National College of Ireland

Sentiment Analysis of Tweets to Classify the Box Office Success of Movies

MSc Research Project
Data Analytics

Kapardhi Kumar Guda
x16151054

School of Computing
National College of Ireland

Supervisors:
Dr. Pramod Pathak
Dr. Dympna O Sullivan
Dr. Paul Stynes

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Kapardhi Kumar Guda
Student ID:	x16151054
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr.Pramod Pathak,Dr.Dympna O Sullivan,Dr.Paul Stynes
Submission Due Date:	11/12/2017
Project Title:	Sentiment Analysis of Tweets to Classify the Box Office Success of Movies
Word Count:	7547

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Kapardhi Kumar Guda
Date:	13th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis of Tweets to Classify the Box Office Success of Movies

Kapardhi Kumar Guda
x16151054
MSc Research Project in Data Analytics

13th August 2018

Abstract

The movie making industry is one of the top money-making machines one can say based on the revenues being generated every year from movies, a whopping 11 billion dollars were total box-office collections of Hollywood movies for the year 2017. The main aim objective of the project is to able to predict the movies success whether its going to be a “flop”,“average”, “hit” or “superhit” which would be beneficial for the movie distributors and production companies. Tweets related to multiple movies were extracted, sentiments and emotions were extracted for these tweets using R language, and a score is given for each emotion and sentiments, before running the models. All precautions were taken to hide the identifiable information like Tweet Id and Screen name. Machine learning models used in the project were SVM, KNN, Naive Bayes and Random Forrest. The models were run in Python language, they were able to predict the class of the movies whether its a “flop”,“average”, “hit” or “superhit” with an accuracy of 60.76%, based on the scores of emotions and sentiments, so this would be extremely useful for the entertainment industry for boosting their revenues. After comparison to other models performance, SVM and KNN did a good job in predicting the outcome of the movies. After analysing the results, it was understood that the movies with higher budget need larger revenues for their success because of the additional expenditures these movies have like Print and Advertising.

Keywords: *Box office, sentiments, emotions, machine learning models, tweets*

1 Introduction

Six hundred movies are being produced every year on average according to Motion Picture Association of America Dodd (2016). This doesnt seem to be decreasing, as the years pass on this number is steadily increasing. These are just the statistics from Hollywood alone, when we consider other countries like China and India the stats make us mind boggling, and movie industry is one of those which never ends, new movies keep coming every week from all parts of the world. But, no one never actually knows how well a movie performs, it depends on multiple factors like cast, director, number of screens its being released, expectations from the fans and release season. Thanks to machine learning we have multiple algorithms that can predict the fate of movies, with some decent accuracy.

Recently there were several researches owing to predict the success of movies, but theres no solid method to do so and lot of research has yet to be done in this field, which will benefit the people who are investing millions on a single movie.

One can never guarantee a movies success, there are several movies which had some star cast and good directors, but the movies turned out to be a disaster. One such example is the recently released Star Wars Solo movie,Rubin and Rubin (2018) which had a budget of around 275 million for production this doesnt include Print and Advertising which would be around 150 million additional expense to the existing budget, and the movie earned just 390 million in collections worldwide, that made huge dent for the production house and distributors. But, if one can predict this out come on before hand they can save millions. Previous researchers used Linear Regression, Logistic Regression, SVM, ANN, Naive bayes methods and few other methods were also used for predicting the success at box-office, but much needs to be done because the accuracy for most of the models were just around 45% (all the reference for models are provided in literature review section) this might be due to several reasons like the method of analysis that were used. Most of the previous research works concentrated on the reviews of critics for movies, IMDB data, Wikipedia movie data,"electronic Word of Mouth", Yahoo movies data, only few researchers used data from twitter but they analysed it in different way by counting the followers of user, and deciding the level of impact these people have revenue of twitter, which is actually good because influential people usually will have more impact on the general users, if a celebrity tweets about a movie saying its really good his or her followers might end up going to movies, but this doesnt happen always.

So, theres a gap in the existing research as previous research work had been concentrating more on the reviews of the users and critics, than the actual content thats being generated on social media like Twitter and Facebook which contains lots of information. If, the data from these platforms can be analysed in a better way the results would be more meaningful and accuracy can be increased.

1.1 Hypothesis

Sentiment analysis of tweets combined with machine learning algorithms will help in classification of movies at box-office much better.

The proposed model falls in to quantitative research, as it tries to solve the problem of predicting the box office success of movies, in this study the data from twitter was extensively utilised, tweets related to multiple movies were extracted and these were used to extract the emotions and sentiments from them, this step helped in giving better results in terms of accuracy. These emotions were given a score, emotions used for the research were anger, anticipation, disgust, joy, fear, sadness, surprise and trust. Score was given for each tweet considering whether its positive or negative tweet. Four models were tried and tested for performance KNN and SVM gave the best results among them, performing better than another researchers accuracy when similar models were run by Subramaniyaswamy et al. (2017) and Vr and Babu Pb (2014). So, from this study we can say sentiment analysis on the tweets would fill in the existing gap present in this area up to an extent. This paper is the first to use sentiment analysis of tweets in predicting box office success, which remained a challenge from many years.

1.2 Research Question

Can sentiment analysis help in better prediction of box-office success categories using machine learning models?

The rest of the paper has been divided in to following sections Literature review, Methodology, Implementation, evaluation and results, conclusion and future work.

2 Related Work

Prediction is a method of forecasting something, this can be categorised in to two ways Quantitative and Qualitative. Quantitative approach has numerical data in it which might be historical data used for analysis, where in Qualitative approach is a kind of observation. Analysing the market situation could be an example for Qualitative research. Most of the related works fall under Quantitative research.

Of all the related works, most of the papers used regression as their primary technic, and few papers used SVM and Neural Networks for classification of movies.

2.1 Classification using (SVM, Regression techniques)

The first paper thats most relevant to this project is Galvao and Henriques (2018) the aim of this project was to estimate a movies profit, using regression, neural nets and decision trees, they achieved an accuracy of 56.1%. To discuss the methodology used for this model, they used SEMMA approach developed by SAS. Opus data along with IMDB data was used for analysis, various variables used in the project were, actor, director, budget, nominations, genre, Oscars, reviews and others. Dependent variable was profit and all others were independent. The good thing in this model is the use of correlation technique while choosing the dependent variables which helps in reduction of unnecessary dimensions. For, the predictions of profit there were 9 different categories which is based on Sharda and Delen (2006). The categories are divided based on their revenues at the box-office. After the analysis, the conclusion was that actor, director and whether the movie is a sequel or not would impact the performance of movies which is obvious.

Subramaniyaswamy et al. (2017) In this paper the researchers studied the influence of trailer views, Wikipedia page views, critics ratings and time of release. All these were independent variables in the study. Time of release will always have huge impact on the collections, a movie released during holiday break will have more ticket purchases than the movie released during non-holiday release. Especially, in countries like India collections tends to increase by a huge margin during holidays, this is the main reason producers look to release movies during holiday weekends. The data was obtained from box-office mojo and Wikipedia movies data. Data was classified based on the budgets in to three categories, and success for these movies was calculated by multiplying the budget twice to be considered a hit for low budget films, 2.5 times for medium budget films, and 3 times for high budget films. After running SVM they were able to classify 56% of movies correctly whether they become a hit or flop based on the profits of the movie. From the results of this study it was evident that budget has huge impact on the success of movies which might be not accurate, because therere proven examples that low budget movies earning huge collections. The data set for this model was very small had only 150 rows.

Nikhil Apte (2011)IMDB data was used for study, it included 2400 movies details. Linear regression was the model used for study, the good thing about this project was they segregated movies based on genre and run the regression model based on intuition that different genres perform differently in different theatres. And they achieved better accuracy i.e., 20% more compared to other models because the data was very small for few genre movies. This means if they had 100 movies in fiction genre, and 250 movies in drama genre, the model when run on drama genre had better accuracy i.e., 20% more compared to fiction genre model.

Vr and Babu Pb (2014) The model uses SVM and regression, but the data completely scrapped from IMDB, the results were low the accuracy was just 42.2% for regression model, the data set was very small had only 1000 rows and they even used ratings for the movies rotten tomatoes. They checked for correlation and removed the unwanted attributes. Linear regression, logistic regression and SVM models were used to predict the success and the model finally tells that among the independent variables actor, director, budget, and gender are significant in the data set.

2.2 Neural Network models

Sharda and Delen (2006) this was the first paper to introduce neural networks to predict the box office success. And they achieved an accuracy of 39%("bingo"), they termed the accuracy as bingo. The multiple layer perceptron neural network was used in this study, the dependent variable is the box office gross revenue in this study. And this was divided in to 9 classes when the box office gross is exceeding 200 million more than budget itll be coded as a block buster. Seven independent variables were used in the project, MPAA rating, competition, star value, genre, special effects, number of screens, whether its a sequel or not. They choose these variables based on the suggestions from industry expert. The issue with this research was theyre trying to predict the category of movie success in to one of the nine categories which might have given very less accuracy, the positive result from the research is the correlation between number of screens and movies has a definitive impact on the box office. The size of the data set was increased and implemented the model again and got the accuracy as 52.6% in the year 2010 by the same publishers. But this, research was done 8 years ago, might be able to get better accuracy with more adjustments to the model.

Ghiassi et al.; 2015 The paper is perhaps the most successful model with an accuracy of 94%, this was published in the year 2015. The same variables were used like the previous model that was discussed as they were trying to increase the accuracy of the model with additional variables like MPAA rating, sequel, screen count, production budget, pre-release advertising expenditures and seasonality were added in this model. Pre-advertising budget was very decisive move, because basically its the extra expenditure incurred but thats not shown in the budget. Higher budgets always result in higher revenues, but this doesnt indicate profits Basuroy et al. (2003) due to the additional expenditures this was the case for Star Wars Solo movie. All the additional variables in this model were had good correlation with success of movies, seasonality factor Einav (2007) had always played a role in Hollywood industry. He suggested the weekends during which the movies will have more success. Dynamic architecture for artificial neural network (DAN2) is based on forwarding the learned knowledge to next layers. This method was previously employed by other authors for predicting and forecasting, They, used this model to replicate the previous method discussed above and increased the accuracy to 74%. They trained the data on 80% and tested it on 20% model. And finally, they implemented DAN2 to forecast the pre-production of movie revenues, "Competition, star value, genre, and special effects have been removed as variables, while production budget, pre-release advertising expenditures, runtime, and seasonality have been added" Ghiassi et al. (2015) they run the model with new movies data set, as previous model had older data set. Pre-advertising data for 354 movies were obtained which was really good, but for the above-mentioned model and this model data was very less, we can actually get acceptable results with more data and again getting the data for pre-released movies advertising costs is not easy as it is not made publicly available. Finally, they got the results with an accuracy of 94% which is good.

Vitelli (2007) This model uses 2-layer neural networks for prediction. The model used IMDB and OMDB data and final data set consisted data related to 3600 movies, the F1 scores were calculated with features being increased and finally the model achieved an accuracy of 61.28%. All the first layered models performed well than the second layered models. These were various proposed models that used neural networks to predict the box office success.

2.3 Electronic Word of Mouth and Customer Engagement behaviour model

Electronic word of mouth(eWOM) , is the way of spreading opinion either positive or negative review about anything like an object or place through the internet Hennig-Thurau et al. (2004). These kinds of informations spreads rapidly on the social media platforms but the authenticity is always an issue in this.

Baek et al. (2017) The model had three hypotheses, and the data was collected from twitter, blog, yahoo movies and YouTube. "Twitter has stronger impact on box office revenue in the initial stage than later stage. Yahoo movies and blogs have stronger impact in the later stage than initial stage. Impact of YouTube on box office revenue has equal effect in initial or later stages" Twitter has stronger impact in the initial stage due to the option of retweet which makes the spreading of data very easy. Yahoo movies data consists of reviews, so it takes time for the reviews to sink in to the public because a movie with 1000 reviews will have more impact than a movie with one review, this makes yahoo movies more effective in the later stages. All these hypotheses were effectively proved to be crucial in success of a movie.

Oh et al. (2017) This paper studied the effects of "Customer engagement behaviour" in social media. The results were Facebook and YouTube correlations are positively impacting the revenue at box-office. The model has two parts in it personal and interactive. And these will lead to economic performance of a movie. Data was collected from Facebook, YouTube and twitter. Independent variables Fb likes, fb talk (stories related to movies on their wall), this is a good feature for consideration. And there are several other variables related to YouTube also. After the data was ready they run the regression model and model was mainly concentrating to find the relations between independent and dependent variables. They had a controlled variable release type, which is basically the type of released movies, whether wide or limited release and about 55% of the releases were found to be wide release. Another controlled variable was genre type, there were 9 types in them most of the movies were belonging to drama genre.

2.4 Independent Sub Space Method

Hur et al. (2016) This paper proposed a different model by analysing the reviews for sentiments, the data was obtained movies released in Korea, several different variables were used in this study compared to other studies like number of audience that watched the movie in first week, picture factors like popularity of directors and actors, power of distributors and sales power, whether movie has been from us or korea, rating, screen ratio and seasonality, finally sentiments from the reviews. In this model, all the algorithms are trained and tested together to find the right algorithm. They did 2 experiments first one to find the important variables that are required for analysis, and second experiment to forecast the revenues. They trained 90% of the dataset, after running the models they found that machine learning models can predict the more accurately when the data is insufficient. The main setback in this model is the movies data primarily consists of Korean movies which is biased.

2.5 Other Models

Kennedy (2008) The data set had 220 movies released in the year 2007 and movies with at least 10 reviews were only selected for analysis, the reviews scores were again converted on a scale of 100, 5 rated movies would get 100 on the scale. Firstly, user and critic scores were compared, and the conclusion was theres no relation between critics and users rating. And there was no correlation between Metacritic score and box office revenue of the movies, but after the first weekend of the release theres a good correlation between critic score and box office revenue which indicates that public tend to read reviews and its impacting the revenues. And a linear model was run to study the impact of critic reviews on limited and wide releases. Results say that production and marketing are critical for a movies success.

Baek et al.; 2014 They studied the effects of positive tweets and negative tweets on box office revenue, based on effects of tweets on product sales. They used expectation confirmation Theory for this approach. In this model basically, the level of satisfaction is studied on basis of expectation towards a product and its performance and the difference leads to disconfirmation. Data was collected from twitter related to movies that were released at the time analysis, tweets are classified in to positive, negative and neutral. The irrelevant tweets are not considered which is a good way of eliminating non-related tweets, after the analysis they found box office revenue is directly proportional to number of tweets. And expectancy confirmation theory didnt have any impact on box office revenue.

Yoo et al. (2012) In this project data for 4052 movies were extracted from IMDB since 1931 to 2011, later they choose only english movies that were released in United States. Linear and logistic regression were run to classify the movies. Three set of variables were considered theyre sentiment features, simple and complex features, they concluded that these features were insufficient to predict the box office revenues.

Ericson and Grodman (2011), Mestyn et al. (2013), Cocuzzo and Wu (2013) all these papers proposed similar models for prediction of box office revenue or to categorise movies whether they are hit or flop. Ratings and gross revenues were predicted using models like Naive bayes, SVM, Regression models. All these papers results were similar to each other.

After analysing all the related works, we can say all the papers were trying to concentrate more on the reviews and data obtained from IMDB, critic scores and few models used twitter data but analyse whether their volume had any impact on the revenue or not. No paper concentrated on sentiments of the tweets, but sentiment analysis was done on the reviews to study the impact combining them with other variables and tried to predict the gross of the movies.

3 Methodology

The aim of the project is to predict the box office success of the movie, whether the movie is going to be a “flop”,“average”, “hit” or “superhit” which comes under classification. Sentiment analysis of tweets to extract emotions and sentiments from each tweet to predict the box office success of the movies was never done prior this. This is the first paper to utilise this technique for predicting and classifying the movies by using sentiment analysis.

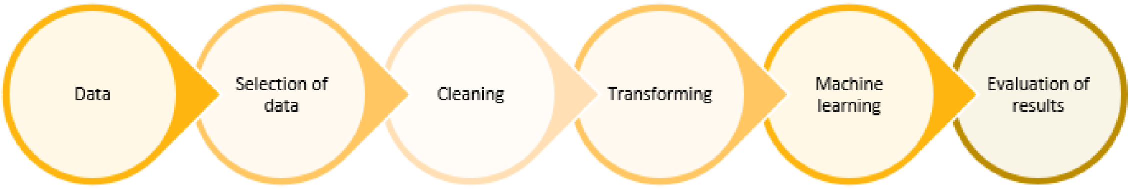


Figure 1: Workflow for the proposed model of classifying movies

Movies were the chosen domain, a good amount of research was done in this domain trying to predict the success of movies but the accuracys were less. Researches tried to concentrate more on reviews and finding the correlation between the variables that impact the box office revenues.

Data related to movies can be extracted from multiple sources, like Twitter, Facebook, YouTube, blogs, IMDB, Wikipedia and other sources. But, for this research twitter was used as the source for data because this wasnt much explored earlier for movie analysis.

The data was extracted using R language with the help of Twitter API, and the cleaning of data like removing unwanted characters, and removing identifiable information was also done in R, when data was extracted several variables were imported which were not required like screen name, screen id, followers and others.

Sentiment analysis was done on the cleaned data using syuzhet package, all the emotions and sentiments were extracted for each tweet. Budget and Gross revenue for each movie were extracted from the Wikipedia. Profit was calculated for the movies by subtraction Gross revenue from budget. Then all the movies were grouped in to four categories based on their profits.

Data was then transformed in to required format, for the analysis I dont need the tweets anymore so dropped the column of tweets but kept the rest columns for further analysis.

Machine learning models were run on the transformed data, SVM, Random forrest, Naive Bayes and Knn were used for machine learning part. This part was done in Python. Results were then evaluated by checking the accuracy and pattern from each model, calculation of accuracy is explained below

To evaluate the results for the proposed model Accuracy was used, accuracy is calculated by using precited class and actual class. There are totally four different classification classes there are False positive, True Positive, True Negative, False Negative. By using all the classes accuracy will be obtained for the model, accuracy is given by

$$Accuracy = \frac{TruePositive + TrueNegative}{FalseNegative + TruePositive + FalsePositive + TrueNegative} \tag{1}$$

The model calculates all the values when its run, and it gives final accuracy for all the models.

The above explained workflow would fall under KDD (Knowledge discovery and data mining) Fayyad (1996)

4 Implementation

4.1 Platform check for Data extraction

The first thing that needs to be decided before starting the project is to decide the platform from which data can be extracted. The domain being movies data can be from various platforms like Facebook, Twitter, YouTube, Blogs, IMDB, Wikipedia, box office mojo, rotten tomatoes and others. The initial plan was to extract the data from Facebook, but due to the new GDPR restrictions data extraction from Facebook has become impossible. YouTube data is not much useful because it has only trailers related to movies.

Twitter was the best to extract the opinions of public related to movies, as everyone tweets these days, and Twitter has 335 million users according to Statista by the end of 2018 second quarter cycles and Text (n.d.). As, twitter is the platform where one can tweet about anything, and its easier to retweet. So, people tend to tweet about what they feel, or retweet about something they like with out any restrictions. And word limit has been increased to 280 characters.

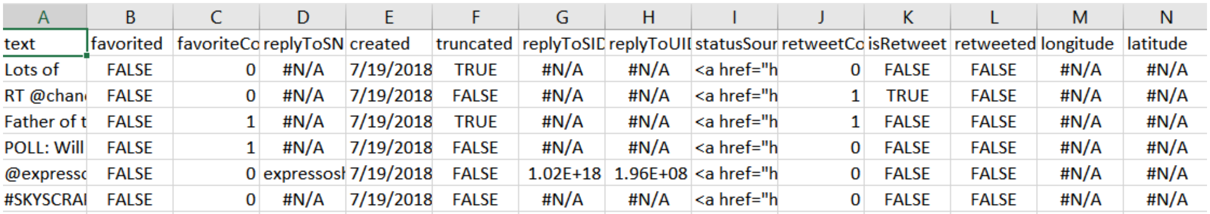
Wikipedia data related to movies was also extracted for budget and box office revenues as this is the best open source from where data can be extracted. And the revenue is given for Worldwide collections.

4.2 Data Extraction

Once the platform is decided for data extraction, had to start thinking about the software that needs to be used for data extraction. So, choose R language for this task.

Created connection with Twitter and R, by using Twitter API to extract data. So, once the connection is made with Twitter and R, could start extracting the data. Initially, planned to extract the data related to sales of a product (Iphonex) but it was not possible as the Iphonex released around November 2017, and twitter API allows to access the tweets only until recent 7 days, and there is twitter RestAPI, which will allow to get the data for past 30 days, but again we can extract only 3200 tweets from the past thirty days. So, had to think about alternatives and tried to purchase the data but all the data related to Iphonex sales has been erased from twitter and they do not have an official page on Twitter. And more over the data purchase plans were too expensive.

So, decided to extract the data related to movies, as movies get released every Friday and there is abundant of data that can be extracted. The catch here is rate limit on the extraction of tweets, if one tries to extract too many tweets with no time gap, the account gets blocked once the rate limit is exceeded. Data extraction took quite a lot of time could extract around 20,000 tweets a day, around 90,000 tweets were extracted for analysis related to fifteen different movies. Below is the screen capture of the extracted data using twitter API with all the variables(Identifiable information has been removed).



A	B	C	D	E	F	G	H	I	J	K	L	M	N
text	favorited	favoriteCo	replyToSN	created	truncated	replyToSID	replyToUII	statusSour	retweetCo	isRetweet	retweeted	longitude	latitude
Lots of	FALSE	0	#N/A	7/19/2018	TRUE	#N/A	#N/A	<a href="h	0	FALSE	FALSE	#N/A	#N/A
RT @chan	FALSE	0	#N/A	7/19/2018	FALSE	#N/A	#N/A	<a href="h	1	TRUE	FALSE	#N/A	#N/A
Father of t	FALSE	1	#N/A	7/19/2018	TRUE	#N/A	#N/A	<a href="h	1	FALSE	FALSE	#N/A	#N/A
POLL: Will	FALSE	1	#N/A	7/19/2018	FALSE	#N/A	#N/A	<a href="h	0	FALSE	FALSE	#N/A	#N/A
@expressc	FALSE	0	expressosl	7/19/2018	FALSE	1.02E+18	1.96E+08	<a href="h	0	FALSE	FALSE	#N/A	#N/A
#SKYSCRAI	FALSE	0	#N/A	7/19/2018	FALSE	#N/A	#N/A	<a href="h	0	FALSE	FALSE	#N/A	#N/A

Figure 2: Twitter extracted data

In twitter there are two types of location settings, one being the tweet location and other being the home location from where the account was created, most of the users tweets locations give NA due to the default settings very few enable the location. But the home location must be given at the time of account creation so initially tried to extract the home location and succeeded with this but few of the location names were just random like Hogwarts School so these does not make any sense. But, the rest other locations were correct, extracted these details with the help of screen name. The reason for doing this was to split the box office revenue region wise and compare it with the tweets to analyse whether tweets from a region make any difference on the revenues. But, this did not work because all the locations should be converted to their respective latitude and longitude co-ordinates to plot them on a map which has highly impossible as there were hundreds of individual locations in the data set. For example Newyork city, was given as NY, (NY,Usa), Newyork and so on, so recoginizing these cities individually is not possible.

4.3 Dataset explanation

Text is the tweet content, usually when the tweets are extracted we get only 140 characters, so had select the option Extended = True, to get all the 280 characters without missing the data. Latitude and Longitude were mostly NA, as the users dont enable the location while tweeting. And two other variables were removed due to identifiable information, screen name and Id from which the tweets are coming.

The data that was extracted from Wikipedia which contains movies name, budget and its profits were combined to the existing data set.

And a new row named profits was given for all the tweets, by subtracting gross revenue of the movie from budget.

So, these were all the variable that were in the basic data set, before cleaning or transforming.

4.4 Cleaning dataset

When the data imported to a csv file with function write.csv lot of unknown characters, special characters and unknown numbers reflect in the csv file. For the sentiment analysis of the text content all these had to be cleaned before running the sentiment analysis, had to remove all the stop words, unnecessary spaces, screen names while retweeting, all the emojis, all the abbreviations.

Gsub function could not be used because there were too many unknown characters, so all the characters which are non-alpha numeric were removed with a single function str_replace_all.

A loop function was written for this to check each tweet and clean everything at a single go. Rest of the variables which were not needed for sentiment analysis were removed as explained above.

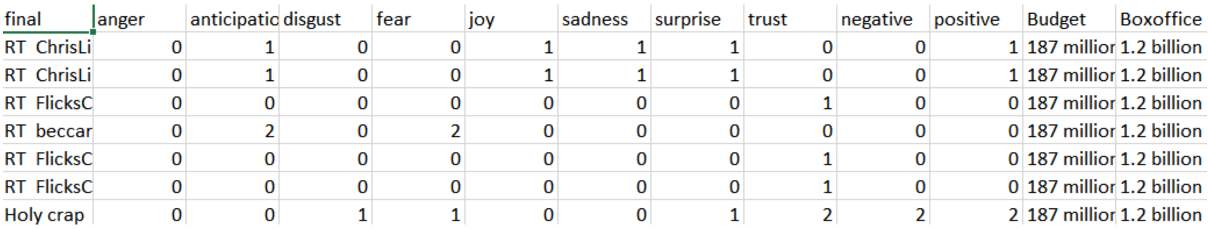
4.5 Sentiment analysis with “Syuzhet” package

The cleaned data is ready for performing the sentiment analysis, installed all the necessary packages, other packages tm and snowballc were used for text mining of the data.

The best thing about Syuzhet package is it breaks the tweets based on the emotions and sentiments and gives each emotion a score (anger, anticipation, distrust, joy, sadness, fear, surprise,trust) positive and negative sentiments. Sagar (2018)

Initially planned to utilise IBM Watson natural language understanding for the same analysis but the free version of the package allows to analyse only 30,000 tweets but the data set used was 3 times what has being analysed so started to look for other packages and found this excellent package “Syuzhet”.

Below is the image of a sentiment analysed tweets with their respective scores, we can see scores being distributed based on the tweet content, and movies budget and box office can also be seen next to emotions columns.



final	anger	anticipatic	disgust	fear	joy	sadness	surprise	trust	negative	positive	Budget	Boxoffice
RT ChrisLi	0	1	0	0	1	1	1	0	0	1	187 millior	1.2 billion
RT ChrisLi	0	1	0	0	1	1	1	0	0	1	187 millior	1.2 billion
RT FlicksC	0	0	0	0	0	0	0	1	0	0	187 millior	1.2 billion
RT beccar	0	2	0	2	0	0	0	0	0	0	187 millior	1.2 billion
RT FlicksC	0	0	0	0	0	0	0	1	0	0	187 millior	1.2 billion
RT FlicksC	0	0	0	0	0	0	0	1	0	0	187 millior	1.2 billion
Holy crap	0	0	1	1	0	0	1	2	2	2	187 millior	1.2 billion

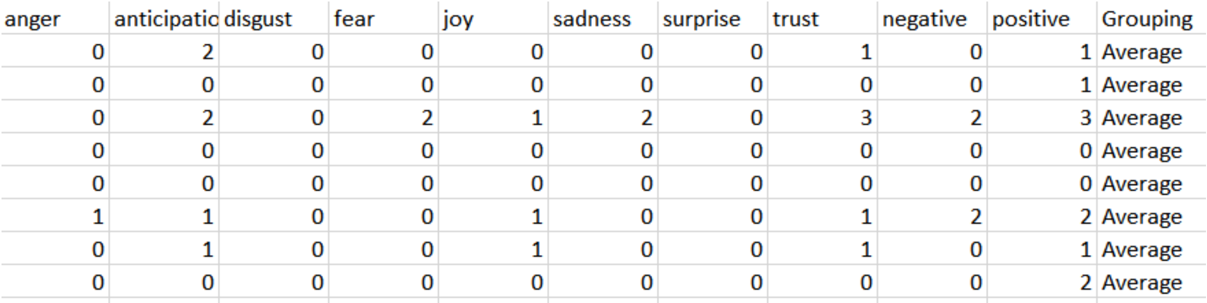
Figure 3: Sample with emotions and sentiments score

All the tweets were analysed and given a score, but there were few tweets which did not make any sense to the data like a person posted a tweet about some random thing in between these movie tweets so had to clear all such tweets as this would create unnecessary noise in the data.

4.6 Transformation

There was no much transformation needed for the data set, but for the next step to run the machine learning models, had to group the movies based on their profits whether the movies were hit, superhit, average or flop.

Grouping of the movies was done using if else function, movies were grouped in to four different categories based on the revenues, movie will be grouped in to flop if the revenues does not exceed its budget plus an extra 30% percent is not grossed, suppose if movies budget is 100 million and it earns 125 million or 130 million at box office it will be treated as flop. If the movie grossed above 150 million it would fall in to the category of average movie, and if it grossed above 180 million it would be treated as hit movie, if crossed above 200 million it would be treated as super hit.



anger	anticipatic	disgust	fear	joy	sadness	surprise	trust	negative	positive	Grouping
0	2	0	0	0	0	0	1	0	1	Average
0	0	0	0	0	0	0	0	0	1	Average
0	2	0	2	1	2	0	3	2	3	Average
0	0	0	0	0	0	0	0	0	0	Average
0	0	0	0	0	0	0	0	0	0	Average
1	1	0	0	1	0	0	1	2	2	Average
0	1	0	0	1	0	0	1	0	1	Average
0	0	0	0	0	0	0	0	0	2	Average

Figure 4: Transformed data set after grouping

The difference is due to the advertising and printing cost of the movies which would not be included in the budget, and the ranges were considered based on other previous papers indication. An additional 10 million was added for the additional costs that incur for movie production. The ranges can be changed if the model is implemented and if one knows the exact costs of the additional expenses to classify in which range it is going to be.

The final data set before running the machine learning model consists of all the emotions, sentiments and grouped output, the model does not require tweet text or name of the movie, budget and profit.

The final data set which cleaned and transformed is shown above in figure 4.

4.7 Machine learning models

The term machine learning was initially coined by Arthur Samuel and its history goes back to 1959. Its a process of teaching machine(computers) to learn about data with out programming explicitly by Samuel (1959). The machine learning models are again split in to two types theyre supervised machine learning models and unsupervised models. Classification and regression tasks come under supervised models. When theres no labelled data it comes under unsupervised models. The projects aim is to predict the class of the movies box office revenue whether its going to be a hit or flop. These models come under classification, so the models thatll be used fall under supervised machine learning algorithms.

Four different models were used for analysing the data they are Naive Bayes, SVM, KNN and Randomforest. All these models were run in Python because of the flexibility compared to usage of R language. All the required packages like panda and numpy were imported. The path of csv file was given to read the data.

4.7.1 Naive Bayes Classifier

For this model to run, the data was first split in to train and test at 80/20 ratio. Gaussian naive bayes was implemented from sklearn in this the likely hood of the features are assumed to be gaussian.

After running the model, it was able to correctly classify for 59% of the instances. Which means the accuracy for the model is 59%, this is a decent accuracy compared to recent papers which tried to predict the box office success as mentioned in the literature review and evaluation part.

4.7.2 Support Vector Machines

This model can be used both for regression and classification purposes. This was the second model that was used to predict the class, SVM can handle even non-linear data, the model tries to achieve maximum distance between the classes. But for the analysis as this is a classification problem in Python used SVC which is just a different type of implementation for the same algorithm. "The SVM module (SVC, NuSVC, etc) is a wrapper around the libsvm library and supports different kernels while linear SVC is based on liblinear and only supports a linear kernel" from stackover flow.

The kernel which was used in the model was linear kernel which is faster and scale the data a lot better, different kernel functions can be specified and custom kernels can also be used. For the case in project used linear kernel which separates the hyperplane with the maximum possible distance. And linear kernels are generally used when the number of features is greater than number of observations.

After implementing the model, the accuracy was found to be 60.39 which was slightly better than Naive bayes model. The data was split with same ratio as for the above model.

4.8 KNN(k Nearest Neighbours)

The third model used for predicting the class is KNN, even this model can be used for both classification and regression, K was set as 3 for the study purpose, Knn model is simple and effective at the same time.

After running this model, the accuracy was 60.76 which was slight improvement over the above two models. This was the highest accuracy achieved from all the four models that were used for machine learning part. This model was able to predict 60.76% of the classes accurately from the given data.

4.9 Random Forest

Random forest is an ensemble classifier which is a combination multiple decision tree. Even this model can be used for both classification and regression, but for the study purpose it was used for classification. Two thirds of the data are selected to train the data set, and rest is used to estimate the error. The advantages of these model are there is no need to prune the data, overfitting is not an issue, parameters can be set easily in the model.

Before the model was up and running small changes were required the classes had to be changed to categorical variables, to run the model so the grouping column was changed in to categorical variables. Training and testing sizes were set, decision trees were set as 1000 but unfortunately the accuracy was just 22% which was significantly very low compared to other models.

So, of the all above models that the data was tested and trained KNN gave highest accuracy and next SVM (svc) gave almost same accuracy. Considering the recent outputs from other models our models performed quite decently.

5 Evaluation

5.1 Accuracy results from machine learning models that were used in this paper

Table 1: Accuracy for the proposed models

Models	Naivebayes	SVM(svc)	KNN	Randomforest
Accuracy	59.07%	60.39%	60.76%	22.97%

These were the results from all the models that were used to predict the class of box-office and got some good results compared to other papers which tried to predict the classes of movies at box-office.

5.2 Comparing results with recently developed approaches

Table 2: Accuracy results from other recently proposed models

Models	Accuracy	Authors
SVM	56.25%	Subramaniyaswamy et al. (2017)
SVM	39%	Vr and Babu Pb (2014)
ANN	57%	Vitelli (2007)
ANN	56.1%	Galvao and Henriques (2018)
DAN2	94.1%	Ghiassi et al. (2015)
Logistic regression	49%	Yoo et al. (2012)

The below table gives the results of accuracy from various papers published recently, comparing the results with the proposed model using sentiment analysis to predict the class of box office gave the result 60.76% which is better improvement when compared to the previous models. In the table only DAN2 has 94.1%, all other models were below 57.

The main reason for DAN2 performing better than other models is that the data set which was used had contained a variable which has information about the pre-advertising budget of about 354 movies. As, advertising costs act as crucial decider for a movie success, when this variable was used it gave much better accuracy compared to all other models. When the advertisement costs and actual budget are added up, the profits of the movie might change, for example *Insane Studio Accounting: Warner Bros Claims $167 Million Loss Over Harry Potter and the Order of the Phoenix* (2010) Warner Brothers studios reported a loss of around $167 million when the movie Harry potter and the order of phoenix ended its screening in theatres, even after earning $938 million at the box office worldwide and standing as one of the years top grossers. This was due to the distribution, advertising and print costs which were excluded from the movies budget when these combined the results vary with a huge difference. This indicates the importance of that single variable pre-advertising budget. Refer the article here

If this is excluded as the model which was proposed in this paper doesnt contain that information because its not available openly on any websites , it can be said easily that using sentiment analysis, gives better results compared to other models of using critics reviews.

Lets recall the hypothesis Sentiment analysis of tweets combined with machine learning algorithms predicts success of movies at box-office much better, so from the accuracy we can say that our hypothesis is true, sentiment analysis of all the tweets has a better impact on predicting the revenue class of movies. So, when the model used in real time with better conditions and increase in the size of data set and if the historical data can be accessed for analysis, the model can be expected to perform better and reach the levels of above 70%.

6 Conclusion

For a movie to perform well at the box office, a lot of time and millions must be spent on a single movie, but the return investment for most of the movies would be disappointing. The ratio of the well-received movies would be very less. Most of the previous papers have proposed models that decide the fate of the movies after releasing, so for these models to able to predict the class, data should be available which might take a week to gather everything and then run the model to predict but in this time the might have been declared a disaster already due to the reviews and word of mouth talk.

The proposed model predicts the class (i.e.,“flop”,“average”, “hit” or “superhit”) based on the tweets collected, which can be for a movie which has been released or which has not yet been released unlike previous research. A movie thats going to be released this week, will have tweets with the hashtag or tweets in their official accounts, so once these tweets are extracted sentiment analysis will be done on them, and these sentiments can be used for predicting the class based on their budget. So, this model is not only for movies which are already released but can be used on the movies that are going to be released this is the main difference in the proposed model compared to the other works.

Tweet content was the major variable used in the model, along with movies budget, box office collections, and grouping variable of the movies. So, it can be clearly understood that sentiment analysis combination to predict the class of the movies performs well. And the accuracy has been increased to 60.76% using KNN, and 60.39% using SVM the benchmark for predicting the success of a movie was 39% by Vr and Babu Pb (2014). The proposed model will help the producers and distributors with a competitive advantage being a novel method. If more details regarding the pre-production costs are available, the model might do its best.

6.1 Scope

The data consisted of 83,768 rows in total and it contained tweets regarding 15 different movies, including retweets. The overall scope was to able to predict the class of the movie, depending on sentiment analysis of tweets. There was an issue while trying to extract the location of the tweets because most the location would be disabled by default and very few people enable it, the alternate to this was to extract the home location of the user which has to be given mandatorily while setting up the twitter account. A plan is in place regarding the location identification and mapping them to the revenues of box-office which has been explained in the future work.

6.2 Generalization

The proposed model, can be generalized and used for several purposes like predicting the success class of a novel but the difference would be in predicting number of copies sold a range can be predicted, a music album sales for this model previous album sales of the composer would be an excellent variable, if a new clothing line or a newer version of iPhone model is being introduced in the market the sales can be predicted. In this way the proposed will work efficiently in multiple domains involving sales.

7 Future work

For further research the model can be expanded to predict the ticket sales region wise based on collecting the number of tweets and positioning them on a map, with this the box office revenues can be separated from region wise. This can be done, if the Home location of the users can be extracted, as the tweet location for most of the users is disables, the only other way is to extract the location of Home user, which was actually done in beginning of the project, but the issue with this would be extracting the coordinates for the location of home locations, and then they can be separated, and the box office collections for region wise data can be extracted and these both columns can be compared to find if the tweets and revenues have impact on region wise collections or not.

Another experiment can also be conducted, the budget of the movies that was considered was excluded with details like printing and advertising costs which was additional expense incurred by the producers, if such details can be found, the grouping of the movies can be done exactly depending on the combined budget and class of the box office performance can be predicted based on the tweets, and can get more accurate results there by increasing the accuracy of the model.

The only limitation of the model is regarding the data that is available in the Wikipedia, if the data related to movies budget and actual revenues can be obtained from the distributors and producers along with the tweets for a longer period for the analysis (i.e., greater than 7 days), the model can perform much better. This would improve performance.

7.1 Feature selection

For a movie to become successful the sentiment of the tweets should be positive, and emotions can be anything sad, happiness, frustrated and others. This means for example after watching the movie Titanic every one were sad because of the heros demise but the overall review of the movie was positive even though the emotions in the movie contained sadness in it. So, a might perform exceptionally well at the box-office no matter if the ending is sad or happy. Few movies after performing well at the box office might incur losses due to high advertising costs, few might become flops due to poor box office performance. Overall the emotions and sentiments of a movie are most important for successful classification of a movie.

References

Baek, H., Ahn, J. and Oh, S. (2014). Impact of Tweets on Box Office Revenue: Focusing on When Tweets are Written, *ETRI Journal* **36**(4): 581–590.
URL: *http://doi.wiley.com/10.4218/etrij.14.0113.0732*

Baek, H., Oh, S., Yang, H.-D. and Ahn, J. (2017). Electronic word-of-mouth, box office revenue and social media, *Electronic Commerce Research and Applications* **22**: 13–23.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S1567422317300054*

Basuroy, S., Chatterjee, S. and Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets, *Journal of Marketing* **67**(4): 103–117.
URL: *http://www.jstor.org/stable/30040552*

Cocuzzo, D. and Wu, S. (2013). Hit or Flop: Box Oce Prediction for Feature Films, *MACHINE LEARNING* p. 5.

cycles, T. t. p. g. i. S. a. n. l. f. t. i. g. b. c. o. c. D. t. v. u. and Text, S. C. D. M. u.-t.-D. D. T. R. i. t. (n.d.). Topic: Twitter.
URL: *https://www.statista.com/topics/737/twitter/*

Dodd, C. J. (2016). Theatrical Market Statistics, p. 31.

Einav, L. (2007). Seasonality in the U.S. motion picture industry, *The RAND Journal of Economics* **38**(1): 127–145.
URL: *http://doi.wiley.com/10.1111/j.1756-2171.2007.tb00048.x*

Ericson, J. and Grodman, J. (2011). A Predictor for Movie Success, p. 5.

Fayyad, U. (1996). From Data Mining to Knowledge Discovery in Databases, p. 18.

Galvao, M. and Henriques, R. (2018). Forecasting model of a movie's profitability, IEEE, pp. 1–6.
URL: *https://ieeexplore.ieee.org/document/8399184/*

Ghiassi, M., Lio, D. and Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network, *Expert Systems with Applications* **42**(6): 3176–3193.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0957417414007088*

Hennig-Thurau, T., P. Gwinner, K., Walsh, G. and Gremler, D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?, *Journal of Interactive Marketing* **18**: 38–52.

Hur, M., Kang, P. and Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and Independent subspace method, *Information Sciences* **372**: 608–624.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0020025516306016*

Insane Studio Accounting: Warner Bros Claims $167 Million Loss Over Harry Potter and the Order of the Phoenix (2010).
URL: *https://www.slashfilm.com/insane-studio-accounting-warner-bros-claims-167-million-loss-over-harry-potter-and-the-order-of-the-phoenix/*

Kennedy, A. (2008). Predicting Box Office Success: Do Critical Reviews Really Matter, p. 7.

Mestyn, M., Yasseri, T. and Kertsz, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data, *PLoS ONE* **8**(8): e71226.
URL: *http://dx.plos.org/10.1371/journal.pone.0071226*

Nikhil Apte, Mats Forssell, A. S. (2011). Predicting movie revenue, *stanford.edu* (cs229).

Oh, C., Roumani, Y., Nwankpa, J. K. and Hu, H.-F. (2017). Beyond likes and tweets: Consumer engagement behavior and movie box office in social media, *Information & Management* **54**(1): 25–37.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0378720616300271*

Rubin, Rebecca, B. L. and Rubin, Rebecca, B. L. (2018). Solo: How Big a Box Office Dud Is the Star Wars Spinoff?
URL: *https://variety.com/2018/film/news/solo-a-star-wars-story-box-office-losses-1202825432/*

Sagar, C. (2018). Twitter Sentiment analysis using R.
URL: *http://dataaspirant.com/2018/03/22/twitter-sentiment-analysis-using-r/*

Samuel, A. L. (1959). Some studies in machine learning using the game of Checkers, *Ibm Journal of Research and Development* pp. 71–105.

Sharda, R. and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks, *Expert Systems with Applications* **30**(2): 243–254.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0957417405001399*

Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V. and Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM, IEEE, pp. 182–186.
URL: *https://ieeexplore.ieee.org/document/8389394/*

Vitelli, M. (2007). Predicting Box Office Revenue for Movies, p. 5.

Vr, N. and Babu Pb, S. (2014). Predicting Movie Success Based on IMDB Data.

Yoo, S., Kanter, R., Cummings, D. and Maas, A. (2012). Predicting Movie Revenue from IMDb Data, p. 5.