Received 17 April 2015
Revised 8 September 2015
13 September 2015
14 September 2015
Accepted 14 September 2015

Using Twitter data to predict the performance of Bollywood movies

Dipak Damodar Gaikar and Bijith Marakarkandy
Department of Information Technology,
Thakur College of Engineering and Technology, Mumbai, India, and
Chandan Dasgupta
SBM, NMIMS University, Mumbai, India

Abstract

Purpose – The purpose of this paper is to address the shortcomings of limited research in forecasting the power of social media in India.

Design/methodology/approach – This paper uses sentiment analysis and prediction algorithms to analyze the performance of Indian movies based on data obtained from social media sites. The authors used Twitter4j Java API for extracting the tweets through authenticating connection with Twitter web sites and stored the extracted data in MySQL database and used the data for sentiment analysis. To perform sentiment analysis of Twitter data, the Probabilistic Latent Semantic Analysis classification model is used to find the sentiment score in the form of positive, negative and neutral. The data mining algorithm Fuzzy Inference System is used to implement sentiment analysis and predict movie performance that is classified into three categories: hit, flop and average.

Findings – In this study the authors found results of movie performance at the box office, which had been based on fuzzy interface system algorithm for prediction. The fuzzy interface system contains two factors, namely, sentiment score and actor rating to get the accurate result. By calculation of opening weekend collection, the authors found that that the predicted values were approximately same as the actual values. For the movie *Singham Returns* over method of prediction gave a box office collection as 84 crores and the actual collection turned out to be 88 crores.

Research limitations/implications – The current study suffers from the limitation of not having enough computing resources to crawl the data. For predicting box office collection, there is no correct availability of ticket price information, total number of seats per screen and total number of shows per day on all screens. In the future work the authors can add several other inputs like budget of movie, Central Board of Film Certification rating, movie genre, target audience that will improve the accuracy and quality of the prediction.

Originality/value – The authors used different factors for predicting box office movie performance which had not been used in previous literature. This work is valuable for promoting of product and services of the firms.

Keywords Prediction, Sentiment analysis, Twitter, Social media, Social network, Fuzzy inference system

Paper type Research paper

1. Introduction

The emergence of the web and online social media has represented a fundamental shift as it has added new dimensions to the production and dissemination of news and information. Social media is defined as media designed to disseminate information through social interaction which has been created using highly accessible and scalable publishing techniques. Users usually generate content, access information to reach a

Industrial Management & Data Systems
Vol. 115 No. 9, 2015
pp. 1604-1621

0263-5577
DOI 10.1108/IMDS-04-2015-0145

large audience. Social media has replaced the traditional one-way mass media to consumer communication channel with an interactive dialogue, which helps in creation and exchange of user-generated content. Companies analyze social media data to perform analytics and sentiment analysis. The massive growth of online social networks like Twitter, Facebook and other social networking portals have created a need to determine people's opinion and moods. Users can browse information and opinions from diverse sources that help them tap into the crowds while making more informed decisions (Asur and Huberman, 2010).

Social media is a promising link which helps to build connection on social networks, personal information channels and mass media. Social media data in the form of user-generated content on blogs, blog reviews, microblogging like Twitter, discussion forums, different types of social sites, product review and multimedia sharing web sites present many new opportunities and challenges to both producers and consumers of information (Asur and Huberman, 2010). Although mass media have created a new type of marketing and communication that connects the bridge between simple word-of-mouth (eWOM) and ideas which helps business to run in a profitable manner. Posting user feedback on products has become increasingly popular for people to express their opinions and sentiments toward products and services. Analyzing the immense online reviews available, would yield to a database which could be of economic value to producers, vendors and other interested parties. In the movie domain, a single movie can experience sales margins between millions of rupees in profits or losses for a movie house in a particular year. Therefore, that movie studio is immensely involved in forecasting the performance of the upcoming movies (Xiaohui *et al.*, 2010).

Microblogging services in recent times have been a popular communication tool among internet users. It generates millions of daily messages for popular web sites. Due to a free format of messages and an easy accessibility of microblogging platforms for internet users, users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to microblogging services. If more and more users post about products and services which they use or express their political and religious views about them through microblogging web sites, then it becomes very valuable source of understanding public mood. Such data can be efficiently used for marketing or social studies (Pak and Paroubek, 2012).

Microblogging is online eWOM branding like Twitter, is now serving as electronic eWOM, forming a eWOM branding which is based on social networking and trust. Twitter has been swamped with active users during the last few years and much attention has been given in analyzing the social behavior and opinions of users. The wide-spread popularity of online social networks and the resulting availability of data have enabled the investigation of new research questions, such as the analysis and estimation of public opinion on various subjects (Charalampidou, 2012).

People's sentiment toward a particular matter when expressed online, can be very useful in many cases whose classification and estimation arises to a crucial point. The volume of discussion about products on Twitter can be correlated with the product's performance. It is also known that social network users represent the aggregate voice of millions of potential consumers, especially for products designed for the target-group of young-aged technology users. This reveals a brand new aspect that companies should consider closely, and this free and high-scale feedback can give them the opportunity to understand consumer needs and take proper action (Charalampidou, 2012). Additionally, a lot of effort has been made in social media analysis, regarding its power of predicting real-world outcomes. In the recent years, some of the research that

has been already made has shown that the information gained from social networks can be indeed useful to make quantitative predictions on some specific domains (Bindra *et al.*, 2012).

2. Literature review

With the growth of blogs and social networks, opinion mining, sentiment analysis became a field of interest for many researches. The topic of using social media to predict the product future of movies is one of the most popular topics among the masses. Experts are able to come out with substantially accurate predictions about the movie with help of readily available information about the movie and its related revenues. The different authors own research is also based on the field of cinematography. The reason for this is the popularity of movies in society and the many publicly available metrics for success, such as box-office performance.

Seminal work on predicting power of social media was done by (Gruhl *et al.*, 2005). They were able to predict spikes in book sales based on blog posts. Mishne and Glance (2006) were able to predict the box office performance using positive sentiment polarity about the movie.

Asur and Huberman (2010) were the pioneers in using Twitter data for predicting box office revenues of movies. They investigated quantitative and qualitative approaches to analyze large amounts of data from Twitter web sites to predict the collection of the box-office movies. The approach is interesting because it looks both at the volume of tweets and at the sentiment expressed in those tweets. They look first at the pre-launch buzz and try to find a correlation between the number of messages the promotion efforts have generated and the actual box-office success. Second the look at the sentiment expressed by the different tweets posted on Twitter web sites, when bad review will only discourage new visitors from seeing the movie. The volume of tweets is useful and used to build a powerful model for future predicting box-office performance and sentiment analysis can improve the predictions. They used correlations and linear regressions model, relatively simple techniques, to show the relation between their data and the box-office values. Then interesting result from their research is that they managed to obtain better results than by the Hollywood stock exchange, which is an artificial online market and results are valuable. They conclude the social media can be used as a real-world performance indicator.

Leskovec (2011) investigated and presented techniques for social media modeling and sentiment analysis optimization. This paper also investigates methods for data extraction from social media web sites on large scale and discusses modes of replicating and rectifying for the defects arising from incomplete and missing data. Moreover, discussing methods for extracting and show how to quantify and enlarge the impact of media giants on the popularity and attention given to particular content, and to build predictive models.

Vasu Jain (2013) present prediction of movie success using sentiment analysis of tweets social media content contains rich information about people's preferences. An example is that people often share their thoughts about movies using Twitter. We did data analysis on tweets about movies to predict several aspects of the movie popularity. The main results we present are whether a movie would be successful at the box office or not.

Bollen *et al.* (2010) were able to correlate semantic scores obtained from tweets over a ten-month period with the Dow Jones Industrial Average (DIJA). Their research indicated that the mood of the society is an important link to predict the stock market.

They used Granger causality along with a self-organizing fuzzy neural network and were successful in predicting the stock market with an accuracy of 86.7 percent.

Nassirpour *et al.* (2012) used a machine learning framework based on Recursive Auto Enconders. The method used vector space representations for phrases. Their method outperformed the traditional pre-defined sentiment lexica approaches.

O'Connor *et al.* (2010) used Twitter data to predict presidential elections. They used the subjectivity lexicon from Opinion Finder for classifying the tweets as positive or negative. They used this score to calculate a sentiment score. They found that a correlation existed between the presidential approval polls and Twitter sentiment data. Tumasjan *et al.* (2010) found that the count of tweets referring to a party or candidate on Twitter was in tune with the election results and the mean absolute error of the prediction in this case was close to the poll results. Skoric *et al.* (2012) have shown that Twitter data can to some extent be used for prediction of elections. They studied 2011 Singapore General Election and concluded that predictive power also depends on democratic maturity of the country, competitiveness of the election and media freedom.

3. Methodology

The major work done in this paper was to predict the box office movie performance and opening weekend box office collection of any Bollywood movie. We had considered movie-related tweets as an input factor for the experimental purpose, which was to extract information from Twitter. For movie performance, we had used two components, first component was a set of movie tweets from Twitter web sites and second component corresponding to movie actor/actress rating.

The research had investigated the important problem of mining opinions and sentiments from movie reviews, which had predicted the box office movie performance by classifying movie into three categories: hit, flop and average. For calculating box office opening weekend collection, it was predicted by pre-release hype factor, total number of shows per day on all screens and the average price of all tickets per screen. Pre-release hype factor is an important factor as far as estimating the openings of the movie was concerned.

The flow proposed system is shown in Figure 1. Proposed system initially selected any Bollywood movie for prediction before release. The project work was broadly classified into different modules for application development.

3.1 Extracting data from Twitter

We extract only filtered data from Twitter, i.e. data that corresponds to a particular movie before release, which will be the system input. Accessing tweets from Twitter is the primary requirement for building a database to get processed and extract information. It uses Twitter4j Java API (an unofficial third party Java library for the Twitter API) to extract the tweets through authenticating connection with Twitter web sites.

The proposed method is used for extracting Bollywood movie data from Twitter which is shown in Figure 2. Given below are detailed steps that describe the method of extracting tweets from Twitter web sites:

(1) Initially for experimental setups, we had selected movie (in GUI) with a specific hashtag (#) which was related to that particular movie. An online service called HashTags.org was used, that identifies the trending hashtag over the entire Twitter network. This constituted the first means, to identify which keywords should be adopted in order to filter the initial volume of tweets and track other related hashtag on the same topic.

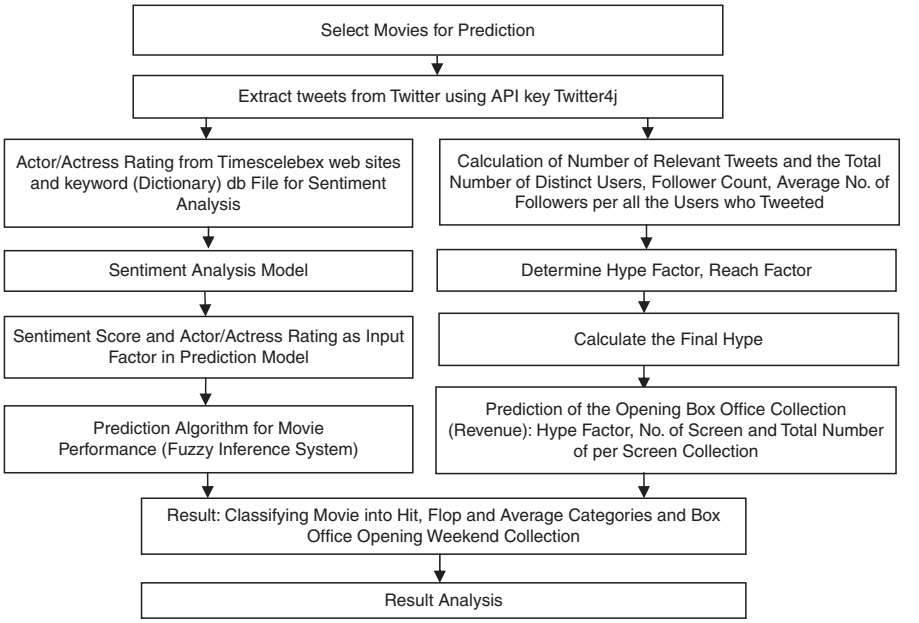

Figure 1.
Block diagram of proposed system

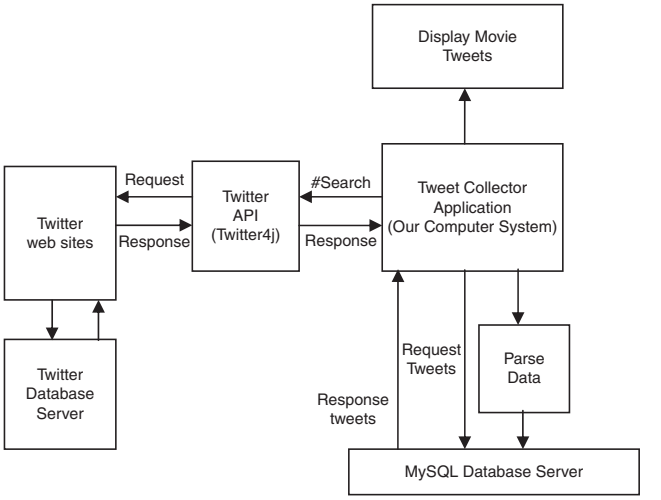

Figure 2.
Fetching data from Twitter

(2) We used Twitter4j API in our application which helped us to connect to Twitter. Twitter4j is a feasible and flexible library for getting connected to Twitter and to communicate with the customer application. For connecting to Twitter, we setup consumer key, consumer secret key, access tokens and access token secret key in application code. Requests to Twitter API are authenticated based on the open authentication 1.0 standard in which two pairs of keys are used in the authentication, namely, consumer key and access token and both can be obtained from the Twitter web site application section.

(3) Twitter sends request to Twitter database server and server responds to tweets within a particular period that is already setup in application.

(4) Extracted tweets were sent to an application and stored in a database in tabular format. Movie tweets and retweets are stored in different database files (i.e. tables).

(5) The user can access movie tweets anytime that are stored in MySQL DB just by sending a request to DB and DB will respond to related movie tweets via application GUI.

(6) The extracted tweets were stored into the database which was used in sentiment analysis, which is shown in Figure 1. Further the downloaded movie database file (excel format) and then that file is used as input in the sentiment analysis classification model.

3.2 Sentiment analysis

Sentiment analysis is an important type of text analysis that supports decision making by extracting and analyzing opining-oriented text to identify positive, negative and neutral opinions. Also it measures how positively or negatively an entity regarded that based on people, organizations, events, locations, products and topics. When different users express their social, religious and political views on Twitter, then their tweets are one of the most valuable sources of people's opinions. Figure 3 shows how sentiment analysis is done and how sentiment score is used for prediction.

To perform sentiment analysis of Twitter data, the Probabilistic Latent Semantic Analysis (PLSA) classification model is used to find the sentiment score in the form of positive, negative and neutral.

Many existing models and algorithms for sentiment mining are developed for the binary classification problem, i.e., to classify the sentiment of a review as positive or negative. However, sentiments are often multi-faceted and can differ from one another in a variety of ways, including polarity, orientation, graduation, etc. Therefore, for applications it is necessary to understand the opinions accurately. Here extraction of ratings starts with modeling sentiments in online reviews, which presents unique challenges that is not possible to be easily addressed by conventional text mining methods by classifying reviews as positive or negative, as most current sentiment mining approaches are designed for not to provide a comprehensive understanding of the sentiments reflected in blog reviews (Pavlou and Dimoka, 2006). To organize the model of a variety of natures of complicated sentiments. Sentiments are analyzed which is embedded in reviews as a result of the combined role of a number of hidden factors.

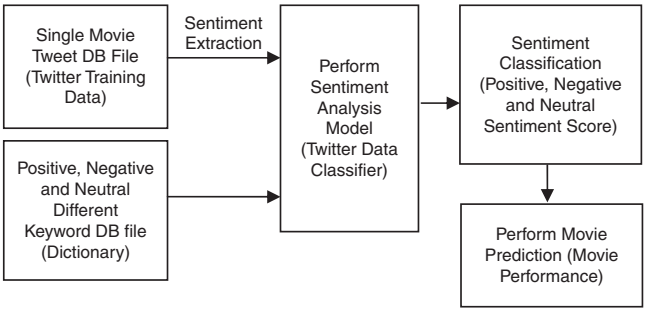

Figure 3.
Tweets sentiment analysis

To evaluate hidden factors which are present in reviews posted by customers. A new approach is used to review mining based on PLSA. The use of hidden factors in PLSA provides the model the ability to accommodate the intricate nature of sentiments with each hidden factor focussing on one specific aspect (Hofmann, 1999). This model will serve as the basis for predicting movie performance. It would be too simplistic to just classify the sentiments expressed in a review as positive, negative and neutral.

The research methodology used here is to train a classifier to classify tweets (reviews written in English for any movie) in the test set as positive, negative and neutral. Using above to perform sentiment analysis on Twitter data the rationale behind this is that for sentiment analysis, sentiment-oriented words, such as good or bad are more indicative than other words. This analysis results in a sentiment score which reflects the overall sentiment, or emotional feeling of the review entered by the customer.

We had implemented a PLSA model in Matlab to train and test our data. Sentiment analysis in Matlab programming is a great language for prototyping ideas. Initially we extracted one single movie database file in CSV format using string concatenation function. Then we imported keyword (Dictionary) database file and did the sentiment analysis of each tweet with all keywords. The dictionary has three subsection column, i.e., positive, negative and neutral words. Using inbuilt Matlab command line function regexpi is used to perform sentiment classification. We scan all rows of the tweet database, read second (tweet text) column of tweets database to read tweets and also read database where keywords (dictionary) are stored. After which we do sentiment analysis of each tweet with all keywords. For each sentence, each word has to be matches with the keyword inside the database file. If there's a match between the word of the sentence and the word of the positive, negative and neutral dictionary, it will return non-zero value and if stored keyword does not exist in tweet then it will give an empty matrix. Using direct function cell2mat in Matlab, it is used to calculate the total number of positive, negative and neutral sentiment scores. Different sentiment analysis score is used for prediction model (fuzzy inference system (FIS)) as input.

3.3 Prediction of box office movie performance and box office opening weekend collection

For prediction of movie box office performance, we used data mining algorithm and FIS. To perform FIS model we use sentiment scores and actor/actress rating as input factor. The movie performance result was classified into three categories: hit, flop and average. We also calculated box office opening weekend collection using hype factor (unique users, followers, average number of followers), number of shows per days in all screens and average price of all tickets per screen per show. The prediction accuracy is measured by using the mean square error (MSE) method.

3.3.1 Prediction using FIS (movie performance). For movie performance using a FIS model we have used two inputs and one output. The first input is positive, negative or neutral sentiment score and the second is actor/actress rating that uses movie performance using FIS prediction model. The final prediction output result will be display hit, flop and average categories. Figure 4 shows working on movie performance prediction algorithm.

For actor and actress rating we have used a zoom web sites name are Times Celebex. That is official Bollywood stars rating web sites in India and worldwide. Zoom (2015) Times Celebex is a rating system that assesses the top Bollywood actors, both male and female, on a monthly basis based on a pre-set point system that determines the "*T* score"

of every selected Bollywood actor/actress. *T* score was calculated by considering the factors like popularity, performance and visibility and box office performance of that indusial actor/actress. *T* score calculation consists of the following terms:

- box office returns and recognition;
- ability to stay in the news across print and TV media;
- visibility through brand endorsements on print and TV media; and
- promotions of their upcoming movie releases on print and TV media.

Popularity among fans across mediums including the internet and social media. Times Celebex is the most robust and factual ratings index for Bollywood celebrities as it is based on comprehensive data and hard facts rather than the subjective perception of a jury alone. Data are sourced from credible external agencies that service the media industry and collated from 60+ publications, 250+ TV channels, 10,000+ cinema halls and millions of users across the internet. Calculated month-on-month, Times Celebex is a dynamic index and hence a very current measure of a celebrity's power as opposed to any other one-time annual report (Zoom, 2015).

A FIS is a system that uses fuzzy set theory to map different inputs to outputs. The FIS will be basically implemented using two models Mamdani and the Sugeno. We use Mamdani model for implementation, using two types of fuzzy logic toolbox, i.e. command line function toolbox and graphical interactive tools (GUI). We used following function to display the result in GUI format: fuzzy(), mfedit(), ruleedit() (Mathworks, 2014).

3.3.2 Movie box office prediction using hype factor (opening weekend collection). The prediction of movie box office opening weekend collection method was based on pre-release hype, total number of shows per day on all screens and the average price of all tickets per screen. Which drives a layman to a theater to watch a movie and this decides the openings of a movie that is the occupancy of theaters playing that particular movie. Different users have their own opinion, which they express through simple tweets. We setup analyzes the opinion mining in these tweets with respect to a movie prior to its release in theaters. Estimate the hype of movie surrounding it and also predict the box office openings of the movie (Reddy *et al.*, 2012) .Working of method used for implementation is shown in Figure 5 and detailed steps are discussed below:

(1) For movie box office opening weekend collection initially we calculated final hype factor. For final hype factor need to find out movie hype factor and movie reach factor.

(2) For initial movie hype factor need to find the total number of relevant tweets, number of distinct users those posted the tweets. The number of distinct users

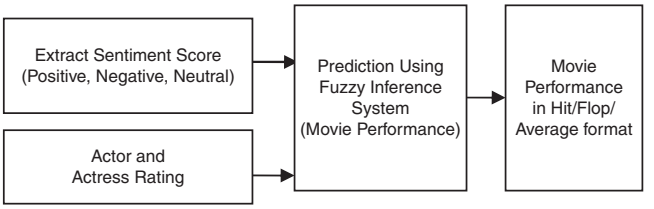

Figure 4.
Prediction using fuzzy inference system

can be calculated by counting user-id of the users. This process starts one before the release of the any movie. The following formula we used for the hype factor (α) calculation (Reddy *et al.*, 2012):

$$\alpha = \frac{\text{Number of distinct users}}{\text{Number of tweets by all users}} \tag{1}$$

(3) For reach factor (σ) we, need follower count of a particular user who referred to the movie in his tweet if the follower count is above a particular threshold value (τ). Where $\tau =$ average number of followers per all tweeted users. σ can be scaled down to a scale of 0.1-1 with 0.1 being assigned to the threshold value, assuming cases where the follower count being more than ten times the threshold value as a rare case and assigning it the value 1. The reach factor can be given as (Reddy *et al.*, 2012):

$$\sigma = \frac{\text{follower count} - \tau}{\text{follower count}} \tag{2}$$

(4) The final hype can be given as:

$$\text{Hype} = \frac{\alpha + \sigma}{2} \tag{3}$$

The hype factor (α) gives values which may be an integer or decimal (Reddy *et al.*, 2012). Therefore, the concept of ratio is used to calculate the hype and thus the chances of movie success. If the value is closer to Number 1 then there are high chances of movie generating high opening collections similarly, if the value is farther from number 1 there are less chances of getting a good opening. As per above methodology, if we get a value closer to number 1, it represents the chances of the movie to be a big hit in box office. The reason to use the ratio-based approach is to get the closest approximate hype which would be difficult had it been without the number of users. Total number of distinct users is an important factor because hype can be best known through the number of users being interested in a particular movie. The success of a movie at the box office can be best determined by the ratio of the number of users to the number of tweets rather than taking only number of tweets in consideration which would not give the best possible approximation.

The opening weekend collection can be calculated using the hype factor and the knowledge of how many screens the movie is going to release in the occupancy of each movie theater is analogous with the hype surrounding the movie. The opening box office collection (O) can be predicted as (Reddy *et al.*, 2012):

$$O = \mu \times \text{Hype} \times \varphi K \tag{4}$$

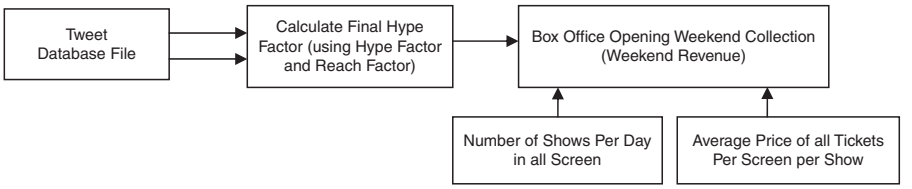

Figure 5.
Movie box office prediction using hype factor (Box Office Revenue)

where O is the opening box-office collection is the number of shows per day on all screens together for the weekend and φ is the average price of all tickets per screen per show.

4. Results

In the current study a total of 10,269 tweets and retweets were extracted for 14 different movies a week before its release as this is a critical period as major promotional campaigns for Bollywood movies begin a week before release. Tweets were extracted for 14 different movies released over a period of six months. The time period for data acquisition was preset in the data extraction application a priori. Movies are normally showcased on Fridays, with some rare exception of movie which are showcased on Wednesday. Since an average of two new movies is released each week, we collected data over a time period of six months from June to December to have sufficient data to measure predictive behavior for consistency. Data about the actual box office collection were obtained from Koimoi a popular web site. Table I gives the list of movie names and date of release that have been used in the study.

4.1 Data search using the specific hashtag

In this study, we extract the tweets using the search term "SinghamReturns" which is the Ajay Devgan film released on August 14, 2014. Table II shows that few tweets searching hashtag word that's used in our application to search *Singham Returns* movie.

4.2 Sentiment analysis result

For sentiment analysis we have used *Singham Returns* movie tweets for experimental purpose. We used PLSA classification model for finding different sentiment score,

Sl No.	Movie name	Release date
1	*Ek Villain*	2014-06-27
2	*Kick*	2014-07-15
3	*Entertainment*	2014-08-15
4	*Singham Returns*	2014-08-15
5	*Mardaani*	2014-08-22
6	*Desi Kattey*	2014-08-29
7	*Raja Natwarlal*	2014-08-29
8	*Dawaat-e-Ishq*	2014-09-05
9	*Mary Kom*	2014-09-05
10	*Finding Fanny*	2014-09-12
11	*Khoobsurat*	2014-09-19
12	*Happy New Year*	2014-10-23
13	*Ungli*	2014-11-28
14	*PK*	2014-12-19

Table I. Names and release dates for the Bollywood movies

Movie name	Movie release date	Top hashtag	Other hashtag
Singham Returns	August 15, 2014	SinghamReturns	AataMajhiSatakli AjayDevgan

Table II. Movie data search using specific hashtag

which in the form of positive, negative and neutral. Table III shows that the total number of tweets used for experiment. This also shows positive, negative and neutral score which are calculated using classifier model. The sentiment analysis score will be used in FIS algorithm as input.

4.3 Box office movie performance using FIS model

The result of box office movie performance we used FIS prediction algorithm. We used sentiment score and actor/actress rating as input factor and then we converted it into a different fuzzy set and the output was displayed in different categories: hit, flop and average. We used fuzzy logic toolbox command line function in Matlab. The result was displayed using a GUI function that has been explained below.

4.3.1 Actor/actress rating training data set. The second input was actor/actress rating of particular movie which is collected from the Times Celebex web sites. The actor/actress rating was taken form month when movie was released. Table IV show that movie names and particular month ratings of actor/actress (Zoom, 2015).

In this research we categories rating from superstar to flop and rating range is from 100-0 for FIS that is considered as fuzzy set. Table V shows actor/actress category and rating range.

4.3.2 FIS model implementation

(1) FIS editor result: the FIS editor for movie prediction handles the high-level issues for the systems like how many input and output variables, what are their names. The fuzzy logic toolbox does not limit the number of inputs.

Table III.
Total number of tweets and positive, negative, neutral sentiment score

Category	Sentiment score
Positive	298
Negative	35
Neutral	6
Total number of samples (tweets)	263
Total number of samples (retweets)	774

Table IV.
Actor/actress rating

Movie name	Actor/actress name	Rating
Ek Villain	Riteish Deshmukh	55
Kick	Salman Khan	77
Entertainment	Akshay Kumar	45
Singham Returns	Ajay Devagan	65
Mardaani	Rani Mukharji	55
Desi Kattey	Sunil Shetty	35
Raja Natwarlal	Imran Hashami	45
Dawaat-e-ishq	Aditya Kapoor	50
Mary Kom	Priyanka Chopra	46
Finding Fanny	Deepika Padukone	45
Khoobsurat	Sonam Kapoor	33
Happy New Year	Sharukh Khan	68
Ungli	Imran Hashami	34
Pk	Aamir Khan	62

Figure 6 shows the two input variables as sentiment score and actor/actress rating and only one output variable as movie prediction. We need to write fuzzy word in command line for displaying FIS editor.

(2) Membership function editor: the membership function editor is used to define the shape of all the membership functions associated with each variable. Here all the membership functions used are of triangular type and trapezoidal. The membership functions for input score and actor/actress rating are defined in separate membership function editor. The membership function for output parameter coefficient of movie prediction in this case is also defined in separate membership function editor. The entire two separate membership function for input and output variables has a different range which depends on sentiment score vs final prediction. For displaying MF editor need to write, mfedit word in the command line function used.

(3) Rule editor: rule editor is used for adding the list of rules that defines the behavior of the system. These are basically computer programming like "if then structure." Here three rules are used to define the behavior of the box office performance. In the rule editor shown above, we can increase or decrease the rule which depends upon the requirement of the result. We can delete, add and change the rules in the rule editor.

4.4.1 Predicted movie performance vs actual movie performance. Using fuzzy inference model shows, movie prediction result in hit, flop and average categories. For comparison of predicted result, blogtobollywood.com and koimoi.com web sites were referred. Table VI shows actual vs predicted movie box office performance (verdict). We also used Internet Movie Data Base (IMDB) rating for comparison with final result that is shown in Table VI (Brook, 2006).

Sl No.	Actor/actress category	Rating range
1	Super Star	100-60
2	Star	59-40
3	Flop	39-0

Table V. Actor/actress category and rating range

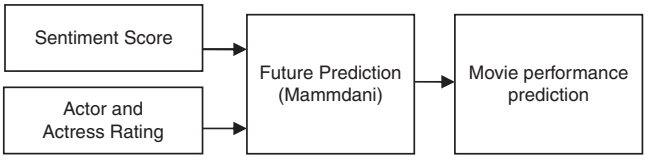

Figure 6. Mamdani-fuzzy system for the proposed model

Movie name	Release date	IMDB rating (out of 10)	Actual box office performance (komoi.com)	Movie performance (prediction using FIS model)
Singham Returns	2014-08-15	6.3	Hit	Hit

Table VI. Predicted vs actual movie performance

4.5 Prediction of weekend opening box office collection using hype factor

For predicting the opening weekend earnings of a Bollywood movie we use following points:

(1) For experimental result, we consider extracted tweets that are search using term "SinghamReturns" which is the Ajay Devgan film released on August 14. Below are few sample tweets that are extracted using hashtag *SinghamReturns*:

@SinghamReturns I'll joined with #SinghamReturnsTomorrow and also enjoy a nation anthem in theatre.

@ajaydevgn you are too good in #SinghamReturns! No one can do action scenes like you ‼ #ekdum #Rockstar.

@KareenaMania totally gonna watch #SinghamReturns it's gonna be a blockbuster can't wait to see #kareenakapoorkhan on the big screen.

Singham Returns: Advance Booking Starts In Full Swing www.koimoi.com/box-office/singham-returns-advance-booking-starts-in-full-swing/ â€¦ via @koimoi.

Every body talking abt #SinghamReturns must say @ajaydevgn's biggest blockbuster till date. #SinghamReturnsIn2Days.

(2) Using collection of tweets, we calculated final hype factor based on Equation (3). The model for predicting the opening weekend box office collection of Bollywood movie *Singham Return*, based on the hype factor and other factor using Equation (4). We extracted tweets approximately a month and half prior to the release. A more accurate result can be obtained on performing the analysis a week prior to the release.

(3) Through this collection of tweets the numbers of relevant tweets were calculated to be 27, tweeted by 22 distinct users having an average follower count of 93 which can be taken as the thresh-hold. Hence, Equation (1) is used for calculating the hype factor (α):

$$\alpha = \frac{208}{263} = 0.7908$$

using Equation (1)

Considering a user with 114 followers, the value of σ is obtained using Equation (2):

$$\sigma = \frac{114-93}{114} = 0.79$$

using Equation (2)

Hype is then given as:

$$\text{Hype} = \frac{115+20}{2} = 0.89$$

using Equation (3)

(4) With this data the film and the assumption that the film releases in 3,600 screens with 85,000 being the approximate mean full house collection of each

screen, the film is estimated to earn:

$$O = 0.8935 \times 10{,}800 \times 85{,}000$$

using Equation (4)

$$O = 820{,}233{,}000$$

in the opening weekend

This model predicted the opening weekend collection at the box office for the Bollywood movie *Singham Returns*, based on the hype factor calculated approximately a month and half prior to the release. A more accurate result can be obtained on performing the analysis a week prior to the release.

(5) Table VII shows that the *Singham Returns* movie predicted opening weekend box office collection and actual box office collection (data used in web sites).

(6) Test data: we also calculated MSE using four different movies, including *Singham Returns*. The different results are compared with the help of training errors. Here the employed errors are the MSE of the input data set. If *Yt* is the actual observation for time period *t* and *Xt* is the forecast for the same period. In Table VII, predictor values are computed by using hype factor and other input values. The formula used for calculation of *MSE* is as shown in Equation (5). Where *N* is number of days of weekend box office collection, *Yt* is the actual observation and *Xt* is the forecast, in this case we have taken *N* as three days Table VIII shows mean square value based on t actual vs predicted box office collection:

$$MSE = \frac{1}{N}\sum (Y_t - X_t)^2 \tag{5}$$

X_t is the box office collection predicted using hype factor, number of shows per day at all screens taken together for the weekend and average price of all tickets per screen

Sl No.	Movie name	Release date	IMDB rating (out of 10)	Movie budget (in crores)	Box office opening weekend collection (predicted in crores)	Box office opening weekend collection (actual) in crores
1	*Singham Returns*	2014-08-15	6.3	50	82.02	77.65

Table VII. Predict vs actual box office opening weekend collection

Sl No.	Movie name	(*Xt*) predicted box office collection in crores	(*Yt*) actual box office collection in crores	Mean square error
1	*Kick*	95.00	86.00	27.00
2	*Singham Returns*	82.02	77.65	6.36
3	*Happy New Year*	121.70	113.86	20.48
4	*PK*	95.00	104.24	17.86

Table VIII. Actual vs predicted value and mean square error value

per show and Y_t is the actual box office collection. Figure 7 shows that MSE value of movie box office weekend collection using prediction method. In this graph, y axis shows error value (in crores of rupees) and x axis shows serial number of movie number as per Table VIII. Also Figure 8 shows actual vs predicted movie collection graph based on Table VIII values.

5. Conclusion

In earlier research historical data had been often used to predict the results of movie performance. In this research the box office collection was predicted prior to the release of the movie with help of social media data. In this study sentiment information mined from current movie tweets from Twitter was used for predicting movie performance. In this study, prediction result is calculated on the basis of FIS model that is applied on Bollywood movies which are released in the year 2014. FIS output had been compared with actual category of the movie using movie-based web sites. A set of experiments conducted on movie data sets confirmed the accuracy and effectiveness of the model. It was found that the predicted results are approximately same as the actual results. In this research tweets were also used for calculating hype factor for predicting opening weekend box office collection of movie before its release. The prediction accuracy is measured by using MSE method.

The outcome of the proposed model leads to a fruitful knowledge that can be readily used by movie distributors to make critical business decisions better.

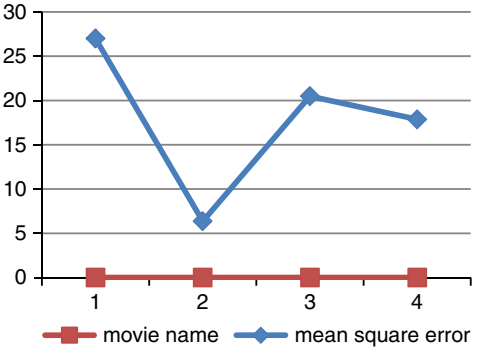

Figure 7.
Means square error of final movies box of weekend collection using prediction method

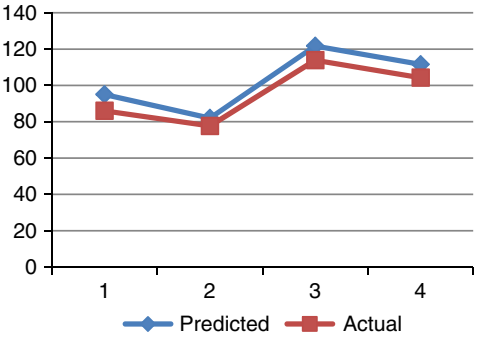

Figure 8.
Actual vs predicted movie collection graph

Individuals usually decide on watching a movie due to its positive WOM. In the current work an attempt has been made to predict box office collection prior to release of a movie. The findings of our research indicate that prerelease sentiment and hype factor play an important role in the success and box office collection. The insights gained from this research can be used to develop a prototype system which may be useful to marketers in course correction of marketing campaigns to garner positive sentiments before release of the movie.

6. Limitations of the study and scope for further work

Our study has some limitations. The current study suffers from the limitation of not having enough computing resources to collect the data. For predicting box office collection, there is no correct availability of ticket price information, total number of seats per screen and total number of shows per day on all screens. In this study only input factors are used for movie prediction but other factors like box office revenue, actor, actress, director, banner and producer also determine success of a movie. In the future researchers can enhance the current system can by considering few more inputs like budget of the movie (production budget and promotional budget), movie genre, Central Board of Film Certification rating, targeted audience of the movie as other independent variables to get a more accurate prediction.

References

Asur, S. and Huberman, B.A. (2010), "Predicting the future with social media", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, *Vol. 1*, pp. 492-499.

Bindra, G.S., Kandwal, K.K., Singh, P.K. and Khanna, S. (2012), "Tracing information flow and analyzing the effects of incomplete data in social media", *IEEE Fourth International Conference*, pp. 235-240.

Bollen, J., Mao, H. and Zeng, X.J. (2010), "Twitter mood predicts the stock market", 1010.3003, *Conference on Artificial Intelligence*, *October*, pp. 1-8.

Brook, D. (2006), "Online database websites for movies, television, and video games", available at: www.imdb.com (accessed April 5, 2014).

Charalampidou, K. (2012), "Estimating popularity by sentiment and polarization classification on social media", doctoral dissertation, TU Delft, Delft University of Technology, Delft.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005), "The predictive power of online chatter", *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 78-87.

Hodeghatta, U.R. (2013), "Sentiment analysis of Hollywood movies on Twitter", *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1401-1404.

Hofmann, T. (1999), "Probabilistic latent semantic analysis", *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 289-296.

Leskovec, J. (2011), "Social media analytics: tracking, modeling and predicting the flow of information through networks", *ACM 22nd International Conference on World Wide Web*, pp. 277-228.

MATHWORKS (2014), "Fuzzy logic toolbox: building a fuzzy inference system", the Math Works Inc.", Natick, MA, available at: http://in.mathworks.com/products/fuzzy-logic (accessed April 20, 2013).

Mishne, G. and Glance, N. (2006), "Leave a reply: an analysis of weblog comments", third annual workshop on the Weblogging Ecosystem, Edinburgh, May 22-26.

Nassirpour, S., Zargham, P. and Mahalati, R.N. (2012), "Electronic devices sales prediction using social media sentiment analysis".

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010), "From tweets to polls: linking text sentiment to public opinion", *Time Series. ICWSM*, Vol. 11, pp. 122-129.

Pak, A. and Paroubek, P. (2010), "Twitter as a corpus for sentiment analysis and opinion mining", *Proceedings of LREC*, pp. 1320-1326.

Pavlou, P.A. and Dimoka, A. (2006), "The nature and role of feedback text comments in online marketplaces: implications for trust building, price premiums, and seller differentiation", *Information Systems Research*, Vol. 17 No. 4, pp. 392-414.

Reddy, A.S.S., Kasat, P. and Jain, A. (2012), "Box-office opening prediction of movies based on hype analysis through data mining", *International Journal of Computer Applications*, Vol. 56 No. 1, pp. 1-5.

Skoric, M., Poor, N., Achananuparp, P., Lim, E.P. and Jiang, J. (2012), "Tweets and votes: a study of the 2011 Singapore general election", *System Science (HICSS), 2012 45th Hawaii International Conference on IEEE*, pp. 2583-2591.

Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M. (2010), "Predicting elections with Twitter: what 140 characters reveal about political sentiment", *ICWSM*, Vol. 10, pp. 178-185.

Vasu Jain, V. (2013), "Prediction of movie success using sentiment analysis of tweets", *The International Journal of Soft Computing and Software Engineering*, Vol. 3 No. 3, pp. 308-313.

Xiaohui, Y., Liu, Y., Huang, X. and An, A. (2012), "Mining online reviews for predicting sales performance: a case study in the movie domain", *Knowledge and Data Engineering, IEEE Transactions*, Vol. 24 No. 4, pp. 720-734.

Zoom (2015), "Zoom times celebex is official Bollywood stars rating websites in India and worldwide", available at: http://timescelebex.com (accessed December 11, 2014).

Further reading

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011), "Sentiment analysis of Twitter data", *Proceedings of the Workshop on Languages in Social Media*, pp. 30-38.

Apala, K.R., Jose, M., Motnam, S., Chan, C.C., Liszka, K.J. and de Gregorio, F. (2013), "Prediction of movies box office performance using social media", *Advances in Social Networks Analysis and Mining IEEE/ACM International Conference*, pp. 1209-1214.

Broniatowski, D.A. (2012), "Extracting social values and group identities from social media text data", *IEEE Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10 No. 5, pp. 557-570.

Choi, S.J. and Jeong, O.R. (2013), "SNS information extraction for social search", *Information Science and Applications International Conference on*, pp. 1-2.

Doshi, L.L.P. (2010), *Using Sentiment and Social Network Analyses to Predict Opening Movie Box Office Success*, Department of Electrical and Computer MIT, Cambridge, MA.

Dung, N.V. (2013), *A Framework to Analyse and Visualise Public Sentiment Using Twitter Data*, University of St Andrews School of Computer Science, Fife.

Georgiou, A. (2013), "Are TV Ratings Possible with Twitter?", Department of Computer Science University of Bristol, Bristol.

Jang, J.S. (1993), "ANFIS: adaptive-network-based fuzzy interface system", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 23 No. 3, pp. 665-685.

Kumar, S., Nikumbh, P.J. and Anuradha, G. (2012), "S-ANFIS: sentiment aware adaptive network-based fuzzy interface system for predicting sales performance using blogs/reviews", *International Journal of Multidisciplinary in Cryptology and Information Security*, Vol. 2 No. 1, pp. 22-23.

Liu, B. (2012), "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies*, Morgan and Claypool, University in Toronto, Ontario, pp. 1-167.

Liu, Y., Huang, X., An, A. and Yu, X. (2007), "ARSA: a sentiment-aware model for predicting sales performance using blogs", *Proceeding 30th Annual International ACM SIGIR Conference Re-search and Development in Information Retrieval (SIGIR)*, pp. 607-614.

MathWorks, Inc. and Wang, W.C. (2001), *Fuzzy Logic Toolbox: For Use with MATLAB: User's Guide*, The Mathworks, Natick, MA.

Mekhilef, S. and Borhanazad, H. (2014), "Fuzzy logic systems by Matlab", IEEE Malaysia Section.

Ngai, E.W.T., Moon, K.-l.K., Lam, S.S., Chin, E.S.K. and Tao, S.S.C. (2015), "Social media models, technologies, and applications: an academic review and case study", *Industrial Management & Data Systems*, Vol. 115 No. 5, pp. 769-802.

Shahheidari, S., Dong, H. and Bin Daud, M.N.R. (2013), "Twitter sentiment mining: a multi domain analysis", *Complex, Intelligent, and Software Intensive Systems 7th IEEE International Conference*, pp. 144-149.

Tang, J., Wang, T. and Wang, J. (2008), "Information flow detection and tracking on web 2.0 blogs based on social networks", *IEEE 9th International Conference for Young Computer Scientists Principles*, pp. 1664-1670.

Tsagkias, M. (2012), "Mining social media: tracking content and predicting behavior", PhD thesis, University of Amsterdam, Amsterdam.

Twitter (2014), "Twitter is an online social networking service websites", available at: https://twitter.com (accessed August 10, 2013).

Valentine, M.M., Kulkarni, V. and Sedamkar, R.R. (2013a), "Fuzzy based SR-ANFIS model for predicting sales performance in movie domain", *International Journal of Advanced Computing (IJAC)*, Vol. 5 No. 3, pp. 81-88.

Valentine, M.M., Kulkarni, V. and Sedamkar, R.R. (2013b), "A model for predicting movie's performance using online rating and revenue", *International Journal of Scientific and Engineering Research*, Vol. 4, pp. 277-283.

Yu, X., Liu, Y., Huang, X. and An, A. (2012), "Mining online reviews for predicting sales performance: a case study in the movie domain", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24 No. 4, pp. 720-734.

Zhang, Y. and Pennacchiotti, M. (2013), "Predicting purchase behaviors from social media", *ACM 22nd International Conference on World Wide Web*, pp. 1521-1532.

Zhou, X., Tao, X., Yong, J. and Yang, Z. (2013), "Sentiment analysis on tweets for social events", *IEEE 17th International Conference on Computer Supported Cooperative Work in Design*, pp. 557-562.

Corresponding author
Dipak Damodar Gaikar can be contacted at: dipgaikar@gmail.com