Prediction of Box Office for Bollywood Movies Using State-of-the-Art SentiDraw Lexicon for Twitter Analysis

Shashank Shekhar Sharma [1]
Gautam Dutta [2]

Abstract

Films are a high-risk industry. Accurate prediction of movie box-office revenues can reduce this market risk and inform the investment decisions regarding promotion of the movie closer to a film's release or right after release. Studies have shown that chatter on social media platforms like Twitter along with certain movie-related factors can be useful in predicting success of movies. Sentiment of tweets for any movie gives important information about the consumer's reaction and the polarity of these sentiments has been shown to have an impact on prediction of box-office revenues. This paper presented a novel Bollywood domain specific sentiment lexicon that delivered state-of-the-art performance for polarity determination of reviews. SentiDraw lexicon was built on movie reviews scraped from IMDB and calculated the sentiment orientation of these words by calculating the probability distribution of words across reviews with different star ratings. The results showed that SentiDraw lexicon delivered a superior performance compared to any other lexicon-based method. This significantly contributed in enhancing the prediction accuracy of box office for movies using textual data from Twitter for analysis. In fact, this study demonstrated an extremely parsimonious regression model that used only budget, hype factor, tweet volume, and polarity of tweets for a robust prediction of box office revenues even before the release of a movie.

Keywords : **sentiment lexicon, box office prediction, SentiDraw method, movie reviews, Bollywood, Twitter**

Paper Submission Date : February 17, 2020 ; Paper sent back for Revision : October 17, 2020 ; Paper Acceptance Date : November 12, 2020 ; Paper Published Online : June 25, 2021

The Indian movie industry has seen a huge upswing in the last decade after opening of multiplexes across cities. Compared to Hollywood, people often forget about the world's other movie capital – India (Dastidar & Elliott, 2019). The Indian film industry is expected to grow at 11.5% year-on-year, reaching total gross realization of ₹ 24,900 crores ($3.9 billion) by 2021 (EY India, 2019). A huge part of these revenues is generated in the first few weeks after the release of the movies.

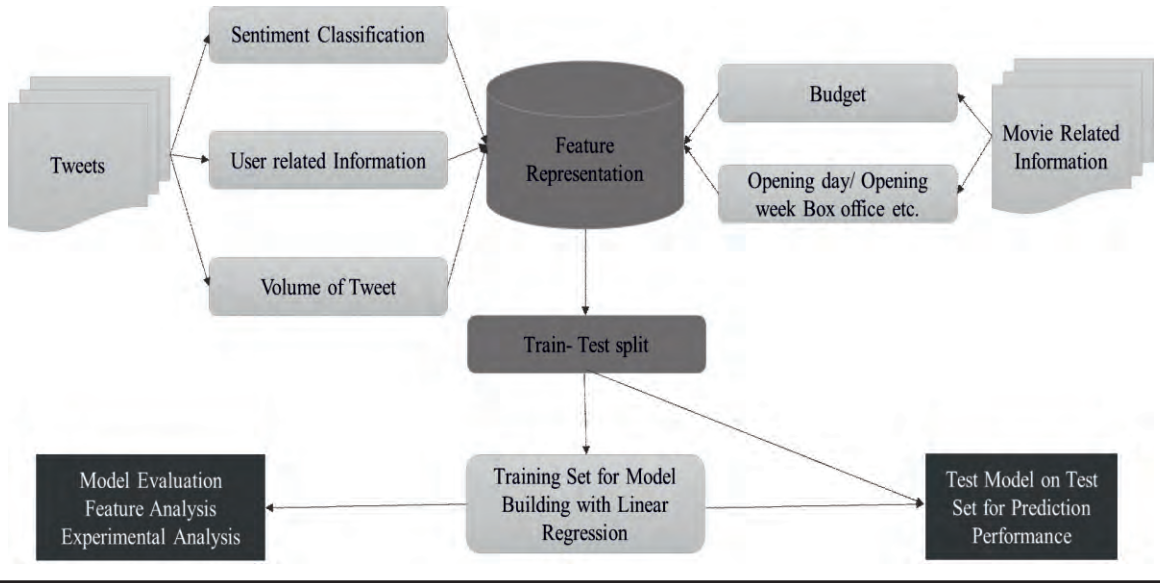
However, predicting movie box-office revenues is a challenging problem. Many studies have been conducted to use the available variables to predict box office revenues of movies. Researchers have consistently proven that promotion and reviews have a strong influence on a film's overall sales (Thomas & Patel, 2020). Most studies use common variables such as genre, star, or director power, season of release, and number of screens. Some studies have also used the distribution power for predicting box office sales.

Several studies have also used word of mouth (WoM) (Dellarocas et al., 2007) as a significant factor for a movie's success. Many insights into the process of word-of-mouth (WoM) generation by consumers' peers have been provided by research studies in the last decade and have illustrated its influence on successes of experiential

[1] *Research Scholar,* Indian Institute of Foreign Trade, IIFT Bhawan, B-21, NRPC Colony, Block B, Qutab Institutional Area, New Delhi - 110 016. (Email : shashank.mick@gmail.com) ; ORCID iD : 0000-0002-2931-2193
[2] *Professor,* Indian Institute of Foreign Trade, IIFT Bhawan, B-21, NRPC Colony, Block B, Qutab Institutional Area, New Delhi - 110 016. (Email : gautam@iift.edu) ; ORCID iD : 0000-0003-1500-7929

DOI : https://doi.org/10.17010/ijom/2021/v51/i5-7/161644

Figure 1. ***Using Social Media and Movie Related Variables for Predicting Box Office Revenue***



products (Venkataraman & Raman, 2016). Quantity of WoM has a significant impact on sales growth (Ghose & Ipeirotis, 2011), which has been proven in multiple studies. However, in these studies, this effect seemed to be moderated by sentiment of reviews as well. Chintagunta et al. (2010) also used user generated reviews but relied on sequence of release for different regional U.S. markets, and found that valence of UGRs mattered while volume did not. Niraj and Singh (2015) also made use of semantics and proved that valence did in fact affect box office revenue significantly. Studies in general have been able to establish strong relationships between online chatter around a movie and movie revenues. However, not many studies have used user generated text on Twitter for forecasting revenues in context of Bollywood movies. Also, the focus has been on the ratings of the review and very few studies have tried to judge the valency of Twitter through semantic analysis of consumer generated reviews like tweets and microblogs in social media. Use of semantics may further refine the model and provide more accurate prediction as shown schematically in Figure 1.

Most models use linear regression, while some have used diffusion theory (Dellarocas et al., 2007) to capture the unique patterns of entertainment goods marketing where heavy pre-release marketing is done which declines rapidly post release. Reddy et al. (2012) proposed a hype factor to determine pre-hype for a movie before its release using the number of tweets and number of distinct users who tweeted about a movie and used the same to predict box office figures for movies. The halo-effect of reviewers based on their online "clout" may also be an important factor (Bhāle & Tongare, 2018). The total number of tweets for a given movie determines the hype factor. Number of unique users and quantity of tweets generated per second are other important variables. Hype factor (α) is then calculated by using the following formula :

$$\alpha = \frac{\text{No of tweets by all users}}{\text{No of distinct users}} \qquad (1)$$

Recently, there have been some studies that used machine learning algorithms to predict a movie's success (Dhir & Raj, 2018 ; Jaiswal & Sharma, 2017) but these studies focused on classifying movies as hit or flop and did not attempt to predict the revenues as such. Also, they used ratings available on sites like IMDB and Rotten

Figure 2. ***General Steps of Feature Processing and Representation for Tweets***

Tomatoes after the release of movies instead of using social media to predict the outcome even before the release of the movies.

While online reviews, especially tweets and comments, are very useful for predicting success of a movie, several practical limitations stand in the way. The rate at which tweets get generated right before and after the release of a movie is very high and this makes it very difficult to manually label the sentiment polarity of movies. NLP (natural language processing) based methods use both supervised and unsupervised methods and make it possible to process millions of tweets quickly and predict the polarity to a high level of accuracy which can then be used for prediction. This process of determining sentiment polarity and the strength of this sentiment is called Sentiment Analysis (SA). Machine learning (ML) methods are often employed to enhance the performance of NLP methods. Machine learning algorithms include support vector machines (Abbasi et al., 2011), logistic classification (Bai, 2011), naïve bayes (Pang et al., 2002), and rule based classifiers (Prabowo & Thelwall, 2009). Shaukat et al. (2020) used ANN trained on Stanford movie dataset and achieved an accuracy of 91% in polarity classification. Figure 2 shows some of the common pre-processing steps which can have the most significant impact on the prediction performance.

Most often, bag of words (BoW) is used as a feature where the words or phrases present in a text document are represented either based on their presence or absence or by the frequency of their appearance. The syntax structure is not accounted for in this technique. Pang and Lee (2004) used BoW with unigrams and bigrams together and found unigrams to have a high level of predictive ability. Availability of labeled dataset is a consistent issue in developing ML models for classification that can be easily used across domains. Also, supervised models are generally more computationally intensive (Pang & Lee, 2004). Sentiment lexicons overcome these problems. These lexicons provide sentiment score for thousands of words or even phrases. These sentiment scores can then be used for calculating the sentiment polarity and intensity of this polarity for a given document (Musto et al., 2014). Several generic sentiment lexicons like SentiWordNet (Baccianella et al., 2010), MPQA Subjectivity Lexicon (Taboada et al., 2011), SO-CAL (Khoo & Johnkhan, 2018), Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001), SentiStrength (Thelwall, 2017), SentiWords (Gatti et al., 2016), and Affective norms for English word (ANEW) (Utomo et al., 2018) developed in recent years have been able to demonstrate a decent performance in sentiment prediction across domains. One of the popular concepts that is widely used for automated development of lexicon is called PMI (pointwise mutual information) (Musto et al., 2014).

Supervised methods generally perform much better when trained within the same domain. The performance, however, deteriorates quickly as an attempt is made to apply the models across domains (Gatti et al., 2016). Domain specific sentiment lexicons tend to perform better (Jiménez-Zafra et al., 2016). Many studies have focussed on different techniques to build domain specific lexicons. Some studies (Saif et al., 2014) have advocated techniques that build domain specific lexicons from scratch, while others have attempted to recalibrate popular generic lexicons as per requirements of a given domain (Khan et al., 2015). While both types of techniques show a considerable increase in performance, they have not been shown to match ML based classification in prediction tasks consistently. An interesting approach in recent studies has been to calculate sentiment scores on the basis of conditional probabilities of a word's occurrence across documents of different sentiment polarities (Almatarneh &

Gamallo, 2018). Use of probability distribution seems promising as these methods have shown a significantly better performance for domain specific classification tasks compared to PMI based methods.

Lexicons built using conditional probability (Khan et al., 2016) of words across reviews with different ratings (Almatarneh & Gamallo, 2018) have a significantly better performance in polarity determination. Khan et al. (2016) built on SentiWordNet 3.0 to develop SentiCS lexicon. They used part of speech along with usage-based ranks and calculated sentiment scores using feature weight which is chi-square-based. Further, these weights are normalized to give them a score between –1 to +1. Lee et al. (2014) also used a technique based on conditional probability where they defined the labelled buckets and then attributed a sentiment score to each word on the basis of these distributions across these labelled buckets. They obtained a high performance on polarity determination (*F*-score of 82%) on a movie review dataset. Some other studies have recently employed this technique in social media domains (Labille et al., 2017) by using information theoretic techniques based on conditional probabilities to develop social media specific lexicons. SPLM (Almatarneh & Gamallo, 2018) made use of relative frequency of each word for a given category which can be a rating point for a movie or other reviews that are accompanied by a rating. The average scores across positive range (AVP) and negative range (AVN) for each word are calculated by taking the average of RF values for each range across the number of rating points in each of the two ranges. The SO score of the word is given as the difference between these two values – AVN and AVP. SentiDomain uses the probabilistic score for each word *'w'* and is computed using posterior probabilities by calculating the difference of the probability of *'w'* being positive, $p(\text{pos}|w)$, and its probability of it being negative, $p(\text{neg}|w)$. SentiPosNeg method simply uses the probability of a word being positive or negative based on its relative frequency of occurrence in positive and negative reviews and uses that as the sentiment score after minmax normalization.

The method advanced in this paper builds on conditional probability based technique and suggests a novel way to enhance the performance of similar techniques by making use of dispersion for each term across the rating point. For both movie reviews datasets in the movie domain (Hollywood and Bollywood), the *F*-scores and accuracy have been shown to be greater than 85% in the polarity determination task. This is either comparable or superior to other lexicon-based methods known to us.

Key Objectives and Contributions

This study provides a comprehensive method for predicting movie box-office revenues using Twitter data. This paper aims to enhance the sentiment classification accuracy to determine the polarity of tweets more accurately by building a Bollywood domain specific SentiDraw lexicon so as to help increase prediction performance of box-office success of movies. The key contributions of this study are :

(1) To introduce a new technique of developing sentiment lexicon for sentiment classification of opinions and reviews which enhances performance of classification compared to other lexicons. In fact, the results are, in many cases, comparable to supervised methods.

(2) To demonstrate the superiority of SentiDraw method by developing a lexicon for Hollywood along with Bollywood and comparing the performance with other methods on some of the most experimented datasets like CMRD, Cornell movie reviews data (Pang & Lee, 2004), and LMRD - Large movie review data set (Mass et al., 2011).

(3) To use movie related variables and tweets for Bollywood movies released between 2017 and 2019 and build a regression model for prediction of box office revenues.

(4) To demonstrate a parsimonious method that uses only Twitter data to predict box office revenues of a movie

even before its release. The accuracy performance obtained using SentiDraw lexicon ranged from 80.5% to 84.5% across datasets. The parsimonious regression models were trained using data for 64 movies using the ratio of positive to negative tweets determined through SentiDraw, which demonstrated that this positive to negative ratio of tweets had a significant impact on prediction of box office revenues. The models, thus trained, achieved a mean absolute error (MAE) of 11.6 crs and root mean squared error (RMSE) of 19.4 crs for test data.

Materials and Method

SentiDraw Framework

SentiDraw uses ratings of the reviews which are generally given as number of stars by reviewers on each review that they post online. The distribution of star ratings calculated as conditional probabilities are then used for determining sentiment scores across words. SentiDraw advances the SPLM method by making use of distribution of star ratings of words across reviews in the labelled training data for generating SO of the words present in this dataset. The scores calculated by these probabilities are weighted with a scoring schematic and then normalized to produce final SO scores instead of simply using the difference of probabilities for positive and negative frequencies as done in most methods like SPLM. Reviews of movies and other product categories like electronic goods, restaurants, etc. are available on customer review sites like IMDB, Tripadvisor, Yelp, Amazon, etc. For this paper, 80,000 movie reviews were scraped from IMDB for 200 movies released between the years 2012 and 2018 using 'Scrapy,' which is a Python library for web scraping. Reviews that contained less than five words were dropped. Further, reviews that did not contain star ratings were also dropped since they couldn't be labelled ; 40,000 reviews each were finally selected for Bollywood and Hollywood movie domains for developing respective SentiDraw lexicons. Star ratings are mostly given on a scale of 5 or 10. Higher ratings indicate higher intensity of positive opinion. For data processing, each review is first POS-tagged using Stanford-POS tagger and then tokenized. All stop words and punctuations are subsequently removed from the bag of words. Only nouns, adverbs, adjectives, and verbs were retained in the bag of words labelled with their respective POS tags (as *'a'*, *'n'*, *'r'*, *'v'*) as these POS have been shown to show more subjectivity than others (SentiWordNet). The taxonomy for these tokens is given the following SentiWordNet method.

The count of each term or token *'t'* across reviews in the dataset is denoted by C_t. The count frequency of *'t'* across all star ratings is denoted as $f_{t,r}$ for each rating *'r'*. The probability of a given token *'t'* for each rating point *'r'* is then calculated by using the ratio of its frequency count for a rating point over its total frequency count across the dataset.

$$P(t,r) = \frac{f_{t,r}}{C_t} \qquad (2)$$

After this, a weighted average sentiment score for each word is calculated by using the sum of product of the probability of its occurrence at each rating point and a prior determined sentiment value of each rating point as given in Table 1. A sentiment score between '–5'and '+5' is ascribed to each rating point, where rating points higher than '5' are given positive scores to ascribe positive sentiment and this score is higher for higher star ratings. Rating points '5' and below '5' are scored negatively with decreasing value for lower ratings. This scoring pattern follows other research papers in this domain that considered '5' and '6' as neutral ratings for reviews (Sharma & Dutta, 2018) with 10 scale star rating and '3' as neutral rating for 5 scale star rating. For reviews with 10 scale star ratings, a low value of –1 and +1 are ascribed for rating points of '5' and '6,' and for reviews with 5 scale rating, the value ascribed to rating point of '3' is 0. Table 1 shows the sentiment score for each rating point.

The weighted average sentiment scores (WS_t) for each token *'t'* is calculated by :

Table 1. ***Sentiment Scores for Rating Points***

10 Scale Star Rating	5 Scale Star Rating	Rating Sentiment Score (R_s)
1 and 2	1	–5
3 and 4	2	–3
5	-	–1
-	3	0
6	-	+1
7 and 8	4	+3
9 and 10	5	+5

$$Ws_t = \Sigma^{r}_{i=o} P(t, r_i) \times R_{s,i} \quad (3)$$

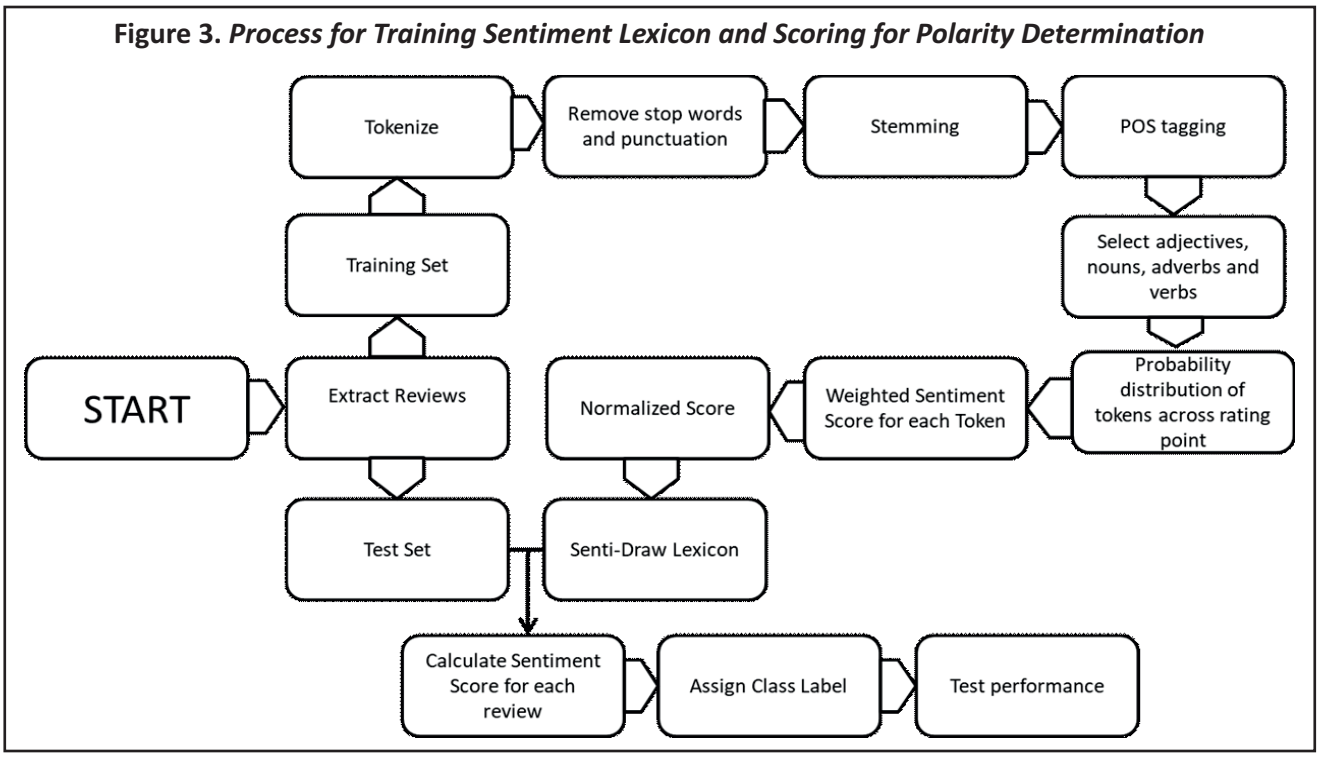
where, $'r'$ is the star rating label, $'N'$ is the total number of start rating labels (5 or 10 in most cases), and $R_{s,r}$ is the rating sentiment score for that rating point $'r'$ as per Table 1.

The weighted average sentiment scores for the words are then normalized between –1 and 1 using MinMax normalization.

$$SO_t = \frac{WS_t}{WS_{t,\,Max} - WS_{t,\,Min}} \quad (4)$$

Figure 3 gives an overview of the steps followed in this study for building the lexicon and testing its performance.

Figure 3. ***Process for Training Sentiment Lexicon and Scoring for Polarity Determination***



The reviews extracted for both these datasets were each divided into training set and testing set in an 80 : 20 ratio. The lexicons were then built using the training data with SentiDraw framework explained above for all datasets. The reviews were classified as positive or negative on the basis of the mean or median value score of each review, where a SO score of 0 or greater than 0 is classified as positive and others are classified as negative. All the above steps were then repeated for other benchmark methods (SPLM, SentiPosNeg, and SentiDomain) which also make use of star rating of labels to compare the performance of SentiDraw against them.

Koimoi.com and Bollywoodboxoffice.com act as a credible source for box office updates of Bollywood movies. Using python, data were scraped from these websites for 80 movies released between 2017 and 2018

Table 2. ***List of Movies on Which Model Training and Prediction Performance Has Been Tested***

Film	Lifetime Box Office	Box Office Verdict
102 Not Out	52.04	Hit
1921	15.94	Average
Aiyaary	18.22	Super-Flop
AndhaDhun	74.59	Super-Hit
Baazaar	24.77	Flop
Batti Gul Meter Chalu	37.73	Flop
Bhaiaji Superhit	6.25	Super-Flop
Blackmail	20.35	Average
Dhadak	74.19	Hit
Fanney Khan	10.55	Super-Flop
Hate Story IV	22.38	Average
Helicopter Eela	4.13	Super-Flop
Hichki	46.21	Super-Hit
Kaalakaandi	6.34	Super-Flop
Karwaan	19.22	Average
LoveYatri	11.24	Super-Flop
Mohalla Assi	1.64	Super-Flop
Mukkabaaz	10.51	Average
Mulk	21.10	Average
October	39.06	Flop
Pad Man	81.82	Average
Pari	28.96	Flop
Parmanu	65.89	Average
Pataakha	6.95	Super-Flop
Race 3	166.40	Hit
Raid	103.07	Average
Satyameva Jayate	80.50	Super-Hit
Sui Dhaaga	79.02	Super-Hit
Tumbbad	13.57	Flop
Yamla Pagla Deewana : Phir Se	9.60	Super-Flop

A Gentleman	20.59	Super-Flop
Baadshaho	78.10	Flop
Badrinath Ki Dulhania	116.68	Super-Hit
Bareilly Ki Barfi	34.55	Hit
Begum Jaan	20.91	Average
Commando 2	25.09	Average
Firangi	10.27	Super-Flop
Fukrey Returns	80.32	Super-Hit
The Ghazi Attack	20.30	Super-Flop
Guest iin London	10.64	Super-Flop
Half Girlfriend	60.30	Average
HaseenaParkar	8.03	Super-Flop
Hindi Medium	69.59	Super-Hit
Indu Sarkar	4.95	Super-Flop
Ittefaq	30.21	Average
Jagga Jasoos	54.16	Flop
Jolly LLB 2	117	Super-Hit
Judwaa 2	138.61	Hit
Kaabil	103.84	Hit
Lipstick Under My Burkha	19.21	Super-Hit
Lucknow Central	11.20	Super-Flop
Meri Pyaari Bindu	9.59	Super-Flop
Mubarakan	55.59	Flop
Munna Michael	32.89	Flop
Newton	22.80	Hit
Phillauri	27.10	Average
Poster Boys	12.73	Super-Flop
QaribQaribSinglle	17.08	Flop
Sachin	50.89	Hit
Sarkar 3	9.93	Super-Flop
Secret Superstar	63.40	Average
Shubh Mangal Saavdhan	43.11	Hit
Simran	17.26	Flop
Tubelight	119.26	Average
Tumhari Sulu	36.15	Hit
Romeo Akbar Walter	38.83	Average
Junglee	24.70	Average
Kesari	154.41	Average
Badla	87.99	Average
Luka Chuppi	94.75	Average
Sonchiriya	6.60	Average

Total Dhamaal	154.23	Average
Gully Boy	140.25	Average
EkLadki Ko DekhaTohAisaLaga	20.28	Average
Thackeray	18.19	Super-Flop
Manikarnika The Queen of Jhansi	92.19	Flop
Why Cheat India	8.66	Average
Fraud Saiyyan	0.53	Super-Hit
Rangeela Raja	0.19	Super-Flop

Table 3. ***Summary Statistics of Budget and Box Office Performance of Sample Movies***

	Lifetime Box Office	Opening Week Box Office	Opening Weekend Box Office	Opening Day Box Office	Budget
Count	80	80	80	80	80
Mean	44.6	31.4	20.9	5.4	39.8
Std	41.9	29.8	20.4	5.7	32.0
Min	0.19	0.19	0.15	0.05	0.1
25%	12.4	10.4	6.9	1.8	20.0
50%	24.9	19.9	13.0	3.3	30.0
75%	70.7	42.8	31.3	7.8	48.5
Max	166.4	140.7	103.0	28.5	154.0

(as shown in Table 2). Along with budget, the data (given in Table 3) included details of box office revenues generated by each movie on the opening day, opening weekend, opening week, and lifetime. Next, using movie related hashtags for each movie, Twitter API was accessed using open source python libraries that can scrape tweets from Twitter for each of the assigned hashtags along with a bunch of tweet related information such as name of the user who had tweeted, number of followers for that user, number of likes for the user, and date when the tweet was done. The tweets were all labelled as pre-release tweets and post-release tweets based on whether the tweet was done before or after the release of the movie. Pre-release tweets included tweets made up to a month before the release of the movie, while post-release tweets only included tweets made on the day of the release.

The tweets were then scored using both SentiDraw and SentiWordNet independently and they were classified as positive or negative accordingly. Ratio of positive to negative tweets was calculated for each case and used as a feature along with the average sentiment value of tweets for each movie for each of the lexicons.

Three regression experiments using ordinary least squares (OLS) regression have been presented in this study. The key variables are listed in Table 4. OLS estimates the relationship between one or more independent variables and a dependent variable ; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line. The first experiment uses budget and opening day revenues as variables along with tweet related information for tweets made both before and after the release of a movie. This model can be used for making prediction of box office success of movies right after the release by quickly assimilating the reaction through the tweets that get generated on the day of release along with tweets generated prior to release. The hype factor, described above, is also taken as one of the variables along with volume of tweets, number of distinct users, their followers, and ratio of positive and negative tweets classified using SentiDraw lexicon. All the variables are scaled using minmax normalization

Table 4. ***Set of Features Extracted from Twitter Data***

Feature Description	Coded_Feature
Number of distinct users who tweeted for each of the movies after release	post_distinct_users
Hype factor determined from tweets generated after release	Post_Hype
Number of followers of distinct users who tweeted for each of the movies after release	post_distinct_users_followers
Average sentiment value across tweets determined by SentiDraw lexicon after release	post_sentidraw_value_movie
Average sentiment value across tweets determined by SentiWordNet lexicon after release	post_sentinet_value_movie
Total tweets generated for a movie after release	post_total_tweets
Number of distinct users who tweeted for each of the movies before release	prior_distinct_users
Hype factor determined from tweets generated before release	Pre_Hype
Number of followers of distinct users who tweeted for each of the movies before release	prior_distinct_users_followers
Average sentiment value across tweets determined by SentiDraw lexicon before release	prior_sentidraw_value_movie
Average sentiment value across tweets determined by SentiWordNet lexicon before release	prior_sentinet_value_movie
Total tweets generated for a movie before release	prior_total_tweets
Ratio of positive to negative tweets as determined using SentiDraw lexicon before release	Pre_SentiDraw_Ratio
Ratio of positive to negative tweets as determined using SentiWordNet lexicon before release	Pre_SentiNet_Ratio
Ratio of positive to negative tweets as determined using SentiDraw lexicon after release	Post_SentiDraw_Ratio
Ratio of positive to negative tweets as determined using SentiWordNet lexicon before release	Post_SentiNet_Ratio

before running regression for building the prediction models. Table 5 shows that the ratio of positive and negative tweets using SentiWordNet lexicon is not correlated with lifetime box office revenues of movies and is not used for building the regression model.

Table 5. ***Correlation of All Features Under Evaluation with Lifetime Box Office Revenues***

	Correlation with Lifetime Box Office
First Week Box Office	0.97
Opening Weekend Box Office	0.94
Opening Day Box Office	0.90
Budget	0.81
post_total_tweets	0.75
post_distinct_users_followers	0.64
prior_total_tweets	0.59

prior_distinct_users_followers	0.44
prior_distinct_users	0.40
post_distinct_users	0.40
Pre_SentiDraw_Ratio	0.38
Post_SentiDraw_Ratio	0.25
Post_SentiNet_Ratio	0.09
Pre_SentiNet_Ratio	–0.12
Post_Hype	–0.58
Pre_Hype	–0.66

Table 6. ***Selected Features or Independent Variables for Each of the Three Experiments***

Experiment 1 : Post Release	Experiment 2 : Pre-Release with Budget	Experiment 2 : Pre-Release only Twitter
Budget	Budget	Pre_Hype
Opening_day	Pre_Hype	prior_total_tweets
post_distinct_users	prior_total_tweets	Pre_SentiDraw_Ratio
Post_Hype	Pre_SentiDraw_Ratio	
post_distinct_users_followers		
post_total_tweets		
prior_distinct_users		
Pre_Hype		
prior_distinct_users_followers		
prior_total_tweets		
Pre_SentiDraw_Ratio		
Post_SentiDraw_Ratio		

The other two experiments (as shown in Table 6) are done for predicting the box office performance even before the release of the movies. Both are parsimonious methods. While the first one uses only pre-hype, tweet volume, and positive to negative ratio of tweets along with budget for building the regression model, the second model drops budget as a variable and only uses the other three features extracted from Twitter to assess if a robust regression model can be developed using only tweets before the release of a movie for prediction.

Analysis and Results

It is evident from Table 7 that SentiDraw lexicons developed using Hollywood and Bollywood datasets demonstrate an improvement in performance over comparable methods in polarity determination when used on test datasets for each of the respective domains. Furthermore, to test if these lexicons can have a more generalized utility across a similar dataset, experiments are carried out on with popular CMRD and LMRD datasets and compared with SentiWordNet. Table 8 shows the relative performance of these lexicons with SentiDraw lexicon made with Hollywood dataset for testing the robustness of the SentiDraw method compared to others when used for other datasets in the same domain.

The results demonstrate the superior performance of SentiDraw method. Lexicon made with SentiDraw

Table 7. ***Comparison of Classification Performance of Each Lexicon Method on Hollywood and Bollywood Movie Reviews Dataset***

Performance Metrics	Dataset	Bollywood	Hollywood
Accuracy	SentiWordNet	66.60%	66.20%
	SentiDomain	82.90%	79.80%
	SentiDraw	**84.50%**	**80.50%**
	SentiPosNeg	81.60%	70.90%
	SPLM	83.30%	79.30%
F Score - Positive	SentiWordNet	72.50%	74.20%
	SentiDomain	83.00%	80.40%
	SentiDraw	**85.20%**	**81.50%**
	SentiPosNeg	83.00%	79.30%
	SPLM	84.10%	80.10%
F Score - Negative	SentiWordNet	57.30%	51.20%
	SentiDomain	82.10%	79.00%
	SentiDraw	**84.00%**	**80.80%**
	SentiPosNeg	80.10%	79.00%
	SPLM	82.50%	79.10%

Table 8. ***Comparison of Classification Performance of Each Lexicon Method on LMRD and CMRD***

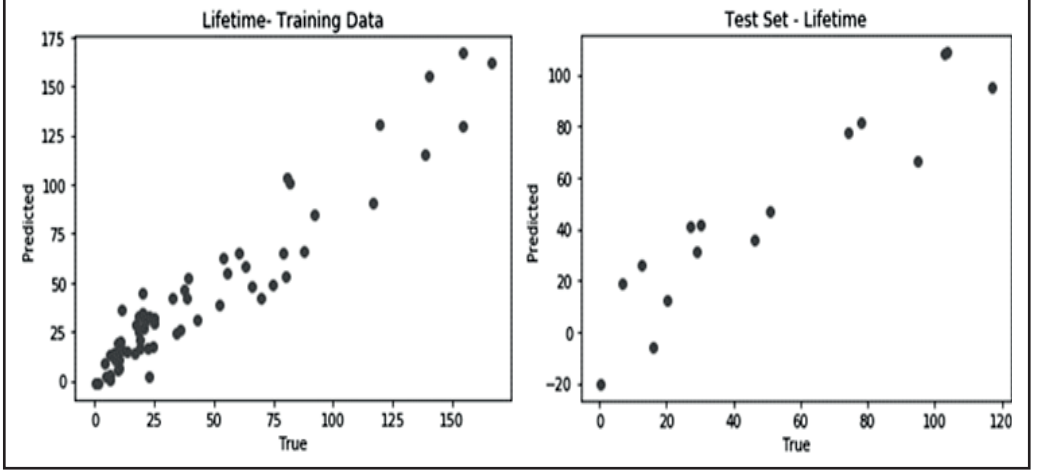
	CMRD			LMRD		
Lexicon	**Accuracy**	**Positive *F*-Score**	**Negative *F*-Score**	**Accuracy**	**Positive *F*1 Score**	**Negative *F*1 Score**
SentiWordNet	61.60%	54.50%	66.80%	56.90%	69.00%	29.40%
SentiDomain	80.30%	79.80%	80.40%	74.10%	**74.00%**	73.00%
SentiDraw	**83.00%**	**82.50%**	**83.30%**	**74.60%**	**75.00%**	**74.20%**
SentiPosNeg	78.00%	79.30%	77.50%	73.20%	74.60%	72.30%
SPLM	77.80%	78.00%	77.30%	73.60%	74.80%	72.00%

Table 9. ***Summary of Key Metrics Obtained from OLS Regression Experiments***

		Experiment 1 : Post Release	Experiment 2 : Pre - Release with Budget	Experiment 3 : Pre - Release only Twitter
	R^2	0.907	0.799	0.657
	Adjusted R^2	0.887	0.786	0.640
Train	**RMSE**	12.9	19.0	24.8
	MAE	10.4	13.4	18.9
Test	**RMSE**	13.9	19.4	29.9
	MAE	11.6	14.7	24.7

method performs better on accuracy and *F*- scores for both positive and negative classification compared to any other sentiment lexicon based method (as per our knowledge). This experiment illustrates that SentiDraw method

Figure 4. ***Experiment 1 : True Versus Predicted Lifetime Box - Office for Both Training and Test Sets***



is an ideal candidate for polarity determination task when building regression models for prediction of box office success of movies.

After classification of tweets using SentiDraw lexicons, regression models are built for predicting box office revenues of movies with OLS regression. The three experiments differ from each other conceptually. While the first experiment takes in variables like opening day box office and information from Twitter after the release of the movie, the other two experiments are for estimating box office revenues even before the release of the movie. As is expected, Experiment 1 has a high adjusted coefficient of determination (R^2) of 0.89 and achieves the lowest mean absolute error (MAE) and root mean squared error (RMSE) for both training and test sets as summarized in Table 9 and is also evident from Figure 4 that plots true and predicted values of lifetime box office revenues for training and test sets. However, this model suffers from multicollinearity. Table 11 shows that several independent

Table 10. ***Experiment 1 : Comparison of Independent Variables and Their Impact***

Features	Coef	Std Eror	*P*>|*t*|
const	97.34	6.29	0.000
Budget	22.20	8.85	**0.015**
Opening_day	62.14	10.31	**0.000**
post_distinct_users	–0.06	16.25	0.997
Post_Hype	–11.88	16.40	0.472
post_distinct_users_followers	18.18	9.28	0.056
post_total_tweets	–5.17	21.50	0.811
prior_distinct_users	–0.06	16.25	0.997
Pre_Hype	–3.31	11.33	0.771
prior_distinct_users_followers	0.59	7.10	0.933
prior_total_tweets	–4.48	39.92	0.911
Pre_SentiDraw_Ratio	11.74	5.43	0.035
Post_SentiDraw_Ratio	1.92	5.78	0.741

Table 11. ***Experiment 1 : Correlation of Features Within Themselves and With Dependent Variable***

	Lifetime	Budget	Opening _day	Post_ distinct _users	Post_ Hype	Post_ distinct _users_ followers	Post_ total_ tweets	Prior_ distinct _users	Pre_ Hype	Prior_ distinct _ users_ followers	Prior_ total_ tweets	Pre._ Senti Draw _Ratio	Post_ Senti Draw _Ratio
Lifetime	1.00	0.81	0.90	0.40	–0.58	0.64	0.75	0.40	–0.66	0.44	0.59	0.38	0.25
Budget	0.81	1.00	0.77	0.41	–0.41	0.51	0.64	0.41	–0.55	0.45	0.55	0.19	0.18
Opening _Day	0.90	0.77	1.00	0.44	–0.44	0.52	0.68	0.44	–0.68	0.41	0.66	0.21	0.08
post_distinct _users	0.40	0.41	0.44	1.00	0.09	0.53	0.64	1.00	–0.32	0.55	0.92	0.06	0.02
Post_Hype	0.58	–0.41	–0.44	0.09	1.00	0.51	0.61	0.09	0.54	–0.21	–0.13	0.41	–0.34
post_distinct_ users_followers	0.64	0.51	0.52	0.53	–0.51	1.00	0.79	0.53	–0.64	0.68	0.65	0.32	0.33
post_total _tweets	0.75	0.64	0.68	0.64	0.61	0.79	1.00	0.64	–0.64	0.54	0.78	0.39	0.25
prior_distinct _users	0.40	0.41	0.44	1.00	0.09	0.53	0.64	1.00	–0.32	0.55	0.92	0.06	0.02
Pre_Hype	0.66	0.55	–0.68	0.32	0.54	–0.64	–0.64	–0.32	1.00	–0.61	–0.62	0.02	–0.15
prior_distinct_ users_followers	0.44	0.45	0.41	0.55	–0.21	0.68	0.54	0.55	–0.61	1.00	0.66	0.04	0.10
prior_total _tweets	0.59	0.55	0.66	0.92	–0.13	0.65	0.78	0.92	–0.62	0.66	1.00	0.06	0.03
Pre_SentiDraw _ Ratio	0.38	0.19	0.21	0.06	–0.41	0.32	0.39	0.06	–0.02	0.04	0.06	1.00	0.53
Post_SentiDraw _Ratio	0.25	0.18	0.08	0.02	–0.34	0.33	0.25	0.02	–0.15	0.10	0.03	0.53	1.00

variables have a high correlation between them. Table 12 also confirms the same using variable inflation factor (VIF) which is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least square's regression analysis. Table 12, when considered along with correlations among independent variables, cautions usage of features like number of distinct users and number of tweets together for building the OLS regression model. Due to these issues and presence of several features, only two variables – 'Budget' and 'Opening Day' revenues – appear to have a significant relationship with the dependent variable.

Experiment 2, which takes in budget along pre-hype, total volume of tweets, and ratio of positive to negative tweets based on scoring done using SentiDraw proves to be a robust model for prediction of movie revenues. With adjusted R^2 at 0.79 and MAE for training and test set at 13.4 and 14.7 respectively, the performance of prediction task is only slightly inferior compared to Experiment 1 which estimates lifetime revenues post release after taking in opening day collection as well as one of the variables along with bunch of other post release tweet related variables as shown in Table 10. Performance drops significantly for Experiment 3, highlighting the impact of budget as a key variable. Ratio of positive to negative tweets, as determined by SentiDraw, exhibits a strong relation with lifetime box office revenues in both cases with higher positive ratio having a positive impact of the success of a movie at the box office. The *P*-value shows that budget, pre-hype, and pre_sentidraw_ratio have a

Table 12. ***Experiment 1 : VIF Factor Across Features to Diagnose Multicollinearity***

Features	VIF Factor
Budget	67.07
post_distinct_users	inf
Post_Hype	32.46
post_distinct_users_followers	24.77
post_distinct_users_tweets	10.02
post_total_tweets	228.87
prior_distinct_users	inf
Pre_Hype	31.02
prior_distinct_users_followers	11.82
prior_distinct_users_tweets	62.39
prior_total_tweets	478.18
Pre_SentiDraw_Ratio	10.96
Pre_SentiNet_Ratio	5.85
Post_SentiDraw_Ratio	11.80
Post_SentiNet_Ratio	4.99

Table 13. ***Experiment 2 & 3 : Comparison of Independent Variables and Their Impact***

	Experiment 2			Experiment 3		
	Coef	Std Eror	*P*>|*t*|	Coef	Std Eror	*P*>|*t*|
const	86.17	4.73	0.000	79.03	5.97	0.000
Budget	50.85	7.85	0.000			
Pre_Hype	–31.38	7.83	0.000	–48.24	9.58	0.000
prior_total_tweets	9.58	8.61	0.270	29.95	10.39	0.005
Pre_SentiDraw_Ratio	22.78	5.41	0.000	30.07	6.86	0.000

significant impact on the movie box office performance (Table 13). This demonstrates the viability of the SentiDraw method and the lexicon built using star rating of Bollywood reviews as a robust lexicon method for polarity determination of tweets which can significantly improve box office prediction of movies even before the release of a movie. Pre-hype as an independent variable is also a strong predictor of the outcome, although the relationship with revenues is inverse, suggesting that lesser number of distinct users producing more tweets helps fortunes of a movie more positively.

Figure 5 depicts the true versus predicted lifetime box-office for both training and test sets and the accuracy of prediction is evident from the Figure. From Table 14 that summarizes the VIF factor across features, we can conclude that the independent variables in Experiment 2 do not exhibit any multicollinearity. The test for normal distribution and randomness of residues, when plotted against the dependent variable, shows that these important assumptions (normal distribution and absence of heteroscedasticity) hold true for these models. Further detailed investigation on the impact of each variable on the model (refer to Figure 6) reveals that all four variables used in Experiment 2 produce random distribution of residues suggesting that there is little or no unexplained error directly impacting the effect of these variables on the outcome. Figure 7 shows the correlation between all the

Figure 5. ***Experiment 2 : True Versus Predicted Lifetime Box - Office for Both Training and Test Sets***

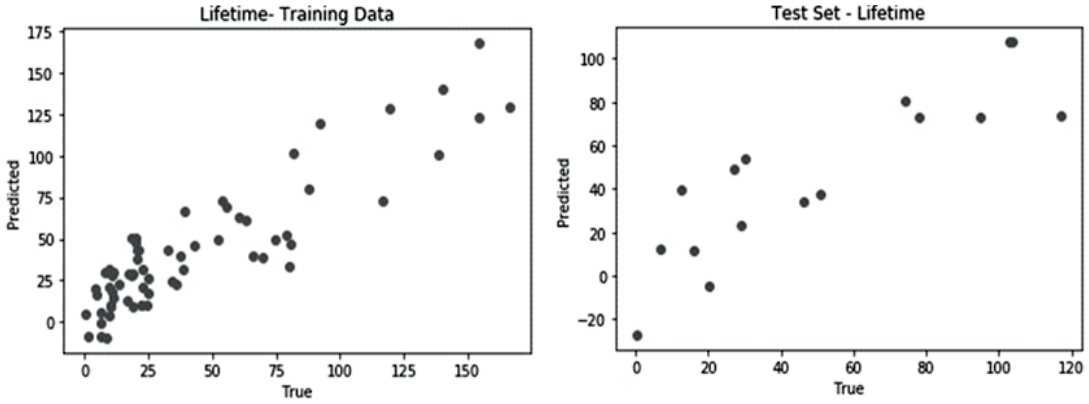
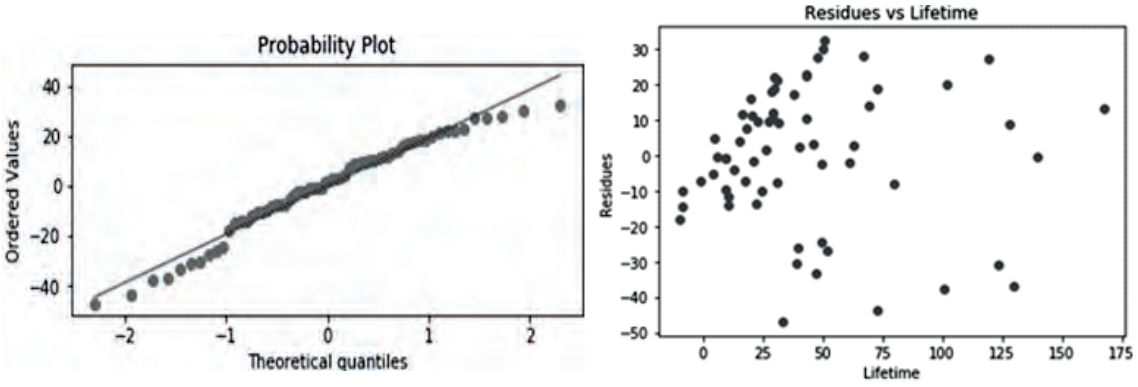


Table 14. ***Experiment 2 : VIF Factor Across Features to Diagnose Multicollinearity***

Features	VIF Factor
Budget	3.97
Pre_Hype	2.18
prior_total_tweets	3.99
Pre_SentiDraw_Ratio	1.40

Figure 6. ***Experiment 2 : Residue Probability Plot to Check for Normality and Random Distribution of Residues with Lifetime Revenues***



variables. Partial regression plots also confirm the correlation findings, demonstrating a very high positive relation between budget and lifetime revenues. Impact of polarity of tweets is not as influential but has a significant positive impact on increasing the performance of prediction as revealed in the partial regression and CCPR plot of polarity ratio (Figure 8), which provides a way to judge the effect of one regressor on the response variable by taking into account the effects of the other independent variables.

Figure 7. ***Experiment 2 : Correlation of all Features and Dependent Variables***

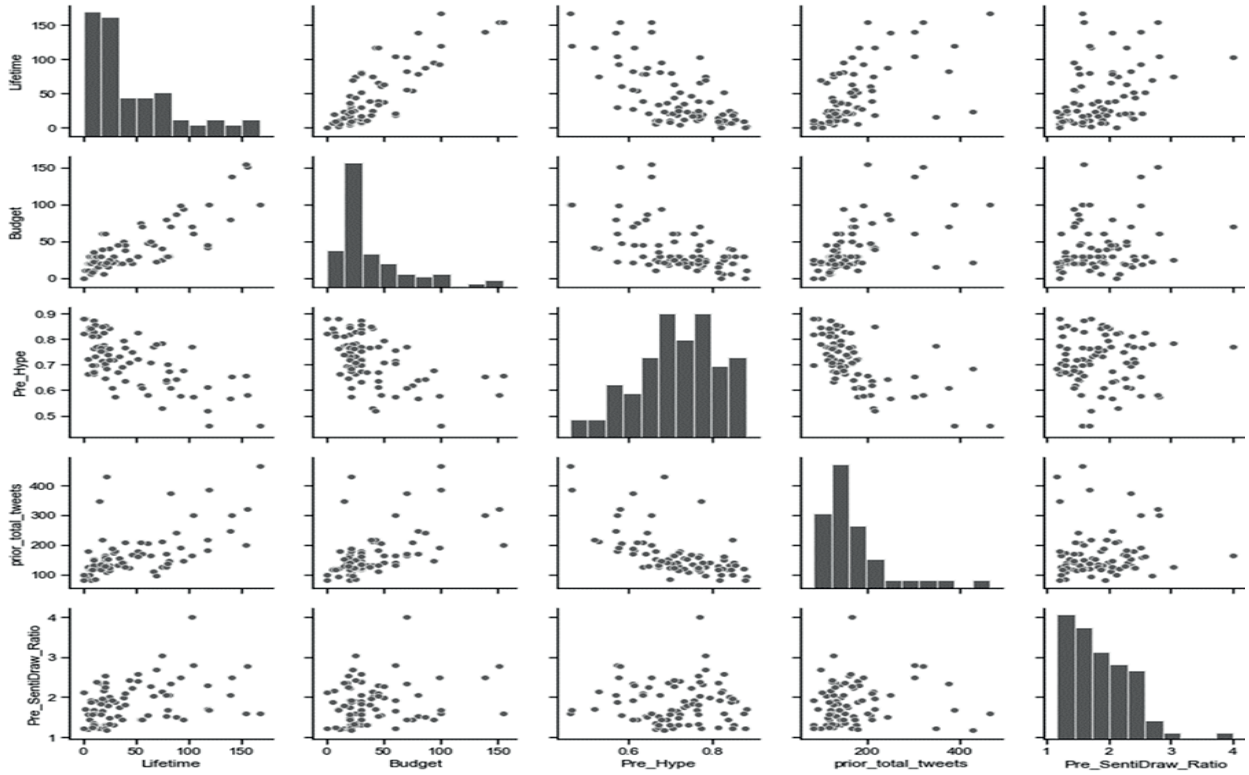
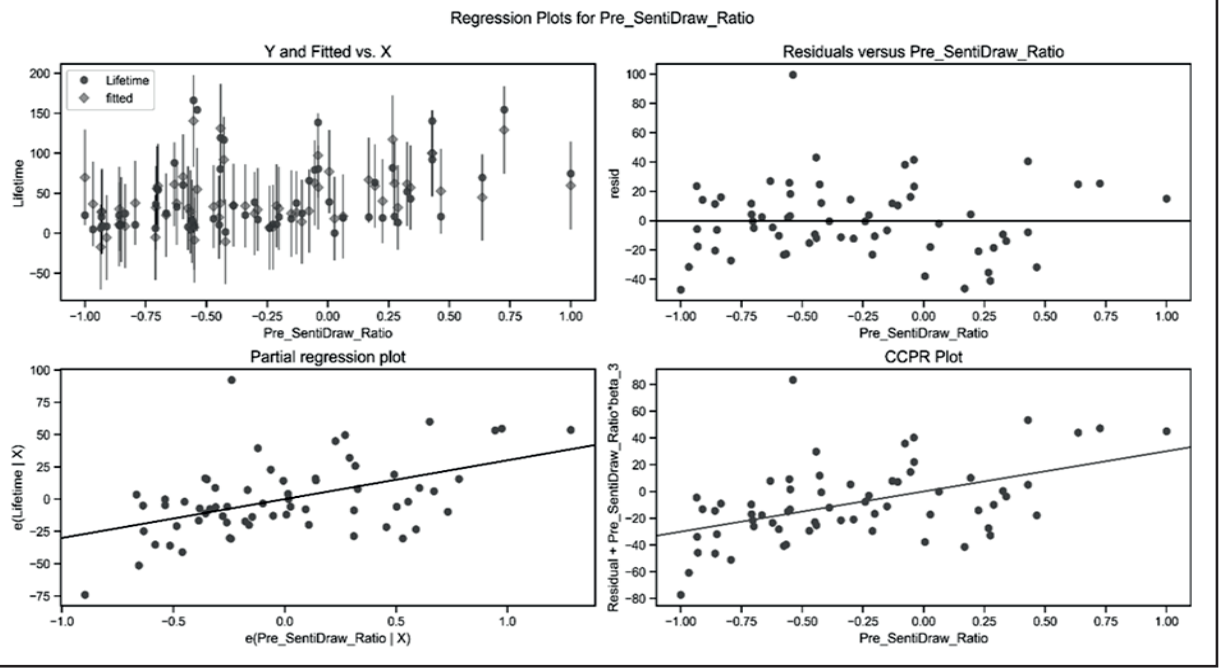


Figure 8. ***Experiment 2 : Exploration of Impact of Positive to Negative Ratio of Tweets as Determined by SentiDraw Lexicon on the Prediction Performance***



Conclusion

Previous studies had not used star rating dispersion of words to arrive at sentiment scores for them. This technique appears to be very likely to produce an effective way to determine sentiment scores of terms as words with higher positive intensity will be likely to appear more often in positive reviews with higher ratings and vice versa. Also, customer reviews data with labelled star ratings are also available readily across different platforms and can provide a steady stream of voluminous data for building domain specific lexicons for assessing WOM around movies or other products using the rating dispersion method enhanced by SentiDraw framework.

This paper attempts the same and tests SentiDraw method for building lexicons for movie domain and tests it against other methods of creating sentiment lexicons. Further, a robust set of experiments are done to assess how polarity determination of tweets can help in predicting box office revenue of Bollywood movies before the release of the movie using as few independent variables as possible to build a parsimonious model for prediction. This study provides evidence that the SentiDraw method has high demonstrated ability for building sentiment lexicons for polarity determination of WoM in domains like movies where reviews labelled with start rating are available for training the lexicon using probability distribution of words across rating point to determine their SO scores. This study also presents an elegant regression model using only a movie's budget along with pre-hype, volume of tweets, and ratio of positive and negative tweets as independent variables that are together able to predict box office revenue of movies with low error even before the release of a movie. This study concludes that the SentiDraw method is a powerful way to create domain specific lexicon for movie domains, which can positively impact the performance on box office prediction when used along with other relevant features.

Managerial and Theoretical Implications

A huge part of movie revenues is generated in the first few weeks after the release of the movies. With such a small window, it is important to predict how well a movie may fare post release as several important managerial decisions are based on it. The distributors may be better off if they could analyze the trend accurately and decide on the number of screens they should keep in the later weeks once they can project the foot fall accurately. The producers of movies can cut down or increase advertisement spends if they feel it can further enhance revenues based on available data. Marketers can make better decisions for allocating funds for movie advertisement inside the theatre as most such decisions are made prior to the release of the movie. Since a lot of online buzz and movie reception can be gauged through social media, using this data to enhance the prediction of box office success of movies can help several stakeholders to take timely monetary decisions that can help them increase possible profits and cut losses in time.

Limitations of the Study and Scope of Further Research

There are certain limitations in the study which may pave way for future work as well. Reliable sources of data on number of screens released and advertisement budget was not available for many Bollywood movies ; so, these variables could not be used. There is a lot of noise in Twitter data and there is possibility that some of the tweets may have suffered from this as even after carefully choosing hashtags, it is not possible to remove all noise without manual intervention. Also, very low volume of tweets for some of the lesser known low budget movies reduced performance prediction for these movies.

Going forward, more experiments can be performed that use sentiment value obtained from SO values from SentiDraw lexicon and apply it in other prediction methods like diffusion model suggested in studies that have achieved a high level of accuracy in prediction of movie reviews using WoM and determine the extent to which

using SentiDraw adds to the overall performance. Also, deep learning algorithms (Narayanaperumal, 2020) are quickly evolving and some of them using recurrent neural networks have been shown to outperform SVM on text classifications task lately, and hybrid models can be included as well (Du et al., 2019 ; Iqbal et al., 2019). Ensemble methods can also be used to increase accuracy of classification and subsequent box office prediction. The SentiDraw method employed above can benefit from use of advanced NLP methods to detect sarcasm. This has been shown to further refine sentiment scores and is worth attempting to compare the results. An interesting result of this study has been that tweets and their polarity are a much more significant predictor or movie success than IMDB ratings for Bollywood movies. This can be investigated further as IMDB ratings have a very high impact on Hollywood movie success. Lastly, Word Sense disambiguation also helps in identifying the meaning of the word in each context. If word sense disambiguation is employed at the time of developing lexicon and used when the sentiment scoring is being done on a given text, the performance may improve further.

Authors' Contribution

Mr. Shashank Sharma conceived the idea and undertook the literature review for finding gaps in the area. Mr. Shashank Sharma also devised the lexicon creation method and deployed the qualitative and quantitative design to undertake the empirical study. Mr. Shashank Sharma also extracted all the data used in the study from the Internet using scraping libraries from Python. Dr. Gautam Dutta verified the analytical methods and supervised the study. The experiments were done by Mr. Shashank Sharma using Python with Sci-kit learn and NLTK libraries. Mr. Shashank Sharma wrote the manuscript in consultation with Dr. Gautam Dutta. Dr. Gautam Dutta refined certain sections of the manuscript like Literature Review and Conclusion.

Conflict of Interest

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter, or materials discussed in this manuscript.

Funding Acknowledgment

The authors received no financial support for the research, authorship, and/or for the publication of this article.

References

Abbasi, A., France, S., Zhang, Z., & Chen, H. (2011). Selecting attributes for sentiment classification using feature relation networks. *IEEE Transactions on Knowledge and Data Engineering, 23*(3), 447 – 462. https://doi.org/10.1109/tkde.2010.110

Almatarneh, S., & Gamallo, P. (2018). Automatic construction of domain-specific sentiment lexicons for polarity classification. In, F. De la Prieta et al. (eds), *Trends in cyber-physical multi-agent systems. The PAAMS Collection - 15th International Conference, PAAMS 2017. Advances in Intelligent Systems and Computing* (Vol. 619). Springer, Cham. https://doi.org/10.1007/978-3-319-61578-3_17

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (Vol. 10, No. 2010, pp. 2200 – 2204). https://doi.org/10.1109/mis.2010.94

Bai, X. (2011). Predicting consumer sentiments from online text. *Decision Support Systems, 50*(4), 732 – 742. https://doi.org/10.1016/j.dss.2010.08.024

Bhāle, S., & Tongare, K. (2018). A conceptual model of helpfulness of online reviews in a blink. *Indian Journal of Marketing, 48*(2), 7 – 22. https://doi.org/10.17010/ijom/2018/v48/i2/121331

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance : Accounting for sequential rollout and aggregation across local markets. *Marketing Science, 29*(5), 944 – 957. https://doi.org/10.1287/mksc.1100.0572

Dastidar, S. G., & Elliott, C. (2019). The Indian film industry in a changing international market. *Journal of Cultural Economics, 44*(1), 97 – 116. https://doi.org/10.1007/s10824-019-09351-6

Dellarocas, C., Zhang, X. (Michael), & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales : The case of motion pictures. *Journal of Interactive Marketing, 21*(4), 23 – 45. https://doi.org/10.1002/dir.20087

Dhir, R., & Raj, A. (2018). Movie success prediction using machine learning algorithms and their comparison. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC),* 385 – 390. https://doi.org/10.1109/icsccc.2018.8703320

Du, Y., Zhao, X., He, M., & Guo, W. (2019). A novel capsule based hybrid neural network for sentiment classification. *IEEE Access, 7,* 39321 – 39328. https://doi.org/10.1109/access.2019.2906398

EY India. (2019, January 14). *The Indian film tourism industry has potential to generate US$3b by 2022* [press release]. https://www.ey.com/en_in/news/2019/01/indian-film-tourism-industry-has-potential-to-generate-usd-3-billion-by-2022

Gatti, L., Guerini, M., & Turchi, M. (2016). SentiWords : Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing, 7*(4), 409 – 421. https://doi.org/10.1109/taffc.2015.2476456

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews : Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498 – 1512. https://doi.org/10.1109/tkde.2010.188

Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. (2019). A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. *IEEE Access, 7,* 14637 – 14652. https://doi.org/10.1109/access.2019.2892852

Jaiswal, S. R., & Sharma, D. (2017). Predicting success of Bollywood movies using machine learning techniques. In, *Proceedings of the 10th Annual ACM India Compute Conference (Compute'17).* Association for Computing Machinery. https://doi.org/10.1145/3140107.3140126

Jiménez-Zafra, S. M., Martin, M., Molina - González, M. D., & Urena - Lopez, L. A. (2016). Domain adaptation of polarity lexicon combining term frequency and bootstrapping. In, *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 137 – 146). Association for Computational Linguistics. https://doi.org/10.18653/v1/w16-0422

Khan, F. H., Qamar, U., & Bashir, S. (2015). SentiMI : Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection. *Applied Soft Computing, 39,* 140 – 153. https://doi.org/10.1016/j.asoc.2015.11.016

Khan, F. H., Qamar, U., & Bashir, S. (2016). Senti - CS : Building a lexical resource for sentiment analysis using subjective feature selection and normalized chi - square based feature weight generation. *Expert Systems, 33*(5), 489 – 500. https://doi.org/10.1111/exsy.12161

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis : Comparative evaluation of six sentiment lexicons. *Journal of Information Science, 44*(4), 491 – 511. https://doi.org/10.1177/0165551517703514

Labille, K., Gauch, S., & Alfarhood, S. (2017, August). *Creating domain-specific sentiment lexicons via text mining. WISDOM' 17*. http://www.csce.uark.edu/~sgauch/5543/F17/notes/wisdom17.pdf

Lee, H., Han, Y., & Kim, K. (2014). Sentiment analysis on online social network using probability Model. In, *AFIN 2014 : Proceedings of the Sixth International Conference on Advances in Future Internet* (pp.14–19). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.671.6392&rep=rep1&type=pdf

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies* (Vol. 1, pp. 142 – 150). https://www.aclweb.org/anthology/P11-1015.pdf

Musto, C., Semeraro, G., & Polignano, M. (2014, December). A comparison of lexicon-based approaches for sentiment analysis of microblog posts. Information Filtering and Retrieval. In, *DART@ AI* IA* (pp. 59–68). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.664.7765&rep=rep1&type=pdf#page=66

Narayanaperumal, M. (2020). *Deep neural networks for sentiment analysis in tweets with emoticons* (Doctoral Dissertation). Nova Southeastern University. https://nsuworks.nova.edu/gscis_etd/1117

Niraj, R., & Singh, J. (2015). Impact of user-generated and professional critics reviews on Bollywood movie success. *Australasian Marketing Journal, 23*(3), 179 – 187. https://doi.org/10.1016/j.ausmj.2015.02.001

Pang, B., & Lee, L. (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. In, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04)*. Association for Computational Linguistics, USA. https://doi.org/10.3115/1218955.1218990

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up ? Sentiment classification using machine learning techniques. In, *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (EMNLP '02)*. Association for Computational Linguistics, USA. https://doi.org/10.3115/1118693.1118704

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count : LIWC 2001.* Mahway : Lawrence Erlbaum Associates.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis : A combined approach. *Journal of Informetrics, 3*(2), 143 – 157. https://doi.org/10.1016/j.joi.2009.01.003

Reddy, A. S., Kasat, P., & Jain, A. (2012). Box - office opening prediction of movies based on hype analysis through data mining. *International Journal of Computer Applications, 56*(1), 1 – 5. https://doi.org/10.5120/8852-2794

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). SentiCircles for contextual and conceptual semantic sentiment analysis of Twitter. In, V. Presutti, C. D'Amato, F. Gandon, M. D'Aquin, S. Staab, & A. Tordai (eds), *The semantic web : Trends and challenges. ESWC 2014. Lecture Notes in Computer Science* (Vol. 8465). Springer, Cham. https://doi.org/10.1007/978-3-319-07443-6_7

Sharma, S. S., & Dutta, G. (2018). Polarity determination of movie reviews : A systematic literature review. *International Journal of Innovative Knowledge Concepts, 6*(12), 43 – 55.

Shaukat, Z., Zulfiqar, A. A., Xiao, C., Azeem, M., & Mahmood, T. (2020). Sentiment analysis on IMDB using lexicon and neural networks. *SN Applied Sciences, 2*(2), 1 – 10. https://doi.org/10.1007/s42452-019-1926-x

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267 – 307. https://doi.org/10.1162/coli_a_00049

Thelwall, M. (2017). The heart and soul of the web ? Sentiment strength detection in the social web with SentiStrength. In, J. Holyst (eds), *Cyberemotions. Understanding complex systems.* Springer, Cham. https://doi.org/10.1007/978-3-319-43639-5_7

Thomas, F. C., & Patel, N. K. (2020). Determining the effectiveness of promotion and reviews of Bollywood films from audiences : An empirical study. *Indian Journal of Marketing, 50*(4), 7 – 24. https://doi.org/10.17010/ijom/2020/v50/i4/151570

Utomo, T. S., Sarno, R., & Suhariyanto. (2018, September). Emotion label from ANEW dataset for searching best definition from WordNet. In, *2018 International Seminar on Application for Technology of Information and Communication* (pp. 249 – 252). IEEE. https://doi.org/10.1109/isemantic.2018.8549769

Venkataraman, N., & Raman, S. (2016). Impact of user-generated content on purchase intention for fashion products : A study on women consumers in Bangalore. *Indian Journal of Marketing, 46*(7), 23 – 35. https://doi.org/10.17010/ijom/2016/v46/i7/97125

About the Authors

Shashank Shekhar Sharma is a part-time Research Scholar at Indian Institute of Foreign Trade (IIFT), Delhi. He is a B. Tech engineer (VIT, Vellore) and an MBA from IIFT, Delhi. He has worked as a Marketing Manager in Nestle and Dabur.

Dr. Gautam Dutta is presently working as a Professor at Indian Institute of Foreign Trade (IIFT) for both Delhi and Kolkata campuses. He is a mechanical engineer with masters in business management. He completed his doctoral degree from Indian Institute of Technology (IIT).