International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Classification and Forecasting of Bollywood Movies by Commercial Success using Back-Propagation Neural Network model

Partha Shankar Nayak

Abstract*: Forecasting commercial success of motion pictures remained challenging for producers, critics and other industry leaders in this changing world of web and online media. In this study, the author has explored a back-propagation neural network model with 23 numeric input (BPNN-N23) for classification of Bollywood movies released during the years 2014 through 2017. The proposed model classifies movies in three classes namely "HIT", "AVERAGE" and "FLOP". Common procedures like data filtering, data cleaning and data normalization have been followed prior to feeding those data to the neural network. After comparing the performance of the proposed model with the benchmark models and works, the results show that the said model shows performance that is comparable to the published ones with respect to the assumed Indian empirical settings. This research reveals the extent of the effects and roles of the considered factors as well as the proposed model in predicting the fate of a Bollywood movie in India.

***Keywords*: Artificial Neural Network, Back-propagation, Movie Review, Sentiment Analysis, Bollywood, SMOTE.**

I. INTRODUCTION

Indian film industry is globally the largest among its kind in terms of number of film production. At the end of 2010, it was reported that, in terms of total number of annual film production, India ranks first, followed by Hollywood and China [1]. In 2009, there was a total production of 2961 films in India on celluloid among which a staggering figure of 1288 feature films were present. The Indian media and entertainment industry is predicted to reach Rs 2260 billion by 2020 at a CAGR (Compound Annual Growth Rate) of 14.3% [2]. Relevant websites, social media and news data are the major and effective data-banks for collecting data for forecasting movie grosses [3]. Reviews generally start to appear on websites and social media after one week of release of a movie. Critic reviews by eminent movie analyzers are presented after a few days more though pre-release critic reviews are also available sometimes [4].

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Partha Shankar Nayak*, M. Tech, Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur PSN Academy, Bhadreswar, India. Email: psnayak.it.iiests@gmail.com.

A good farsightedness from the perspective of future status and circumstances has the ability of assisting a business in taking effective and corrective measures in advance [19].

AI (Artificial Intelligence) and Machine Learning are being adopted in almost all business organisations to achieve growth and profitability [11]. Likewise, these technologies have become an inevitable wing in the movie industry. Since reviews also come in text format besides numeric ones, text classification is an important part of research that is an essential feature in today's marketing strategy [5]. On the Internet, the reviews by the customers and buyers play a major role in assessing the quality of a product or service. Several research papers have been published on prediction of box office success of movies [6-9, 20-29]. In this paper, the author presents a 2 layered back-propagation neural network model with 23 numeric inputs (BPNN-N23) developed for prediction of the success status among 3 success classes. Data on the movies released during the years 2014 through 2017 had been collected from relevant websites [10-14]. The present paper is organized as follows. Section 2 of this paper describes the existing research works in this field. Section 3 mentions about the sources of data, data filtering and preparation procedures. Section 4 describes the proposed back propagation neural network model under consideration. Section 5 presents the results of the investigations with the said model and discussion on those results. Section 6 presents the possible future scope and further exploration of the current research work. This paper ends with a conclusion in Section 7.

II. LITERATURE REVIEW

Nagamma, et al (2015) [20] used fuzzy-clustering technique for their autoregressive model for isolating the outlier reviews. They performed their analysis on 165 reviews of the movie titled "Annabelle" for predicting the revenue collection for the current day. They also applied SVM for predicting the increment/increment in sales. Their experiment achieved an accuracy of 62% with SVM (Support Vector Machine) for classification without fuzzy clustering and 89.65% with fuzzy clustering. With Naïve Bayes classifier, they achieved an accuracy of 72.41%.

Ouyang, Zhou, Li and Liu (2015) [21] developed a 7 layer model on CNN (Convolution Neural Network) with conjunction to Word2Vec.

Their framework has 3 pairs of convolution and pooling layers with normalization and dropout technology with the activation function as PReLU (Parametric Rectified Linear Unit).

The source of their dataset was www.rottentomatoes.com. They performed comparison task with Naïve Bayes, SVM, BiNB (Naïve Bayes with bag of bigram features), VecAvg (Averages neural word vectors), RNN (Recursive Neural Network) and MV-RNN (Matrix Vector RNN) where their model achieved a test accuracy of 45.4%.

Joshi & Tekchandani (2016) [22] performed an analytical study on the 17000 Hindi movie reviews from Twitter with unigram, bigram and hybrid (unigram + bigram) features among Naïve Bayes, SVM and Maximum Entropy. In their study they found that SVM showed the best performance with 84% accuracy for classification of data.

Rhee & Zulkernine (2016) [4] proposed a model of ANN with back-propagation algorithm to predict the revenue of movies. The sources of data were IMDb, Metacritic, OpusData and Rotten Tomatoes. Their model has 14 input variables, one hidden layer with 25 nodes and an output layer with 2 nodes. Their model achieved an overall accuracy of 88.8%. SVM was also considered for their study where 100% hit movies were properly classified and 87.9% flop movies were misclassified.

Gaikar and Marakarkandy (2016) [29] used FIS (Fuzzy Inference System) for determining the box-office collection of Bollywood movies. They extracted 10269 tweets for 14 Bollywood movies released between June 2016 and December 2016 on which they applied PLSA (Probabilistic Latent Semantic Analysis) to compute the sentiment score. This score and the Actor/Actress score were used as input to the FIS. The resulting box-office collections were compared to the actual ones of 4 selected movies with MSE (Mean Squared Error) values ranging between 6.36 and 27.00.

Tripathy et al (2017) [23] conducted a study on sentiment analysis of movie reviews from IMDb and Polarity dataset. A hybrid system of SVM and ANN has been used for higher accuracy of results in which SVM did the job of selecting features, which are then used as input to ANN. Instead of using popular Word2Vec, they transformed the text document into vector through CV and TF-IDF functions. They achieved an accuracy of 95% with 600 hidden neurons on IMDb data and 96.4% accuracy on Polarity data.

Dubey & Agarwal (2017) [8] collected 306 tweets on the reviews of the movie "Civil War" on Twitter. They considered two classes viz. positive and negative and used Random Forest algorithm was used. Their study achieved a score of 87.03% in accuracy, 91.55% in F-Score and 79.06% in precision respectively. Wang et al (2018) [9] considered 439 Chinese movies for their study using 2-layered Deep Belief Network with 50 nodes. Their model achieved 0.182 and 0.108 as scores as MAE and RMSE respectively. It is to be noted that most of the studies mentioned above, are related to sentiment analysis on movie reviews. Rhee and Zulkernine (2016) [4] however considered numeric data through scoring different features. Still there are features that can be added to the input variable dataset for a better accuracy of a model. Also, their dataset was limited to top 100 movies based on gross revenue. In this paper the author considered the features taken by the above works and add new ones to the dataset that are correlated to the performance status of a movie. This work concentrates on the Bollywood movies only released during the years 2014 through 2017.

III. RESEARCH METHODOLOGY

The research methodology followed has the following steps:
Step 1: Collection of data and creation dataset

Step 2: Data filtering and normalisation

Step 3: Identification and computation of feature score

Step 4: Interpretation of data distribution

Step 5: Designing the architecture of the model

Step 6: Implementation of the model on the prepared dataset

Step 7: Evaluation and comparison between the performances of the model with the existing models

A. Data Sources

The number of considered Bollywood movies released during 2014 through 2017 is 499. The data sources are Box Office India [10], IMDb [30], E Times [12], BOLLYWOOD hungama [13] and Bollywood Dadi [14]. Box Office India labels the movies into the 8 categories based on net revenue from India and abroad while others follow their own procedure. Data comes in two forms: Text and Numeric. Descriptive reviews are in text format while score or star ratings as well as some other features are in numeric format. The author intends to use numeric data in BPNN-N23.

B. Data Filtering and Normalisation

From the viewers' point of view, the production houses have little effect on the revenue (i.e. success) of a film. People love to watch good movies that play a good story with remarkable performances of the crew. Since the production company is indirectly involved and contributes with those features that directly play role in yielding revenue, this feature has been discarded from the feature vector set.

The data ranges of different features differ widely resulting in a wide variation due to higher scale that will be difficult for a neural network to learn efficiently. To overcome this problem, the data has been normalized by min-max normalization technique so that those lie in the range of 0 and 1:

$$v' = \frac{v - \min_{current}}{\max_{current} - \min_{current}} (\max_{new} - \min_{new}) + \min_{new} \quad (1)$$

where v is the input value of the movie feature and $\max_{new} = 1$ and $\min_{new} = 0$ (in this case). So the equation becomes:

$$v' = \frac{v - \min_{current}}{\max_{current} - \min_{current}} \quad (2)$$

C. Identification and computation of feature score

The correlation coefficients in Table-I between the features considered in this experiment and the success status of movies show both positive and negative correlations.

Among them, the User ratings, Budget Size and Screens show comparatively higher positive correlations with the success status. It can also be noticed that the presence of festival/event has a negative impact in the revenue of a film.

The input feature vectors consist of the individual numeric values of the selected features that are termed as Scores. The computations of these scores are described below.

1) Lead Actor Score

Movie viewers of India generally focus on the lead male and female actors of the Bollywood films. Though the supporting roles are important in a film, people do care less about them. Thus only the lead male and female actor scores has been considered. For determining the scores of the actors and directors, total count of their roles and their salaries per role have been used. The role count during lifetime of a personnel has been considered while the salary has been considered during the last 5 years approximately due to change in currency exchange rate and inflation.

a) Lead Male Actor Score

The lead male score is determined by the following equation:

$$RoleCount_{MActor_i} = \sum_{r=1}^{R} role_r$$

$$Salary_{MActor_i} = \sum_{s=1}^{S} salary_s$$

$$Score_{MActor_i} = RoleCount_{MActor_i} + Salary_{MActor_i} \quad (3)$$

for the i-th male actor.

b) Lead Female Actor Score

A similar method is applied in determining the lead female score.

2) Director Score

A similar method is applied in determining the director score.

3) Production House Score

The quality of a film and its probable commercial success largely depends on a production house. A film is produced by the production house by investing money. A larger production house in terms of resources and fund is capable of spending more money for a film production. The players and directors ranked in higher positions are in demand and deserve more salary. Certain scenes appear unconventional and need to be taken in expensive locations. Multimedia effects add more amount to the expenditure figure. These components find importance in taking production house into account. But since all these components have been specifically taken as participants in the feature vector set, this feature has been discarded as it becomes trivial.

4) Music Director Score

A similar method is applied in determining the music director score.

5) Production Budget

The production budget is the amount of total expenditure (in INR) spent to produce a film.

6) Box Office India Average User Rating

This is the average user rating for a movie provided by the visitors over the website of Box Office India.

7) IMDb Score Average User Rating

This is the average user rating for a movie provided by the visitors over the website of IMDb.

8) E Times Score Average User Rating

This is the average user rating for a movie provided by the visitors over the website of E- Times.

9) Genre

The feature vector will contain all the major 10 genres viz., Biographical, Horror, Crime Thriller, Comedy, Drama, Romantic, Suspense, Action, Animated and Adventure. For a movie, a particular genre will get 1 and all the other genres will be 0. It may be noted that a movie may belong to a combination of more than one genre. In that case, all those matching genres of that movie will get 1.

10) Number of Screens Count

This is the total number of exhibition halls the film has been screened simultaneously.

Table- I: Correlation between the Features and Revenue

Sl No.	Features	Revenue
1	Biographical	0.146731
2	Horror	-0.06902
3	Crime	-0.01754
4	Thriller	0.0054
5	Comedy	0.089574
6	Drama	-0.02759
7	Fantasy	0.167432
8	Romantic	0.046288
9	Musical	0.040919
10	Family	-0.02133
11	Mystery	-0.01347
12	Suspense	-0.01088
13	Sci Fi	0.002509
14	War	-0.0453
15	Action	0.240035
16	History	0.004478
17	Sports	0.031706
18	Animated	-0.03824
19	Adventure	-0.00345
20	User Rating	0.181391
21	Budget Size	0.846119528184634
22	Academy Award	0.231322258
23	Festival / Event	-0.02483
24	Sequel	0.239043
25	Remake	0.097229
26	Lead Actor 1	0.250041262825677
27	Lead Actor 2	0.0148562467403042
28	Director	0.0078676167016279
29	Music Director	0.180403867682183
30	Novel	-0.04863

Sl No.	Features	Revenue
31	True Event	0.083679
32	Number of Screens	0.79117019538891

11) Release Time

Scheduling release time of a film is crucial since films belong to entertainment industry and the viewers must not engage themselves with other more popular events like cricket tournament, world cup football tournament, book fair, religious festivals etc. in which case, they will be busy watching and debating on the said events. It is least likely that during these events, a good film will draw enough attention of the viewers and consequently good revenue to be classified as HIT. 1 has been assigned for presence and 0 for absence of such events.

12) Sequel

This indicates the presence of a sequel or not. If it is a sequel, there will be 1 otherwise 0.

13) Remake

If the film is a remake of another film, there will be 1 otherwise 0.

14) True Event

If the film is based on true event, there will be 1 otherwise 0.

D. Data Distribution

Based on the box plots and scatter matrices (Fig. 1 and Fig. 2) of the features, massive outliers in the Main_Director_Score and Lead_Actor_2_Score are visible. The decision of including or excluding outliers is crucial for observing the model performance. According to the discussion over ResearchGate [33], the following two responses have been focused regarding the outliers in the dataset:

1. Erroneous/mistaken data: No
2. It is likely that the outlier(s) reappear in other datasets: Yes

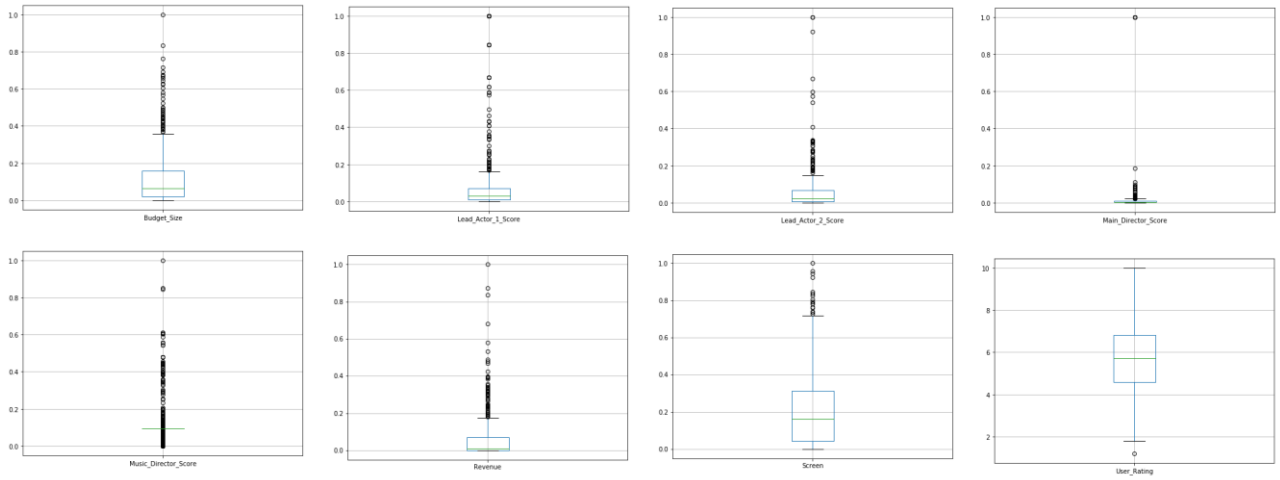

Fig. 1. Box Plots

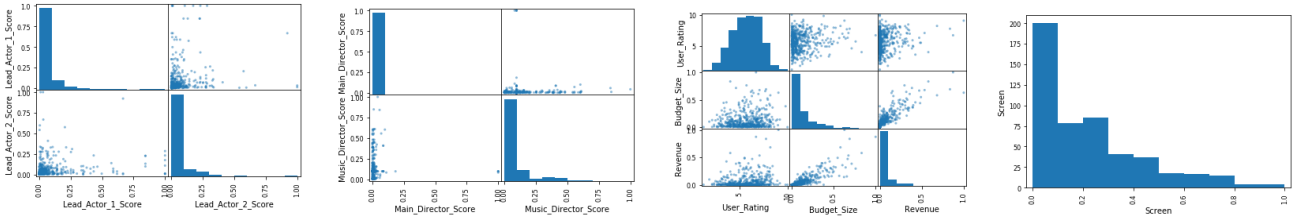

Fig. 2. Scatter Matrices

Marcel M. Lambrechts and Linas Balciauskas [33] have rightly pointed out that the outliers considered as "unusual" depends on the domain. In the current dataset, the data-points regarding Main_Director_Score and Lead_Actor_2_Score are reasonable since the "unusual" ones appear due to new participants or poor performers in the entertainment industry who do not have sufficient history of performances for scoring and may reappear. For these reasons, the author has decided not to delete the records related to these outliers.

IV. PROPOSED MODEL: BACK-PROPAGATION NEURAL NETWORK WITH NUMERIC 23 INPUTS (BPNN-N23)

A. Success classification of Bollywood movies

There is no definite procedure for the classification of success in the Indian film industry [15-16]. Box Office India classifies the success of a movie into 8 categories viz., All Time Block Buster, Block Buster, Super Hit, Hit, Semi Hit, Average, Flop and Disaster [10] based on some factors that it has not revealed.

CONSPROS MEDIA mentioned the basis on which it has classified the success of movies into 8 categories [17]. By visiting other data sources, no specific method has been found to classify a movie. Moreover, the data available for the movies released during the mentioned years is much less that leads to lesser number of data for a specific class. So the author considered 3 classes by the following combination:

HIT - All Time Block Buster, Block Buster, Super Hit, Hit, Semi Hit

AVERAGE - Average

FLOP - Flop and Disaster

Different data sources have labeled the movies differently based on their own method. So the author decided to mark success category based on majority.

In case of equal distribution, one success category of a movie has been randomly chosen. For example, if a movie is marked "HIT" at two places and "SEMI HIT" at another place, "HIT" has been considered due to majority. When all the categories differ, then the success class has been tagged in a random fashion.

B. The architecture of BPNN-N23

The proposed BPNN-N23 consists of 2 layers as shown in Fig. 3. There are 23 feature inputs to the network that implies 33 nodes in the input layer. The hidden layer contains 132 nodes with ReLU activation function. Since there are 3 classes and they are dependent on each other, the output layer contains 3 nodes with Softmax activation function.

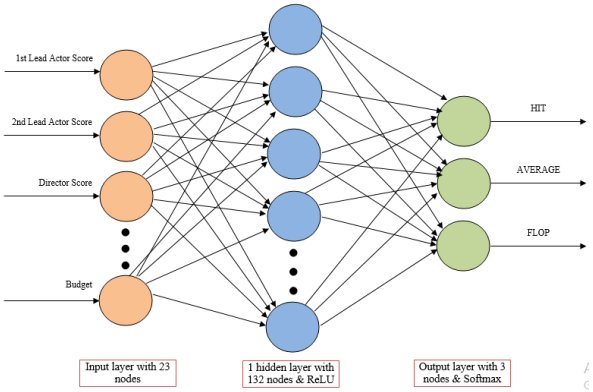

Fig. 3. Back-Propagation Neural Network with Numeric 23 Inputs (BPNN-N23)

V. RESULTS AND DISCUSSIONS

A. Implementation

The computer program has been written in Python using Keras and scikit-learn libraries. Adam [18], a popular stochastic optimizer has been used for the back-propagation algorithm and a k-fold cross validation was performed with the value of k as 10. For measuring the performance of the model, cagorical_crossentropy [31], a logarithmic loss function in Keras has been used. The dataset has 25 movies as AVERAGE, 399 as FLOP and 75 as HIT, which is highly imbalanced (please see Fig. 4) that eventually affects the performance of an ML model [32].

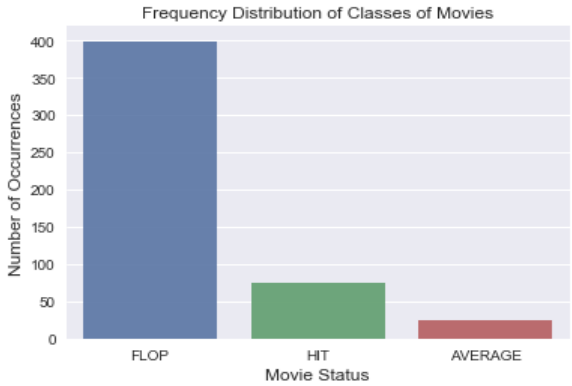

Fig. 4. Frequency Distributions of classes

The author applied SMOTE (Synthetic Minority Over-sampling Technique) [32] in order to balance the dataset. The new size of the dataset becomes 1063. The new dataset is as follows:

- Training Set – that contains 963 data (80%)
- Testing Set – that contains 100 data (20%)

Validation set was not considered for this experiment since the sample size is small. Though computationally expensive, the author opted for cross-validation to save data for the training set.

All data have been normalised according to (2) and stored in CSV files.

B. Evaluation of BPNN-N23

The mean and variance of the cross validation score of the model is 85.78% and 4.84% respectively while the AUC (Area under the ROC Curve) score is 0.653.

1) Confusion Matrix

The confusion matrix displays the summary of the correct and incorrect prediction counts of the classes by the model in tabular form.

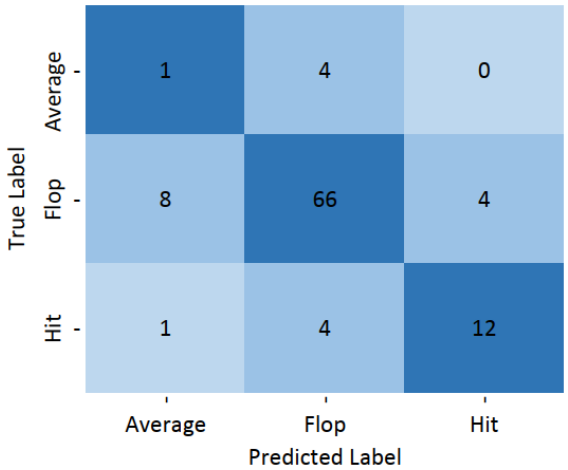

Fig. 5. Confusion Matrix with accuracy

The diagonal cells of the matrix in Fig. 5 show the True Positives of the 3 classes. Out of 100 movies in the Test set, 12 of those labelled HIT, 1 of those labelled AVERAGE and 66 of those labelled FLOP have been classified correctly. The model has misclassified 5 HITs as 1 AVERAGEs and 4 FLOPs while 4 AVERAGE movies have been wrongly classified as 4 FLOPs.

In the FLOP line, 12 movies have been misclassified as 4 HITs and 8 AVERAGES. These counts result in a satisfactory accuracy level of 79.0%. The confusion matrix also shows that the number of FLOPs identified as AVERAGESs is higher than the other misclassifications. This is probably due to the higher number of movies labeled "FLOP" in the data set.

2) Precision, Recall and F1 Score

In BPNN-N23, the Precision and Recall values in FLOP class is high with 0.87 and 0.86 respectively while in HIT class, the Precision scores 0.79 and the Recall score is 0.65. This indicates a good relevancy to the expected results (please see Table-II). The classes HIT and FLOP have F1-scores of 0.71 and 0.86 while AVERAGE scored 0.00 due to the same reason of highly unbalanced data as stated above.

TABLE- II: PRECISION, RECALL AND F1 SCORE

AUC: 0.653

	precision	recall	f1-score	support
AVERAGE	0.00	0.00	0.00	5
FLOP	0.87	0.86	0.86	78
HIT	0.79	0.65	0.71	17
accuracy			0.78	100
macro avg	0.55	0.50	0.52	100
weighted avg	0.81	0.78	0.79	100

3) Model Accuracy

Fig. 6 shows the history of performance of the training and testing procedures over the data sets. The plots initially enjoys a steep rise upto 20 epochs from where both of those tend to flatten. Though the Test plots have flickers, it runs almost in parallel to the Train curve. After a little over 250 epochs, the flickering on the Test plots starts diminishing and becomes almost stable after crossing 320 epochs approximately.

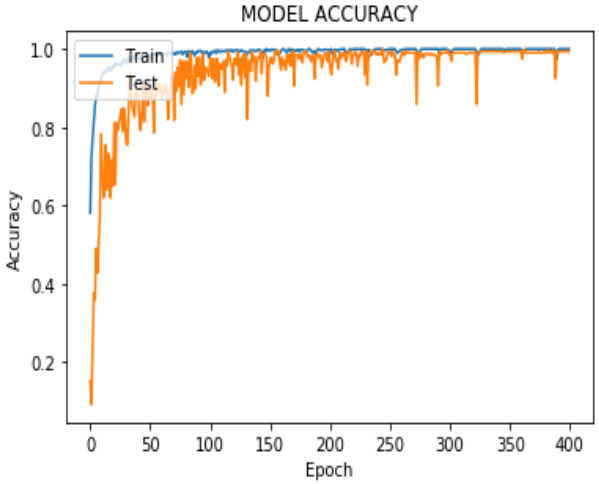

Fig. 6. Model (BPNN-N23) Accuracy

Thus it is obvious that both the lines are attaining a steep slope and almost in parallel, thus indicating that there is a gradual improvement in accuracy upto 50 epochs after which there are some minor deflections at places. Since there is no rise in accuracy from 50 epochs onwards and both the lines are in parallel indicates that there is no overfitting.

4) Model Loss

Model Loss graph indicates the behaviour of the model during optimisation. If the training loss is much greater than the testing loss, the model faces underfitting while in case of the reverse situation, the model faces overfitting. Fig. 7 shows the Model Loss history that shows minimisation of loss upto 70 epochs after which the lines are almost parallel. Since there is no considerable bifurcation between the lines after 70 epochs, it may be assumed that the model is not suffering from overfitting or underfitting. The minimum gap between the curves indicates that the variance is low, that enables the model to generalize on data. The gradual decrease in error in the training curve also indicates low bias signifying a well-fitted training data.

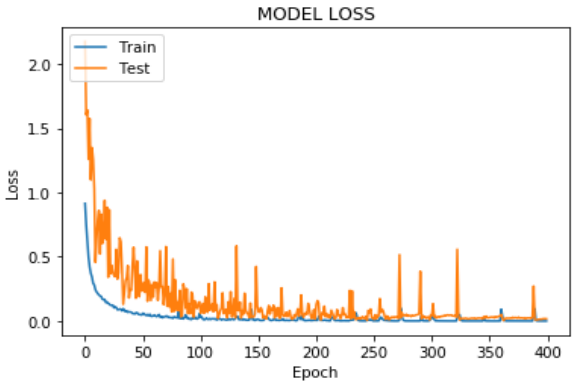

Fig. 7. Model (BPNN-N23) Loss

5) Comparison with Existing Works

BPNN-N23 shows a quite satisfactory result in performance as shown in Table-III when compared to the works done by the other researchers. The first four works are related to sentiment analysis where reviews in textual format were collected from various sources, out of which three have 2 target classes and one is binary as well as multiclass work.

TABLE-III: COMPARISON BETWEEN PROPOSED AND PREVIOUS MODELS

Author	ML Technique used	No of Attributes	No of Hidden Nodes	Data size	No of Classes	Accuracy
Pang & Lee [24]	Naïve Bayes (NB) and SVM	N/A	N/A	2000	2	NB=86.4% SVM=87.2%
Narendra et al [25]	Naïve Bayes	N/A	N/A	2000	2	NB (Stopwords Removal) =72.6 NB (Bigram) =81.6 Hadoop =98.02%
Pouransri & Ghili [26]	Random Forest Logistic Regression SVM Ensemble Averaging	N/A	N/A	50000	2	Random Forest =84.0% SVM =85.8% Logistic Regression =86.6%
				9645	5	Ensemble Averaging =85.5%
Dubey & Agarwal [8]	Random Forest	N/A	N/A	3006	2	87.03%
Zhang et al [27]	BP Neural Network	11	40	241	6	68.1% (Pinpoint) 97.1% (1-Away)

Author	ML Technique used	No of Attributes	No of Hidden Nodes	Data size	No of Classes	Accuracy
Sharda & Delen [28]	MLP Neural Network	7	34	834	9	36.9%
Rhee & Zulkernine [4]	BP Neural Network	14	25	100	2	NN=88.8% SVM=84.2%
BPNN-N23 (Proposed)	BP Neural Network	23	132	499	3	82.0%

The accuracy levels of these works are quite high revolving around 85% except Narendra et al.'s work that achieved a remarkable figure of 98.02% through Hadoop framework. L. Zhang et. al. has achieved 97.1% in 1-Away but their Pinpoint accuracy had a sharp decline to 68.1% probably due to low number of attributes and input data count. Comparatively BPNN-N23 achieves a quite satisfactory accuracy level of 79.0% from which it can be inferred that a larger dataset following back to last 7 to 10 years of productions will throw a light on the pattern of success status.
This will help to develop a more robust model with accuracy of higher order.

It is to be noted that this comparison do not follow a universal metric since there is no benchmark dataset based on which the referenced models present their results and outcomes.

VI. LIMITATIONS AND FUTURE SCOPE

It is needless to say that there is a large scope of research in field of work.

A. Amount of data

To develop an effective prediction of classes, more data on the Bollywood movies down the years should be included. Though it is not feasible to include all the data from 1950s or so, at least last 15 years must be arranged.

B. Number of classes

More number of success classes will help to judge a movie more efficiently. The success classifications done by the referenced websites are in detailed manner and should be followed. More data will eventually help to increase the number of movies in each success class category.

C. Number of the predictors and their types

Though the predictors considered in this research are quite important and relevant, there are more that should be included in research like socio-economic status of the people, political situation etc. to build a more robust model.

D. Number of data sources

This research considered only a few relevant websites. Data from other websites, social media, newspapers, production houses etc. are to be included in future research for developing a more fine-tuned model for classification of movies.

E. Architecture of model

Last but not the least, the architecture of the model is to be carefully designed through rigorous trial and error experiments so that the much higher accuracy can be obtained. The number of layers and nodes, momentum and learning rate, number of epochs and hybrid system with other algorithms are few of the features that should be taken care of.

VII. CONCLUSION

In this research work, the author has proposed a multilayer back-propagation neural network model (BPNN-N23) for determining the success class of a Bollywood movie. The model has 23 inputs that have been carefully chosen by establishing a relevancy with the success status through correlation values and other research works with promising results. A comparison has been done on the performance of the proposed BPNN-N23 with the other existing research works in a tabular form. The model has achieved 79.0% accuracy that clearly shows that there is a large scope of further research.

The decision makers of the Bollywood film industry could find out how much different specific scores of independent variables could be important to the commercial success of a film they are interested in producing with considerably higher efficiency level. It may be argued that changing the quantity, variety and source of the dataset may provide more insight of the present model under consideration. Number of success classes considered here may also be increased to judge a movie more efficiently and in a granular level. The number of layers and nodes, momentum and learning rate, number of epochs and hybrid system with other algorithms are few of the considerations that should be taken care of as far as model characterization is concerned.

It may be concluded that exploring the parameters of the films with more reliable and increased size of dataset will reveal a clearer view of the underlying flow of information while a more superior design of the architecture of the proposed model will assist in more powerful learning capability. This will eventually give the movie personnel a lead in prediction and to perform commercially better.

REFERENCES

1. S. Bhattacharya, “Over The Years”, Educreation Publishing, 2017, page 17.
2. G. Laghate, “Indian media & entertainment industry to reach Rs 2,260 billion by 2020: FICCI KPMG Report”, March 2016. [Online]. Available: https://economictimes.indiatimes.com/industry/media/entertainment/indian-media-entertainment-industry-to-reach-rs-2260-billion-by-2020-ficci-kpmg-report/articleshow/51612183.cms
3. W. Zhang, and S. Skiena, “Improving Movie Gross Prediction through News Analysis”, IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, 2009, pp. 301-304. DOI: 10.1109/WI-IAT.2009.53.
4. T. G. Rhee, F. Zulkernine, “Predicting movie box office profitability: a Neural Network approach”, IEEE International Conference on Machine Learning and Applications, 2016. DOI: 10.1109/ICMLA.2016.0117.
5. S. Lai, L. Xu, K. Liu and J. Zhao, “Recurrent Convolutional Neural Networks for Text Classification”, AAAI Conference of Artificial Intelligence, 2015, pp. 2267-2273. DOI: 10.5555/2886521.
6. S. K. Trivedi and A. Tripathi, “Sentiment Analyis of Indian Movie Review with Various Feature Selection Techniques”, IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 181-185. DOI: 10.1109/ICACA.2016.7887947.
7. J. Singh, G. Singh and R. Singh, “Optimization of sentiment analysis using machine learning classifers”, Human-centric Computing and Information Sciences, vol. 7 issue 32, SpringerOpen 2017. DOI: 10.1186/s13673-017-0116-3.
8. K. P. Dubey and S. Agrawal, “A critical analysis of Twitter data for movie reviews through Random Forest’ approach”, Information and Communication Technology for Intelligent Systems, vol. 2, 2017, pp. 454-460. DOI: 10.1007/978-3-319-63645-0_52.
9. W. Wang, J. Xiu, Z. Yang and C. Liu, “A deep learning model for predicting movie box office based on Deep Belief Network”, International Conference on Sensing and Imaging, June 2018, pp. 530-541. DOI: 10.1007/978-3-319-93818-9_51.

10. Box Office India, 2018. Retrieved from https://boxofficeindia.com.
11. Deepali Kamthania, Ashish Pahwa and Srijit S. Madhavan, "Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business", CIT. Journal of Computing and Information Technology, vol. 26(1), March 2018, pp. 57–68. DOI: 10.20532/cit.2018.1003863.
12. E-Times, 2018. Retrieved from https://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews.
13. BOLLYWOOD hungama, 2018. Retrieved from http://www.bollywoodhungama.com.
14. Bollywood Dadi, 2018. Retrieved from http://www.bollywooddadi.com.
15. INDICINE, 2015. Retrieved from http://www.indicine.com/movies/bollywood/box-office-india-what-makes-a-film-hit-flop-super-hit-or-blockbuster/.
16. FINANCIAL EXPRESS, 2013. Retrieved from https://www.financialexpress.com/archive/hit-super-hit-blockbuster-figure-this-out/1204144/.
17. CONSPROS MEDIA, 2018. Retrieved from https://consprosindia.com/movies-hit-flop-classification/.
18. D. P. Kingma, J. L. Ba, "Adam: A Method for Stochastic Optimization", Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2014. arXiv:1412.6980
19. Ramesh Sharda, Daniel Adomako Asamoah and Natraj Ponna, "Research and Pedagogy in Business Analytics: Opportunities and Illustrative Examples", Journal of Computing and Information Technology - CIT 21, 2013, vol. 3, pp. 171–183. DOI: 10.2498/cit.1002194
20. P. Nagamma, H. R. Pruthvi, K.K. Nisha and N. H. Shwetha, "An improved Sentiment Analysis of online movie reviews based on clustering for box office prediction", IEEE International Conference on Computing, Communication and Automation, 2015. DOI: 10.1109/CCAA.2015.7148530.
21. Xi Ouyang, Pan Zhou, Cheng Hua Li and Lijun Liu, "Sentiment analysis using Convolutional Neural Network", IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015, pp. 2359-2364. DOI: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.
22. R. Joshi and R. Tekchandani, "Comparative analysis of Twitter data using supervised classifiers", IEEE International Conference on Inventive Computation Technologies, 2016. DOI: 10.1109/INVENTIVE.2016.7830089.
23. A. Tripathy, A. Anand and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach", Knowledge and Information Systems, vol. 53, issue 3, December 2017, pp. 805-831. DOI: 10.1007/s10115-017-1055-z.
24. B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of ACL, 2004, pp. 271--278. DOI: 10.3115/1218955.1218990.
25. B. Narendra, K.U. Sai, G. Rajesh, K. Hemanth, M.V.C. Teja and K.D. Kumar, "Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies", August 2016, I.J. Intelligent Systems and Applications. DOI: 10.5815/ijisa.2016.08.08.
26. H. Pouransari and S. Ghili, "Deep learning for sentiment analysis of movie reviews," Technical report, Standford University, 2014.
27. Li Zhang, Jianhua Luo and Suying Yang, "Forecasting box office revenue of movies with BP neural network", Expert Systems with Applications, Elsevier, vol. 36(3), Part 2, April 2009, pp. 6580–6587. DOI: 10.1016/j.eswa.2008.07.064.
28. R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks", Expert Systems with Applications, Elsevier, vol. 30(2), February 2006, pp. 243–254. DOI: 10.1016/j.eswa.2005.07.018.
29. D.D. Gaikar, B. Marakarkandy and Chandan Dasgupta, "Using Twitter data to predict the performance of Bollywood movies", Industrial Management & Data Systems, vol. 115(9), April 2015, pp. 1604-1621. DOI: 10.1108/IMDS-04-2015-0145.
30. IMDb, 2018. Retrieved from https://www.imdb.com.
31. [Online] https://keras.io/losses/.
32. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321–357. DOI: https://doi.org/10.1613/jair.953.
33. [Online] https://www.researchgate.net/post/When is it justifiable to exclude outlier data points from statistical analyses.

AUTHOR'S PROFILE

Partha Shankar Nayak received his M. Tech (Information Technology) from Indian Institute of Engineering Science and Technology, Shibpur and Master of Computer Application & Bachelor in Computer Application from Indira Gandhi National Open University in 2012, 2006 and 2004 respectively. He has published three research works at international conferences on Network Security and Natural Language Processing. He is presently engaged in research work on Data Science and Machine Learning in the domain of entertainment (movies). He is also attached with PSN Academy, P.O. Mankundu, City Bhadreswar, Hooghly, West Bengal, India providing training and placement assistance to under-graduate and post-graduate students as well as IT professionals.