:

- the salient topics discussed around the film and what each topic primarily concerns,
- relative engagement with these topics,
- the general response to the movie, such as if it was positively received or negatively received.

For our project, we chose the James Bond movie *No Time to Die*. Based on this, our general overview of the audience's response to the movie is that it was a great movie that performed very well in box office and in reputation. Certain key findings were that first, most of the tweets were surrounding ratings both positive or negative from the audience, or discussions surrounding the box office and how it compared to other films. Most reviews were positive as the audience were well acquainted fans of the James Bond series and congratulated the movie for being a solid ending. Another key finding was that the tweets could be grouped into 6 distinct categories being actors, box office, comparisons, fans, ratings, and other.

Data

To understand the audience's response on the film, our analysis will be drawn based on 1000 Twitter posts, collected within a 3-day window to ensure that each tweet is unique – no tweet in the collection was repeated or a retweet. To maximize relevance to the film chosen, we set filters such that all 1000 posts have a high likelihood of being related to the movie and are in English. We then filtered further by hashtag with key words such as 'James Bond', 'Movie'. Our choice to filter further using this word was because through an initial test run, we found that a lot of tweets were irrelevant and not at all related to our movie, even though the movie name "No Time to Die" is quite unique on its own. Therefore, we chose to use additional filters such as "James Bond" as the movie is part of the series and it would help narrow down the scope of the tweets. The filter for the word "movie" also helped narrow down the scope of the tweets because although the "James Bond" filter was applied, the James Bond series include the original novels that the movies have names based upon as well, so there may be tweets pertaining to the series but not necessarily talking about the movies. Therefore, this was a necessary filter as well.

Methods

We decided to use Python in our scripts for the data collection process because all three of us are familiar with Python and the language provides many data collecting packages that made the process a lot faster and easier. Our approach to collecting the data was as follows. First, we set up a Twitter Developer account so we could access Twitter's API to collect tweets from users as data. Then, we proceeded to create a Python script, named "collect.py", which aided in fetching the specific tweets pertaining to the information that we wanted. Our collect script first required us to gain access to Twitter's API by using token authentication. We could then access our base URL which has filters for tweets specific to what we needed. To explain, we filtered for no retweets, hashtags such as '#No Time to Die', for tweets to be in English, and lastly for the tweets to be from the most recent postings, as we wanted them to be in order and avoid random duplicates. Furthermore, the tweets were then collected in a json format, as it was the most efficient way to collect such a large amount of data, The json format also

allowed us to easily access specific fields in the data, which was preferrable to us as we only needed certain parts of the tweets.

The next step after collecting the tweets was to conduct an open coding to see what the common topics were amongst the different tweets to find categories to code them in. After this process, we deduced the data could be categorized into the following 6 categories: Actor, Box Office, Comparison, Fan, Ratings, Other. Actor, Box Office, and Ratings were the most obvious as we imagine it would be for any movie analysis. The Comparison category was chosen as we noticed this movie was often compared to either other movies in the James Bond series or to other recently released movies. After deciding on these categories, we manually annotated each of the tweets based on which category they fell under and whether they were of positive, neutral, or negative sentiment.

Lastly, we had to characterize our topics by computing the top 10 highest tf-idf words for each of the categories. We started by creating a script named "divide_category.py". The purpose of this script was to divide the tweets based on each of the different categories. We then created a separate script called "analysis.py" to analyze the output of the 'divide_category.py' file. The analysis consisted of taking all the tweets in each category and computing the top 10 highest tf-idf for that category. We repeated this process for each of the 6 categories that we had and outputted the data into a json file.

Results

From the data collected, we conducted an open coding. The process of open coding consisted of manually going through the first 200 tweets from the data that we had collected and then categorizing the data further into relevant categories. The 6 relevant topics that we chose from the open coding were:

Actor, Box Office, Comparison, Fan, Ratings, Other.

- Actor (A) – Which consisted of any mention of the name of an actor/actress from the movie, such as praise or criticism, or how well they played the role of a character in the film.
- Box Office (B) – Which consisted of any mention of how much or little the movie grossed, or new box office records.
- Comparison (C) – Which consisted of any comparison of the movie to any of its others in the franchise, or any other recently released movie.
- Fan (F) – Which consisted of post movie reactions from the audience such as fan art, podcasts for discussion or criticism, open discussion for the film.
- Ratings (R) – Which consisted of any positive, negative, or just sentiment reviews on the film itself.
- Other (O) – This topic consisted of just random pictures of the movie that included no context, quotes from the movie with no other context, or anything unrelated to the movie itself. None of the findings in this category helped to access any sort of reaction from the audience about the movie which is why we grouped it with anything else that was unrelated to the movie.

Following the selection of the topics, we annotated each tweet based on its sentiment – meaning whether it was appositive, negative, or neutral tweet to further gain insight on the audience's response to the film. Once all the tweets were grouped into topics and categorized based on sentiment, we categorized the topics by computing the 10 words in each category with the highest td-idf scores. The tf-idf was computed using tf = the number of times the category used the word x, idf = log [total number of categories/number of categories that uses the word x].

Discussion

Based on the characterization of our topics, we think that for the most part, the words in each category are fairly accurate to its topic. By looking at Figure 1 below, we can analyze

Figure 1

each of the categories individually and provide reasonings as to how the results are accurate. For the actor category, the top 10 words are coherently representative of the tweets in the category. Many of them pertaining to actors were complimenting the new actors in this film who stood out with their performances. The top tweets include the words "Ana", "de", "Armas" (as in actress Ana de Armas), "Rami" (as in actor Rami Malek), with "praise" as another top word, thus complimenting their performances. Additionally, many of
the names the actors played were also mentioned, such as "Nomi" and "Idris Elba" - two noticeable characters that stood out to the audience. The former being a new and ex -

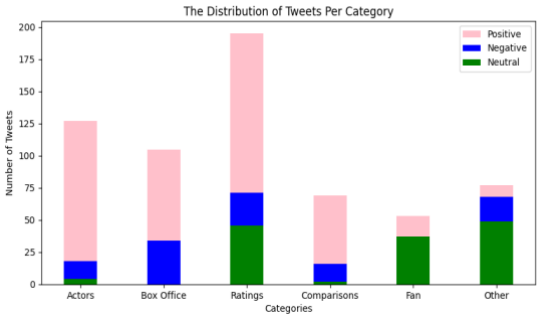

Figure 2

citing character that the audience really enjoyed, and the latter being an old recurring character that met his ultimate demise in the film, causing a lot of discussion. In the box office category, the words primarily relate to discussions surrounding box office earnings and grossing, such as "million" and "lose" both referring to the amount the movie has made or lost in earnings. The negative sentiment was used here to regard any tweet that mentioned loss of money which explains which is shown in Figure 2. The reasoning behind their loss of earring is due to the extended costs, as well as hidden costs that have been accrued due to the pandemic (Southern 2021). For the Comparisons category, the movie *No Time to Die* was being regularly compared to many other top grossing films at the time, hence why films such as the movie *Ghostbusters: Afterlife* alongside *Eternals* were being mentioned frequently in these tweets. For the Fan category, many of the tweets were excited James Bond fans who seemed to really enjoy the final movie. They created additional media to support their love of the series, which included but was not limited to podcasts and fanart. This explains as to why this category contains a lot of top words such as "podcast", "drawing", "artwork", and "episode" (referring to podcast episodes they made about the movie). Furthermore, the mention of telling people to listen to the podcasts were neither positive nor negative which explains why the neutral sentiment is strongly observed in Figure 2. The Ratings category was the most popular category by far with most of the tweets being a rating of the movie. As such, many of the top words are descriptive words that convey either a positive or negative tone for the movie, most of them being positive. Moreover, many comments about the movie are regarding the cinematography, which appeared as the fifth top word in the list. The last word in this category "farewell" is explained by the fact that it is the last movie in the series as well as a farewell to the main character since he died in the movie. In the last category, Others, we included all tweets that don't fit in any of the former categories. These tweets were either completely unrelated to the movie or they were simply a niche enough topic that could not be properly categorized in any of the other categories. For example, many tweets were simple comments mentioning that they were going to watch the movie. While this is talking about the movie *No Time to Die*, it could not be counted as any of the other categories since it was not adding any valuable insight to the movie. Therefore, words such as "watch" and "tonight" are frequent words used in this category because people were mentioning that they would be 'watching the movie tonight'. Consequently, this also explains the significant green in this category since 'watching a movie tonight' provides neither a positive nor negative sentiment. All in all, the categories were made with great consideration in terms of the data, and we were able to categorize all of them with accuracy and with meaning. The categories help provide more insight into the initial question – the favorability of the audience's response regarding the discussions currently happening around the film - and helped define the most pressing reactions to the movie from a handful of different groups. Our audiences' reactions included fans of the series, first time James Bond movie watchers, movie theatre accounts raving about box offices, and movie critics.

Group Member Contributions

There is a group consensus that the work was distributed fairly, and each member contributed their fair share to the project. We collectively met up at the beginning to decide which project we wanted to do as well as split up the work. During that meeting, all three of us worked together to get the twitter API working, pick a movie, code up the file to filter the tweets and to append them into a file once fetched, and conducted the open coding to find the relevant topics. From there, Angele ran the code to collect the tweets over the span of 3 days, as we thought it would be simpler and less room for error if one person took care of this portion. Once we had all 1000 tweets collected, we evenly split them between the three of us to manually annotate each tweet into

a topic and whether the tweet was of positive, negative, or neutral sentiment. We then collectively reviewed all 1000 tweets to make sure that we agreed on the topic and sentiment of each tweet to minimize error. The next step was to compute the td-idf scores, so Vitoria took this on as we could not find time where the three of us could meet to work on it. Lastly, for final report, Angele and Calla worked on it together and gradually pushed so we could review each other's work. At the end, the three of us read over the final report to ensure everything was done properly and that everyone was content with the work produced.

References

Southern, K. 2021. No Time To Die to lose money despite taking $730m at the box office. *The Times*, Los Angeles.