Prediction of Movies Box Office Performance Using Social Media

Vaibhav K Nigam (vnigam@syr.edu)
Student, School of Information Studies
Syracuse University

Abstract – The content of social media contains rich information about people's preference for the things going around the world for instance, opinion of people in the form of tweets for an upcoming movie after watching official trailer released by the production house. I performed data analysis techniques on tweets about movies to predict several aspects of the movie popularity. To generate some interesting patterns for predicting box office performance of movies I collected data from multiple sources including IMDb movie database, Rotten Tomatoes and tweets from twitter. The outcome of this research is to predict whether a movie would be successful at the box office.

Keywords – Text Mining, Sentiment Analysis, Twitter, IMDb, Rotten Tomatoes, K-Means Clustering, Decision Tree Classification, Prediction, Box Office Success

Introduction

Social media such as Twitter is used as a common platform to share preferences, comments and content on all types of subjects by millions of people on a daily basis using free-format, limited-length texts, and these texts (often called "tweets") provide rich information to the companies/institutes who are interested in knowing whether people like a certain product, movie or service. Twitter is a micro blogging website which plays an important role in the research of social network. Film industry business has a strong interest in tapping into these huge data sources to extract information that might help them in getting a better indication of whether their movie is going to be liked by the audience or not also they can plan their moves accordingly by making a better decision-making process. For instance, predictive models derived from social media for successful movies may facilitate filmmakers making more profitable decisions.

The strategy was to first identify the number of tweets about the movie prior to its releasing date, and the to apply clustering model to cluster the movies into two categories of hit or flop and then implement classification model to get an idea whether the model created is generating affirmative results or not. The result can be trusted on the basis of the accuracy of the model which in this research comes out to be 71%. The former was more challenging task because of the restrictions implied on fetching data from social media sites, as matter of fact data back to two weeks could only be fetched from twitter. I used hashtags to gather information about the upcoming movies and to get a better idea of how much the specific movie is talked about, this will be further discussed in the data collection phase of the research paper. However, how to build an engine to detect and summarize user preference accurately remains a challenging problem.

Related Work

The topic of using social media to predict the success of a movie at box office has become vary popular in recent years. Some research papers have worked in showing that twitter-based prediction of box office revenue performs better than market-based prediction by analyzing various aspects of tweets sentiment prior to the movie being released. I have taken help of these research works and implemented a similar approach in creating a prediction model.

Sentiment analysis of twitter data is also hot research topic in recent years. While sentiment analysis of documents has been studied for a long time, the techniques may not perform very well for the data collection from twitter because of the characteristics of tweets. The following are a few difficulties in processing twitter data: the tweets are usually short up to 140 words. The text of the tweets is often ungrammatical some of the research papers have investigated features of sentiment analysis on tweets data.

However, few works directly used sentiment analysis results to predict the success of the movie at the box office. Some of the related work did sentiment analysis but did not used it's results explicitly.

Methodology

I have used sentiment analysis results of tweets prior to the movie being released to predict the box office success of the movie. My methodology consists of three steps:

I. Data Collection

A. IMDb movie database

I used IMDb movie database to get the dataset of upcoming movies along with their release date. I have focused my research on the upcoming movies to be released in the month of December'18, January'19 and Feburary'19. The below table represents the upcoming movies for the following months.

Movie Title	Release Date
Mary Queen of Scots	December 7, 2018
Dumplin'	December 7, 2018
Vox Lux	December 7, 2018
Ben Is Back	December 7, 2018
Spider-Man: Into the Spider-Verse	December 14, 2018
Mortal Engines	December 14, 2018
The Mule	December 14, 2018
Roma	December 14, 2018
Backtrace	December 14, 2018
Capernaum	December 14, 2018
Mary Poppins Returns	December 19, 2018
Aquaman	December 21, 2018
Bumblebee	December 21, 2018
Cold War	December 21, 2018
Holmes & Watson	December 21, 2018
Welcome to Marwen	December 21, 2018
Zero	December 21, 2018
Second Act	December 21, 2018

Vice	December 25, 2018
On the Basis of Sex	December 25, 2018
Destroyer	December 25, 2018
Escape Room	January 4, 2019
Animal Crackers	January 4, 2019
Replicas	January 11, 2019
The Upside	January 11, 2019
A Dog's Way Home	January 11, 2019
Ashes in the Snow	January 11, 2019
Dragon Ball Super: Broly	January 16, 2019
Glass	January 18, 2019
Girl	January 18, 2019
Serenity	January 25, 2019
The Wild Pear Tree	January 30, 2019
Arctic	February 1, 2019
Miss Bala	February 1, 2019
Cold Pursuit	February 8, 2019
Everybody Knows	February 8, 2019
The Lego Movie 2: The Second Part	February 8, 2019
What Men Want	February 8, 2019
The Prodigy	February 8, 2019
Alita: Battle Angel	February 14, 2019
Fighting with My Family	February 14, 2019
Isn't It Romantic	February 14, 2019
Happy Death Day 2U	February 14, 2019
How to Train Your Dragon: The Hidden World	February 22, 2019
All-Star Weekend	February 22, 2019

Table 1

Apart from fetching the names and release dates of the upcoming movies, I scraped the data of the production house rating which I have considered as a feature in predicting the box office success.

B. Twitter Data

I have gathered the tweets for the month of November'18 from twitter using the streaming API of Tweepy to get the tweets relevant to my task. The API, tweepy.streaming.stream, continually retrieves data relevant to some topics from Twitter's global stream of tweets data. I used the unique consumer_key, consumer_secret, access_token and access_token_secret of the twitter application to access the tweets of various people around the world. I searched for the respective

movies using hashtags feature of Twitter to extract the specific tweets about the specific movie. I used tweepy.Cursor function to pass a query to search for respective hashtag, language of the movie as English and the date since when I need the tweets. A separate CSV file was created for each movie which was further processed for fetching the sentiment score of the specific movie, this will be explained further in the data pre-processing part. Date of tweet and the content of tweet is stored in the CSV file, the date column is used for getting an idea of the duration of tweets scraped. The below table represents the sentiment percent of the respective movies,

Movie	Percent Positive Tweets	Percent Negative Tweets
Mary Queen of Scots	0.73913	0.26087
Dumplin’	0.877384	0.122616
Vox Lux	0.802632	0.197368
Ben Is Back	0.851485	0.148515
Spider-Man: Into the Spider-Verse	0.635719	0.364281
Mortal Engines	0.738739	0.261261
The Mule	0.650685	0.349315
Roma	0.836232	0.163768
Backtrace	0.5	0.5
Capernaum	0.756757	0.243243
Mary Poppins Returns	0.717276	0.282724
Aquaman	0.645855	0.354145
Bumblebee	0.92562	0.07438
Cold War	0.58011	0.41989
Holmes & Watson	0.630769	0.369231
Welcome to Marwen	0.96875	0.03125
Zero	0.915501	0.084499
Second Act	0.871744	0.128256
Vice	0.597765	0.402235
On the Basis of Sex	0.267816	0.732184
Destroyer	0.780142	0.219858
Escape Room	0.842276	0.157724
Animal Crackers	0.846154	0.153846
Replicas	0.880795	0.119205
The Upside	0.918138	0.081862
A Dog’s Way Home	0.87931	0.12069
Ashes in the Snow	1.0	0.0
Dragon Ball Super: Broly	0.967195	0.032805
Glass	0.858902	0.141098
Girl	0.829484	0.170516
Serenity	0.742254	0.257746
The Wild Pear Tree	0.666667	0.333333
Arctic	0.758503	0.241497
Miss Bala	0.968811	0.031189

Cold Pursuit	0.590909	0.409091
Everybody Knows	0.348624	0.651376
The Lego Movie 2: The Second Part	1.0	0.0
What Men Want	0.7	0.3
The Prodigy	0.706897	0.293103
Alita: Battle Angel	0.541667	0.458333
Fighting with My Family	0.969052	0.030948
Isn't It Romantic	0.511628	0.488372
Happy Death Day 2U	0.6	0.4
How to Train Your Dragon: The Hidden World	1.0	0.0
All-Star Weekend	1.0	0.0

Table 2

Along with calculating the sentiment score of the respective movies I used web scraping to get the follower count of director, leading actor and actress. I used it as a feature for predicting the success of the movies as many people might want to watch the movie just because it is created by their favorite director or their favorite actor or actress is portraying in that specific movie. These factors also affect a lot on the revenue collected by the movie at the box office. The below table represents the follower count gathered from the twitter pages of the respective director, leading actor and actress,

Movie	Director	Actor	Actress
Mary Queen of Scots	23884	69	NA
Dumplin'	152	2274	39138
Vox Lux	225	NA	44788
Ben Is Back	5	459	517380
Spider-Man: Into the Spider-Verse	4893	NA	1032458
Mortal Engines	178	2752	4324
The Mule	73089	73092	NA
Roma	120251	NA	38636
Backtrace	3136	2801180	NA
Capernaum	740833	NA	NA
Mary Poppins Returns	331	NA	9279
Aquaman	211653	16450	73713
Bumblebee	686	2021	1032458
Cold War	1	14	137
Holmes & Watson	4384	10008	NA
Welcome to Marwen	NA	5606341	1429836
Zero	151492	36877567	1847777
Second Act	NA	586173	7232662
Vice	1014175	146	2465

On the Basis of Sex	438	2021	19225
Destroyer	NA	1321	77369
Escape Room	8671	51576	242207
Animal Crackers	8580	3973491	9279
Replicas	NA	5338	624
The Upside	8407	2126444	77369
A Dog's Way Home	338	NA	196394
Ashes in the Snow	NA	NA	41888
Dragon Ball Super: Broly	NA	158384	NA
Glass	155683	50342	47739
Girl	317	NA	NA
Serenity	NA	NA	1592
The Wild Pear Tree	478882	1219838	NA
Arctic	239636	87027	NA
Miss Bala	16986	794300	497777
Cold Pursuit	NA	3314	873979
Everybody Knows	264	21209	513
The Lego Movie 2: The Second Part	67412	NA	1119942
What Men Want	1360	NA	140374
The Prodigy	16150	1609	662214
Alita: Battle Angel	314810	NA	1429836
Fighting with My Family	623571	13029220	18230
Isn't It Romantic	5199	NA	23700666
Happy Death Day 2U	1547	NA	10650
How to Train Your Dragon: The Hidden World	13391	523591	3299
All-Star Weekend	4702043	11962102	NA

Table 3

C. Rotten Tomatoes

I used rotten tomatoes to scrape "User ratings", "Percentage of users who want to see the movie" and "Average rating on a scale of 10". "User ratings" and "Percentage of users who want to see the movie" features were used in predicting and getting an idea of converting these features into predicting the success of the movie at the box office, "Average rating on a scale of 10" was removed as it contained mostly NaN values. The below table represents the data collected from the rotten tomatoes website,

Movie	Rotten Tomatoes User Ratings	Percentage of Users who want to see (%)	Average Rating on a scale of 10
Mary Queen of Scots	495	97	7.4
Dumplin’	25	88	NA
Vox Lux	237	95	7.4
Ben Is Back	145	92	6.3
Spider-Man: Into the Spider-Verse	711	94	NA
Mortal Engines	452	94	NA
The Mule	212	98	NA
Roma	654	98	9.1
Backtrace	NA	NA	NA
Capernaum	108	89	7.6
Mary Poppins Returns	1169	97	NA
Aquaman	10549	97	NA
Bumblebee	851	94	NA
Cold War	353	96	8.2
Holmes & Watson	258	93	NA
Welcome to Marwen	338	97	NA
Zero	14	86	NA
Second Act	81	85	NA
Vice	196	93	NA
On the Basis of Sex	186	94	6.6
Destroyer	203	94	7.3
Escape Room	30	97	NA
Animal Crackers	NA	NA	NA
Replicas	319	96	NA
The Upside	186	89	6.2
A Dog’s Way Home	16	88	NA
Ashes in the Snow	246	95	NA
Dragon Ball Super: Broly	2	NA	NA
Glass	678	99	NA
Girl	NA	NA	NA
Serenity	184	84	NA
The Wild Pear Tree	58	94	8.4

Arctic	50	97	7
Miss Bala	13	92	NA
Cold Pursuit	15	100	NA
Everybody Knows	123	96	6.2
The Lego Movie 2: The Second Part	22990	97	NA
What Men Want	61	84	NA
The Prodigy	17	94	NA
Alita: Battle Angel	546	96	NA
Fighting with My Family	25	92	NA
Isn't It Romantic	24	100	NA
Happy Death Day 2U	39	95	NA
How to Train Your Dragon: The Hidden World	46381	98	NA
All-Star Weekend	NA	NA	NA

Table 4

II. Data Pre-processing and Sentiment Analysis

The data collected is noisy and is in huge amount, I processed them using distributed computing techniques. I further filtered data and got the tweets talking about the movies via regular expression. In the tables mentioned above there are many columns having NA values, in most of the cases the NA values were replaced with the median of the respective columns. In table 4 the column "Average Rating on a scale of 10" was removed as it was mostly having NA values and they couldn't be replaced with the median. The below table represents the features that I have considered in predicting the success of the movie at the box office,

Type	Feature
Nominal	Actor, Actress, Director
Numerical	Percentage of positive tweets, Percentage of negative tweets, Rotten Tomatoes user ratings, Percentage of Users who want to see, Production house ratings

Table 5

The Nominal values were converted into Numerical values by getting the followers count of the director, leading actor and actress for the respective movies. Below table (Table 6) represents the format in which tweets for the movie "Spider-man" were fetched,

2018-11-23 10:23:28	b"RT @SpiderMarriage: Most of #SpiderMan's greatest, most memorable enemies were all co-created by Steve Ditko. #Chameleon, #Vulture, my fav\xe2\x80\xa6"
2018-11-23 10:21:40	b'RT @ELTElite: Summer 2019 is LIT!!! \xf0\x9f\x94\xa5\xf0\x9f\x94\xa5\xf0\x9f\x94\xa5\xf0\x9f\x94\xa5\nMay - #Avengers 4\nJune 21 - #ToyStory4\nJuly 5 - #SpiderMan Far From Home\nJuly 19 - #TheLionKing ht\xe2\x80
2018-11-23 10:20:22	b'Had a great time #streaming #SpiderMan ! If you guys want updates on future streams and original videos come check\xe2\x80\xa6 https://t.co/15VHi9JLEm'
2018-11-23 10:20:08	b'#SpiderMan\xc2\xa0The Movie PlayStation 2 #videogame 2002 Activision\xc2\xa0https://t.co/172omXlC2T\xc2\xa0#Playstation #PS2 https://t.co/zoXiMcubNi'
2018-11-23 10:18:23	b"RT @Goshdesigns: Let's have some #FreebieFriday fun. A free #Watch is our #giveaway on this fine #BlackFriday2018 TO WIN LIKE, SHARE, RT &\xe2\x80\xa6"
2018-11-23 10:17:13	b'RT @ETPrime_com: Can investors get tips from #StanLee\xe2\x80\x99s comic heroes? @lonelycrowd thinks so. Here are 5 investing tips picked up from the\xe2\x80\xa6'
2018-11-23 10:15:45	b"RT @Goshdesigns: Let's have some #FreebieFriday fun. A free #Watch is our #giveaway on this fine #BlackFriday2018 TO WIN LIKE, SHARE, RT &\xe2\x80\xa6"
2018-11-23 10:12:10	b"#Marvel's #SpiderMan #PS4 #eBay\n\xe2\x8f\xb0 Ends in 5h\n\xf0\x9f\x92\xb2 Last Price GBP 22.50\n\xf0\x9f\x94\x97 https://t.co/jAzR7AEQFg https://t.co/vPWMDcl04v"
2018-11-23 10:11:55	b"RT @Goshdesigns: Let's have some #FreebieFriday fun. A free #Watch is our #giveaway on this fine #BlackFriday2018 TO WIN LIKE, SHARE, RT &\xe2\x80\xa6"
2018-11-23 10:11:28	b"RT @Goshdesigns: Let's have some #FreebieFriday fun. A free #Watch is our #giveaway on this fine #BlackFriday2018 TO WIN LIKE, SHARE, RT &\xe2\x80\xa6"
2018-11-23 10:10:42	b'Wow!!! Some incredible prices for #blackfriday today on both #Ps4 & #Xboxone games...Dont miss out ! #hmvCrazyDeals\xe2\x80\xa6 https://t.co/RdLUersFEs'
2018-11-23 10:10:30	b'RT @nerdist: #MorbiusTheLivingVampire will be the next #SpiderMan spin-off film https://t.co/SKd5yVkBRm https://t.co/ou4pof4GH6'
2018-11-23 10:09:31	b"RT @Goshdesigns: Let's have some #FreebieFriday fun. A free #Watch is our #giveaway on this fine #BlackFriday2018 TO WIN LIKE, SHARE, RT &\xe2\x80\xa6"
2018-11-23 10:08:45	b'RT @JustJared: Two #SpiderMan spinoff movies have been given release dates for 2020! https://t.co/PGpSQiMC8v'
2018-11-23 10:08:01	b'Get 70% Off on #TomHolland #Homecoming #SpiderMan Red Hoodie with Sleeves this Black Friday. #BlackFriday #Deals\xe2\x80\xa6 https://t.co/yYXWlw4GTw'
2018-11-23 10:07:16	b'So dark...#spiderman #ps4 #PS4share https://t.co/q3Ks1jZilm'
2018-11-23 10:06:35	b'RT @spankzilla85: #Venomized #Godzilla! The #Spidermen have their work cut out for them! #spiderman #spidergeddon #spidergwen #ghostspider\xe2\x80\xa6'
2018-11-23 10:06:23	b"RT @Goshdesigns: Let's have some #FreebieFriday fun. A free #Watch is our #giveaway on this fine #BlackFriday2018 TO WIN LIKE, SHARE, RT &\xe2\x80\xa6"
2018-11-23 10:06:11	b'RT @RussLeachDraws: a few Spidey faces from "Draw The @Marvel Way". Forget which issue but it was early in the run. #SpiderMan #makingcomic\xe2\x80\xa6'
2018-11-23 10:05:52	b'RT @Mauro_Celentano: Marvel\xe2\x80\x99s Spider-Man (ps4) : into the spiderverse #spiderman\xe2\x81\xa0 \xe2\x81\xa0 #SpiderManIntoTheSpiderVerse #spidermanps4 @insomniacga\xe2\x80\xa6'
2018-11-23 10:02:11	b"RT @Goshdesigns: Let's have some #FreebieFriday fun. A free #Watch is our #giveaway on this fine #BlackFriday2018 TO WIN LIKE, SHARE, RT &\xe2\x80\xa6"

Table 6

I removed the date and time column as it was not relevant in generating the sentiment of the movies. I processed the text data and used work tokenizer to extract the words from the text corpus. I then converted all the words into lower case along with removing the stop words and any numeric values. Then I implemented stemming using PorterStemmer to get all the words in their respective root format. I used Vaders' Sentiment Intensity Analyzer to get the percentage of positive and negative tweets. The functionality of Sentiment Intensity Analyzer is to break the sentence into positive, neutral and negative fragments. More about sentiment analysis will be described in the next section.

All the features were integrated into one to create a single data frame. The below mentioned table represents few tuples of the integrated table,

Out[31]:

	Movie	Percent Positive Tweets	Percent Negative Tweets	Rotten Tomatoes User Ratings	Percentage of Users who want to see (%)	Average Rating on a scale of 10	Director	Actor	Actress	Production house rating
0	Mary Queen of Scots	0.739130	0.260870	495.0	97.0	7.4	23884.0	69.0	NaN	6.81
1	Dumplin'	0.877384	0.122616	25.0	88.0	NaN	152.0	2274.0	39138.0	4.3
2	Vox Lux	0.802632	0.197368	237.0	95.0	7.4	225.0	NaN	44788.0	6.8
3	Ben Is Back	0.851485	0.148515	145.0	92.0	6.3	5.0	459.0	517380.0	6
4	Spider-Man: Into the Spider-Verse	0.635719	0.364281	711.0	94.0	NaN	4893.0	NaN	1032458.0	6.57
5	Mortal Engines	0.738739	0.261261	452.0	94.0	NaN	178.0	2752.0	4324.0	8.5
6	The Mule	0.650685	0.349315	212.0	98.0	NaN	73089.0	73092.0	NaN	7.03
7	Roma	0.836232	0.163768	654.0	98.0	9.1	120251.0	NaN	38636.0	8.6
8	Backtrace	0.500000	0.500000	NaN	NaN	NaN	3136.0	2801180.0	NaN	6.41
9	Capernaum	0.756757	0.243243	108.0	89.0	7.6	740833.0	NaN	NaN	6.65
10	Mary Poppins Returns	0.717276	0.282724	1169.0	97.0	NaN	331.0	NaN	9279.0	5.9
11	Aquaman	0.645855	0.354145	10549.0	97.0	NaN	211653.0	16450.0	73713.0	7.79

Table 7

The visualization mentioned below represents the number of NaN values in the integrated table,

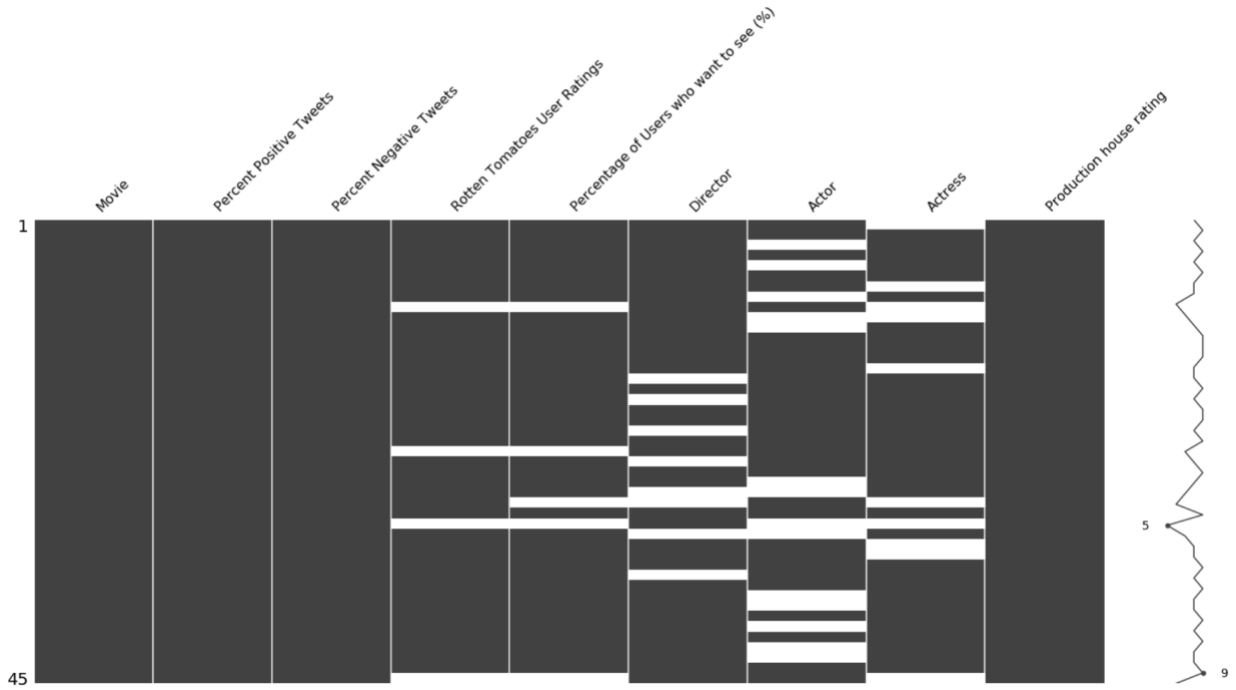

Fig. 1

The line chart on the right represents the missing values and the numbers 5, 9 represent that most of the missing values were between the rows 5 to 9. The visualization below represents the correlation between the missing values,

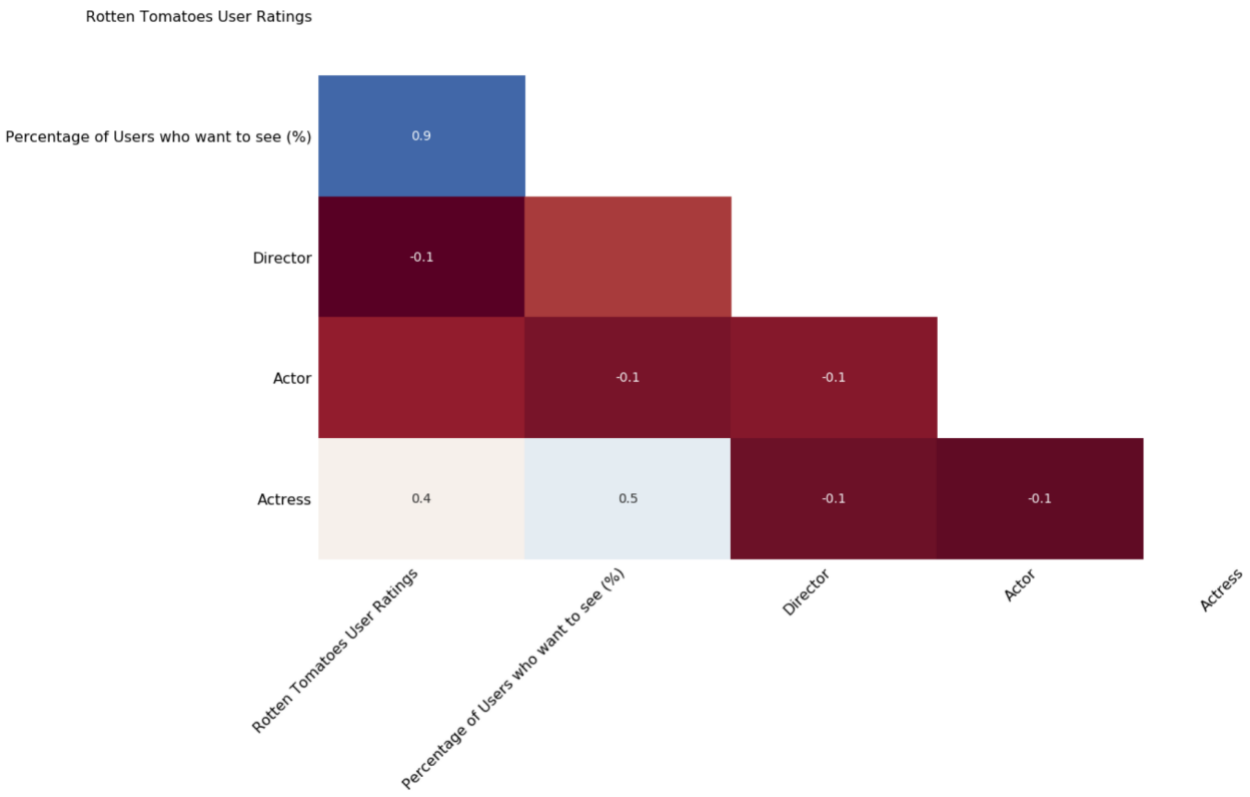

Fig. 2

As mentioned above the column "Average Rating on a scale of 10" was removed as it had too many NaN values and couldn't be replaced with the median of the respective column. Apart from that all the NaN values were replaced with the median values of their respective columns. The table below represents the columns without NaN values,

	Movie	Percent Positive Tweets	Percent Negative Tweets	Rotten Tomatoes User Ratings	Percentage of Users who want to see (%)	Director	Actor	Actress	Production house rating
0	Mary Queen of Scots	73.9130	26.0870	495.0	97.0	23884.0	69.0	73713.0	6.81
1	Dumplin'	87.7384	12.2616	25.0	88.0	152.0	2274.0	39138.0	4.3
2	Vox Lux	80.2632	19.7368	237.0	95.0	225.0	50342.0	44788.0	6.8
3	Ben Is Back	85.1485	14.8515	145.0	92.0	5.0	459.0	517380.0	6
4	Spider-Man: Into the Spider-Verse	63.5719	36.4281	711.0	94.0	4893.0	50342.0	1032458.0	6.57
5	Mortal Engines	73.8739	26.1261	452.0	94.0	178.0	2752.0	4324.0	8.5
6	The Mule	65.0685	34.9315	212.0	98.0	73089.0	73092.0	73713.0	7.03
7	Roma	83.6232	16.3768	654.0	98.0	120251.0	50342.0	38636.0	8.6
8	Backtrace	50.0000	50.0000	186.0	94.5	3136.0	2801180.0	73713.0	6.41
9	Capernaum	75.6757	24.3243	108.0	89.0	740833.0	50342.0	73713.0	6.65
10	Mary Poppins Returns	71.7276	28.2724	1169.0	97.0	331.0	50342.0	9279.0	5.9
11	Aquaman	64.5855	35.4145	10549.0	97.0	211653.0	16450.0	73713.0	7.79
12	Bumblebee	92.5620	7.4380	851.0	94.0	686.0	2021.0	1032458.0	6.2
13	Cold War	58.0110	41.9890	353.0	96.0	1.0	14.0	137.0	7.96
14	Holmes & Watson	63.0769	36.9231	258.0	93.0	4384.0	10008.0	73713.0	6.57
15	Welcome to Marwen	96.8750	3.1250	338.0	97.0	8580.0	5606341.0	1429836.0	7.33

Table 8

III. Prediction

My prediction is based on the statistics of the features engineered from all the data collected from various sources. The data created is unsupervised, so I implemented K-Means clustering using which I created two clusters of buckets 0 and 1. 0 implements the movie won't be successful at the box office and just opposite for 1. The below table demonstrates the clusters I created using K-Means clustering,

	Percent Positive Tweets	Percent Negative Tweets	Rotten Tomatoes User Ratings	Percentage of Users who want to see (%)	Director	Actor	Actress	Production house rating	clusters
0	-0.146363	0.146363	-0.195128	0.766739	-0.247467	-0.298245	-0.243076	0.088300	1
1	0.640983	-0.640983	-0.256079	-1.475349	-0.280473	-0.297878	-0.252555	-2.335992	0
2	0.215276	-0.215276	-0.228586	0.268497	-0.280372	-0.289883	-0.251006	0.078642	0
3	0.493490	-0.493490	-0.240517	-0.478866	-0.280678	-0.298180	-0.121435	-0.694041	0
4	-0.735280	0.735280	-0.167116	0.019376	-0.273879	-0.289883	0.019785	-0.143504	1
5	-0.148590	0.148590	-0.200704	0.019376	-0.280437	-0.297798	-0.262100	1.720593	1
6	-0.650050	0.650050	-0.231828	1.015859	-0.179033	-0.286099	-0.243076	0.300788	1
7	0.406625	-0.406625	-0.174508	1.015859	-0.113440	-0.289883	-0.252693	1.817178	1
8	-1.508190	1.508190	-0.235200	0.143936	-0.276323	0.167651	-0.243076	-0.298041	1
9	-0.045979	0.045979	-0.245315	-1.226228	0.749665	-0.289883	-0.243076	-0.066236	0
10	-0.270820	0.270820	-0.107721	0.766739	-0.280224	-0.289883	-0.260742	-0.790626	1
11	-0.677557	0.677557	1.108711	0.766739	0.013682	-0.295520	-0.243076	1.034837	1
12	0.915683	-0.915683	-0.148960	0.019376	-0.279731	-0.297920	0.019785	-0.500870	0
13	-1.051969	1.051969	-0.213543	0.517618	-0.280683	-0.298254	-0.263248	1.199032	1
14	-0.763470	0.763470	-0.225863	-0.229745	-0.274587	-0.296592	-0.243076	-0.143504	1
15	1.161305	-1.161305	-0.215488	0.766739	-0.268752	0.634220	0.128734	0.590544	0

Table 9

The figure below represents the clusters based on the feature weightage,

```
          Percent Positive Tweets  Percent Negative Tweets  \
clusters
0                        0.690594                -0.690594
1                       -0.863242                 0.863242

          Rotten Tomatoes User Ratings  \
clusters
0                             0.122939
1                            -0.153674

          Percentage of Users who want to see (%)  Director     Actor  \
clusters
0                                       -0.389182  0.084273  0.208395
1                                        0.486478 -0.105341 -0.260494

           Actress  Production house rating
clusters
0        -0.099737                -0.285060
1         0.124672                 0.356325
```

Fig. 3

Then I implemented Decision Tree classification algorithm to classify the movie into hit or flop category. Decision tree classifier is mostly used classification algorithm because of its advantages over other classification algorithms. When we say the advantages it's not about the accuracy of the trained decision tree model. It's all about the usage and understanding of the algorithm. The below figure represents the decision tree classification,

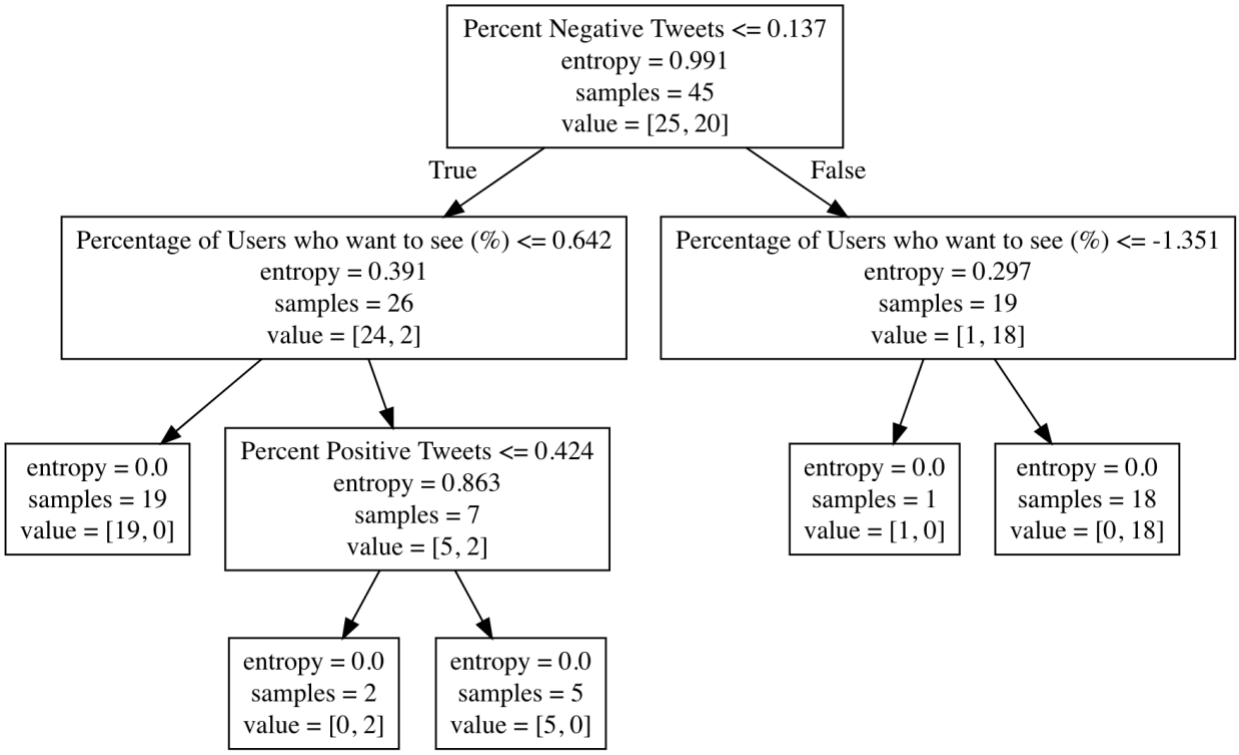

Fig. 4

I used gini index and entropy to understand the accuracy of the model, the figure below represents the accuracy of the model,

```
Results Using Gini Index:
Predicted values:
[1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
Confusion Matrix:  [[7 0]
 [4 3]]
Accuracy :  71.42857142857143
Report :               precision    recall  f1-score   support

        0.0       0.64      1.00      0.78         7
        1.0       1.00      0.43      0.60         7

avg / total       0.82      0.71      0.69        14

Results Using Entropy:
Predicted values:
[1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
Confusion Matrix:  [[7 0]
 [4 3]]
Accuracy :  71.42857142857143
Report :               precision    recall  f1-score   support

        0.0       0.64      1.00      0.78         7
        1.0       1.00      0.43      0.60         7

avg / total       0.82      0.71      0.69        14
```

Fig. 5

Conclusion

I did some preliminary study in using sentiment analysis to predict a movie's box office success. The results show that the box office success can be predicted by analyzing sentiment of the movies with simple metrics and pretty good accuracy.

I understand that there might be more than one factor which might affect the movie box office success, but I concentrate on sentiment analysis in this work. As sentiment analysis on twitter is a challenging topic, I feel that there is a long list of future work. However, this problem itself is an interesting and promising area. Some bottlenecks that I faced were:

- There are limitations of Twitter APIs (e.g. 1500 tweets/day)
- Lot of spam and noise included in randomly picked 100 tweets
- I did not take the total tweets number into account in our prediction metric

Future work which may be done to improve the accuracy and reliability of the prediction model can be, adding more features like genre, holiday season, sports leagues, month of the year and GDP of the country. Using other models and algorithm to get the accuracy of the prediction.

Reference

[1] Darin Im, Minh Thao, Dang Nguyen, Predicting Movie Success in the U.S. market, Dept.Elect.Eng, Stanford Univ., California, December, 2011

[2] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques, 3rd ed.MA:Elsevier, 2011, pp. 83-117

[3] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, 2nd ed.NewYork: Wiley, 1973

[4] Sagar V. Mehta, Rose Marie Philip, Aju Talappillil Scaria, Predicting Movie Rating based on Text Reviews, Dept.Elect.Eng, Stanford Univ., California, December, 2011

[5] Suhaas Prasad, Using Social Networks to improve Movie Ratings predictions, Dept.Elect.Eng, Stanford Univ., California, 2010

[6] The International Movie Database (IMDb). https://www.imdb.com/

[7] Rotten Tomatoes: Movies | TV Shows | Movie Trailers | Reviews - Rotten Tomatoes. *Rotten Tomatoes: Movies | TV Shows | Movie Trailers | Reviews - Rotten Tomatoes.* Retrieved from http://www.rottentomatoes.com/