 doi:10.20944/preprints202206.0246.v1

Data Descriptor

A Dataset Containing Tweets and Their Meta Data for Understanding Social Media Conversations around Movies during Their Release

Joost Michielsen and Francesco Lelli

Tilburg University; j.p.j.michielsen@gmail.com, f.lelli@tilburguniversity.edu

Abstract: In this paper we intend to present a dataset that contain a collection of tweets generated as reactions of the release of 50 different movies. The dataset can be used for gaining useful insights regarding the conversation that is generated around a particular movie. It is particularly suitable for conducting sentiment analysis and other NLP techniques. The dataset contains approximately 2.5 million tweets with their related meta data and cover 50 movies. For each movie, its IMDb rating is included. The movies are the 25 releases with the highest number of votes during 2020 and 2021. The collected tweets represent the reactions of the twitter community during the first week of the release date in US of that particular movie. The tweets per movie ranged from 1.000 to approximately 200.000 tweets with an average of 50.000 per release. We used The Internet Archive Wayback Machine in order to retrieve the IMDb movie rating after one week of the US release date. The tweets and related metadata have been collected using the Tweet Downloader tool.

Keywords: dataset; twitter; tweets; IMDb ratings; movies; sentiment analysis; NLP

1. Introduction

The data that we are releasing intends to offer a benchmark of reference for analysing how movies foster a conversation among their watchers in social media. The dataset contains data in the form of tweets, tweets meta data, and IMDb ratings related to the movies with highest number of votes in 2020 and 2021. The tweets have been collected with a Twitter API tool called Tweet Downloader, made available to Twitter academic research accounts, using a specific filter that we will describe in section 2. The IMDb ratings were collected by the Internet Archive Wayback Machine in order to retrieve the rate of the movie seven days after the release. The use of tweeter for analyse movies is a relatively common practice. For example, a study used tweets to predict the box office revenues of movies [1]. In another case researchers aimed to predict IMDb movie ratings by using surface and textual features of social media [2]. In addition, the study presented in [3] aimed to predict the rating of a movie based on tweets. We believe that this dataset may be useful for ammonizing similar initiatives and foster reproducibility of the results. This dataset, containing roughly 2.5 million tweets, was used to perform sentiment analysis on tweets and analyse the relationship between the sentiment polarity score of a movie and the IMDb movie rating of that movie.

The following specifications table is a summarization of the dataset:

Subject	Computer science, Linguistics, and Artificial Intelligence
Specific subject area	Natural Language Processing is a subpart of the aforementioned subjects and combines all these disciplines. Sentiment analysis, in turn, is part of Natural Language Processing and is able to calculate a sentiment polarity

 doi:10.20944/preprints202206.0246.v1

	score which can be interpreted to see how positive or negative a particular piece of text is.
Type of data	• Date • String (tweet ID, tweet text, author ID, etc.) • Tweet meta data (type of tweet, public metrics of a tweet, verified user or not, etc.) • Float (IMDb rating)
How data were acquired	The IMDb ratings were collected by using the Internet Archive Wayback Machine. Tweet data was acquired by using the Tweet Downloader, which is a Twitter API tool that can be used by Twitter Academic Research accounts.
Data format	Raw
Parameters for data collection	To ensure that enough tweets would be posted about a certain movie it was chosen to select the top 25 movies of IMDb with the highest number of votes from both 2020 and 2021. For each specific movie the query to collect the data, in the form of tweets, existed of the movie name or hashtag used by the official movie account. Next to that each query also specified that either the word “movie” or “film” must be present in the text of the tweet. Lastly, only English tweets were collected. Besides the text numerous fields could be collected as well. See table 2 for the details.
Description of data collection	For each of the 50 movies the US release date and two ratings were recorded: the past and present movie rating. The past rating is the movie rating after 7 days of the US release date. By making use of the Internet Archive Wayback Machine (https://web.archive.org/) it was possible to retrieve the IMDb movie rating in the past of a certain movie. Next to the past movie rating, the present movie rating was also recorded. In order to collect these tweets the Twitter API tool, Tweet Downloader, was utilized. This process applied to each movie and consisted of two steps: 1. Collection period: Starting on US release date and ending 7 days after 2. Query specification based on three elements: I. English Language II. The exact movie name or hashtag as used by the official Twitter account from a movie III. the text of the tweet should always contain the word ''movie'' or ''film'' The selected parameters for each movie were the same as listed in the section above.

 doi:10.20944/preprints202206.0246.v1

Data source location	Data available at the following URL: https://doi.org/10.34894/6WEAUR
Data accessibility	Repository name: DataverseNL Data identification number: 10.34894/6WEAUR Direct URL to data: https://doi.org/10.34894/6WEAUR Instructions for accessing these data: Data are open access

2. Methods for Data Acquisition

The data have been acquired using Twitter and the Wayback Machine. The rest of this section is therefore divided considering these 2 systems in a separated manner.

The Internet Archive Wayback Machine was used to collect all the past ratings of each movie. By copying each link of the concerned movie on the IMDb website, the Internet Archive Wayback Machine searched for all the captured images in the past. For a few movies there were no records of the past movie rating after exactly 7 days of the US release date. In that case, a record for the 8th, 6th, 9th, or 5th day after the US release date was successively looked at. The record of chosen past movie rating was also selected based on that order. All movies had a record within the range of five to nine days after the US release date. Next to the past movie rating, the present movie rating was also recorded. All ratings were stored in a coma separated (.csv) file.

Tweets and tweet meta data were collected by making use of the Tweet Downloader, a Twitter API tool. The same tweet collection process applied for each movie. First the collection period was established which starts on the US release date of the movie and ends seven days after the starting date. Next, the query to extract tweets for each movie was specified which consists of three elements. The first element is the language of the tweet. For all tweets only English tweets were collected. The second element ensures that only tweets are downloaded that contain the exact movie name or hashtag as used by the official Twitter account from a movie. In the case there is no official Twitter movie account then the hashtag that is used by the producer, director, or top actor of the movie was used. If there is no specific hashtag based on the aforementioned criteria, then the movie title is used as hashtag. The third element specifies that the text of the tweet should always contain the word "movie" or "film". This ensures that as few tweets as possible are collected that are not actually about the intended movie. A few examples of the queries that were used:

- Tenet: lang:en (“Tenet” OR #Tenet) (Movie OR Film)
- Onward: lang:en (“Soul” OR #PixarOnward) (Movie OR Film)
- Spider-Man: No Way Home: lang:en (“Spider-Man: No Way Home” OR #SpiderManNoWayHome) (Movie OR Film)
- Encanto: (“Encanto” OR #Encanto) (Movie OR Film)

Based on inspection of the first 100 tweets of each movie, it was concluded that a negligible number of tweets are not about the actual movie. The tweets per movie ranged from 1.100 to approximately 200.000 tweets which were all posted within the first week of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0246.v1

the US release of the movie. On average 50.000 tweets per movie were collected, resulting in a total dataset consisting of roughly 2.5 million tweets about movies. These tweets have been stored in a .csv file

2.1. List of resources used for the collection of data

- Internet Archive Wayback Machine. Available at: https://web.archive.org/
- IMDb top 25 of 2020 based on most number of votes. Available at: https://www.imdb.com/search/title/?title_type=feature&release_date=2020-01-01,2020-12-31&sort=num_votes,desc
- IMDb top 25 of 2021 based on most number of votes. Available at: https://www.imdb.com/search/title/?title_type=feature&release_date=2021-01-01,2021-12-31&sort=num_votes,desc
- Tweet Downloader (Twitter API tool). Available at: https://developer.twitter.com/apitools/downloader

3. Results: Data Description

The CSV file ''Movies.csv'' contains all the data concerning the movies and exists of various columns. See table 1 for these columns and their description. The first column presents all the column names, whereas the second column presents the description of each column of the CSV file containing the IMDb movie ratings. In total ten different columns are used. The dispersion of movie ratings can be seen in figure 1. In this figure the x-axis represents the IMDb movie rating after seven days of the US release date. The y-axis is the count of the IMDb movie ratings. The distribution of IMDb movie ratings is somewhat normally distributed around the 7.5 rating with one outlier being an IMDb rating of 2.

Table 1: Column Structure and Description of Movie Data

Column	Description
Movies with highest number of votes on IMDB	Index based on their place in the top 25 of the highest number of votes in 2020 and 2021
Movie title	Title as stated on the English (United States) version of IMDb
Release date (US)	The US release dates based on the IMDb dates hereby it must be noted that the full release dates are used and not premiere or film festival release dates
Tweets up to	Release date plus seven days
Movie rating (past)	Movie rating as close as possible to the seven-day mark after the US release date
Date rating checked	Date of the movie rating in the past
X days after release	Number of days, after the US release date, that the past rating was captured
Movie rating (present)	Current rating of the movie
Difference in rating	Difference in rating between the past and present ratings
# used	Lists all the hashtags that were used per movie in each query to collect the tweets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0246.v1

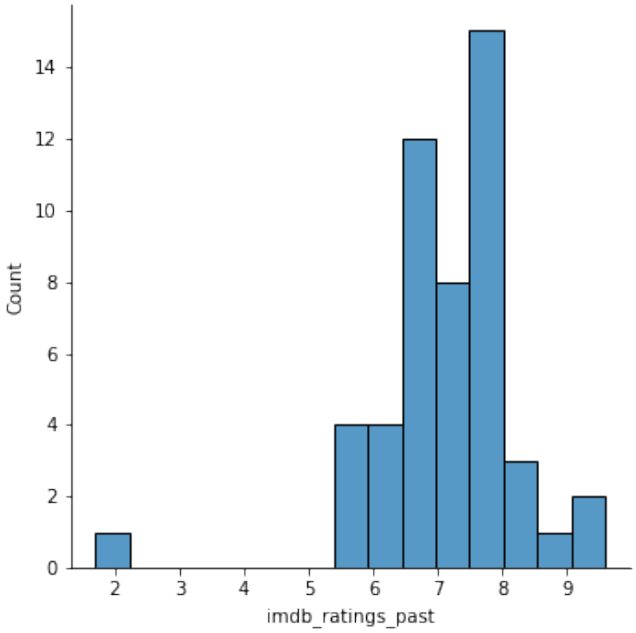

Figure 1: Count of IMDb Movie Ratings

The tweets and tweet meta data are stored in one CSV file. This file is named "Tweets_dataset.csv" and consists of several columns. These columns with description are shown in table 2. The first column presents all the column names, and the second column presents the description of each column of the CSV file containing the tweet data. The total count of tweets per movie can be seen in figure 2. This figure presents two bar charts both with the same x-axis and y-axis. The x-axis is the count of tweets, and the y-axis presents each movie for 2020 and 2021. In both bar charts the tweet count per movie range from roughly 1.000 up to 220.000 tweets.

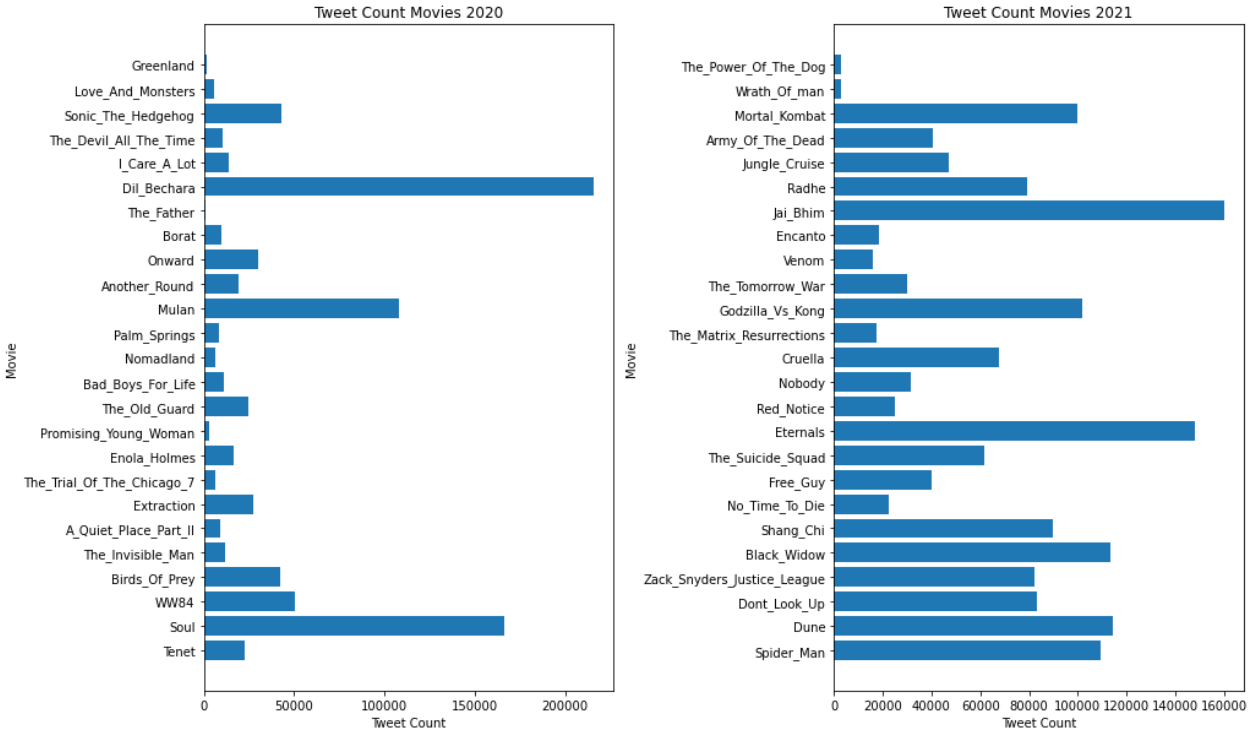

Figure 2: Tweet Count per Movie

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0246.v1

Table 2: Column Structure and Description of Tweet Data

Column	Description
text	the actual text of the tweet
author_id	unique id belonging to a certain Twitter account
entities	lists all the given entities such as usernames, persons, hashtags etc. that are in the text of the tweet
possibly_sensitive	Boolean to clarify if a tweet is possibly sensitive
created_at	the exact date and time the tweet was created
conversation_id	unique ID that refers to the conversation on Twitter (if any)
id	unique ID for each tweet
public_metrics	retweet, reply, like and quote count for the tweet
reply_settings	show who can reply to a certain tweet
author	contains information of the author such as description, username, follower count etc.
referenced_tweets	if the tweet is a retweet, reply or quote this column will show the referenced (original) tweet as well as the author of the referenced tweet with meta data included
geo	shows the added location to a tweet (if any)
in_reply_to_user_id	show the ID of the user that is being replied to
withheld	shows the details of information that is withheld (if any)
movie_title	three of four numbers to indicate about which movie the tweet goes where the first two numbers indicate the year, for example 201 stands for the first movie in 2020 and 2123 for the 23th movie in 2021.
reference	shows the type of a tweet (orginal/retweeted/replied to/quoted which is taken from the referenced_tweet column for easier text handling
actual_text	the text of tweet, in case of a retweet the original text of the retweet is in this column and not the retweet itself
textblob_text	preprocessed text for textblob sentiment analysis
vader_text	preprocessed text for vader sentiment analysis
textblob_sentiment	textblob sentiment polarity score performed on the actual text
textblob_filtered_sentiment	textblob sentiment polarity score performed on the preprocessed textblob text
vader_sentiment	vader sentiment polarity score performed on the actual text
vader_filetered_sentiment	vader sentiment polarity score performed on the preprocessed vader text
verified	boolean which shows if the user that posted the tweet is verified
like_count	amount of likes the tweet has
author_followers_count	number of followers the author of the tweet has

 doi:10.20944/preprints202206.0246.v1

4. Discussion

The collected and presented dataset can be valuable for several purposes:

- With micro blogging social media platforms like Twitter, it is very convenient and fast to express your own opinion about a topic online. Therefore, public opinions about all kind of subjects can quickly be visible if gathered from such social media platforms. Where the collection of data is very fast and relatively easy compared to spreading questionnaires about a certain topic. This data gives insight in the public opinion of Twitter users about 50 different movies.
- In the case that more data is needed, it is possible to add more twitter data about other movies from different years. Also, it would be possible to collect more data about the same movies by using a collection period of more than one week.
- The data can be used for various research activities or studies. Students could use the dataset to learn how to perform sentiment analysis. Next to that, researchers could reuse the data for their own research or experiment.
- In the future this dataset of tweets might be used for other sentiment analysis methods, for example different machine learning techniques, to see how other methods perform. Also, various other analyses can be executed on the tweets, for example an emotion analysis. Even other NLP tasks could be utilized on the dataset.

5. Conclusion

In this preprint a dataset containing roughly 2.5 million tweets about movies was presented. For a total of 50 movies the tweets were collected. The collected dataset was used for sentiment analysis on movie tweets as well as analysing the relationship between the sentiment polarity of movie tweets and the IMDb movie rating of a certain movie. Previous sections described the methods that were used for collection as well as a detailed description of the data files. Next to that, the value of the dataset was discussed, presenting several purposes in which the dataset could be used.

Declaration of Competing Interest: The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

References

[1] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (Vol. 1, pp. 492-499). IEEE.

[2] Oghina, A., Breuss, M., Tsagkias, M., & Rijke, M. D. (2012, April). Predicting imdb movie ratings using social media. In *European conference on information retrieval* (pp. 503-507). Springer, Berlin, Heidelberg.

[3] Schmit, W., & Wubben, S. (2015, September). Predicting ratings for new movie releases from twitter content. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 122-126).