Neuro Fuzzy Approach for Predicting Sales Performance of Movies Considering Sentiments in Online Reviews

Snehal Kulkarni1*, Amarsinh Vidhate (Guide) 2, G.Anuradha 3

1.Lecturer, S.F.I.T, Borivali(w), Mumbai, India,snehalpk@gmail.com
2. Asst. Prof.,R.A.I.T,Mumbai University, India, vidhate.amarsinh@gmail.com
3.Asso. Professor,S.F.I.T,Borivali(w)Mumbai,India. ganusrinu4@yahoo.co.in

Abstract

These “Sentiment without action is the ruin of the soul. — Edward “ An organization has to make the right decisions in a constantly fluctuating business environment. Therefore, predicting the future quantity for the next period most likely appears to be crucial. This work presents a comparative forecasting methodology regarding uncertain customer likings in a movie domain via regression and neuro fuzzy techniques. The main objective is to propose a new future predicting mechanism which is modelled by artificial intelligence approaches including the comparison of both auto regressive method and adaptive network-based fuzzy inference system (ANFIS) techniques to manage the fuzzy demand with incomplete information. The effectiveness of the proposed approach to the demand forecasting issue will be demonstrated using real-world data from a different movie related websites.

Here we are going to extract the information from web and utilizing it for the purpose of sales prediction for movies (like sentiment rating of customer’s reviews, box office revenue of movie, budget of the movie and type of the movie) There are many sales prediction methods but the use of history data will be most efficient way to predict the quality future.

Keywords: ANFIS, regressive model, DSS-Decision support system

1. Introduction

With the increasing use of Web 2.0 platforms such as Web Blogs, discussion forums, Wikis, and various other types of social media, people began to share their experiences and opinions about products or services on the World Wide Web. As an emerging communication platform, Web 2.0 has led the Internet to become increasingly user-centric. People are participating in and exchanging opinions through online community-based social media, such as discussion boards, Web forums, and blogs. Along with such trends, an increasing amount of user-generated content containing rich opinion and sentiment information has appeared on the Internet. Understanding such opinion and sentiment information has become increasingly important for both service and product providers and users because it plays an important role in influencing consumer purchasing decisions [1]. Sentiment-classification techniques can help researchers study such information on the Internet by identifying and analyzing texts containing opinions and emotions [2]. With the flourish of the Web, online review is becoming a more and more useful and important information resource for people.

As a result, automatic review mining and summarizing has become a hot research topic recently. Different from traditional text summarization, review mining and summarizing aims at extracting the features on which thereviewers express their opinions and determining whether the opinions are positive or negative[3].

Posting reviews online has become an increasingly popular way for people to express opinions and sentiments toward the products bought or services received. Analyzing the large volume of online reviews available would produce useful actionable knowledge that could be of economic values to vendors and other interested parties. The idea behind this paper is based on a paper [4][5] where the case study is the movie domain is analyzed and which tackles the problem of mining reviews for predicting movie sales performance. The analysis shows that both the sentiments expressed in the reviews and the quality of the reviews...
have a significant impact on the future sales performance of products in question. For the sentiment factor for that case, Sentiment PLSA (S-PLSA) is used, in which a review is considered as a document generated by a number of hidden sentiment factors, in order to capture the complex nature of sentiments.

In summary, here first time the ratings of the review are calculated by considering the hidden sentiments in it. But up till now for such type of prediction problem the neuro fuzzy approach with sentiment analysis has not implemented, so here the proposed model is — Adaptive Network Based Fuzzy Inference System based on sentiments(S-ANFIS) for the future prediction.

2. Why movie domain

Predicting box-office receipts and category of a particular motion picture has intrigued many scholars and industry leaders as a difficult and challenging problem.

And from the survey regarding writing the reviews, comment, opinion online, the maximum stake is taken by entertainment industry which includes videos, songs, movies, television programs etc.

So one can get to know the clear opinion about different movies after or before its release. Unlike electronic goods of different brands, here for movie domain we can get the exact amount of the box office revenue generation also, so it will help to do the prediction more efficiently.

3. Execution of the problem statement

The work will have a flow mentioned in the block diagram given bellow (see fig 1)

![Fig 1 Block Diagram for Proposed System](image)

In the figure number 1, the input for the process is the reviews/blogs from different web sites, for which we have to do the rating according to the sentiments present in it. Then the second factor i.e. box office revenue will be the next input for the proposed network. The remaining inputs will be budget of the movie as well as type of the movie.

The proposed network is Anfis, so for this the total number of inputs will be review ratings, revenue budget and type of the movies, and the output will be the resulting factor of this four input so it will be the categorization of the movie i.e. whether it will be flop, hit, super hit or blockbuster.

3.1 input data selection and processing

“You don’t have to be a sales manager to appreciate the importance of sales prediction and planning”. Here in this work of the sales prediction, I am considering the example of Movie Domain, as it is a biggest revenue generation industry. And also it is necessary to get the prediction of the upcoming movie related to box office revenue generation and so its category (i.e. whether it’ll be hit, flop or super hit etc...) so that the proper steps can be taken further.

According to procedure, after collecting the reviews/blogs from different web sites [7][8][9][10], it will be analyzed by the sentiment analyzer[11] so that we will get the proper rating of that by considering the sentiment factor present in the review/blog. Here we will get the overall probabilistic sentiment rating of the blog or reviews through the analyzers then and the box-office revenue will be the inputs for the proposed system. Then once we will get the overall sentiment rating of the blog/reviews, the box-office revenue will be collected, simultaneously the budget and type of the movie will be considered and these will act as a input to the proposed ANFIS learning model and the predicted output will be the categorization of the movie in the predefined linguistic category. Input, output factors are shown in fig (see fig 2)

Here in this work the processing is shown as the figure above, fig number 1. Firstly we are choosing the product for prediction; here we have to select any newly released movie or the upcoming movie. The linguistic labels used for this output as Disaster, Flop, Bellow Average, Average, Above Average, Super Hit, Super Duper Hit and Blockbuster.

For above learning model we can take as many as possible training samples, here we have taken the movies from 2009-2012(500 movies).

<table>
<thead>
<tr>
<th>Rating</th>
<th>Revenue</th>
<th>Budget</th>
<th>Type of Movie</th>
<th>Final Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest</td>
<td>Very Low</td>
<td>Low</td>
<td>Drama</td>
<td>Disaster</td>
</tr>
<tr>
<td>Lower</td>
<td>Low</td>
<td>Action</td>
<td>Flop</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Medium</td>
<td>Medium</td>
<td>romance, Musical</td>
<td>Below Average</td>
</tr>
<tr>
<td>Medium</td>
<td>Good</td>
<td>comedy</td>
<td>Average</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Very good</td>
<td>High</td>
<td>Thriller, Horror</td>
<td>Above Average</td>
</tr>
<tr>
<td>Higher</td>
<td>Excellent</td>
<td>Animation</td>
<td>Hit</td>
<td></td>
</tr>
<tr>
<td>Highest</td>
<td>Social</td>
<td>Super Duper Hit</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Fig 2 Linguistic variables for input and output](image)
So here we can predict the sales of the movie if
- We have the review rating before release which we can easily get from the reviews present on different websites.
- We have ratings of the release day.
- We have the rating and revenue of the first week, we can predict for the further weeks.
- And if we have only revenue of the release day and even not Rating present online. (This can be happen if very few or no reviews are present for the movie)
- In addition to this we have budget and type of the movie.

The system can be used in decision support for the movie domain. The decision support system helps in improving the overall movie promotion before the release of the movie itself.

4. Proposed model

Artificial intelligence prediction techniques have been receiving much attention lately in order to solve problems that are hardly solved by the use of traditional methods. They have been cited to have the ability to learn like humans, by accumulating knowledge through repetitive learning activities. Therefore the objective here is to propose new fore-casting techniques via the artificial approaches to manage demand in a fluctuating environment. In this study, a comparative analysis based on regression technique and ANFIS is presented for prediction of the movie performance in future. The artificial techniques used in this study are explained as follows.

4.1 Adaptive network-based fuzzy inference system (ANFIS)

Adaptive network-based fuzzy inference system (ANFIS) [12] can construct an input–output mapping based on both human knowledge in the form of fuzzy if-then rules with appropriate membership functions and stipulated input–output data pairs. It applies a neural network in determination of the shape of membership functions and rule extraction. ANFIS architecture uses a hybrid learning procedure in the framework of adaptive networks.

The architecture of ANFIS can be given, as shown in the diagram bellow, (see fig 3)

![Fig 3 the architecture of proposed model-ANFIS [12]](image)

Layer 1 Every node I in this layer is a square node with a node function

\[ O^1_I = \mu_A(x) \quad \text{Eqn 1} \]

Where \( x \) is the input to node I, and \( A_i \) is the linguistic label (high, low, etc) associated with this node function. In other words, \( O^1_I \) is the membership function of \( A_i \) and it specifies the degree to which the given \( x \) satisfies the quantifier \( A_i \). Usually \( \mu_A(x) \) is choose as bell-shaped with maximum equal to 1 and minimum equal to 0, such as,

\[ \mu_A(x) = \frac{1}{1 + \left(\frac{x - c_i}{a_i}\right)^b_i} \quad \text{Eqn 2} \]

Where \( \{a_i, b_i, c_i\} \) is parameter set. As the values of these parameters change, the bell-shaped functions vary accordingly. Parameters in this layer are referred to as premise parameters.

Layer 2 Every node in this layer is a circle node label as II which multiplies the incoming signals and sends the product out.

\[ w_{ij} = \mu_A(x) \times \mu_B(y), i = 1, 2 \quad \text{Eqn 3} \]

Each node output represents the firing strength of a rule. Also T-norm operators that perform generalized AND can be used as a node function.

Layer 3 Every node in this layer is a circle node label N. the i-th node calculates the ratio of the i-th rule’s firing strength to the sum of all rules firing strength.

\[ \frac{w_{ij}}{w_j + w_{ij}} = 1, 2 \quad \text{Eqn 4} \]

Outputs of this layer are called as normalized firing strength.

Layer 4 Every node I in this layer is a square node with a node function

\[ O^4_I = w_i \cdot f_i = w_i (p_i x + q_i y + r_i) \quad \text{Eqn 5} \]

Where \( w_i \) is the output of layer 3 and \( \{p_i, q_i, r_i\} \) is the parameter set. Parameters in this layer will be referred as consequent parameters.

Layer 5 The single node in this layer is a circle node labelled \( \sum \) that computes the overall output as summation of all incoming signals

\[ O^5_{overalloutput} = \sum_{i} w_i f_i = \frac{\sum_{i} w_i f_i}{\sum w_i} \quad \text{Eqn 6} \]

Thus we have constructed adaptive network which is functionally equivalent to type-3 fuzzy inference system.

4.1 Purpose For Using Adaptive Neuro Fuzzy Inference System

The usage of artificial intelligence has been applied widely in most of the fields of computation studies. Main feature of this concept is the ability of self learning and self-predicting some desired outputs. The learning may be done with a supervised or an unsupervised way. Neural Network study and Fuzzy Logic are the basic areas of artificial intelligence concept. Adaptive Neuro-Fuzzy study
combines these two methods and uses the advantages of both methods. Another reason for using Anfis is the hybrid algorithm used in ANFIS structure consists of the least squares method and the back propagation gradient descent method for training FIS membership function parameters to emulate a given training data. The hybrid algorithm is composed of a forward pass and a backward pass. In the forward pass of the hybrid learning algorithm, the least squares method is used to optimize the consequent parameters with the premise parameters fixed. After the optimal consequent parameters are found, the backward pass starts immediately. In the back-ward pass of the algorithm, the gradient descent method is used to adjust optimally the premise parameters corresponding to the fuzzy sets in the input domain. The output of the ANFIS is calculated by employing the consequent parameters found in the forward pass. The output error is used to adapt the premise parameters by means of a standard back propagation algorithm.

Here the employed training errors are the mean squared error (MSE) of the training data set at each epoch and the mean absolute percentage error (MAPE) of the checking data set at each time. If \( Y_t \) is the actual observation for time period \( t \) and \( F_t \) is the forecast for the same period, then MSE and MAPE are defined as in Eqs 7 and 8

\[
MSE = \frac{1}{N} \sum_{t=1}^{N} (Y_t - F_t)^2 \quad \text{Eqn 7}
\]

\[
MAPE = \frac{1}{n} \sum_{t=1}^{n} \left( \frac{Y_t - F_t}{Y_t} \right) \times 100 \quad \text{Eqn 8}
\]

5. Result and conclusion

As mentioned earlier in this work we are going to predict the sale of particular movie with the help of different factor like, past box office performance, box office collection ,budget of movie, type of movie and important factor is online review’s sentiment factor which are present on different movie web-sites. Then the results will be validated using statistical techniques. It is the extension of the work done earlier in paper [14]

5.1 Results using ANFIS

The architecture and learning rules of adaptive networks have been described in previous chapter. Functionally, there are almost no constraints on the node functions of an adaptive network except piecewise differentiability. Due to these minimal restrictions, the adaptive network's applications are immediate and immense in various areas. So here in ANFIS, we performed the prediction using neural network and fuzzy inference system which combining gives much more good results than existing one. Again there is use of different type of membership functions as it gives more precision in results.

So the ANFIS results will be compared through different errors, i.e. training, checking and testing errors and then we will come up with the results for best combination of Membership functions. Then the evaluation parameters will be MAPE and MSE errors which give the error measures for predicting the movie category.

5.1.1 Computation of Input and Output Membership Functions

A membership function for a fuzzy set \( A \) on the universe of discourse \( X \) is defined as \( \mu_A: X \rightarrow [0, 1] \), where each element of \( X \) is mapped to a value between 0 and 1. This value, called membership value or degree of membership, quantifies the grade of membership of the element in \( X \) to the fuzzy set \( A \). Membership functions allow us to graphically represent a fuzzy set.

The x axis represents the universe of discourse, whereas the y axis represents the degrees of membership in the \([0,1]\) interval (see fig 4).

Here all types of membership functions are evaluated on the basis of different types of errors like training, checking and testing error for estimating the accurate prediction values.

The training error is the difference between the training data output value, and the output of the fuzzy inference system corresponding to the same training data input value, (the one associated with that training data output value). The training error \( \text{trnError} \) records the root mean squared error (RMSE) of the training data set at each epoch.
The checking error is the difference between the checking data output value, and the output of the fuzzy inference system corresponding to the same checking data input value, which is the one as located with that checking data output value. The checking error chkError records the RMSE for the checking data at each epoch. The training, checking and testing errors can be shown in fig (see fig 5).

The above plotted graphs show the training, checking and testing errors with respect to different numbers of input membership function. Here I have measured these errors for different types of membership function. From the Fig no 6, it can be seen that the error rate for the gbell type membership function is at lower side. This is because there is a difference between the tuning parameters for the different types of membership function. The parameters associated with the membership functions will change through the learning process. The computation of these parameters (or their adjustment) is facilitated by a gradient vector, which provides a measure of how well the fuzzy inference system is modelling the input/output data for a given set of parameters. Once the gradient vector is obtained, any of several optimization routines could be applied in order to adjust the parameters so as to reduce some error measure (usually defined by the sum of the squared difference between actual and desired outputs). ANFIS uses either back propagation or a combination of least squares estimation and back propagation for membership function parameter estimation.

The bell membership function has one more parameter than the Gaussian membership function, so it can approach a non-fuzzy set if the free parameter is tuned. Because of their smoothness and concise notation, Gaussian and bell membership functions are popular methods for specifying fuzzy sets. Both of these curves have the advantage of being smooth and nonzero at all points.

Because of their smoothness and concise notation, Gaussian and bell MFs are becoming increasingly popular for specifying fuzzy sets. Gaussian functions are well known in probability and statistics, and they possess useful properties such as invariance under multiplication (the product of two Gaussians is a Gaussian with a scaling factor) and Fourier transform (the Fourier transform of a Gaussian is still a Gaussian). The bell MF has one more parameter than the Gaussian MF, so it has one more degree of freedom to adjust the steepness at the crossover points.

This can be again supported with the following results in which we are getting the lowest error parameters for the Gbell membership.

However, since the triangular and trapezoidal MFs are composed of straight line segments, they are not smooth at the corner points specified by the parameters.

For sigmoidal membership function, this type of MF is although extremely flexible in specifying fuzzy sets, is not used often in practice because of its unnecessary complexity. Hence for further prediction, Gbell membership function will be preferred. Again the results are evaluated by considering the different number of membership functions.

The errors for the different types of membership function can be compared as given in the table (see table 1).

<table>
<thead>
<tr>
<th>Types Of MF</th>
<th>Training Error</th>
<th>Checking Error</th>
<th>Testing Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>G Bell</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.35e-007</td>
</tr>
<tr>
<td>Gaussian1</td>
<td>0.0003</td>
<td>0.0005</td>
<td>0.40e-007</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>0.0003</td>
<td>0.0006</td>
<td>0.8e-007</td>
</tr>
</tbody>
</table>

Here from the results it indicates that the Gbell type membership functions gives the comparatively less errors than others because the generalized bell membership function
function is specified by three parameters and has the function name `gbellmf`. So from the above analysis we can conclude that the results given by the membership function of type GBell are minimum amongst all.

### 5.2 Accuracy Measurements

“To have an appropriate measure for assessing the accuracy of a forecast is perhaps one of the most contentious issues among researchers over the last fifty years”.[16]

The quality (accuracy) of a model can be estimated by examining the inputs (assumptions) to the model, or by comparing the outputs (forecasts) from the model. It claims that testing outputs is the only useful approach to evaluating forecasting methods. But the testing of input is the only worthwhile way to test models. But it is more reasonable to test both inputs, for improvement of a model, and outputs, for selection of the best model. Forecasting accuracy is regarded as an “optimist’s term for forecast errors”. A forecast error, on the other hand, represents the difference between the forecast value and the actual value.

In this case for checking the precision accuracy we will be considering the error factors like MAPE, MSE and RMSE which will be comparing the actual and predicted values.

#### 5.2.1 Mean Squared Errors (MSE)

The mean squared error is an accuracy measure computed by squaring the individual error for each item in a data set and then finding the average or mean value of the sum of those Squares[16]. MSE can be given as,

\[
\text{MSE} = \frac{1}{N} \sum_{t=1}^{n} e^2_t
\]

Where MSE = mean squared error

\(n = \text{time periods}\)

\(e^2 = \text{forecast error}\)

The MSE is having the advantage of being easier to handle mathematically. By using Eqn(9).Since the ability of a forecasting method to detect large errors is often regarded as one the most important criteria, the MSE method has been popular for years.

#### 5.2.2 Root Mean Squared Errors (RMSE)

The term root mean square error (RMSE) is the square root of mean squared error (MSE). RMSE measures the differences between values predicted by a hypothetical model and the observed values. In other words, it measures the quality of the fit between the actual data and the predicted model. RMSE is one of the most frequently used measures of the goodness of fit of generalized regression models.

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} e^2_t}. \quad \text{Eqn 10}
\]

The calculation of RMSE is done separately for training and checking error so that we can select the range of training and checking data appropriately. (See fig 6)

![Fig 6 Training and Checking Errors](image)

#### 5.2.3 Mean Absolute Percentage Error (MAPE)

The percentage error is given by \(p_t = \frac{100e_t}{Y_t}\).

Percentage errors have the advantage of being scale independent, so they are frequently used to compare forecast performance between different data series. The most commonly used metric is Mean Absolute Percentage Error

The MAPE is given by

\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left|\frac{PE_t}{Y_t}\right|
\]

\[
\text{where PE} = \text{percentage error}
\]

\(Y_t = \text{actual observation for time period t}\)

\(F_t = \text{forecast for the same period}\)

\(\text{MAPE = mean absolute percentage error}\)

\(n = \text{time periods}\)

Here in this prediction case of movie as it was mentioned that the results can be taken in different scenarios like in presence of either of the input or when all the inputs will be present, and then the prediction accuracy will be calculated by the different measures described above. The values of MAPE and MSE for ANFIS can be compared as,(see table 2)

<table>
<thead>
<tr>
<th>Different Scenarios</th>
<th>Without Sentiment</th>
<th>With Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANFIS MEAN</td>
<td>ANFIS</td>
<td>ANFIS</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>0.0129</td>
<td>1e-03</td>
</tr>
<tr>
<td>MSE</td>
<td>8.6e-007</td>
<td>6.2471e-007</td>
</tr>
</tbody>
</table>

So here from the above observations in table 2 we can conclude that, the mean average percentage error of the second scenario i.e. with the sentiment rating, the prediction accuracy is more than the effect of no rating.
factor. But it is also giving very less percentage error for another option.

From the values we can conclude that the precision error is lowest in the presence of the sentiment rating than only having any of the factors. In the presence of either of the factors also the model is giving the good precision as the error factor very less.

5.3. K Fold Cross Validation or Averaging model

Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.

So in k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k – 1 sub samples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter. So here in this case we used 4 fold cross validation ,i.e. partitioned the data sets and done training, testing using different combinations of data set and seen the effect on prediction accuracy for the taken data set.

From the table (see table 3), the inference can be done that the set 1 parameters which we have decided for training, testing purpose giving the best result comparatively other fold.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Actual value</th>
<th>Predicted value</th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2108</td>
<td>0.2105</td>
<td>0.0356</td>
</tr>
<tr>
<td>2</td>
<td>0.2108</td>
<td>0.1851</td>
<td>3.0479</td>
</tr>
<tr>
<td>3</td>
<td>0.2108</td>
<td>0.1851</td>
<td>3.0479</td>
</tr>
<tr>
<td>4</td>
<td>0.2108</td>
<td>0.1689</td>
<td>4.9692</td>
</tr>
</tbody>
</table>

6. Comparison with the ARSA Model

In this model the autoregressive models is implemented with sentiments incorporated with it. So first the sentiments has been calculated with the probabilistic latent sentiment analysis model, i.e. SPLSA , then this probabilistic rating and the box office revenue and budget and finally type of movie will act as the input to the ARSA i.e. autoregressive sentiment analysis model . In this case the Multiple Regression model is developed as there are more than one predictors and one final prediction.

Regression analysis is a statistical technique for determining the relationship between a single dependent (criterion) variable and one or more independent (predictor) variables. The analysis yields a predicted value for the criterion resulting from a linear combination of the predictors.[13]

6.1 Multiple Regression Model

Multiple linear regression analysis is one of the most commonly used statistical modelling techniques in the business world for predictions.

Using more than two training inputs may produce a better prediction, because it reduces mainly two kinds of noises. One is brought by the reuse distance measurement. The second kind of noise is brought by the input size. These noises reduce the accuracy of the prediction. According to the regression theory, more data can reduce the of noises and reveal a pattern closer to the real pattern. Accordingly, we apply a regression method on more than two training inputs. In this case, since there are four predictors and one dependant output factor, the preference to the multiple regression model is given.

In general, the multiple regression Eqn of Y on X₁, X₂, ..., Xₖ is given by:

\[ Y = a + b₁x₁ + b₂x₂ + ... + bₖxₖ + e \]  Eqn 12

Here ‘a’ is the intercept and b₁, b₂, b₃... bₖ are analogous to the slope in linear regression Eqn and are also called regression coefficients. They can be interpreted the same way as slope.

The appropriateness of the multiple regression models as a whole can be tested by the F-test in the ANOVA table. A significant F indicates a linear relationship between Y and at least one of the X's.

As given in Eqn number 12, note that we have k independent variables and a slope for each. We still have
one error and one intercept. Again we want to choose the estimates of \( a \) and \( b \) so as to minimize the sum of squared errors of prediction. The prediction Eqn is:

\[
Y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \quad \text{Eqn 13}
\]

Finding the values of \( b \) is tricky for \( k \geq 2 \) independent variables, and will be developed after some matrix algebra. It’s simpler for \( k = 2 \) IPs, here in this case there are 4 inputs and one output so the further steps in the Eqns are given as below.

For the one variable case, the calculation of \( b \) and \( a \) was:

\[
b = \frac{\Sigma xy}{\Sigma x^2} \quad ; \quad a = \bar{y} - b \bar{x} \quad \text{Eqn 14}
\]

This is used by the existing method [4][5]

Now since here as mentioned earlier there are 4 input variables, the Eqn for 2 input variables will be as bellow,

\[
b_1 = (\Sigma x_1 y)^2 (\Sigma x_1 x_2) - (\Sigma x_1 x_1) (\Sigma x_2 y)^2 / (\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2 \quad \text{Eqn 15}
\]

\[
b_2 = (\Sigma x_2 y)^2 (\Sigma x_1 x_2) - (\Sigma x_1 x_2) (\Sigma x_2 x_2)^2 / (\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2 \quad \text{Eqn 16}
\]

At this point, see that all the terms from the one variable case appear in the two variable case. In the two variable cases, the other \( X \) variable also appears in the Eqn. For example, \( X_2 \) appears in the Eqn 15 for \( b_1 \). The terms corresponding to the variance of both \( X \) variables occur in the slopes. Also a term corresponding to the covariance of \( X_1 \) and \( X_2 \) (sum of deviation cross-products) also appears in the formula for the slope.

So the Eqn for \( a \) with two independent variables is:

\[
a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - b_3 \bar{x}_3 - b_4 \bar{x}_4 \quad \text{Eqn 17}
\]

Where \( x, y \) variables are the mean of \( x \) and \( y \) from the training data or existing data. So this Eqn 17 is a straight-forward generalization of the case for independent variable.

So now using the above equations we can further compute the predicted values for the given problem statement.

Here for the same data set, the regression model is done in matlab with the help of ‘regress’ function which takes the value for all inputs and calculates the value of predictors i.e. values of \( b_1,b_2,b_3,b_4 \) are find out to \( b_0 \), (when the prediction is done using sentiment rating)(see table 4)

<table>
<thead>
<tr>
<th>Predicators</th>
<th>With Sentiments</th>
<th>Without sentiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0.74034000000000</td>
<td>1.50970000000000</td>
</tr>
<tr>
<td>B2</td>
<td>0.81127000000000</td>
<td>1.50970000000000</td>
</tr>
<tr>
<td>B3</td>
<td>0.07120000000000</td>
<td>0.08140000000000</td>
</tr>
<tr>
<td>B4</td>
<td>0.17090000000000</td>
<td></td>
</tr>
</tbody>
</table>

After getting these values of predictors, the value of ‘a’ i.e. the regression coefficient is find out with the equation no 17.

The values of ‘a’ for the considered both cases are, \( A_1 = -0.016422630066422 \) (with sentiment) \( A_2 = 0.001500887188235 \) (without sentiment)

So with the help of above values for \( a,b_1,b_2,b_3 \) and \( b_4 \), the final regression equation for both the cases are written as,

\[
Y_{\text{withSenti}} = 0.01642263 + 0.740*x_1 + 0.811*x_2 + 0.0712 + 0.170*x_4 \quad \text{Eqn 18}
\]

\[
Y_{\text{withoutSenti}} = 0.001500885 + 1.509*x_1 + 1.5097*x_2 + 0.081*x_3 \quad \text{Eqn 19}
\]

Where \( x_1 , x_2 , x_3 , x_4 \) are the input values for the given model.

Now with the above found values we will get the predicted output for the given input data. The results are compared with the help of MAPE as well as MSE.(see table 5)

### Table 5 ARSA MAPE,MSE Values

<table>
<thead>
<tr>
<th>Different Scenarios</th>
<th>Without Sentiment</th>
<th>With Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARSA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>0.043</td>
<td>0.04</td>
</tr>
<tr>
<td>MSE</td>
<td>6.8921e-007</td>
<td>8.0981e-007</td>
</tr>
</tbody>
</table>

### 7. ARSA and ANFIS Comparison

Here the final comparison is based on the different error evaluation factors like MAPE, MSE and RMSE errors.

ARSA uses simple multivariate regression technique based on the value of predictors and the existing values. It uses the simple regression values for further prediction. ARSA gives précised results for single predictor, i.e. the predicted output is only depend on the one input

ANFIS uses the approach of neuro fuzzy technique which incorporates the rule based part from fuzzy logic and learning algorithms from neural network. It uses hybrid algorithm for calculating the error factor and giving final output. As it used the combine concept of neuro-fuzzy it is proved to be efficient than ARSA for predicting results, which can be justified with the results mentioned throughout this.

It can be represented as bellow. (see table 6)

### Table 6 ARSA/ANFIS MAPE,MSE Values

<table>
<thead>
<tr>
<th>Different Scenarios</th>
<th>In the absence of sentiment</th>
<th>In Presence Of Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARSA</td>
<td>ANFIS</td>
<td>ARSA</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.0043</td>
<td>0.002</td>
</tr>
<tr>
<td>MSE</td>
<td>6.82e-006</td>
<td>8.6831e-007</td>
</tr>
</tbody>
</table>

The table number 6 indicates the efficiency of ANFIS against ARSA. The mean average percentage error and mean square errors are very less in ANFIS than ARSA.
The final prediction accuracy can be seen in the fig 7 where it indicates that the predicted values from ANFIS are much closer to actual output than in ARSA.

So from the figures (see fig 7) it clearly indicates that the prediction accuracy given by ANFIS is much closer to actual output than the existing approach of using auto regression. Again the prediction done using one important factor like sentiment gives closer values than the prediction done without the same values and it can be verified with the help of MAPE and MSE errors.

7.1. Data Validation using Analysis of Variance

Covariance, correlation deal with the study of two or more variables and their relationships to one another. Covariance and correlation will help us determine if any relationships exist among the variables, if they exist. So finding the correlation coefficient and covariance between the input, output factor gives the validation of the selected input and its effect on output. So there are different two ways to find and proving the relationship between input and output.

7.1.1 Covariance and coefficient of correlation

Covariance and correlation describe how two variables are related.
- Variables are positively related if they move in the same direction.
- Variables are inversely related if they move in opposite directions.

Both covariance and correlation indicate whether variables are positively or inversely related. Correlation also tells you the degree to which the variables tend to move together.

A correlation coefficient of -1 means that the numbers are perfectly inversely correlated. If one grows the other falls. A correlation coefficient of zero means that the two numbers are not related. A correlation coefficient of one or near to one means that the two numbers are perfectly related.

The formula given for finding covariance is given as bellow,
\[
\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

\[\text{Eqn 20}\]

Where \(x\) is independent variable  \\
\(y\) is dependent variable, \(n\) is total samples, \(x', y'\) are mean of \(x\) and \(y\)

And for deriving the correlation coefficient is given by,
\[
\rho(x, y) = \frac{\text{cov}(x, y)}{s_x s_y}
\]

\[\text{Eqn 21}\]

Here \(S_x\) is standard deviation of \(x\) and \(S_y\) is standard deviation of \(y\)

So by applying this formula to verify our data set, the estimated results are given as(see table 7)

<table>
<thead>
<tr>
<th>Different dependant factors</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment Rating and O/P</td>
<td>0.7</td>
</tr>
<tr>
<td>Revenue and O/P</td>
<td>0.8</td>
</tr>
<tr>
<td>Budget and O/P</td>
<td>0.3</td>
</tr>
<tr>
<td>Type of Movie and O/P</td>
<td>0.04</td>
</tr>
</tbody>
</table>

The correlation coefficient of rating and output i.e. category is found to be 0.7 and that with the revenue and output is 0.8. Then the effect of budget and type is also analysed by finding the correlation coefficient between those factors and are found to be lesser then sentiment rating and box office revenue.
So from the above dependency values one can clearly get the picture that the most effective factors in the prediction are sentiment rating and revenue. A correlation coefficient of 0.7 and 0.8 tells you two important things:

- Because the correlation coefficient is a positive number, effect of rating and box office revenue on the final output i.e. of category of the movie, growth is positively related.
- But 0.7 is relatively far from indicating no correlation, the strength of the correlation between sentiment rating and the category of the movie is strong. Same way the correlation coefficient between box office revenue and category is also 0.8, it also indicates the positive growth of both factors and stronger relationship between these factors.

Besides this the values for other two factors are not that much impact able on the final outcome, the relationship between the budget of the movie and the final outcome of movie is comparatively low but show the positive growth. But the last factor which is type of the movie doesn’t affect more on the category of the movie, so whatever the type of the movie, the main factors which are important for success of the movie are rating, revenue and budget of the movie.

Hence we have validates our data through different statistical measures and proven to be best selection of the factors of the data set and the values of the data set as well.

8. Conclusion

The wide spread use of online reviews as a way of conveying views and comments has provided a unique opportunity to understand the general public’s sentiments and derive business intelligence. Here, we have explored the predictive power of reviews using the movie domain as a case study, and studied the problem of predicting sales performance using sentiment information mined from reviews.

The outcome of the proposed models leads to actionable knowledge that can be can readily employed by decision makers. A center piece of the work is ARSA and ANFIS model and used SPLSA for sentiment analysis that helps us move from simple “negative or positive” classification toward a deeper comprehension of the sentiments in blogs. Using SPLSA as a means of “summarizing” sentiment information from reviews developed ANFIS, model for predicting sales performance based on the sentiment information and the products past sales performance. The use of other effective factors is also considered here like the budget of the movie, the type of the movie. The accuracy and effectiveness of the proposed models can be confirmed by the experiments on movie data sets. As the final graph of actual versus predicted shows that the proposed method ANFIS gives much more perfect result than existing one in both the cases i.e. considering the effect of sentiment and ignoring the sentiments. The MAPE error is less when the sentiment factor is considered. So for prediction of movie the sentiment factor is as important as revenue.

But comparatively with the author work[4][5] we can see that in Indian Bollywood movie market, sentiment only doesn’t play the main deciding factor in prediction, opening box office revenue as well as the type of the movie also decides the future of the movie. Equipped with the proposed models, companies will be able to better harness the predictive power of reviews and conduct businesses in a more effective way. So the proposed S-ANFIS(input processed with sentiment analysis) model is general frameworks for sales performance prediction as it is a self learning model and would certainly benefit from the development of more sophisticated models for sentiment analysis and future quality prediction. Sales prediction is management's primary tool for predicting the volume of attainable sales. Therefore, the whole budget process hinges on an accurate, timely sales prediction [15].

References

[3] Li Zhang “Movie Review Mining and Summarization”, Microsoft Research Asia Department of Computer Science and Technology, Tsinghua University Beijing.
[5] Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain, Xiaohui Yu ; Sch. of Comput. Sci. & Technol., Shandong Univ., Jinan, China ; Yang Liu ; Xiangyi Huang ; Aijun An
[16] The Validity of Forecasting, Dr Garrick R. Small and Raymond Wong, Pacific Rim Real Estate Society International Conference Christchurch, 2002