Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057 OPEN ACCESS 2018 13 (SP): 441-445

A Survey on Box-Office Opening Weekend Prediction Using Twitter Data

Apratim Tripathi*, Akhil Jha, Prof. A.M.J. Muthukumaran

Received 13 November 2018 ▪ Revised: 30 November 2018 ▪ Accepted: 06 December 2018

Abstract: We live in times where about 1,986 feature films are produced annually and this amounts to the box office being worth $41.7 billion, making the film industries one of the biggest markets in the world. In out paper we are only concerned with films produced in the United States of America, which produces nearly 600 movies annually. We aim to find out the "buzz" around the movies and therefore predict the opening weekend for each movie. The opening weekend ticket sales depend upon the buzz the movie has generated since the announcement, trailers, TV spots and various other factors.Now, the impact of these trailers, TV spots, interviews on the general audience can be seen in terms of twitter reactions. This buzz in fact may be negative leading to a lower opening weekend. The buzz of the movie is the number of people talking about the movie. More the number of people talking about the movie, greater is the buzz surrounding the movie. To calculate the number of people talking about the movie we take twitter data into account. Now, people could be talking positively about the movie or negatively about the movie and we can use sentimental analysis algorithms to predict that. Finally based on these factors we try to form a model with best fits our data in order to make future predictions possible.

Keywords: Tweets, Box-Office, Opening Weekend.

INTRODUCTION

We focus on two main relations in our paper. The first relation is the hype vs. sale. He we aim to find out what amount of hype leads to a particular amount of opening weekend box office earnings. The second relation is the Market Strategy for promotion vs. sale. As different companies use different strategies to promote the movie, we aim to find out what market strategies are more successful than others. This is useful not only for the companies promoting a particular movie but also for a production studio as they can better judge what type of hype a particular type of marketing can generate therefore leading to greater number of sales. Different studios adopt different market strategies and some are successful while others are not. We will try to distinguish the types that succeed from the types that don't. With a large industry such as the film industry on which billions of dollars depend, the amount of research from a statistical point of view is very limited and we aim to change this.

TERMINOLOGY

Some technical terms used in this paper are as follows:

1. *Sentimental Analysis:* "Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information."[5]
2. *Regression:* a measure of the relation between the mean value of one variable and corresponding values of other variables often used in Machine Learning.
3. *Opening Weekend*: the box office receipts from Friday through Sunday. In particular, the weekend box office for the initial week of release, or opening weekend.

Apratim Tripathi*, Department of Computer Science Engineering, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, India. E-mail: anujtripathi1997@gmail.com

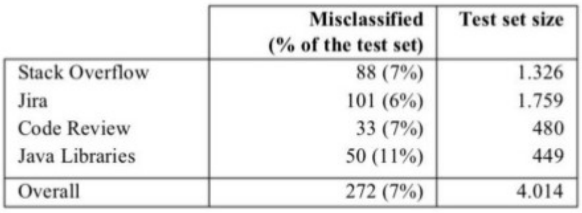
Akhil Jha, Department of Computer Science Engineering, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, India. E-mail: akhil.jha6@gmail.com

Prof. A.M.J. Muthukumaran, Asst. Professor, Department of Computer Science Engineering, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, India. E-mail: muthu.a@ktr.srmuniv.ac.in

RESEARCH CHALLENGES

1. *Data Collection:* Data needs to be collected from social media, and we do not have a particular API or tool which can fetch all the relevant data all at once, from any of the social media. Twitter provides an API but it comes with its limitations, such as we can only collect data for one week from the present time. This makes it impossible to analyze previous trends and therefore perform machine learning. To work around this, we build a web crawler which goes through each and every tweet on twitter and if it finds it relevant, it copies it onto the database. Though this is a reliable way to collect data but this is very slow.

2. *Sentimental Analysis:* A lot of research is going on in the field of sentimental analysis and though we have made a lot of progress but we are still pretty far from achieving perfect result. Things such as, "This made me cry." in context of the movie might be positive reviews in case of a movie, but sentimental analysis algorithms cannot still get around such phrases and therefore give negative polarity to the result. This leads to errors creeping into the results.



	Misclassified (% of the test set)	Test set size
Stack Overflow	88 (7%)	1.326
Jira	101 (6%)	1.759
Code Review	33 (7%)	480
Java Libraries	50 (11%)	449
Overall	272 (7%)	4.014

Fig. 3.1: Manual inspection of text misclassified [1]

3. *Linear Regression*: In this, we do not have an upper limit, that is, according to this algorithm, if we have an unlimited amount of hype surrounding a movie, the movie is going to earn an unlimited amount of money on the opening weekend. This is obviously not possible, because even if the movie is playing on all the screens in the world and all the shows are full, the amount of audience that can watch the movie on the opening weekend is limited by the number of screens available. Therefore, we need to use better regression models to overcome this (logarithmic regression).

TAXONOMY

1. *Data Collection* can be done from multiple sources, the ones covered in the papers we survey are:

   a.) *YouTube* **-** Taking into count the views, likes and comments on a video (trailer, TV spots etc.).

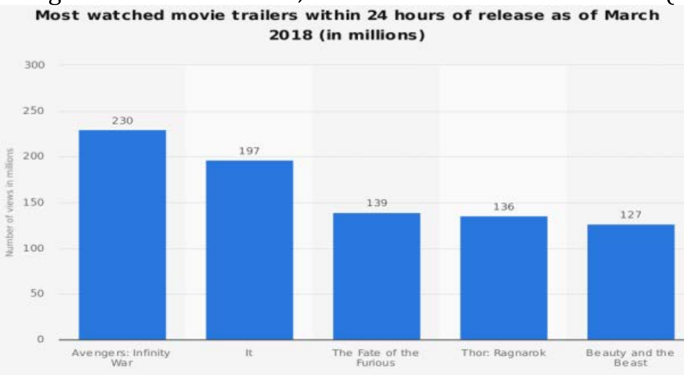
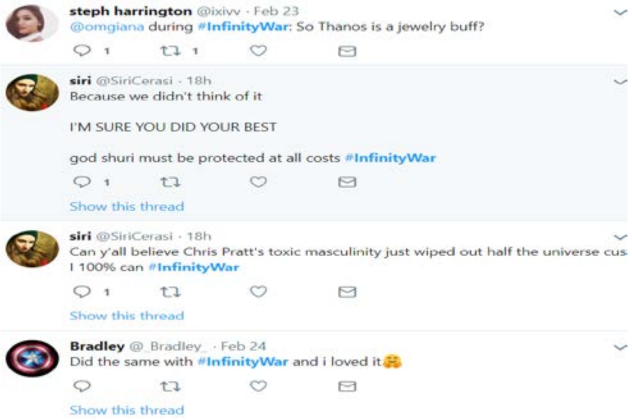


Fig. 4.1: Most watched movie trailers within 24 hours of release (March 2018)

   b.) *Twitter* – Taking into count the tweets, retweets and their likes.



steph harrington @ixivv · Feb 23
@omgiana during #InfinityWar: So Thanos is a jewelry buff?
1 1

siri @SiriCerasi · 18h
Because we didn't think of it

I'M SURE YOU DID YOUR BEST

god shuri must be protected at all costs #InfinityWar
1
Show this thread

siri @SiriCerasi · 18h
Can y'all believe Chris Pratt's toxic masculinity just wiped out half the universe cus
I 100% can #InfinityWar
1
Show this thread

Bradley @_Bradley_ · Feb 24
Did the same with #InfinityWar and i loved it🤗
Show this thread

Fig 4.2: Tweets on twitter for Avengers: Infinity War

2. *Sentimental Classification* can be done in the following ways:
   a.) Machine Learning - we further have two main types:
   - Supervised - algorithms such as Decision Trees, Naive Bayes etc.
   - Unsupervised - algorithms such as SOM, Neural Networks etc.

   b.) Dictionary Based - examples of dictionaries include wordNet, SentientWord.
   c.) Ontology Based (theory of existence) - mostly used in the feature extraction phase of sentimental analysis.

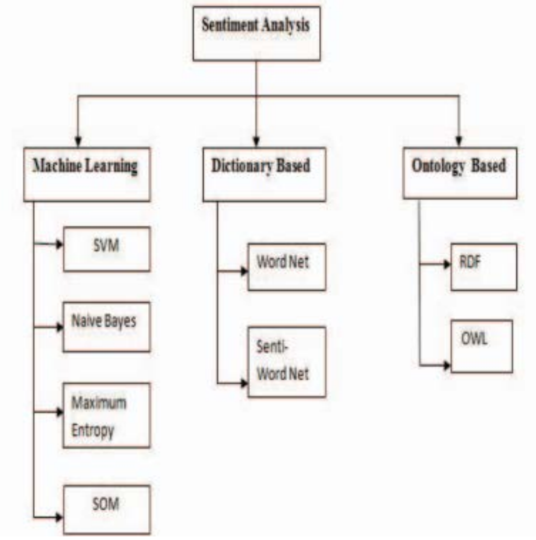


Fig. 4.3: Sentiment analysis methodology taxonomy [2]

3. *Regression* - We have many different types of regressions and we need to find the type that best fits our data.

- Linear
- Logistic
- Polynomial
- Stepwise
- Logarithmic

THE PAPERS SURVEYED

1. *"A Survey on Sentiment Analysis Methods and Approach"* (*2016 IEEE Eighth International Conference on Advanced Computing) [2]*

The main research direction of this paper is the betterment in the understanding of sentiment by a machine.

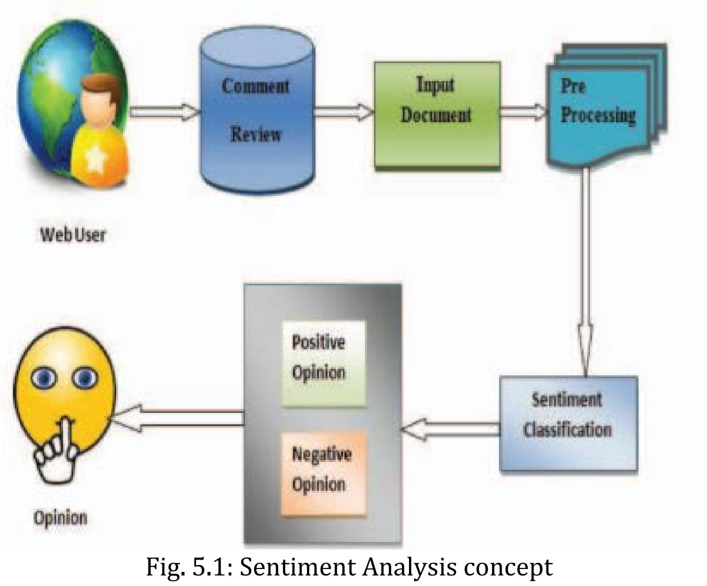


Fig. 5.1: Sentiment Analysis concept

The paper establishes an algorithm which used all the above-mentioned algorithms and derives its result from all of the results combined.

This is mainly a theory paper and the results are based on the mixture of the above-mentioned algorithms which therefore gives the best accuracy.

This paper is better than the individual methods adopted for sentiment analysis, because it gives an accuracy of sentiment up to 89.6% which is significantly higher than the others.

"This survey gives the knowledge about the sentiment analysis issues such as Polarity shift problem, data sparsity, binary classification briefly and how they are handled in different domains." [2]

2. *"Prediction of Movies Box Office Performance Using Social Media" (2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining)[3]*

The main research direction of this paper is using data collected from social media sources such as Twitter, IMDb movie database and YouTube to generate interesting models in order to predict box office performance of movies with the help of various data mining tools. The prediction involves using features extracted from several sources which include Tweets, retweets and follower count from twitter, historical movie database and sentiment analysis of Youtube comment section for trailers. The predictions are classified in three main classes: Flop, Neutral and Hit.

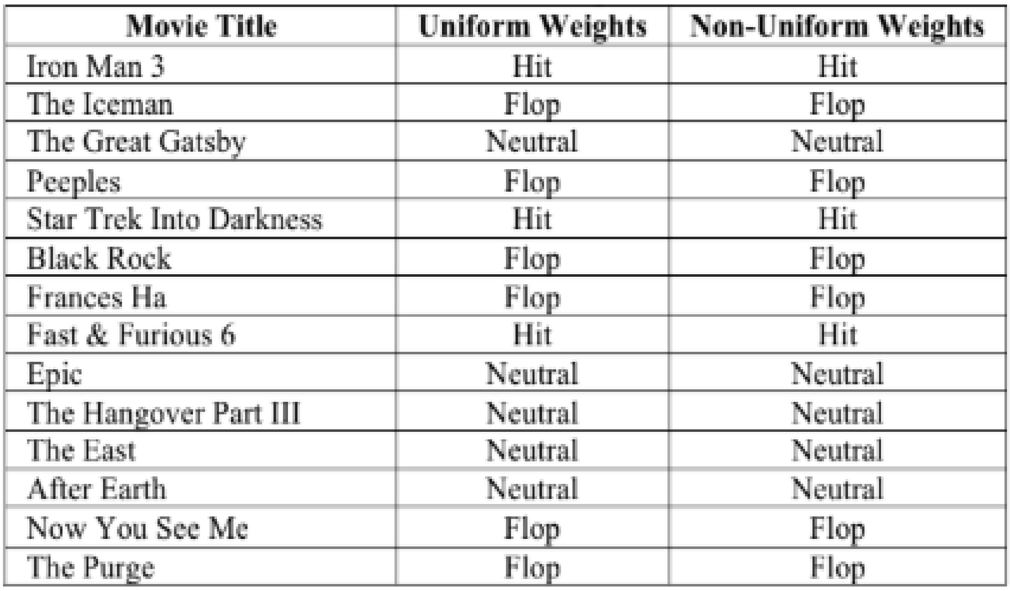
The procedure here involves three main steps:

a.) *Normalize training data* - This enables us to get relevant features and discard everything else, to faster processing of the data.
b.) *K-Means Clustering* - Putting the data into relevant chunks, and making a particular sample belong with a set of samples closely related to it.
c.) *Generating a prediction model* - Choosing the model which best fits the data which therefore enables us to make future predictions.

This paper does the analysis based on two main approaches:

a.) *Weighted Attributes* – "In this approach, equal weights are given to all the attributes. Out of the 35 movies, 7 movies were clustered as a Hit, 12 movies were labelled as a Flop and 16 movies were grouped under Neutral, meaning prediction is not certain. The labelled dataset is then tested using Weka's Naïve Bayesian and J48 classifiers predictive models."[3]
b.) *Unweighted Attributes* – "In this approach, more weight is given to the sentiment attribute. Out of the 35 movies tested, 8 are clustered as Hits, 12 as as Flop and 15 as Neutral."[3]

The approach giving the best results is therefore chosen.



Movie Title	Uniform Weights	Non-Uniform Weights
Iron Man 3	Hit	Hit
The Iceman	Flop	Flop
The Great Gatsby	Neutral	Neutral
Peeples	Flop	Flop
Star Trek Into Darkness	Hit	Hit
Black Rock	Flop	Flop
Frances Ha	Flop	Flop
Fast & Furious 6	Hit	Hit
Epic	Neutral	Neutral
The Hangover Part III	Neutral	Neutral
The East	Neutral	Neutral
After Earth	Neutral	Neutral
Now You See Me	Flop	Flop
The Purge	Flop	Flop

Fig. 5.2: Clustering Results

This paper uses methods which provide useful insights such as:

a.) Popularity of actors in the movie plays a key role in determining the success of the movie.
b.) Genre/Prequel - Certain genre's sell more than others and if the movie has had a prequel, it is set to earn more.
c.) Surprisingly sentiments play a very little role in the success of a movie.

The models used in this paper can be improved further by refining the neutral class and finding the actual opening weekend performance in terms of box office collection.

3. *"Business Intelligence from Social Media: A Study from the VAST Box Office Challenge" (IEEE Computer Graphics and Applications, Volume: 34, Issue: 5, Sept.-Oct. 2014) [4]*

The main research direction of this paper was, deriving Business intelligence from social media.

The math/algorithm used in this paper is mainly regression. We have a multivariable dataset with a single response variable (i.e. opening weekend prediction).

Analysis of a large set of potential linear regressions each of varying weightages was done.

The features used for regression are:

a.) Number of tweets.
b.) Number of screens the movies was released in.
c.) Sentiment scores for the movie.

$$Y = X\boldsymbol{\beta} + \varepsilon$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Fig. 5.3: Linear Regression model in matrix form

This is a theory paper which is based on an application paper. The application paper is the work of a team which participated in the VAST Box Office Challenge.

This paper uses Weighted Linear regression Models Averages as compared to other linear regressions which therefore make the results more accurate as compared to others.

CONCLUSION

This is a new and booming field and we need sophisticated methods to derive useful insights and therefore work must be done in this relatively new field of study which despite driving millions of dollars in revenue, is relatively statistically, unstudied. Proper studies can lead to better models in advertising and marketing which can therefore lead to increase in the Opening Weekend Box Office Revenue.

REFERENCES

[1] Nicole Novielli, Daniela Girardi, FilippoLanubile. "A Benchmark Study on Sentiment Analysis for Software Engineering Research". *arXiv:1803.06525*

[2] Ms. A.M. Abirami, Ms. V. Gayathri."A Survey on Sentiment Analysis Methods and Approach."*2016 IEEE Eighth International Conference on Advanced Computing.*

[3] 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." .*2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[4] Yafeng Lu, Feng Wang, and Ross Maciejewski. "Business Intelligence from Social Media: A Study from the VAST Box Office Challenge". *IEEE Computer Graphics and Applications, Volume: 34, Issue: 5, Sept.-Oct. 2014*.

[5] "https://en.wikipedia.org/wiki/Sentiment_analysis"