UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Risko Ruus

Wisdom of the Crowd Vs. Reviews of the Experts: A Case Study Regarding Predicting Movie Box-Office Results

Master's Thesis (30 ECTS)

Supervisor Rajesh Sharma, PhD

Tartu 2018

Wisdom of the Crowd Vs. Reviews of the Experts: A Case Study Regarding Predicting Movie Box-Office Results

Abstract:

Predicting movie sales figures has been a topic of interest for research for decades since every year there are dozens of movies which surprise investors either in a good or bad way depending on how well the film performs at the box-office compared to the initial expectations. There have been past studies reporting mixed results on using movie critics reviews as one of the sources of information for predicting the movie box-office outcomes. Similarly using social media as a predictor of movie success has been a popular research topic. In this thesis, we perform a case study to evaluate out of two – the (wisdom of the) crowd or the movie critics reviews, which one can predict the outcome of the movies more accurately. We analyze the Hollywood and Bollywood movies from the last three years, which belong to two different geo as well as cultural locations. We used Twitter for collecting the wisdom of the crowd and used movie critics review scores from movie review aggregator sites Metacritic and SahiNahi for Hollywood and Bollywood movies respectively. To perform our evaluation, we extracted various features and used them to build prediction models using different machine learning algorithms. After measuring the performance of prediction models using features from both Twitter and movie critic reviews, we did not find conclusive evidence to declare a clear-cut winner.

Keywords:
Box-office forecasting, machine learning, Twitter

CERCS: P170 Computer science, numerical analysis, systems, control

Kollektiivse Tarkuse Võrdlemine Filmikriitikute Arvustustega: Uurimustöö Filmide Kassatulu Ennustamise Kohta

Lühikokkuvõte:
Teadlased on aastakümneid tegelenud filmide kassatulu ennustamisega, sest iga aasta linastub suur hulk teoseid, mille tulemused üllatavad nende rahastajaid kas heal või halval viisil sõltuvalt esialgsetest prognoosidest. Eelnevad uurimustööd on avaldanud vastakaid tulemusi filmikriitikute arvustuste kasutamise kohta filmide kassatulu ennustamiseks. Niisamuti on kaasatud sotsiaalmeedia ühe võimaliku andmeallikana filmide müügiedu prognoosimiseks. Käesolevas töös uurime, milline neist kahest erinäolisest allikast on kasulikum ennustamaks parema täpsusega filmide kasumlikkust. Uuritavateks andmeteks oleme kogunud viimase kolme aasta jooksul linastunud Hollywoodi ja Bollywoodi filmid, mis on erineva geograafilise asukoha ning kultuurilise taustaga. Kollektiivse tarkuse näitena uurime sotsiaalvõrgustiku Twitteri andmeid ning võrdleme neid filmikriitikute arvustustega Hollywoodi ning Bollywoodi filmiportaalidest Metacritic ja SahiNahi. Kaasame mitmeid erinevaid tunnuseid ning rakendame erinevaid masinõppe algoritme ennustusmudelite ehitamiseks. Meie vaatluste tulemused näitavad, et võrreldes filmikriitikute eksperthinnangutega pole kollektiivsete teadmiste abil võimalik filmide kassatulu paremini ennustada ega vastupidi.

Võtmesõnad:
Filmide kassatulu prognoosimine, masinõpe, Twitter

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Acknowledgements

I would like to thank Rajesh Sharma from University of Tartu's Institute of Computer Science for being an excellent mentor to me on this thesis. Rajesh was always very enthusiastic and helped me by providing ideas to experiment with and motivational support from the very beginning of our collaboration. I also would like to thank my dear friend, Ormes Liivak, for proofreading and pointing out some of my weird grammar constructions. Finally, I am grateful to my girlfriend, Johanna Nisu, for all her support during this whole process and helping me to identify hashtags for upcoming movies.

Contents

1 Introduction **7**
- 1.1 Past Research . . . 7
- 1.2 Goal . . . 8
- 1.3 Approach . . . 8
- 1.4 Contributions . . . 9
- 1.5 Outline . . . 9

2 Related Work **10**
- 2.1 From Social Web Platforms . . . 10
- 2.2 From Expert Movie Reviews . . . 13
- 2.3 Summary . . . 14

3 Dataset collection **16**
- 3.1 Movie Selection . . . 16
- 3.2 From Twitter . . . 16
  - 3.2.1 Why it is Difficult to Gather a Large Set of Movie Tweets . . . 17
  - 3.2.2 Gathering Realtime Tweets . . . 17
  - 3.2.3 Gathering Historical Tweets . . . 19
  - 3.2.4 Extracting Tweets From Historical Data . . . 21
- 3.3 From Expert Review Aggregator Sites . . . 21
- 3.4 From Movie Revenue Information Sites . . . 22
- 3.5 Data Cleaning . . . 22

4 Approach **24**
- 4.1 Feature Engineering . . . 24
  - 4.1.1 From Movie Metadata . . . 24
  - 4.1.2 From Movie Critics Review Data . . . 25
  - 4.1.3 From Twitter Data . . . 25
- 4.2 Sentiment Analysis . . . 25
  - 4.2.1 Tweets . . . 25
  - 4.2.2 Movie Reviews . . . 26
- 4.3 Dependent and Target Variables . . . 26
  - 4.3.1 Movie Metadata . . . 27
  - 4.3.2 Critic Reviews . . . 27
  - 4.3.3 Twitter . . . 28
  - 4.3.4 Target Variable . . . 28
- 4.4 Exploratory Data Analysis . . . 28
  - 4.4.1 Hollywood . . . 29
  - 4.4.2 Bollywood . . . 32

4.5 Predictive Modelling . . . 36
4.5.1 Ensemble Learning Methods . . . 37
4.5.2 Bias-Variance Tradeoff . . . 38
4.6 Building a Machine Learning Pipeline . . . 38
4.6.1 Reading the Data . . . 39
4.6.2 Encoding Categorical and Ordinal Features . . . 39
4.6.3 Tuning Model Hyperparameters . . . 40
4.6.4 Saving the Model and Loading it Back from Disk . . . 40
4.6.5 Reporting and Comparing Results . . . 40
4.6.6 Visualizing Predictions and Feature Importances . . . 41

5 Empirical Results **42**
5.1 Wisdom of the Crowd Vs. Reviews of the Experts . . . 42
5.2 Hollywood Vs. Bollywood . . . 46

6 Conclusion **49**
6.1 Summary . . . 49
6.2 Limitations . . . 49
6.3 Future work . . . 49
6.3.1 Using Additional Realtime Tweets . . . 49
6.3.2 Using Aspect Level Sentiment Analysis . . . 50
6.3.3 Additional Sentiment Analysis Approaches . . . 50
6.3.4 Additional Variables . . . 50
6.3.5 Time Effects . . . 50

References **55**

A Appendices **56**
A.1 List of Notable Libraries Used . . . 56
A.2 Hyperparameter Values . . . 57
A.2.1 Random Forest Default and Custom Hyperparameter Values . . . 57
A.2.2 XGBoost Default and Custom Hyperparameter Values . . . 58
A.2.3 Hyperparameters for achieving the Best Score with Random Forest Algorithm on Hollywood Movies . . . 58
A.2.4 Hyperparameters for Achieving the Best Score with XGBoost Algorithm on Hollywood Movies . . . 59
A.2.5 Hyperparameters for Achieving the Best Score with Random Forest Algorithm on Bollywood Movies . . . 59
A.2.6 Hyperparameters for Achieving the Best Score with XGBoost Algorithm on Bollywood Movies . . . 60
II. Licence . . . 63

1 Introduction

Hundreds of movies are released every year in the world. However, not every movie turns out to be a commercial success. For example, only three or four major movies out of every ten major Hollywood movies produced are profitable [Vog14]. Forecasting the box-office results has been a big concern for the movie industry as early box-office predictions help to make vital decisions regarding marketing budget allocation and distribution. Equally important is determining the best screen allocation for a movie in each country since empty seats mean bad business for movie studios and cinemas alike. However, past studies have shown it is difficult to predict the tastes of moviegoers [SE96, BCR03, Liu06] and subsequently forecasting the box-office results has been a big concern for the movie industry.

1.1 Past Research

Litman was the first to study multivariate regression models [Lit83, LK89] for predicting the box-office outcome of movies. Predictor variables considered in such research include the number of theaters the movie is scheduled to be released in, parental rating and the budget of the film. Many researchers consider predicting commercial movie success a classification problem. For example in [SD06, AJM$^+$13, QGCA17] movies are classified into different categories usually ranging from a flop to a blockbuster. These segments are created by using the movie production budget as an estimated figure for calculating how much a movie should make to earn its production costs back. The problem with this approach is that while recently many studios have started to reveal their film production budgets, the money spent on marketing is not disclosed and can influence the actual profitability of the movie significantly. Also as mentioned in [SS00], star actors are often paid a percentage of the movie profits and their salaries might not be included in the movie production budget figures making the movie budget deceptively low. For these reasons we have followed the example of studies such as [JDGS10, ADAH$^+$10, Hon14] and consider predicting commercial movie success as a regression problem and predict the amount of money a movie is expected to earn after its opening weekend.

Most of these studies have looked at mainly at Hollywood movies [AH10, MYK13, BCR03], apart from a few other regional studies such as from Korean [KHK15] , Chinese [LDC$^+$16] and [NS15] Bollywood. In this work, we studied both Hollywood and Bollywood movies to understand if the regional or cultural aspects play any role for prediction.

Most of the previous studies involving predicting movie success ahead of its release have either worked on social media platform such as Twitter [AH10], wikimedia [MYK13], Facebook coefficient of determination (r^2), google search queries [PC13] or have only analysed movie expert's reviews [BCR03, ES97, Kin07, NS15].

Historically since the rise of the movie industry, movie critics reviews have been

published in daily newspapers, magazines and more recently in online news portals. In comparison, presently, movie enthusiasts, often use online social platforms such as Twitter to express their opinions about the movies. We particularly analyzed "wisdom of the crowd" on Twitter for movies, which refers to the collective opinion of a community or a group. Although each tweet in Twitter might sound like a weak chirp and carries only a little amount of information, but a steady stream of expressed opinions makes up a strong signal, which as shown by previous studies on predicting the stock market [BMZ11] or upcoming election results [TSSW10] could indicate the intent of the general public.

1.2 Goal

Social media content can be thought of as a very large collection of collective wisdom. When asking the right questions from such data, it is possible to make predictions about future outcomes and the question we will be asking is about predicting the box-office outcome of upcoming movie releases [AH10]. In comparison, movie critics reviews refer to the views expressed by a smaller group of domain experts. We are interested in finding out if it is the experts or if it's the wisdom of the crowd, which emits a stronger signal which enables to predict the box-office outcome of the movies better. Understanding this would help the stakeholders, including distributors and movie theatre operators to make improved financial decisions when promoting the film at the *critical period*[1] of its release.

1.3 Approach

This thesis is an empirical study, which involves collecting all the necessary data for building prediction models for the Hollywood and Bollywood movies released between April 2015 and April 2018. For comparing, who is the better predictor, the wisdom-of-the-crowd or the movie critics, we examine models, which have been created using feature available before the release of the movie. All models we build use general movie information e.g. budget and opening theatre count as base variables. We call this set of features the movie *metadata*. In addition to the meta features, the Twitter-based models use the hourly tweet rate from two weeks before the film's release and the sentiment score of the movie tweets as additional dependent variables. Similarly, for building the regression model based on movie expert reviews, we combine the metadata features with scores and review counts from film review aggregator sites. Last, we evaluate prediction results using Random Forest [Bre01] and XGBoost [CG16] machine learning algorithms.

[1] We use the same definition for the critical period as [AH10]. It is defined to be between a week before the movie is released until two weeks from its release date. This is usually the time when most of the promotional budget is being spent on various forms of advertising.

We do not find conclusive evidence to recommend features from Twitter over movie critic reviews and vice versa.

1.4 Contributions

In this thesis, we make the following contributions:

1. **Wisdom of the crowd vs. experts:** Our empirical study shows that people's collective wisdom (gathered from Twitter) can help to predict movie opening weekend box office results, but does not always achieve a higher accuracy than models using features from movie expert's aggregated review scores.

2. **Large scale study:** To the best of our knowledge this research is made on largest amount of Hollywood and Bollywood movies with the related tweets to date and shows that a strong movie box-office predictor variable can be extracted from only 1% of random tweet sample.

3. **Hollywood & Bollywood:** The work offers a unique cross-cultural comparison of box-office predictions for Hollywood and Bollywood - the two of the world's biggest movie markets.

1.5 Outline

Rest of the thesis is organized as follows. In chapter 2 we will give a brief overview of previous related research that has been regarding predicting movie box-office results using social media as a source of information. Chapter 3 will focus on describing the movie tweet data collection process as well as gathering all the data regarding general movie information and aggregated critics scores needed for predicting the final results. Next in chapter 4 we will look at how the collected data was processed and which machine learning algorithms and tools we used for conducting our work. An overview of our empirical results is in chapter 5, where we compare whether it is the wisdom of the crowd or critics who can predict the box-office outcomes better. We also look at more closely, which features are most important for models to predict the future box-office results. Finally chapter 6 is for describing our overall contribution and proposes some topics for future research.

2 Related Work

Even before the rise of movie community websites and social networks, predicting movie sales has been a popular topic. Researchers have been trying for decades to capture the "magic" elements that really drive people to go to the cinema and discover the ingredients needed for making a blockbuster movie. Various different features have been explored to determine how much predictive power sources like movie critics reviews might possess. In this chapter we provide an overview of the past research done on predicting the success of movies. We look at papers which in similar to our work are using either social media as a source for predicting box-office revenue and or critics movie reviews.

2.1 From Social Web Platforms

Before the rise of the internet most of the dependent variables used for predicting movie box-office outcome, have been based on movie metadata e.g., its genre, parental rating and actors which as reported by [CK05] can explain approximately 60% of the variances. With the rise of dedicated communities for movie lovers, blogs and various web services, researchers have been looking for additional sources of information, which could help predict the movie economical success even better.

[ADAH+10] were able to predict box-office revenue from 600,000 blog entries obtained from *Spinn3r*[2], an API for social media information, with a relative error of 26.21%. Authors of [WSC12] have compared the predictive power of tweet sentiment analysis and online movie review sites such as Internet Movie Database (IMDb) and Rotten Tomatoes[3] and find that Twitter users are more positive in their reviews compared to the dedicated review site's ratings. This could indicate that people are more inclined to tweet about a movie when they feel positive about it and in case of a negative experience they will not bother to tweet about it.

Some studies like [AJM+13] have compared the prediction sources of different web resources and social networks, namely IMDb, Twitter, and Youtube. They find that the popularity of the leading actress estimated by the followers count the actress has on Twitter is a strong predictor, but the sentiment score from movie trailer comments does not help to determine the financial success of a movie.

[2] https://www.spinn3r.com

[3] https://rottentomatoes.com

Paper	Problem Investigated	Data source (Period)	# of Movies
[AH10]	Prediction of movie box-office results using tweet rate and sentiment analysis	2.89 million tweets (November 2009 to February 2010)	24
[RLW13]	Analyzing the effect of tweeter's follower count and tweet valence on movie sales	4.2 million tweets (June 2009 to February 2010)	63
[WSC12]	Whether Twitter user's movie reviews can predict movie's box-office success	1.77 million tweets (February 2012 to March 2012)	34
[Jai13]	Prediction of movie box-office results using only tweet sentiment	Same as [AH10] + (8 Movies from 2012) 200 tweets each)	32
[AJM$^+$13]	Prediction of movie box-office performance using data from multiple social media sources	Twitter, Youtube, IMDb (May 2013 to July 2013)	35
[GMD15]	Predicting opening weekend box-office results of Bollywood movies using the number of tweets, tweet sentiment, and actor/actress star rating	10269 tweets (June 2014 to December 2014)	14

Table 1. Summary of related papers which use Twitter as a source for Hollywood or Bollywood box-office predictions

In a novel study, [AH10] have shown that data from Twitter, in particular, the average hourly tweet rate and sentiment analysis of the tweets can be used to predict movie box-office outcomes using a simple linear regression model (r^2(t) = 0.98 at the release night of the movie). They find that for predicting the box-office outcome, sentiments from tweets after the movie is released have a stronger effect. To evaluate their tweet-based model they compare it to the Hollywood Stock Exchange (HSX) index[4], a website where players can trade virtual stocks of latest movies and find that hourly tweet-rate for movies is a significantly better predictor than historical HSX prices. However [MYK13] does point out in Fig. 5 of their work that the paper of [AH10] achieves such a high score because most of the 24 movies considered are commercial successes, which the model is capable of predicting better than movies with low or moderate success.

In their work [MYK13] show that movie box-office performance can be estimated from the activity levels of Wikipedia articles about the movie before its release. They consider features like the number of views, the number of edits and number of different users the page has had before the film is released. In comparison, their work includes

[4]https://www.hsx.com

312 movies, which made their debut in 2010, which is a considerably more substantial amount than 24 movies investigated in the work of [AH10]. What they show is that the model based on Wikipedia activity data can make predictions with quite a good coefficient of determination, $r^2 > 0.925$ even one month before a movie is released. Our work includes slightly more Hollywood movies and instead of using Wikipedia activity levels as dependent variables for predicting movie revenue, we use information from movie tweets and film critic review scores.

Similarly to Wikipedia activity levels, Facebook official movie fan page activity is used as a prediction feature in [TYL14]. When using only the number of screens on the opening week and the Facebook official movie page activity features before the release, the study reports r^2 increase from 0.68 to 0.88.

Predictions from social media can be made not only about movie's financial success as [OBTdR12] were able to rate movies very close to their IMDb star rating using tweets from Twitter and comments from YouTube. For predicting Academy Award nominations and movie box-office results, [KNS+08] show successful results using movie comments from IMDb users as a possible source of information.

A whitepaper from Google [PC13] on 99 movies released in 2012 shows that Google search volume explains 70% of the variance in the opening weekend box-office performance of the film. However when they looked at the movie trailer title search volume four weeks before the release, together with seasonality and movie franchise status information, the same explained variance reached a high 94%.

Research involving predicting movie profitability is not only limited to Hollywood releases. Korean researchers in [KHK15] have studied their local market and demonstrated using 212 domestic movies released between September 2011 and December 2013, that prediction success of movie revenue increases using metadata and features from multiple social media networks. Similarly predicting movie box-office success on the Chinese domestic market has been researched by [LDC+16] using 57 movies with 5 million tweets collected from the Sina Weibo microblog[5]. They were able to achieve an adjusted r^2 value 0.94 for their model, which uses a custom purchase intention feature. The score is higher than using the model proposed by [AH10], which achieved r^2 of 0.89 on the test dataset. The only previous study on predicting the box-office results of Bollywood movies that uses features from social media we were able to find, [GMD15], unfortunately, looks at only 14 movies and reports prediction results from this small sample. They report Mean Squared Error (MSE) for four movies separately instead of measuring the predictive capabilities of the model as a whole.

[5]https://www.weibo.com

2.2 From Expert Movie Reviews

Predicting movie box-office outcome using critic reviews as a source has been studied already in the early 1990's as seen from Table 2.

The authors of [BCR03] look at expert reviews and find confirmation to the common belief that positive reviews help box-office performance and bad reviews have a negative impact on the sales. However their findings show that the effect of negative reviews wears off after some time, but the positive impact does not. This observation could indicate the role of critics being more influencers rather than predictors for a movie's performance.

In his study on movies released in 2003 in the U.S. [Kin07] finds that Metacritic scores do not have a strong relationship with the gross earnings of the films, however, he does report that movies released in over 1000 screens have a positive correlation of 0.33. This may suggest that regarding more popular films the critics and the audience have a more shared understanding and that critics also act as influencers and bring people to see more highly rated movies.

Some research has also done on the textual data of critic's movie reviews like [JDGS10] who use movie earnings text analysis on pre-release reviews and metadata features available before movie's release for predicting the opening weekend box-office results. What they find is that the textual data can improve predictions when combined with seven movie metadata features such as the number of screens, genre, budget and parental rating.

Rotten Tomatoes ratings are used by [BKJ09] with the Ordinary Least Squares (OLS) method to find the importance of many variables such as the production budget, the previous gross revenue (if the movie had a prequel) and the release period besides the movie critic ratings. They find the critic scores to have a positive and significant effect on the movie box-office revenue although it is much smaller when compared to independent variables like the number of opening screens and the budget of the movie.

The aggregate movie critic score impact on movie box-office revenue is studied by [ES97], and they find it to have a small positive effect. However, they do report that the impact is more influential on the total gross revenue of the movie and weaker for predicting the opening weekend earnings. However, authors of [BBK07] find in similar to [BCR03] and in opposite to [ES97] in their study focusing on individual movie critics, that critics act as more influencers rather than predictors.

For Bollywood movies, there have been fewer studies on the impact of movie critics on movie box-office revenues than for Hollywood. Authors of [NS15] look at both the online user-generated and the expert reviews from daily newspapers and find that volume and valence from both sources have had a positive effect on the financial success of movies. However, they do note that the user-generated content valence score is more effective when it is not blatantly positive and contains a few negative comments as well. This finding could mean that people may find a bit more critical reviews to be more credible.

Paper	Problem Investigated	Data source (Time period)	# of Films
[JDGS10]	Whether text features from pre-release reviews can substitute for and improve over a strong metadata-based first-weekend movie revenue prediction	7082 reviews from various newspapers (Movies from 2005 to 2009)	1718
[BCR03]	How critical are critical reviews? The box-office effects of film critics, star power, and budgets	Baseline in California [6] website and *Variety* magazine (Movies from 1991 to 1993)	200
[Kin07]	Does film criticism affect box-office earnings?	Metacritic (Movies from 2003)	273
[BKJ09]	Which independent variables are significant in predicting the total domestic box-office.	Rotten Tomatoes (Movies from 1997 to 2001)	466
[ES97]	Whether critics act as predictors or influencers in terms of box-office revenue	2104 reviews from *Variety* magazine (Movies from 1990 to 1993)	172
[BBK07]	Impact of individual critics influence on the market performance of movies	46 distinct reviewers from *Variety* magazine (Movies from 1997 to 201)	466
[NS15]	The impact of professional and word-of-mouth movie reviews on Bollywood movie success	Aggregated expert reviews from daily newspapers on movie ticket website[7]	48

Table 2. Summary of papers which evaluate the relationship between movie box-office results and movie critic reviews

2.3 Summary

We have seen from the results of related literature that predicting movie box-office results is a difficult task researchers have been studying for decades. Movie success can be predicted from different sources with limited and sometimes also with quite promising results. For performing the regression analysis using machine learning algorithms the sample size of the movies is a crucial factor for building a model capable of learning from the training data. Most studies using Twitter data in predictions so far have been

[6] http://www.baseline.hollywood.com
[7] http://bookmyshow.com

limited to a small number of movies, report the coefficient of determination (r^2) as the metric and use linear regression models. Movie critic reviews alone do not seem have much predictive power, but in combination with other features they may explain more variance and contribute towards creating a strong predictor. Studies on both movie tweets and professional movie critic reviews have shown the features from both sources to have a positive effect on the financial success of the movie. We are interested in exploring whether movie metadata features together with features from Twitter or movie critic reviews can to make up a better prediction model. We attempt to build our models using more movies and evaluate different machine learning algorithms than previous studies involving Twitter data. In the following chapter, we will take a closer look at the dataset we collected for our research.

3 Dataset collection

3.1 Movie Selection

This thesis considers Hollywood and Bollywood movies released between April 10th, 2015 and April 6th, 2018. In most cases, films in Hollywood and Bollywood premiere on Fridays(nearly 85% according to [DVW99]). However, sometimes movies are also released on Wednesdays. For the sake of consistency, we focused only on the films that are released on Fridays. For Hollywood movies, we only included movies, which had a wide release from its first release day. A film in Hollywood is considered to be in a wide release when it is running in 600 or more cinemas [Box]. If a movie had a limited release initially, but later went into a wide release then we did not include the film in our work. For Bollywood movies, since we did not find any definition for a wide release, thus, we did not apply any such selection criteria for them.

We needed to start gathering tweets to find out how frequently a movie is mentioned on Twitter and what sentiment the tweets carry. We used popular movie information sites Box Office Mojo[8] and IMDb[9] to find the upcoming Hollywood movies. In addition to IMDb for finding the Bollywood movie release dates we also used Wikipedia articles about Bollywood release dates for years 2017[10] and 2018[11].

3.2 From Twitter

Twitter is an online social platform where people tweet from around the globe about almost every imaginable topic 24/7. It is essential that we identify only tweets about upcoming movies we are interested in collecting. Earlier studies on Twitter data [AH10, WSC12], have looked for the movie title in the tweet text mainly because hashtags were not so popular back then in Twitter as mentioned by [BGM12]. This approach has a drawback when a movie title is a simple common word or a phrase like in the case of 2015 Hollywood movie *Sisters* which can come up in many tweets not referring to the movie. Because of this limitation, most studies exclude films with such titles from their work. More recent papers like [SP17, GMD15] however, use the unique hashtags people use in their tweets to match a tweet to a movie. This approach has the benefit of still being able to find tweets about a movie with a non-unique title like *Sisters* when people have marked them with a hashtag such as #SistersMovie[12]. When inspecting the official Twitter pages of such films, we found that the movie studios often pick the main hashtag for the movie and use it consistently in their marketing campaigns. When such tweets

[8] http://www.boxofficemojo.com

[9] http://www.imdb.com

[10] https://en.wikipedia.org/wiki/List_of_Bollywood_films_of_2017

[11] https://en.wikipedia.org/wiki/List_of_Bollywood_films_of_2018

[12] https://twitter.com/sistersmovie

reach their audience, then they tend to use the same hashtag in their own tweets. In our work, we also decided to identify tweets by the hashtags that were used most often to refer to the movie the tweet was about.

3.2.1 Why it is Difficult to Gather a Large Set of Movie Tweets

As can be seen from Table 1 the amount of films included in research papers about movie box-office returns using Twitter data is limited. There are a few reasons, which could help to explain why more movies are not included:

1. **Twitter's privacy policy does not permit hosting public datasets.** In 2010 Twitter updated their privacy policy and does not allow publicly hosting datasets. Since then they have asked researchers to stop hosting their datasets for the public. Because Twitter allows people to delete their tweets it is understandable that such deleted tweets should not be available in the public domain inside research datasets.

2. **Twitter Search API[13] is limited to searching back in history for only about a week.** Thus, gathering data about movies released in past years is simply not suitable due to this restriction.

3. **Paid services for gathering historical tweet data are too expensive.** There are paid services like Twitter's premium or enterprise API offering, however, the pricing[14] of these services is too expensive to consider using it to collect millions of tweets.

4. **Collecting real-time tweets takes much time and effort.** Twitter offers gathering of real-time tweets using the Streaming API[15]. Most researchers use this method for finding the tweets for movies by filtering the stream returned by the movie title or relevant hashtags. Still, to use films from past several years, they must first gather the tweets and wait until they can proceed with their work. We did not find previous studies, where tweets about movies had been collected for more than one year.

3.2.2 Gathering Realtime Tweets

Similar to the authors of most related papers, we began collecting real-time tweets about Hollywood and Bollywood movies using the Twitter Streaming API. For obtaining the tweets, we use a Python library called Tweepy[16] that itself uses Twitter's Streaming API

[13] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets
[14] https://developer.twitter.com/en/pricing/search-fullarchive
[15] `https://developer.twitter.com/en/docs/tweets/filter-realtime/overview`
[16] `http://www.tweepy.org`

under the hood. Twitter Streaming API returns a live feed of a maximum of 1% of the total real-time tweets when no search filter is applied. However, in our experiment, we did not hit this limit since we restricted the search only with particular movie hashtags. If a tweet contained hashtags from more than one movie, then we discarded such tweets since such tweets on closer observation tended to be promotional giveaways or other types of advertisements. Also, for tweets with multiple hashtags, we would not be able to clearly identify from the tweet, which movie was it mainly about. Because tweets are limited to only 140 characters[17] then most often people would not express their opinion about multiple films in a single tweet.

Every Thursday we monitor the IMDb and Box Office Mojo release schedule pages to identify wide release Hollywood movies about to be released in two weeks from now. Similarly, for Bollywood movies, we check the IMDb release schedule and Wikipedia pages[10] and[11] for finding the upcoming films. Next, we start looking for hashtags that people are using to tweet about the upcoming movies. To validate our usage of hashtags, we wrote a script that uses Twitter Search API to look for the movie title or some keywords from past week's tweets. The script returns an ordered list of popular hashtags and how many times in total a hashtag was found. We also visit the official Twitter pages of the movies and look, which hashtags are being used by the movie studio and the movie's followers most often in their tweets about the movie. Generally, the top hashtag found by our script matches the one most used on the Twitter page of the movie. However, sometimes we identify more hashtags that are being used quite often so we included those as well in our search. For example, the movie *Father Figures* had only one popular hashtag (#FatherFigures), but some had more, like in the case of the movie *Disaster Artist* (#DisasterArtist, #TheDisasterArtist). Picking the most popular hashtags is a laborious manual process, but it is essential for capturing the right tweets for the upcoming movies.

We start collecting tweets two weeks before the movie's release date and stop collecting after it had been in the cinemas for two weeks. In total for Hollywood and Bollywood movies released between November 2017 and May 2018, we collected at least four weeks worth of tweets.

During the data collection, we found that some movie release dates were not always set in stone even a couple of weeks before the release. For example, the release of Bollywood movie *Padmavaat* was postponed due to political reasons and this, in turn, caused other movies such as *Padman*, *Firangi* and, *Tera Intezaar* to change their release dates. Another example from Hollywood is the film *Gotti*, which had a confirmed release date, but its distributor *Lionsgate* sold the film back to its producers and studio just ten days before the release date and the movie release got postponed. Such changes in movie release schedules make collecting realtime data before a movie's release more difficult.

[17]During our data collection process, Twitter updated the tweet limit from 140 to 280 characters. https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

To counter this problem we checked whether movie release dates were still the same before we would finish gathering the tweets. If the release had shifted a week, then we would stop collecting the tweets a week later and corrected the movie's release date. When the movie release was indefinitely postponed, then we stopped fetching tweets for the film and started collecting again when a new date was confirmed.

Fig. 1 shows the process of evaluating whether a tweet is about a movie we were interested in gathering tweets for. If a tweet did not contain any hashtags or did not contain hashtags about films, then we skipped processing it. Further, if the tweet included any movie hashtags we were interested in, then the number of movies the tweet was about was calculated. If the tweet had hashtags for multiple distinct films, then we discarded the tweet since we could not determine, which movie the tweet was mostly about. Finally, the tweet referring to a single film was stored and assigned to the movie.

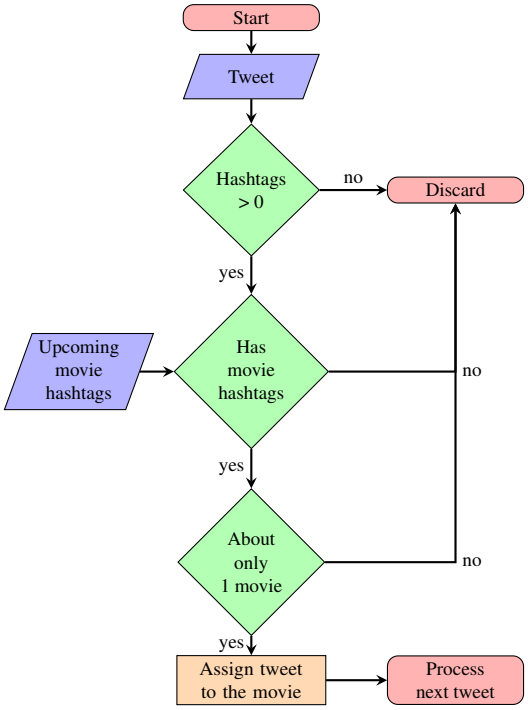

Figure 1. Flowchart of the process for extracting relevant movie tweets

3.2.3 Gathering Historical Tweets

In January 2018 we realized that we would have only around 60 Hollywood and 40 Bollywood movies available for further research in April when we planned to begin building a prediction model from the collected data. However, we realized this data would not have had enough movies to create a reliable prediction model. We could have followed the example of papers like [AH10, BKJ09], which use the r^2 value to measure

how much variance is explained in the target variable by the features used for predicting. Instead, we wanted to weigh the predictive power of Twitter against critics reviews using metrics like Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), MSE and Root Mean Squared Error (RMSE) in addition to the r^2. Unfortunately due to reasons described in chapter 3.2.1 we were not able to quickly come up with a solution to the problem of not having enough historical data about movie tweets. However, after an extensive search for Twitter datasets still hosted in public, we found that the Archive Team[??] had been posting monthly dumps of the *Spritzer* version of the Twitter Streaming API on archive.org. A Spritzer version of the grab is collected by not defining any search keywords to filter by, meaning that a random sample of 1% of all the tweets being posted to Twitter would be fetched.

To validate whether we could use the 1% random tweet sample from the hosted dataset to estimate, how many tweets in total would be in 100% of data, we experimented using six movies released in December 2017. Because we started also gathering realtime tweets in October 2017 and had the 1% sample available from a separate data source, we were able to check our estimated number of tweet count against the actual number of tweets for the movies. Table 3 lists the predicted and actual tweet numbers and predicted difference percentage. As expected the estimated difference between actual tweet count and count from 1% sample size is minimal in case of blockbuster movies such as *Star Wars: The Last Jedi*, a difference of mere 0.38%. For less popular films such as *All the Money in The World*, the estimated tweet count is less accurate, but still quite reasonable, a difference of 18.18%. Authors of [WCZ15] have studied the Spritzer version of the Twitter stream on a number of datasets to see if there is any sampling bias in the stream because Twitter has not revealed, how the data is sampled and does not guarantee a constant sampling rate. Overall they find that the stream is suitable for conducting research experiments and the sampling ratio measured on their datasets was on an average of 0.95%.

Movie	**Estimated hourly tweet rate**	**Actual hourly tweet rate**	**Difference**
Ferdinand	59	57	+3.50%
Star Wars: The Last Jedi	1856	1863	-0.38%
Pitch Perfect 3	117	115	+1.74%
Downsizing	49	50	-2.00%
All the Money in The World	13	11	+18.18%
Father Figures	9	11	-18.18%

Table 3. Estimated tweet rate from 1% of Twitter sample data vs. actual hourly tweet rate for 6 Hollywood movies from December 2017

After our experiment confirmed that the 1% of tweets hosted by the Archive Team

fits for our purpose and enables us to include more movies from recent years to our work, we proceeded to download the monthly tweet datasets from March 2015 to December 2017. The total size of the compressed tweet set was 1.41TB containing 4.3 billion tweets. While downloading the monthly data grabs, we noticed that for some months the file sizes were smaller. It looks like there were periods when fetching the tweets was broken for the Archive Team, and as a result, for some days no data had been gathered. To overcome this problem we replaced the missing periods with the average tweet rate for the movie. The following section describes in more detail how the tweets for movies we were interested in were extracted.

3.2.4 Extracting Tweets From Historical Data

After gathering and validating the historical tweets, we had to find the Hollywood and Bollywood movies released during these years and look up the right hashtags for each film from the web. For finding the relevant Hollywood and Bollywood movie release dates we again used the Box Office Mojo and Box Office India websites and collected the movies which release date fitted into our historical tweet set timeline. Finding hashtags for the films was again a manual process of looking at the official twitter pages of the movie and searching for the most popular hashtags people had been using when tweeting about the film.

The general process of filtering tweets for relevant movies was similar to filtering tweets from Twitter's realtime feed described in section 3.2.2 and shown on Fig. 1, was also applied for historical tweets. The only difference was that instead of listening to a real-time stream of tweets the historical tweets were read line-by-line from a total of 1.41TB compressed daily files. As a result, a total of 281322 tweets mentioning hashtags for Hollywood and Bollywood movies were extracted from the historical tweet 1% sample set.

3.3 From Expert Review Aggregator Sites

Critics' movie reviews are usually published a few days before or on the public release date of the movie, which leaves enough time to influence the movie-goers decision whether to go and see the film or not. Similar to previous work done in studies [Kin07, GCV13, HTHW07], we decided to use movie review aggregator scores and review counts as an input variable for predicting the box-office outcome. For both Hollywood and Bollywood, there are many sites that collect the scores of different movie critic sources and use the individual review scores to calculate an aggregate. Such websites usually have their own algorithms for assigning weights to different critics and review sources for calculating the optimal score. The scores carry the general sentiment of movie critics for a particular movie. The history and popularity of sites like Metacritic

and Rotten Tomatoes[18] among movie-lovers, has shown that people find the service of aggregate scores useful in their decision-making process for picking movies to see.

For Hollywood movies we collected movie review scores from the critic score aggregator website Metacritic and for Bollywood review scores we gathered from the movie info portal SahiNahi. The main reason we picked these review sites was that compared to many competitor review sites we investigated, these two had scores available for the most movies in our dataset. Also as mentioned before, Metacritic had been used in a number of past studies. We did not find any articles, which had used SahiNahi scores as an input variable for box-office score predictions, but we did not also find any other Bollywood movie critic aggregate site scores having been used either. In addition to the overall movie score, we collected the number of total reviews the scores were based on. Since both sites also report the count of positively or negatively classified reviews we were able to collect this information as well. We expect movies with a higher number of individual critic reviews to be more popular and attract a broader audience to the cinemas than movies with fewer reviews.

3.4 From Movie Revenue Information Sites

General movie information e.g. runtime, genre and the box-office results for Hollywood movies was collected from Box Office Mojo website which is often used as a source of financial movie information in similar studies to ours [AH10, MYK13]. In case of Bollywood we collected the data from movie information portal Box Office India. Since for Bollywood movies the parental rating information was not available from Box Office India, we gathered the information from Times of India daily news website[19] which includes movie reviews for most of the Bollywood movies. For us the most interesting data points were the number of theatres the movie were released in, the opening weekend gross domestic income and the budget of the movie.

3.5 Data Cleaning

Unfortunately we did not end up having all the features for every movie we collected available. For example for some Hollywood and Bollywood movies the budget info had not been disclosed. Because we use the budget as one of the predictor variables then movies with no budget information were discarded from further study. Also for a few movies like *The Bounce Back*, the Metascore was not available because there were not enough critic reviews about the movie available for Metacritic to generate an aggregated score. Table 4 shows the number of movies remained after cleaning was applied. The amount of movies is divided into movies we used for building the model and the movies

[18] https://www.rottentomatoes.com

[19] https://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews

Info	Hollywood	Bollywood
Movies in Train/Test Set		
Time range	April 2015 - December 2017	
Movies before cleaning	337	267
Movies after cleaning	318	170
Movies in Validation Set		
Time range	January 2018 - April 2018	
Movies before cleaning	29	16
Movies after cleaning	28	16

Table 4. Movie and tweet information comparison between train/test and validation datasets

we later used for the validation dataset. The reason why the number of Bollywood movies dropped from 267 to 170 after cleaning was that for several less popular movies we did not have the budget or critic rating information available. Also, movies like these might not have had enough tweets in the 1% tweet sample data matching the hashtags we were looking for. One Hollywood movie was removed from the validation dataset because we had used a wrong hashtag for a movie and therefore it was an obvious outlier with too few tweets.

4 Approach

After we had gathered the data about movie tweets, critic scores and general information about the movies described in Chapter 3, there was still work to be done before we could start predicting the movie box-office outcome. This chapter covers various data preparation steps taken and an overview of our approach to predictive modeling.

4.1 Feature Engineering

Feature engineering in machine learning is a process of inventing and discovering new features as input to machine learning models. Having domain knowledge about the field of research is needed to come up with such new features. In addition to the initial features we gathered in Chapter 3 which we could directly apply to our models, there were also additional features we could prepare.

4.1.1 From Movie Metadata

Movie release date by itself is not very likely to be a useful predictor variable, however using this information we can extract similar to previous studies [CK05, Fet10] the *seasonality aspect* of the release, which has shown to have a positive effect on predicting the movie revenue.

Using the movie release date we can also calculate, how many movies were *simultaneously* released on the same weekend. Since people have limited time for going to the cinema and will watch one or two movies during the weekend, more competition from other releases could mean a loss of revenue. Movie release dates sometimes shift at the last minute to avoid clashes with big blockbuster movies.

According to [Wik], distributors for Hollywood movies can either be from one of the six major distributors, e.g., Universal, from one of the eight mini-major, e.g., Lionsgate or minor distributors. Instead of using specific distributors for distributor feature values we use one hot encoding to create three dummy features to indicate to which set of distributors a movie belongs to. We do this to reduce the number of features to make the model more general.

From the movie titles, we could identify whether a movie might be a sequel if the title ended with a number or contained a colon e.g. *Kahaani 2, Maze Runner: The Scorch Trials*. Instead of using a boolean type feature to indicate whether the movie is a sequel, we used the sequel number as the value. We expect films with more sequels to have an established fanbase and have proven themselves to be profitable in the past to justify a new release.

4.1.2 From Movie Critics Review Data

For movie review data we had captured the aggregated critic score for the movie, but we also had for Hollywood the individual number of positive, neutral and negative critic review available. For Bollywood films, we did not have the number of reviews in three separate sets and only had the number of positive and negative reviews. We simply summed the number of reviews together to create a new feature, the *total number of critic reviews*.

4.1.3 From Twitter Data

After we had extracted all the tweets belonging to particular movies, we decided to calculate the *hourly tweet rate* from the period of two weeks before movie's release and also calculated the hourly tweet rate for the individual days. For each movie, we also calculated the average sentiment *polarity* and *subjectivity* scores of the tweets. The next section covers our approach to tweet sentiment analysis in more detail.

4.2 Sentiment Analysis

Sentiment analysis, also often referred to as opinion mining, is a process of using natural language processing and text analysis methods to determine and quantify the subjective emotions of the text author. It can be divided into two main sub-tasks, the subjectivity recognition, and polarity detection. Subjectivity information shows how many personal impressions the text contains and polarity reflects the author's favoritism or dislike towards the topic. In our study, we attempt to quantify this information and use it as an input variable to our prediction models.

4.2.1 Tweets

Similar to previous studies [CL17, AH10], we wanted to capture the sentiment expressed in the tweets about the movie and use it as a feature for predicting the movie box-office outcome. A natural expectation confirmed by [AH10] is that when word of mouth about a movie has a positive tone, then it is likely to influence others to go and see the film. Negative feedback about the movie should have the opposite effect and steer people away from watching it.

Before calculating the sentiment scores for films, there were a few preprocessing steps we applied to the tweets.

1. Discarding tweets, not in English
2. Removal of movie title from the tweet
3. Removal of Twitter features

- Removal of Reserved words (e.g., RT, FAV)
- Removal of Twitter mentions
- Removal of URLs
- Removal of hashtags

For detecting, if a tweet is in English, we use a Python library named Langdetect[20] and discard the tweet from further analysis if it is not. Removing the movie title is an important step especially for a movie like *Love, Simon*, where the word "love" is present in tweets quite often and the final score biased more towards positive sentiment. For removing Twitter-specific features from tweets, we used the library Preprocessor[21]. Stop words were discarded using the stop word list from Python's Scikit-Learn[22] library, which provides a list of 318 English stop words compared to the default 153 in Natural Language Toolkit (NLTK).

To get the average tweet sentiment for a movie, we calculate the sentiment score of each tweet individually and then take the mean score of all the tweets. Unfortunately, we did not have the time to manually label positive and negative tweets for training a custom sentiment classifier on movie tweets. Also, we did not find a publicly hosted corpus for such purpose. However, the Python library *TextBlob*[23] we use for sentiment analysis can be configured to use a Naive Bayes classifier from NLTK, which has been trained using movie reviews. This means that the classifier we use has been trained using text from the same domain as our work. As an output from TextBlob, we get the polarity and subjectivity scores for the input tweet text and the average of the scores across all tweets for a movie will be used as sentiment polarity and subjectivity features for the prediction model.

4.2.2 Movie Reviews

As mentioned in paragraph 3.2, in this work we do not use individual movie review texts as a data source for extracting features to apply to our model. Instead, we use aggregated scores from different critic's movie reviews. In our work, the aggregated score itself reflects the sentiment about the movie along with the ratio of positive, negative and neutral reviews.

4.3 Dependent and Target Variables

Finally, after we had performed sentiment analysis and feature engineering, we had a list of all the different variables we could use for predicting the box-office results. There

[20] https://pypi.org/project/langdetect
[21] https://pypi.org/project/tweet-preprocessor
[22] http://scikit-learn.org
[23] http://textblob.readthedocs.io/en/dev/

are also a number of different target variables we could choose to predict including the opening weekend, opening week and total domestic revenue for a movie.

4.3.1 Movie Metadata

- **Budget** - For Hollywood movies, this means the production budget amount. In case of Bollywood movies, the budget also includes the advertising costs.
- **Theaters/Screens** - The difference between theaters and screens in this context is that one theater location might show the film on multiple screens. For Hollywood, we strictly count the number of theaters the movie was playing at during the release week. For Bollywood movies the number of screens is counted instead. A screen for Bollywood movies is at least three shows per day.
- **Parental rating** - The rating, which helps to identify, which movies are suitable for children. For Hollywood movies the Motion Picture Association of America (MPAA) rating and for Bollywood movies the Central Board of Film Certification (CBFC) rating is used.
- **Distributor power** - The company responsible for marketing of the film. It is usually different from the company that produced the movie. In our dataset, this information is only available for Hollywood movies.
- **Runtime** - Movie length in minutes.
- **Genre** - The main genre the film belongs to.
- **Release period** - Four separate boolean value features indicating if the movie was released in the Christmas period (November-December), Summer (May-August), Easter (March-April) and other (the remaining months).
- **Simultaneous releases** - The number of movie releases on the same weekend.
- **Sequel number** - If the movie is a sequel then it would have a value of 1, if it is the third movie in the series, then the value would be 2 and so on.

4.3.2 Critic Reviews

- **Metascore/Critic rating** - For Hollywood movies, we use the Metascore rating from Metacritic website, which ranges from 0 to 100. For Bollywood movies, the average critic score between 0.00 and 5.00 from SahiNahi portal is used.
- **Total number of reviews** - The total number of critic reviews about the movie.

- **The number of positive reviews** - The number of critic reviews about the movie that had a positive sentiment about the movie.
- **The number of mixed reviews** - The number of critic reviews about the movie that had a mixed sentiment about the movie. In our dataset this information is only available for Hollywood movies.
- **The number of negative reviews** - The number of critic reviews about the movie that had a positive sentiment about the movie.

4.3.3 Twitter

- **Hourly tweet rate** - Average tweets for a movie during two weeks before the release.
- **Sentiment polarity** - A float value within the range [-1.0, 1.0] showing movie tweet polarity score. A high negative value close to -1.0 would mean that people are saying bad things about the movie before its release on Twitter. A high positive value close to 1.0 would indicate that people are anxiously anticipating the movie release. A value around 0.0 would mean that people have mixed or neutral feelings about the movie.
- **Sentiment subjectivity** - A float value within the range [0.0, 1.0]. A value of 0.0 would mean the tweets about the movie are very objective, and a value of 1.0 would mean the tweets contain very personal opinions and beliefs.

4.3.4 Target Variable

- **Opening weekend** - The domestic revenue for a movie earned from the opening weekend. According to [SS00] most movies typically make 25% of their income during the opening weekend making it a suitable target variable for estimating the eventual financial success of a film.

4.4 Exploratory Data Analysis

Before starting to build machine learning models it is good to know what kind of data we have in our dataset and how it looks like from a higher level. In this section we are looking at the summary statistics of feature values, their distribution and the correlations between the features and the target variable.

4.4.1 Hollywood

Table 5 lists the summary statistics of numerical variables for Hollywood movies. There is quite a big difference between the film that made the least amount of money during the opening weekend and the highest grossing film (0.39 and 300 million USD). A similar difference can be observed for the movie budgets. This wide gap shows that our sample of movies is quite broad and represents both smaller wide release movies as well as big blockbuster movies. From the Metascore values overview, we can see that the selection includes both critically acclaimed movies (highest Metascore of 94 and lowest 11) and both the mean and 50th percentile are around 50, which shows the average critic score of a film. The estimated hourly tweet rate is another interesting variable where we can see that more popular movie tweet rates affect significantly the mean of 221, which is quite a bit larger than the rate at 75th percentile (168).

Variable	**Min**	**Max**	**Mean**	**Std.**	**25%**	**50%**	**75%**
Opening weekend box-office	0.39	247.97	26.90	37.88	6.20	13.70	28.82
Opening theaters	659	4370	2924	890	2384	3033	3576
Budget (mil. USD)	0.90	300	57.71	59.68	18	35	80
Runtime (in minutes)	80	163	109	16	96	107	120
Simultaneous releases	1	5	2.84	0.88	2	3	3
Metascore	11	94	50.08	16.70	35.25	50	62
Number of critic reviews	4	56	33.26	12.58	25	34	43
Number of positive critic reviews	0	55	14.75	14.08	3	11	24
Number of mixed critic reviews	0	32	12.50	6.97	7	12	18
Number of negative critic reviews	0	27	6.00	5.84	1	4	9
Tweet sentiment polarity	-0.26	0.75	0.17	0.11	0.11	0.16	0.22
Tweet sentiment subjectivity	0.09	0.66	0.29	0.08	0.24	0.29	0.33
Hourly tweet rate	2	4701	220.99	480.34	25.25	68	168.25

Table 5. Summary statistics for variables describing 318 Hollywood movies used for building the prediction model

On Fig. 2 the Hollywood distributor distribution is displayed and Fig. 3 shows the three distributors groups used as features for the eventual prediction models. We can

see that indeed all the major distributors are among the top six and represent roughly two-thirds of all the movies released.

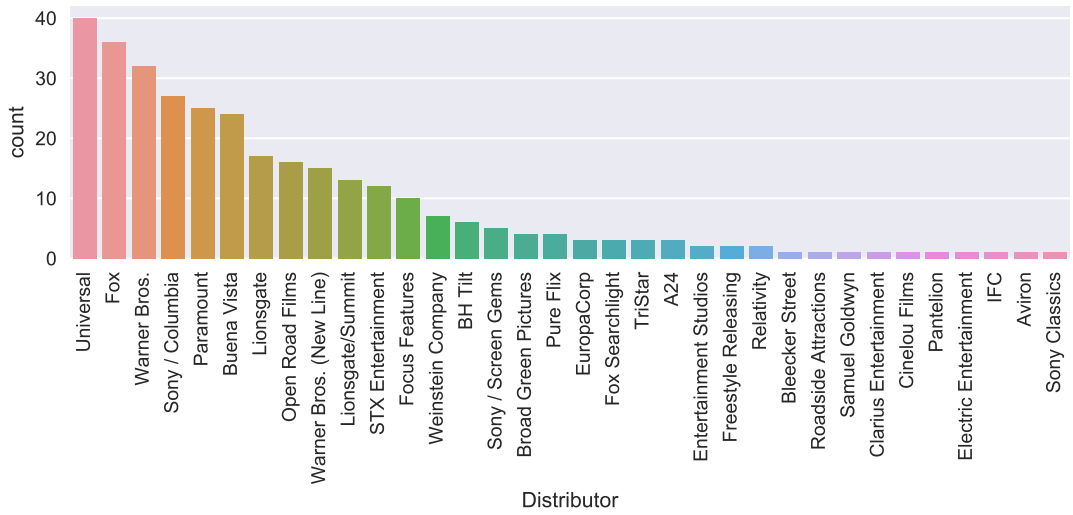

Figure 2. Hollywood movie distributors

.

Over 40% of movies produced are either action movies or comedies with the rest of the less popular genres on Fig. 5. MPAA rating system classifies films by their suitability to children. From Fig. 6 we can see that only one movie was in the G (general audience) category where all the ages are permitted to see the film. No movies belonged to the NC-17 category, which restricts seeing the film for people under 17 years old. Most movies, however, are in the PG, PG-13 and R categories which might contain some material to be inappropriate for children. In the case of R rated movies, the children under 17 have to be accompanied by a parent or an adult guardian.

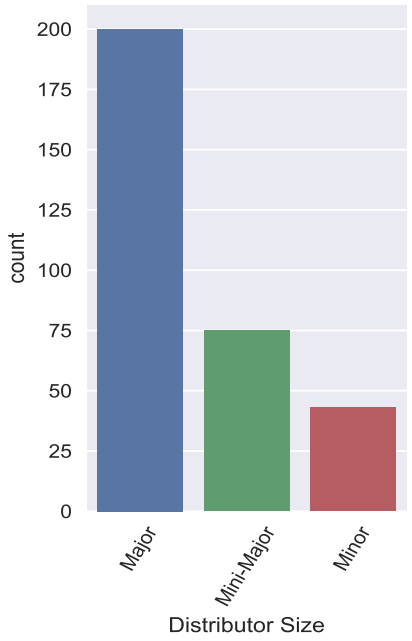

Figure 3. Studio distribution of Hollywood movies

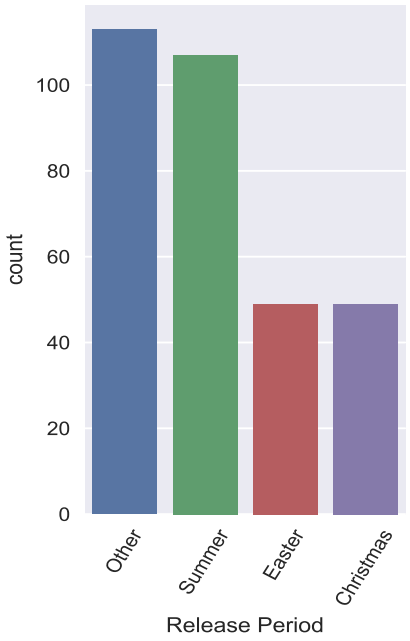

Figure 4. Release period distribution of Hollywood movies

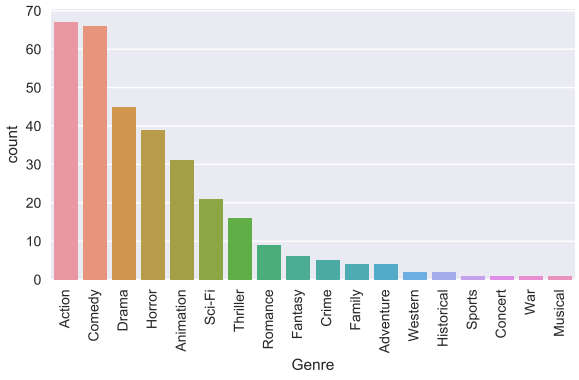

Figure 5. Genre distrib. of Hollywood movies

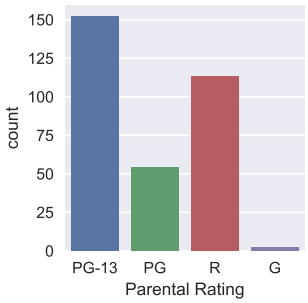

Figure 6. Parental rating distrib. of Hollywood movies

The heatmap on Fig. 7 with feature correlation information can give us strong hints for understanding which variables could be important for predicting the opening weekend box-office. The top three positively correlated features are the number of tweets (0.78), budget (0.72) and the number of theaters (0.61), which all indicate quite strong correlations. We expect these features to be also useful for regression models for predicting the movie revenue. The top three negatively correlated features are the number

of releases on the same weekend (-0.31), the number of negative reviews (-0.2) and tweet sentiment subjectivity (-0.045). The negative correlation here does not necessarily mean that a feature will not be useful for making box-office predictions. On the opposite, the moderate negative correlation of releases on the same weekend variable hints at the expected outcome that more movies opening at the same weekend compete for the same general population to go and see their film and the more movies there are to choose from the less they make on average compared to films that have none or few competitors. It can also hint that sometimes smaller movies try not to compete with big blockbuster movie releases and will release on a different weekend to avoid the strong competition from the hit movies. The weak correlation with the negative review count also shows that the more negative reviews the film has, the less money it is likely to make.

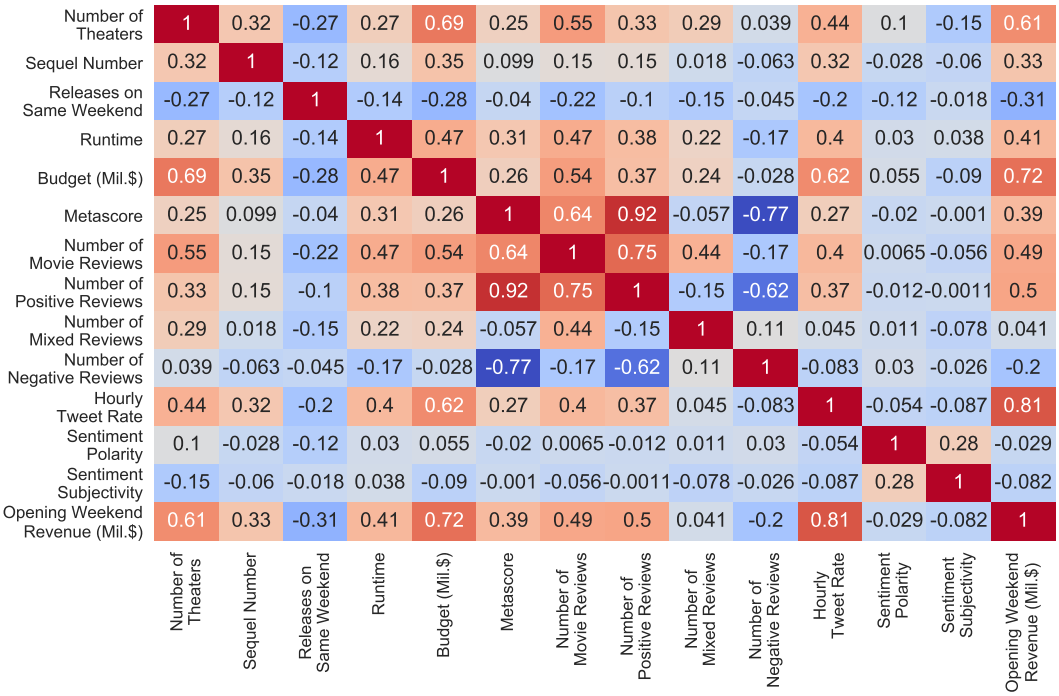

Figure 7. Feature correlations for the Hollywood dataset

.

4.4.2 Bollywood

Table 6 shows the statistical information of the numeric variables in the Bollywood movies train/test dataset, which we can compare the data to the Hollywood dataset overview in Table 5. The highest number of simultaneous releases is the same as in Hollywood, but the mean number of releases is lower (2.20 compared to an average 2.84

films in Hollywood). The total number of Hollywood movies is also larger in our study and hints at a stronger competition in Hollywood. Similar to Hollywood movies, the gap between the lowest and highest budget and opening weekend box-office results is high. As the difference in the number of screens the film released in is also large, this is somewhat expected. On average, the total maximum number of critic reviews collected from Bollywood SahiNahi website is higher than the total movie review count about Hollywood movies on Metacritic. We can also see that the tweet sentiment on average is a bit more positive and subjective for Bollywood movies and the mean hourly tweet rate is similar for both markets.

Variable	**Min**	**Max**	**Mean**	**Std.**	**25%**	**50%**	**75%**
Opening weekend box-office[24]	1.38	1272.80	187.74	237.46	29.40	111.25	247.55
Opening screens	175	4600	1630.15	1001.60	850	1420	2200
Budget[24]	9.50	2100	427.51	410.92	142.50	290	545
Runtime (in minutes)	92	185	132.99	16.60	123	132	142
Simultaneous releases	1	5	2.20	1.02	1	2	3
Critic rating	0.67	3.96	2.45	0.66	2.02	2.44	2.90
Number of critic reviews	3	115	38.06	22.61	19.25	35.50	52.75
Number of pos. critic reviews	0	90	18.61	18.97	4	12	27
Number of neg. critic reviews	0	89	19.46	14.66	8	18	27
Tweet sentiment polarity	-0.04	0.46	0.20	0.06	0.16	0.20	0.24
Tweet sentiment subjectivity	0.09	0.49	0.33	0.06	0.30	0.33	0.37
Hourly tweet rate	2	1592	142.46	242.47	16	46.50	172.25

Table 6. Summary statistics for variables describing 170 number of Bollywood movies used for building the prediction model.

Fig. 8 shows that most Bollywood movies in our dataset (39%) belong to the drama

[24]Large sums of money in India are often represented in crores. One crore is equal to 10^7 rupees.

genre. The ratio of drama movies is substantially different from Hollywood where it has a share of only 14% and is the third most popular genre. However, compared to Hollywood where most movies where action films (20%), in our Bollywood dataset the action genre ranks 4th with a share of 9%.

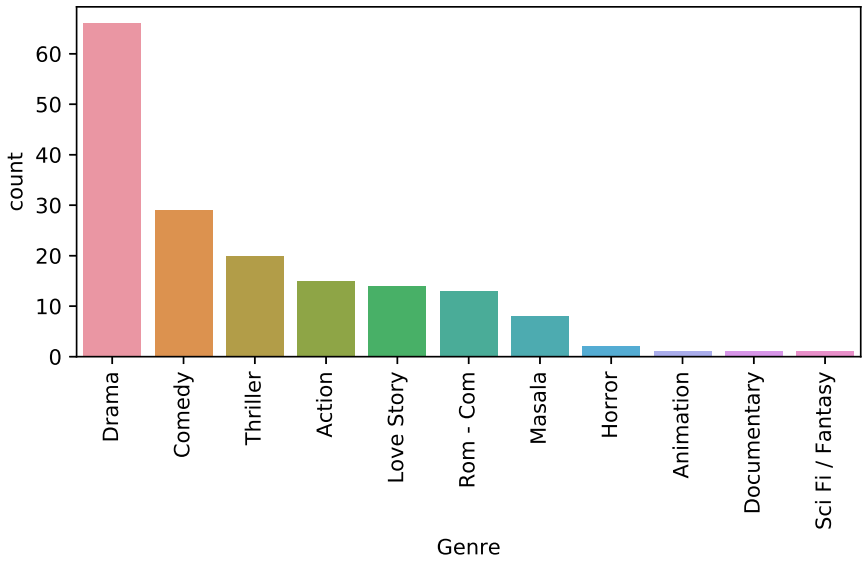

Figure 8. Genre distribution of Bollywood movies

.

Bollywood release period distribution shown on Fig. 9 is similar to Hollywood releases on Fig. 4. There are twice as many movies in the Summer and the Other categories because they last four months instead of two in Easter and Christmas periods. Fig. 10 shows the CBFC rating for Bollywood movies. Predominantly, the UA parental rating has been given to most Bollywood movies, which require children below the age of 12 to be accompanied by an adult. In our Bollywood movie dataset, 19% of the films were restricted to adults only, whereas in Hollywood no such movies were released.

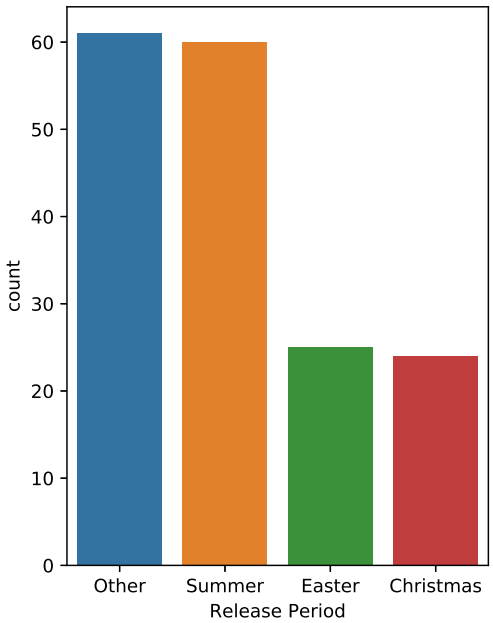

Figure 9. Release period distribution of Bollywood movies

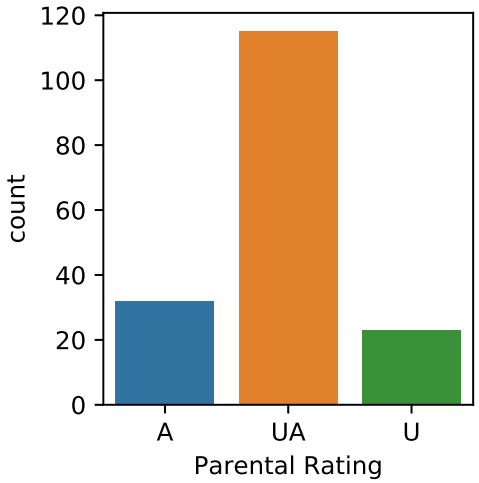

Figure 10. Parental rating distribution of Bollywood movies

According to the heatmap on Fig. 11, the top three features best correlated with opening weekend revenue for Bollywood are: budget (0.88), the number of screens (0.87) and hourly tweet rate (0.69) as they were for Hollywood on 7. The main noticeable difference is that the budget and number of screens are more highly correlated to the opening weekend revenue than they are in Hollywood dataset. Also, the hourly tweet rate has a slightly weaker correlation in the case of Bollywood movies, 0.69 vs. 0.81 in Hollywood. The top features in Hollywood that have a negative relationship are also negatively associated in Bollywood dataset.

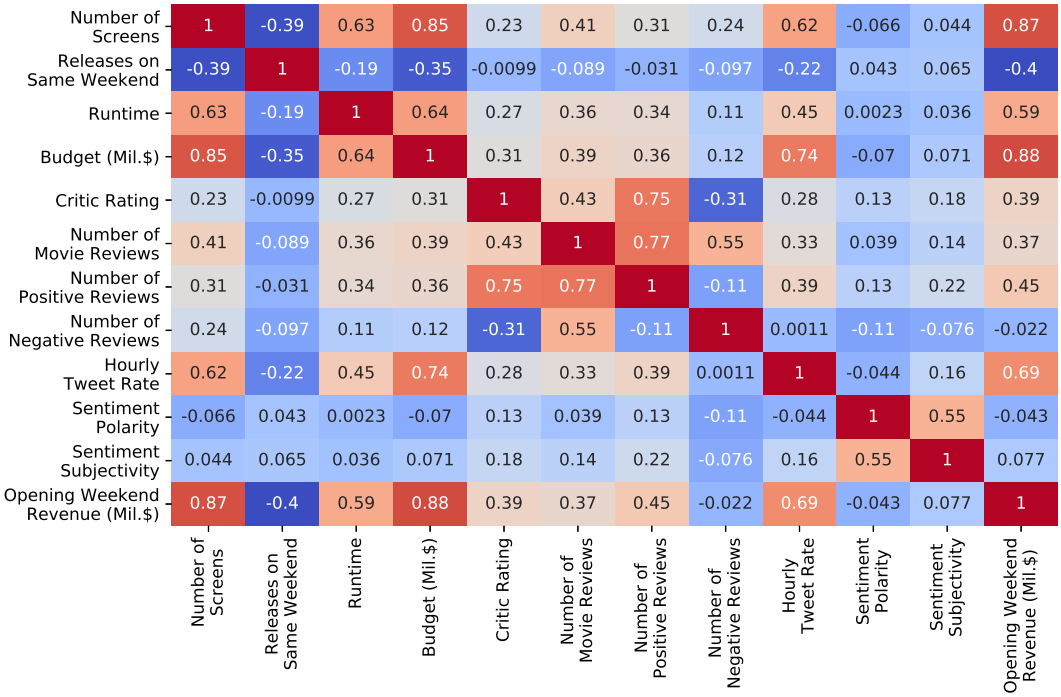

Figure 11. Feature correlations for the Bollywood dataset

.

4.5 Predictive Modelling

Predictive models are used when we need to predict an unknown event based on some previous information. As discussed in the related works section, predicting movie success can be considered either as a classification or as a regression problem. We chose to treat predicting movie box-office results as a regression task because regression algorithm output is a single continuous target variable - the amount of money a movie will make and is more easily interpretable than a category defined by an arbitrary threshold. Also, when needed, the prediction results from a regression model output could be later used to classify movies into categories such as hits or flops. There are several supervised machine learning algorithms that have support for regression problems with each having their strengths and weaknesses. In this section we will briefly look at some of the algorithms we will be using in our work and describe their working principles for predicting an outcome.

4.5.1 Ensemble Learning Methods

Ensemble learning methods combine results from several machine learning algorithms or several different models. These are so-called meta-algorithms that use the average prediction result of models they encompass in case of regression or vote on the output category in case of classification problems. People use them for the lower error rate they generally provide and also they are less prone to overfitting than individual models because they have on average a lower bias. The only downside of using an ensemble learning method is that it takes more time or resources to train multiple models.

Bootstrap Aggregation, also referred to as *bagging* is an example of an ensemble learning method. In practice, it usually uses the same learning algorithm for training multiple models each on a different set of training data. For each subset of data, the training data is picked randomly. If the data is chosen with replacement, then more than one individual instance of the data can end up in the same subset. These subsets are then used to train individual models, and their outputs are averaged to return a single prediction result. The process of training these models can be done in parallel since one model result does not depend on the other.

Random Forest [Bre01], which we will be using in our work, is one example of such ensemble learner, which uses bagging. In a Random Forest learning algorithm, multiple models using the Decision Tree algorithm are built with a different subset of training data for each model. Finally, the average prediction result from each model is returned as the predicted output.

Boosting is another ensemble learning technique. Compared to bagging, where models are created in parallel, the models when using boosting need to be built in series. As with any supervised learning algorithm, the objective is to define some loss function like MSE for regression problems and try to minimize it. Boosting algorithms start by first building a simple weak learner with a prediction performance slightly better than an average guess and calculate the prediction error residuals. They use the information learned from predicted errors and build subsequent simple predictors to learn from the mistakes of the previous simple models. Previous predictors are not changed and only new ones are added, finally, when the stopping criterion is reached the predictors are combined by giving weights to each predictor. It can happen that the performance of the loss function that is being optimized to train the model starts improving, but the performance on the test set starts decreasing. This means that the models start to overfit and it may be better to stop training at that point.

XGBoost (eXtreme Gradient Boosting) [CG16] is a popular machine learning algorithm implementation that uses gradient descent algorithm to minimize the loss when adding new weak learners. It improves on the standard Gradient Boosting decision tree algorithm by improving the speed and memory utilization. In addition to that it adds regularization support to penalize models with too many parameters, which helps to reduce overfitting. We decided to test the algorithm on our dataset because it is reportedly

used by many winning Kaggle[25] competition submissions.

4.5.2 Bias-Variance Tradeoff

The central problem of supervised machine learning is about finding the right balance between bias and variance. Ideally, we want to build a model that has both low bias and low variance, but it is very difficult to achieve both at the same time. Learning methods, which produce high variance can easily lead to *overfitting* when model fits the training data really well, but performs poorly on test data, because the model had learned all the small nuances of the training data, which was representative of all the data. In other words, overfitting occurs when the model captures the noise and the outliers in the data along with the underlying pattern. These models are usually complex like Decision Trees, SVM or Neural Networks which are prone to overfitting.

A learning method with high bias, however, would mean using a simpler model, which does not capture important regularities in the data on the training set and therefore also *underfit* on the test set. Underfitting occurs when the model is unable to capture the underlying pattern of the data. These models usually have low variance and high bias. The Linear Regression statistical method is one example of such a learning method.

We compare the training and test set errors to measure if our model is over- or underfitting to the data. In case of overfitting, the error reported on the training set will be low, but high on the test set. If we are underfitting our model to the data then both the training and test set errors will be quite high.

4.6 Building a Machine Learning Pipeline

For conduction our experiments on predicting the box-office results of Hollywood and Bollywood movies we are using the Python 3.6.4 programming language and many open-source software libraries specifically implemented to help people working in the Data Science industry to do their job more efficiently. Table 11 in Appendix A.1 includes some of the main software libraries we are using in our experiments and briefly describes their use in our context. During the recent years, the support and availability of such tools have grown rapidly, and it is easier for newcomers to enter the field due to the availability of many free learning resources such as Youtube videos, blogs and Massive Online Open Courses (MOOC).

For writing code for gathering and processing the data, we used a traditional IDE, however for visualizing our data, building the machine learning models and evaluating the results we implemented a pipeline using the Jupyter Notebook [KRKP+16] software environment. In our experience, the tool suits very well for our workflow and enables faster feedback and supports a nice visual feedback loop. Fig.12 shows the general flow

[25] https://kaggle.com is a web service for hosting and participating in machine learning competitions

and steps of the machine learning pipeline we set up for evaluating the results of the prediction outcomes.

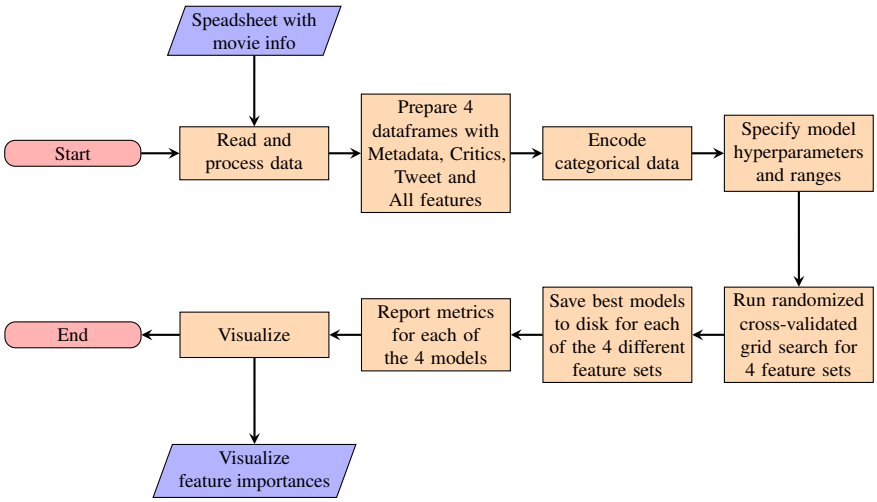

Figure 12. Flowchart of our machine learning pipeline

4.6.1 Reading the Data

We read the collected data previously organized into a spreadsheet format using the pandas library into four separate dataframes each containing a different set of features. The reason we separate the data is to be able to train models with different dependent variables with the goal of comparing the prediction results and determining whether it is the features from Twitter or features from movie critics that hold the better box-office predictive power. Table 7 describes the four different sets of features we evaluate.

Feature groups	**Features available**
Meta	Metadata features listed in section 4.3.1
Critics	Metadata features + movie critic features listed in section 4.3.2
Twitter	Metadata features + Twitter features listed in section 4.3.3
All	Uses all features from Meta, Critics and Twitter feature groups

Table 7. Different feature groups and the features they contain.

4.6.2 Encoding Categorical and Ordinal Features

Unlike XGBoost, the `RandomForestRegressor` implementation of scikit-learn is not able to handle categorical values directly. We need to perform a separate step and encode categorial feature values such as the parental rating, release period and genre. This is

achieved using one hot encoding by creating a binary variable for each possible value of a categorical feature. For example, if the movie belongs to the comedy genre, then it would have a variable *is_comedy* with value 1, and the film would have value 0 for all other genres. The benefit of using one hot encoding over label encoding, which would assign a fixed numeric value for a specific genre e.g. 1 for comedy and 2 for an action movie, is the fact that such label encoding would contain a relationship not present in real life. A comedy genre is not twice as much as an action movie. By using one hot encoding, we avoid this problem, however as a downside, we end up having more columns in our dataset. Even though the parental rating is an ordinal feature (feature values are ranked from general audience to adults only), we also use one hot encoding for the variable because it is difficult to define the difference between the possible values.

4.6.3 Tuning Model Hyperparameters

Usually, it is best to train the first model using the default parameters and then try to improve the score by gathering more data or doing feature engineering. To improve the scores even further, machine learning algorithms have several hyperparameters which we can control to affect the final prediction outcome of the model. Unfortunately, there is not a one size fits all solution here because the outcome heavily depends on the data used for model training. The recommended approach is to start off with a wide set of parameters and quickly try out different combinations and then we should pick the parameter values that showed the best performance and based on that info try to pick new ranges to narrow the values further down. Sci-Kit Learn's `RandomizedSearchCV` method offers a convenient way to pick random samples of hyperparameter combinations and evaluate the model scores by performing k-fold cross-validation. The method returns the model with the best performance and the hyperparameters that were used to train the model. In our experiments, we always use 10-fold cross-validation and then calculate the average scores across all folds for the final result.

4.6.4 Saving the Model and Loading it Back from Disk

To be able to later re-use for predicting and further examinations we can save the models to disk using the Joblib package. It comes installed with Sci-Kit Learn and provides utilities for saving and loading Python objects using the `joblib.dump` and `joblib.load` methods respectively.

4.6.5 Reporting and Comparing Results

There are several scoring metrics we are interested in observing from the model prediction results. Previous research has used different metrics and to be able to compare our results to theirs, we report the r^2, MAPE, MAE, MSE and RMSE scores for each model.

4.6.6 Visualizing Predictions and Feature Importances

Scoring metrics are useful for explaining the results, but a visual representation of the predicted results can offer additional help such as the easier detection of outliers in our data. Visualizing how each feature contributed to achieving the best score gives us an idea, which features are more important than others when making predictions.

5 Empirical Results

In this section, we give an overview of our analysis about both Hollywood and Bollywood movies. To quickly evaluate different algorithm and feature combinations, we used the machine learning pipeline described in section 4.6. For building and validating our prediction models, we decided to split our dataset into two parts (90 % training and 10 % test set). Table 4 shows that 318 Hollywood movies were in the train/test set and 28 movies were left for validating the model performance to make sure the model is not overfitting and is able to generalize well to our problem. For Bollywood, 170 movies were used for building the model and the validation contained 16 movies.

We report the first set of scores using default algorithm hyperparameters and then after evaluating the initial results we try out different combinations of the parameters to see if we can improve the model performance. Although we report many different scoring metrics, during cross-validated randomized hyperparameter search, we always optimize the model to have the lowest MSE. For doing this we use the Sci-Kit Learn's `RandomizedSearchCV` utility for selecting the best parameters that maximizes the score of the held-out data. Table 12 in Appendix A.2.1 lists the hyperparameter grid values we tested with Random Forest algorithm. Similarly Table 13 in Appendix A.2.2 lists the parameters and ranges evaluated with the XGBoost algorithm. We specify the number of iterations (`n_iter`) parameter for `RandomizedSearchCV` to be 100, meaning that not all parameter combinations are exhaustively checked, but only 100 random combinations. The law of diminishing returns applies here, meaning that the more combinations we would check, the better results we would get, but we would be spending much more time on evaluating the combinations. However in practice it makes sense to see what parameter combinations from a limited set give better results and then drill down deeper and explore the neighboring ranges of variables that worked well for the model. First we gathered the results using Random Forest and XGBoost algorithms with default hyperparameters and then try to improve the prediction scores by selecting random combinations of hyperparameters.

5.1 Wisdom of the Crowd Vs. Reviews of the Experts

To compare the predictive performance of the features from Twitter against the features from movie critic aggregator websites, we train models with different subsets of data listed in the Table 7 using Random Forest and XGBoost machine learning algorithms.

Table 8 shows the performance metrics for different models on the Hollywood movies dataset. Since the sample size of training and validation data is different, comparing MAE, MSE, RMSE between the different datasets is not meaningful. However, we can see that the RF model with Twitter features has the best scores on the training set, but does not do well on the validation set. Models with critics features perform worse compared to Twitter on the training set, but have better scores on the validation set. To our surprise the

models using features from only movie metadata perform the best on the validation set. This is an indication that the models with more features are overfitting with the training data. In this case a less complex model is able to generalize better. To overcome this problem, early stopping could be used to measure the model performance on a separate evaluation holdout dataset. The training would be stopped when the performance on the evaluation dataset has not improved after a specified number of iterations.

Model		**Train/Test Set**					**Validation Set**				
		r^2	MAE	MSE	RMSE	MAPE	r^2	MAE	MSE	RMSE	MAPE
Metadata	RF	0.67	11.1	444.3	21.1	72.3	0.22	7.6	101.1	10.1	54.1
	RF*	0.71	11.5	385.9	19.6	73.4	-0.59	9.9	205.1	14.3	69.0
	XGB	0.68	11.0	443.4	21.1	74.1	**0.25**	**7.0**	**96.5**	**9.8**	**47.6**
	XGB*	0.70	11.1	417.6	20.4	77.8	0.01	8.1	126.9	11.3	53.4
Metadata ∪ Critics	RF	0.75	10.1	345.7	18.6	66.8	-0.49	9.3	191.6	13.8	64.1
	RF*	0.74	10.1	355.0	18.8	64.5	-0.19	8.3	153.5	12.4	60.4
	XGB	0.74	10.4	368.7	19.2	69.9	-0.17	8.6	151.8	12.3	63.9
	XGB*	0.74	10.2	301.9	17.4	75.8	0.04	7.7	123.6	11.1	60.6
Metadata ∪ Twitter	RF	**0.80**	**9.3**	**264.8**	**16.3**	64.5	-1.10	10.4	266.6	16.3	56.4
	RF*	0.78	9.50	296.4	17.2	**61.3**	-1.11	9.86	271.9	16.5	53.7
	XGB	0.71	10.1	374.9	19.4	68.0	-1.04	9.7	263.8	16.2	55.4
	XGB*	0.75	10.3	300.7	17.3	74.9	-1.37	11.01	305.7	17.5	64.7

Table 8. Hollywood train/test-set performance for RF (Random Forest), XGB (XGBoost) models using metadata and combinations with critics and Twitter features, measured using five different metrics. Within a column, **boldface** shows the best result for a metric. Models marked with the asterisk symbol (*) indicate that hyperparameter tuning was performed and the results are reported for the best estimator.

Table 9 shows the model performance on the Critics and Twitter features using the Bollywood dataset. Both sets of models, with critics and Twitter features, do equally well on the train/test sets. The models with Twitter features however do better on the validation set in terms of r^2, MAE, MSE and RMSE, but critics models are able to achieve a lower MAPE. Our initial expectation was that lower box-office revenue error metrics would also result in a lower MAPE. However, it could be that the small sample size of 16 movies in the validatation dataset in this instance performed better in terms of MAPE using the models with critics features and there are outliers, which penalize the results for the Twitter-based models more than others. Compared to low r^2 scores on the validation dataset for Hollywood, movies in the Bollywood dataset have a high r^2. This is illustrated by the differences between Hollywood and Bollywood opening weekend revenue predictions for XGBoost models with Twitter features on Fig. 13 and Fig. 14. The Figures show that predictions for Bollywood (Figure 13) follow a linear line (r^2 = 0.86), but are more widely spread for Hollywood (Figure 13) forecasts (r^2 = -1.37).

After evaluating the model performance metrics (see Tables 8 and 9), we could not find a clear winner and therefore cannot declare either movie expert reviews or Twitter a better source for predictions. However, our cross-validated results on train/test sets

show there is value in features from Twitter and movie critic review scores in predicting opening weekend box-office results over a sole metadata-based model.

Model		Train/Test Set					Validation Set				
		r^2	MAE	MSE	RMSE	MAPE	r^2	MAE	MSE	RMSE	MAPE
Metadata	RF	0.75	6.3	98.1	9.9	116.9	0.85	5.8	87.5	9.4	82.0
	RF*	0.78	6.2	89.4	9.5	108.5	0.90	5.6	60.6	7.8	91.5
	XGB	0.76	6.3	87.6	9.4	122.9	0.83	6.8	96.3	9.8	110.3
	XGB*	0.79	6.0	78.9	8.9	111.6	0.87	6.1	74.8	8.7	100.5
Metadata ∪ Critics	RF	0.77	5.9	96.2	9.8	94.7	0.77	6.5	131.8	11.5	**49.4**
	RF*	0.80	5.7	81.6	9.0	103.8	0.86	5.5	80.9	9.0	61.0
	XGB	0.80	5.7	81.1	9.0	**91.5**	0.85	6.0	87.5	9.4	67.6
	XGB*	**0.83**	**5.5**	**69.2**	**8.3**	92.6	0.85	5.9	86.3	9.3	50.0
Metadata ∪ Twitter	RF	0.80	5.9	87.9	9.4	101.0	0.79	7.2	120.8	11.0	66.6
	RF*	0.82	5.6	78.9	8.9	103.2	**0.92**	**4.9**	**46.6**	**6.8**	72.5
	XGB	0.80	6.0	86.3	9.3	92.7	0.86	6.0	78.5	8.9	82.3
	XGB*	**0.83**	5.9	69.4	**8.3**	124.2	0.90	**4.9**	54.6	7.4	86.0

Table 9. Bollywood train/test-set performance for RF (Random Forest), XGB (XGBoost) models using metadata and combinations with critics and Twitter features, measured using five different metrics. Within a column, **boldface** shows the best result for a metric. Models marked with the asterisk symbol (*) indicate that hyperparameter tuning was performed and the results are reported for the best estimator.

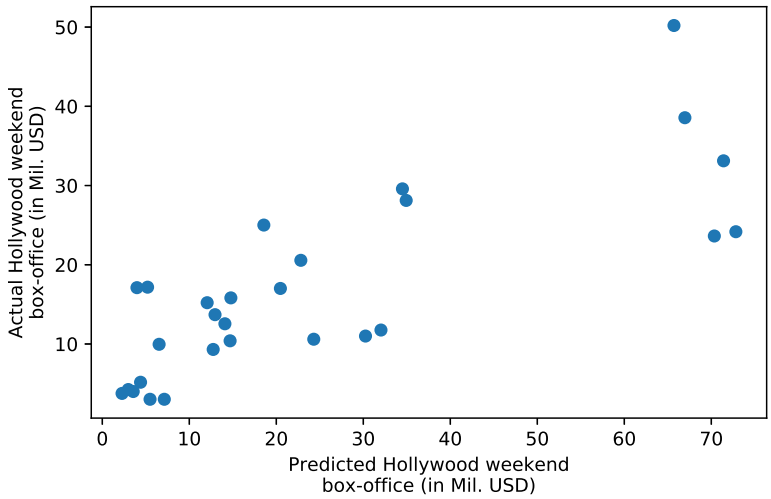

Figure 13. Hollywood validation set predictions compared to actual results using XGBoost model with Twitter features.

.

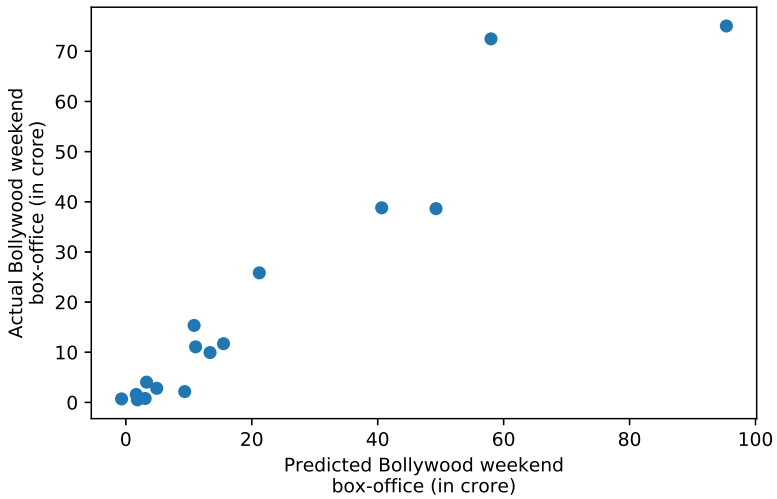

Figure 14. Bollywood validation set predictions compared to actual results using XG-Boost model with Twitter features.

.

From Figures 15 and 16 we can see, which variables were most used to reduce the variance when splitting nodes inside the individual trees of the XGB* and RF* models. These figures indicate that features from Twitter are more important than the features of movie critic reviews.

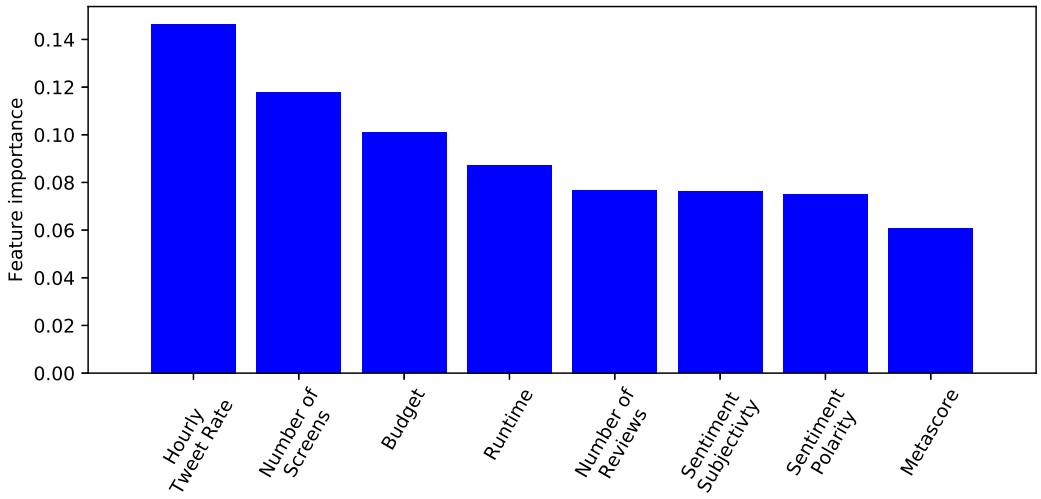

Figure 15. Feature importance for the Hollywood XGB* model in Table 10

.

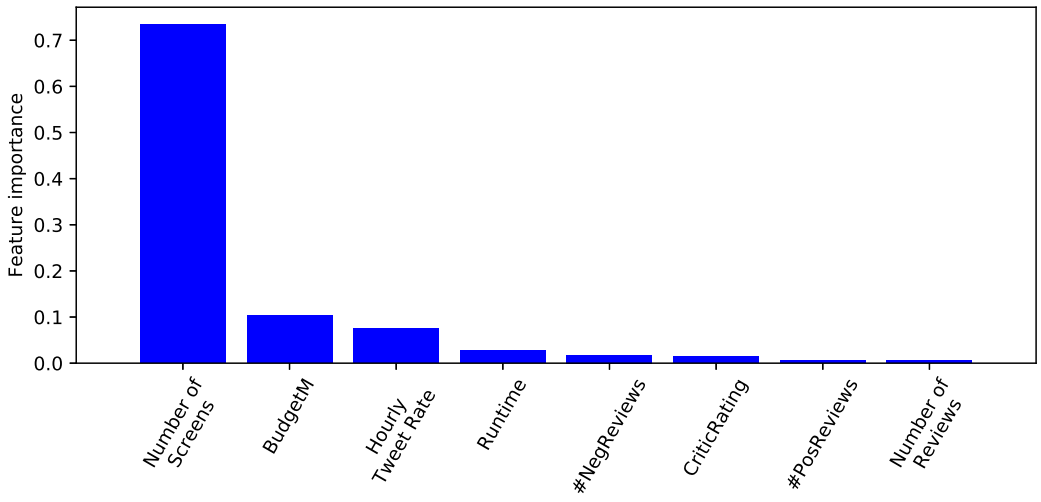

Figure 16. Feature importance for the Bollywood RF* model in Table 10

.

5.2 Hollywood Vs. Bollywood

We are not able to compare Hollywood and Bollywood movie revenue prediction models using scoring metrics MAE, MSE, RMSE, because the output is a monetary value and is different for both markets. Instead, we can compare the r^2 and MAPE values our models were able to achieve. For performing this evaluation, we build the models using the feature group *All*, listed in Table 7. The models consist of combined metadata, critics and twitter features.

The r^2 values for the train/test set are similar for both Hollywood and Bollywood, showing that depending on the model, for Hollywood between 72% and 80% and for Bollywood 79% to 86% of opening weekend box-office variance can be explained by the target variables. However on the validation set, in case of Hollywood movies the r^2 values are very low compared to Bollywood. Figures 17 and 18 show the predicted vs. the actual validation set opening weekend box-office results using XGBoost model trained with all features. We can see that for Bollywood a linear regression line can be fitted in a way, which does not produce large residuals between the predicted values and the regression line. In Hollywood's case the predicted values are more spread around and do not form a straight linear pattern resulting in a low r^2 metric value. However low r^2 does necessarily mean that the model cannot be used to make predictions from. The MAPE values are similar on the validation sets for Hollywood and Bollywood movies, ranging from 50.8% to 63.8%.

Model		Train/Test Set					Validation Set				
		r^2	MAE	MSE	RMSE	MAPE	r^2	MAE	MSE	RMSE	MAPE
All features (Hollywood)	RF	0.80	9.0	254.2	15.9	63.9	-1.30	11.1	302.0	17.4	59.4
	RF*	0.79	8.9	282.3	16.8	57.9	-1.11	10.7	270.9	16.5	61.0
	XGB	0.72	9.5	334.7	18.3	61.4	-1.58	10.7	332.3	18.2	61.3
	XGB*	0.80	8.7	265.7	16.3	**55.1**	-0.46	9.5	188.4	13.7	56.3
All features (Bollywood)	RF	0.79	5.8	87.6	9.4	92.7	0.79	7.2	121.0	11.0	55.9
	RF*	0.83	5.6	78.4	8.9	97.7	**0.92**	4.87	48.3	6.9	57.6
	XGB	0.80	5.8	84.3	9.2	82.9	0.87	5.5	75.4	8.7	63.8
	XGB*	**0.86**	5.4	64.9	8.1	87.3	0.91	4.9	50.7	7.1	**50.8**

Table 10. Hollywood and Bollywood train/test-set performance for RF (Random Forest), XGB (XGBoost) models using all available features, measured using five different metrics. Within a column, **boldface** shows the best result for a metric. Models marked with the asterisk symbol (*) indicate that hyperparameter tuning was performed and the results are reported for the best estimator.

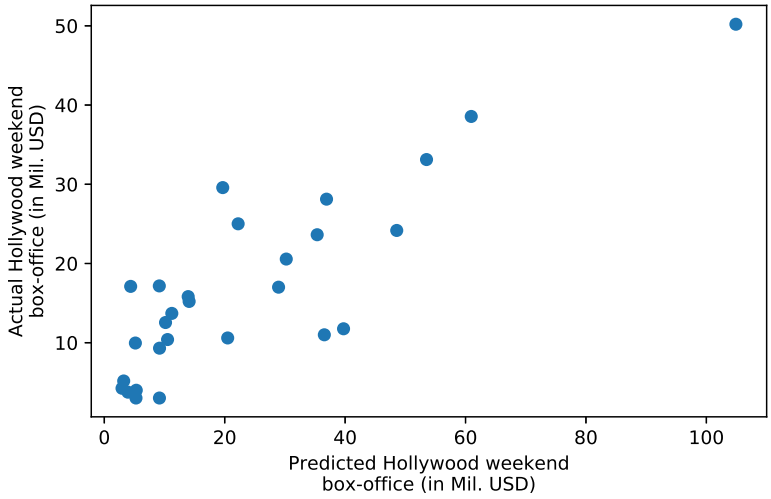

Figure 17. Hollywood validation set predictions compared to actual results using XGBoost model with all the features.

.

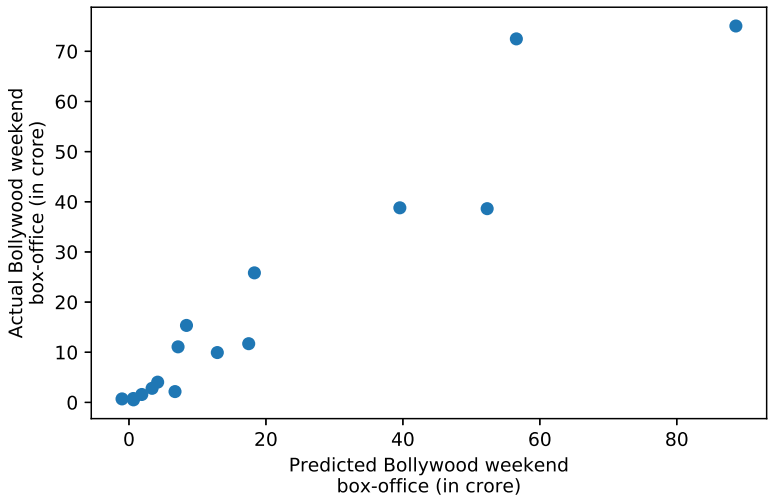

Figure 18. Bollywood validation set predictions compared to actual results using XG-Boost model with all the features.

.

6 Conclusion

6.1 Summary

Movie sales prediction has been an interest to many researchers as they often carry huge investments. In this thesis, we investigated movie sales prediction problem from two different perspectives. Firstly by analyzing the reviews given by movie critics. Secondly, we focus on wisdom of the crowd, collected using Twitter, which is a social media platform. Although, we combined various other metadata features of the movies for example, budget, star cast etc. However, our main focus is to evaluate if it is the movie critics or it is the wisdom of the crowd which can predict the box office collection more precisely. We performed our study using various real datasets consisting of 1) large Twitter dataset about movies, 2) movie ratings given by movie critics and 3) metadata information about the movies itself. In this study, we were not able to find conclusive evidence that the wisdom of the crowd prevails over movie critics in predicting the box office revenues and the other way around.

6.2 Limitations

Although the thesis studies the largest amount of movies with their related tweets to date there are some limitations to our approach. Using 1% tweet sample dataset makes it possible to study a large set of movies and estimate the word-of-mouth volume of tweets, but limits the quality of sentiment analysis we can perform. The problem is that in worst case the movie sentiment is represented by only a few tweets and the random sampling has a big effect on less popular movies. However as shown by previous research [AH10, LIM16] tweet sentiment information is not a very strong predictor variable and using a 1% sample tweet dataset would still be useful for gathering the data to be able to use the strong tweet volume feature in movie box-office prediction models.

Our study was limited to using only tweets in English. Extracting tweet sentiment from tweets in different languages would capture the opinion of a larger demographic. Even though many people in India tweet about Bollywood movies in English, processing Hindi tweets should be a part of future work.

6.3 Future work

6.3.1 Using Additional Realtime Tweets

We stated gathering realtime tweets for upcoming Hollywood and Bollywood movies from October 2017. We are currently planning to keep gathering the data at least until August 2018. Future studies can utilize this dataset for further analysis of movies and their related tweets.

6.3.2 Using Aspect Level Sentiment Analysis

In our current work we used the movie critic aggregator scores as a general sentiment polarity score for the movies. For future work we propose to extract different aspect-level sentiment information from movie reviews similar to [PGS17]. Separate aspec-level sentiment scores e.g. for acting, directing, music could all be used as features for the prediction model.

6.3.3 Additional Sentiment Analysis Approaches

In our work we experimented with a simple sentiment classifier using the bag-of-words method and NaiveBayes classifier. Future work might be conducted using more modern sentiment analysis techniques such as Word2Vec[26].

6.3.4 Additional Variables

Although we used many features, there are additional variables which might be used to build better prediction models. For example from Twitter, the follower counts of actors and movie director could be extracted for estimating the movie popularity.

6.3.5 Time Effects

For our models we used the average hourly tweet rate over two weeks before movie release. Future work could look at the tweet volume on a more granular level e.g. at the daily tweet rate. Some models might be able to capture interesting patterns from either rising or declining tweet rates before the movie release.

[26]https://www.tensorflow.org/tutorials/word2vec

References

[ADAH+10] Fabian Abel, Ernesto Diaz-Aviles, Nicola Henze, Daniel Krause, and Patrick Siehndel. Analyzing the blogosphere for predicting the success of music and movie products. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pages 276–280. IEEE, 2010.

[AH10] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[AJM+13] Krushikanth R Apala, Merin Jose, Supreme Motnam, C-C Chan, Kathy J Liszka, and Federico de Gregorio. Prediction of movies box office performance using social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 1209–1214. IEEE, 2013.

[BBK07] Peter Boatwright, Suman Basuroy, and Wagner Kamakura. Reviewing the reviewers: The impact of individual film critics on box office performance. *Quantitative Marketing and Economics*, 5(4):401–425, 2007.

[BCR03] Suman Basuroy, Subimal Chatterjee, and S Abraham Ravid. How critical are critical reviews? the box office effects of film critics, star power, and budgets. *Journal of marketing*, 67(4):103–117, 2003.

[BGM12] P Thomas Barthelemy, Devin Guillory, and Chip Mandal. Using twitter data to predict box office revenues, 2012.

[BKJ09] Stephanie M Brewer, Jason M Kelley, and James J Jozefowicz. A blueprint for success in the us film industry. *Applied Economics*, 41(5):589–606, 2009.

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[Box] BoxOfficeMojo. Bob Office Tracking By Time. `http://www.boxofficemojo.com/about/boxoffice.htm`, accessed 2018-04-15.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[CK05] Byeng-Hee Chang and Eyun-Jung Ki. Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics*, 18(4):247–269, 2005.

[CL17] Deepankar Choudhery and Carson K Leung. Social media mining: prediction of box office revenue. In *Proceedings of the 21st International Database Engineering & Applications Symposium*, pages 20–29. ACM, 2017.

[DVW99] Arthur De Vany and W David Walls. Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of cultural economics*, 23(4):285–318, 1999.

[ES97] Jehoshua Eliashberg and Steven M Shugan. Film critics: Influencers or predictors? *The Journal of Marketing*, pages 68–78, 1997.

[Fet10] Marc Fetscherin. The main determinants of bollywood movie box office sales. *Journal of global marketing*, 23(5):461–476, 2010.

[GCV13] Shyam Gopinath, Pradeep K Chintagunta, and Sriram Venkataraman. Blogs, advertising, and local-market movie box office performance. *Management Science*, 59(12):2635–2654, 2013.

[GMD15] Dipak Damodar Gaikar, Bijith Marakarkandy, and Chandan Dasgupta. Using twitter data to predict the performance of bollywood movies. *Industrial Management & Data Systems*, 115(9):1604–1621, 2015.

[Hon14] Lee Yoong Hon. Expert versus audience's opinions at the movies: Evidence from the north-american box office. *Marketing Bulletin*, 25:1–22, 2014.

[HTHW07] Thorsten Hennig-Thurau, Mark B Houston, and Gianfranco Walsh. Determinants of motion picture box office and profitability: an interrelationship approach. *Review of Managerial Science*, 1(1):65–92, 2007.

[Jai13] Vasu Jain. Prediction of movie success using sentiment analysis of tweets. *The International Journal of Soft Computing and Software Engineering*, 3(3):308–313, 2013.

[JDGS10] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, 2010.

[KHK15] Taegu Kim, Jungsik Hong, and Pilsung Kang. Box office forecasting using machine learning algorithms based on sns data. *International Journal of Forecasting*, 31(2):364–390, 2015.

[Kin07] Timothy King. Does film criticism affect box office earnings? evidence from movies released in the us in 2003. *Journal of Cultural Economics*, 31(3):171–186, 2007.

[KNS+08] Jonas Krauss, Stefan Nann, Daniel Simon, Peter A Gloor, and Kai Fischbach. Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS*, pages 2026–2037, 2008.

[KRKP+16] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks-a publishing format for reproducible computational workflows. 2016.

[LDC+16] Ting Liu, Xiao Ding, Yiheng Chen, Haochen Chen, and Maosheng Guo. Predicting movie box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications*, 75(3):1509–1528, 2016.

[LIM16] Carlo Lipizzi, Luca Iandoli, and José Emmanuel Ramirez Marquez. Combining structure, content and meaning in online social networks: The analysis of public's early reaction in social media to newly launched movies. *Technological Forecasting and Social Change*, 109:35–49, 2016.

[Lit83] Barry R Litman. Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture*, 16(4):159–175, 1983.

[Liu06] Yong Liu. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, 70(3):74–89, 2006.

[LK89] Barry R Litman and Linda S Kohl. Predicting financial success of motion pictures: The'80s experience. *Journal of Media Economics*, 2(2):35–50, 1989.

[MYK13] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, 8(8):e71226, 2013.

[NS15] Rakesh Niraj and Jagdip Singh. Impact of user-generated and professional critics reviews on bollywood movie success. *Australasian Marketing Journal (AMJ)*, 23(3):179–187, 2015.

[OBTdR12] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke. Predicting imdb movie ratings using social media. In *European Conference on Information Retrieval*, pages 503–507. Springer, 2012.

[PC13] Reggie Panaligan and Andrea Chen. Quantifying movie magic with google search. *Google Whitepaper—Industry Perspectives+ User Insights*, 2013.

[PGS17] Rajesh Piryani, Vedika Gupta, and Vivek Kumar Singh. Movie prism: A novel system for aspect level sentiment profiling of movies. *Journal of Intelligent & Fuzzy Systems*, 32(5):3297–3311, 2017.

[QGCA17] Nahid Quader, Md Osman Gani, Dipankar Chaki, and Md Haider Ali. A machine learning approach to predict movie box-office success. In *Computer and Information Technology (ICCIT), 2017 20th International Conference of*, pages 1–7. IEEE, 2017.

[RLW13] Huaxia Rui, Yizao Liu, and Andrew Whinston. Whose and what chatter matters? the effect of tweets on movie sales. *Decision Support Systems*, 55(4):863–870, 2013.

[SD06] Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.*, 30(2):243–254, February 2006.

[SE96] Mohanbir S Sawhney and Jehoshua Eliashberg. A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, 15(2):113–131, 1996.

[SP17] Steve Shim and Mohammad Pourhomayoun. Predicting movie market revenue using social media data. In *Information Reuse and Integration (IRI), 2017 IEEE International Conference on*, pages 478–484. IEEE, 2017.

[SS00] Jeffrey S Simonoff and Ilana R Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3):15–24, 2000.

[TSSW10] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185, 2010.

[TYL14] Wan-Hsin Tang, Mi-Yen Yeh, and Anthony JT Lee. Information diffusion among users on facebook fan pages over time: Its impact on movie box office. In *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*, pages 340–346. IEEE, 2014.

[Vog14] Harold L Vogel. *Entertainment industry economics: A guide for financial analysis*. Cambridge University Press, 2014.

[WCZ15] Yazhe Wang, Jamie Callan, and Baihua Zheng. Should we use the sample? analyzing datasets sampled from twitter's stream api. *ACM Transactions on the Web (TWEB)*, 9(3):13, 2015.

[Wik] Wikipedia. Major film studio. `https://en.wikipedia.org/wiki/Major_film_studio`, accessed 2018-04-15.

[WSC12] Felix Ming Fai Wong, Soumya Sen, and Mung Chiang. Why watching movie tweets won't tell the whole story? In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 61–66. ACM, 2012.

A Appendices

A.1 List of Notable Libraries Used

Library	Version	Purpose in our work
numpy	1.14.2	Mostly for working with n-arrays. Used as a building block for many other libraries listed here.
scipy	1.0.0	For scientific computing, uses numpy as its basic data structure.
pandas	0.20.3	For reading in the data, transforming and visualising it.
matplotlib	2.0.2	Used for data visualization.
seaborn	0.8.1	Used for data visualization. It is based on matplotib and has support more and prettier plots.
scikit-learn	0.19.1	Used for many machine learning tasks including regression analysis, model selection and preprocessing. It is built on numpy, scipy and matplotlib libraries.
xgboost	0.7	Library that adds support for the XGBoost machine learning algorithm implementation

Table 11. Some notable Python libraries used in our empirical work.

A.2 Hyperparameter Values

A.2.1 Random Forest Default and Custom Hyperparameter Values

Hyperparameter	Default value	Custom Values	Description
max_depth	None	[None, 7, 11, 15]	Maximum depth of a tree. None means the tree will grow in depth by splitting nodes until each leaves have less number of samples than defined by min_samples_split
max_features	None	[auto, sqrt]	Maximum number of features to consider when looking for the best split.
min_samples_leaf	1	[1, 3, 5]	Minimum number of samples required for a leaf node.
min_samples_split	2	[2, 6, 10]	Minimum number of samples required to split a node.
n_estimators	10	[10, 100, 1000]	Number of trees in the forest.
bootstrap	True	[True, False]	Whether bootstrap samples are used when building trees.

Table 12. The default and custom Random Forest hyperparameter values selected for tuning.

A.2.2 XGBoost Default and Custom Hyperparameter Values

Hyperparameter	Default value	Custom Values	Description
max_depth	3	[3, 7, 11]	Maximum depth of a tree.
learning_rate	0.1	[0.01, 0.05, 0.1, 0.2]	Step size shrinkage used in update to prevent overfitting.
n_estimators	100	[10, 100, 1000]	Number of boosted trees to fit.
subsample	1	[0.5, 0.7, 1.0]	Subsample ratio of the training instance.
colsample_bytree	1	[0.5, 0.7, 1.0]	Subsample ratio of columns when constructing each tree.
gamma	0.0	[0.1, 0.2, 0.3, 0.4]	Minimum loss reduction required to make a further partition on a leaf node of the tree.
min_child_weight	1	[1,3,5]	Minimum sum of instance weight needed in a child.

Table 13. The default and custom XGBoost hyperparameter values selected for tuning

A.2.3 Hyperparameters for achieving the Best Score with Random Forest Algorithm on Hollywood Movies

Hyperparameter	Metadata	Critics	Twitter	All
n_estimators	500	100	1000	1000
min_samples_split	7	3	3	7
min_samples_leaf	3	1	5	5
max_features	AUTO	NONE	AUTO	NONE
max_depth	9	13	None	13
bootstrap	FALSE	TRUE	TRUE	TRUE

Table 14. Random Forest hyperparameters used to get the best results for Hollywood movies shown in Tables 8 and 10 (The models marked with 'RF*').

A.2.4 Hyperparameters for Achieving the Best Score with XGBoost Algorithm on Hollywood Movies

Hyperparameter	Metadata	Critics	Twitter	All
subsample	0.5	0.5	0.5	0.5
n_estimators	100	1000	1000	100
min_child_weight	3	5	5	5
max_depth	3	11	11	11
learning_rate	0.05	0.01	0.05	0.05
gamma	0.0	0.1	0.4	0.4
colsample_bytree	1.0	1.0	1.0	1.0

Table 15. XGBoost hyperparameters used to get the best results for Hollywood movies shown in Tables 8 and 10 (The models marked with 'XGB*').

A.2.5 Hyperparameters for Achieving the Best Score with Random Forest Algorithm on Bollywood Movies

Hyperparameter	Metadata	Critics	Twitter	All
n_estimators	1000	1000	1000	1000
min_samples_split	3	3	3	3
min_samples_leaf	3	1	3	3
max_features	AUTO	NONE	NONE	NONE
max_depth	9	5	13	9
bootstrap	TRUE	TRUE	TRUE	TRUE

Table 16. Random Forest hyperparameters used to get the best results for Bollywood movies shown in Tables 9 and 10 (The models marked with 'RF*').

A.2.6 Hyperparameters for Achieving the Best Score with XGBoost Algorithm on Bollywood Movies

Hyperparameter	Metadata	Critics	Twitter	All
subsample	0.5	0.5	0.5	0.5
n_estimators	1000	100	100	1000
min_child_weight	5	1	3	5
max_depth	3	11	7	11
learning_rate	0.01	0.2	0.05	0.05
gamma	0.4	0.2	0.0	0.4
colsample_bytree	0.7	0.7	1.0	0.7

Table 17. XGBoost hyperparameters used to get the best results for Bollywood movies shown in Tables 9 and 10 (The models marked with 'XGB*').

Acronyms

r^2 Coefficient of Determination. 7, 11, 12, 15, 19, 20, 40, 43, 44, 46, 47, *Glossary:* r^2

CBFC Central Board of Film Certification. 27, 34, *Glossary:* CBFC

HSX Hollywood Stock Exchange. 11, *Glossary:* HSX

IMDb Internet Movie Database. 10–12, 16, 18, *Glossary:* IMDb

MAE Mean Absolute Error. 20, 40, 42, 43, 46, *Glossary:* MAE

MAPE Mean Absolute Percentage Error. 20, 40, 43, 46, *Glossary:* MAPE

MOOC Massive Online Open Course. 38, *Glossary:* MOOC

MPAA Motion Picture Association of America. 27, 30, *Glossary:* MPAA

MSE Mean Squared Error. 12, 20, 37, 40, 42, 43, 46, *Glossary:* MSE

NLTK Natural Language Toolkit. 26, *Glossary:* NLTK

OLS Ordinary Least Squares. 13, *Glossary:* OLS

RMSE Root Mean Squared Error. 20, 40, 42, 43, 46, *Glossary:* RMSE

Glossary

r^2 Pearson correlation coefficient, which is a measure of linear correlation between two variables. 7

CBFC The Central Board of Film Certification is a government organization responsible for certifying and classifying Indian Movies. 27

HSX The Hollywood Stock Exchange is a web-based game of trading shares of virtual movie stocks. 11

IMDb The Internet Movie Database is a large online database of various movie information across the world. 10

MAE The average of all absolute errors between the predicted and expected values. 20

MAPE The mean absolute percentage error measures the size of the error in percentage terms and is calculated by taking the average of absolute percentage errors. 20

MOOC An online course, which an unlimited number of students can take via the web usually for free. 38

MPAA The Mean Absolute Error measures the average absolute difference of the errors in a set of predictions.. 27

MSE The average of all squared errors between the predicted and expected values. 12

NLTK A Python software toolkit for natural language processing tasks. 26

OLS Also known as Linear Regression. It attempts to estimate the relationship between dependent and a target variable by fitting a line to minimize the sum of squared errors between predicted and actual values.. 13

RMSE The square root of MSE to be able to represent the original units of target value. Because the errors are squared before they are averaged, RMSE gives a higher weight to large errors than MAE. 20

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Risko Ruus**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   of my thesis

   Wisdom of the Crowd Vs. Reviews of the Experts: A Case Study Regarding Predicting Movie Box-Office Results

   supervised by Rajesh Sharma

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 18.05.2018