A Model for Predicting Movie's Performance using Online Rating and Revenue

Ms.Mary Margarat Valentine, Ms. Veena Kulkarni, Dr.R.R.Sedamkar

Abstract— Blog reviews,discussion forums, different type of social sites have created a new type of marketing and communication that connects the gap between simple word-of-mouth and a viral form of opinions which can move virtual mountains for a business. In the Movie Domain, a single movie can have the variation between millions of rupees of profits or losses for a studio in a given year. It's not surprising, therefore, that movie studios are intensely involved in predicting the performance of the movies

This paper work proposes auto regression method and adaptive networks of fuzzy inference system model for the prediction of movie performance. ARSA model is implemented using two inputs and one output. Two Inputs are online sentiment ratings and box office revenue and the output is category of the movie. Proposed ANFIS System uses Fuzzy logic design: input, Fuzzificcation, inference engine, defuzzification, and output. There are many sales prediction methods but the use of history data will be most efficient way to predict the quality future. Both the regression method and Fuzzy rules are formulated and applied on 145 Bollywood movies which are released in the year of 2010 to 2013. The prediction result is calculated on the basis of auto regression and ANFIS model. Both the ARSA and ANFIS model's output is compared with actual output .The prediction accuracy are measured for both the models using forecast accuracy methods MAPE and MSE.

Index Terms— Adaptive Networks, Artificial Intelligence, Blogs, Fuzzy Inference System, Prediction, Regression, Sentiment Analysis

1 INTRODUCTION

ONLINE reviews are crucial to any business and becoming more imperative every day to manage reputation. Reviewing plays the major role in the face of online marketing since the Internet became a household convenience which allows all businesses to have lively, positive participation from users and gives customers to create a relationship with those businesses.

Sentiment mining techniques can help researchers to study sentiments incorporated in reviews on the Internet by identifying and analyzing the texts containing feelings and opinions [2]. Those online reviews provide a wealth of information on the products and services, and if the information is properly utilized, can present vendor highly valuable network intelligence and business intelligence to make possible the improvement of their business. As a result, online review mining has newly received a great deal of consideration. Since what the people thinks of a product can no doubt control how well it sells, understanding the opinions and sentiments expressed in the relevant online reviews is of high importance [4]. In this paper, actionable knowledge is created by building up models and algorithms that can utilize mined reviews from blogs. Such models and algorithms can be used to effectively predict the performance of products [5]. Previous research was conducted on the predictive power of reviews which considered the huge volume of reviews or link structures to predict the trend of product performance rather than the quality reviews to consider the effect of the sentiments present in the blogs [1], [3]. It has been reported that although there seems to exist strong correlation between the volume of reviews and sales spikes, using the volume or the link structures alone do not provide satisfactory prediction performance[1],[3]. Indeed, the sentiments expressed in the online reviews are more powerful than volumes of reviews.

Prediction of product performance is an extremely domain-driven task, for which a deep understanding of a variety of aspects involved are important. In this paper, the movie domain is taken as a case study, in which the different issues like modeling reviews, producing performance predictions, and obtaining an actionable knowledge. As the result, three different factors are identified which play a vital role in predicting the box office revenues in the movie domain, namely, user's sentiments, past sales performance, and quality of the review. A framework is propsed for sales prediction with all those factors included. First factor is modeling sentiments in reviews, using text mining methods. Simply classifying reviews as positive or negative does not provide a comprehensive understanding of the sentiments reflected in reviews [4]. S-PLSA focuses on sentiments rather than volume of reviews or topics. Therefore, instead of considering volume of reviews primarily focused only on the reviews that are sentiment related [7]. The second factor which is considered in this paper is the past sale performance of the same product, or in the movie domain, past box office performance of the same movie. Based on S-PLSA information, ARSA model is presented for predicting movie sales performance by considering the input factors sentiment ratings and box office revenue. No standard methods exist for converting human knowledge or experience into rulebase and a fuzzy inference system [6]. Fuzzy Inference System is implemented in the framework of adaptive networks with the help of Review mining and sentiment analysis along with box office revenue.

2 RELATED WORK

Movie domain is considered as an input, because information related to the movie and revenue information, are easily available. The movie information which is used for conducting experiments includes two factors.

1. The first factor is a set of blog reviews on movies which are of interest, collected from the Web.
2. Second factor contains the corresponding revenue data for these movies.

2.1 Why Only Movie Domain?

From the recent studies regarding writing the reviews, online opinions, online comments, discussion forums, the most stakes is taken by film industry which includes videos, songs, movies, television programs etc. So it is very simple to get the clear review about various movies after or before its release. If the prediction is focused on electronic goods, then it is required to consider different companies/brands, but here for movie domain it is possible to get exact amount of the box office revenue information. So it will help in predicting sales with the help of earlier data. The movie review data set is obtained from different websites. The publicly accessible website for movie reviews is IMDB Website.

3 Execution Flow for the Proposed Model

Flow for the proposed system is as shown in the Figure 1, i.e.

- Select any newly released movie or the upcoming movie product for prediction.

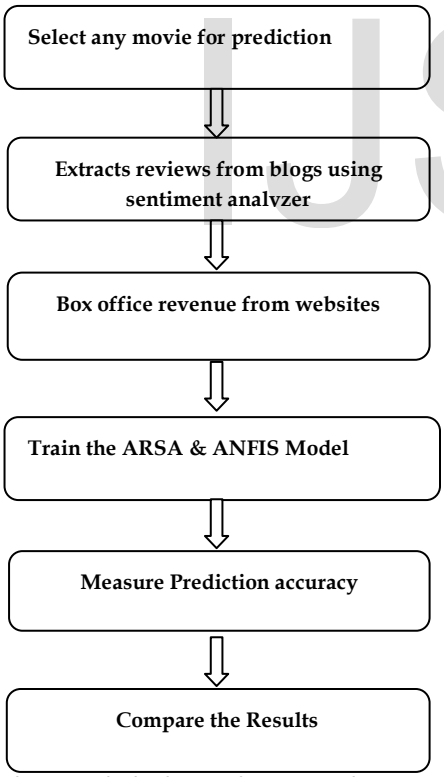

Figure 1. Block Diagram for proposed System

- Extract the reviews from different blogs for sentiment analysis.
- For data analysis, bollywood movies are considerd and its revenue and rating are collected before and after release.
- The first input will be the rating of any movie after sentiment analysis [12].
- The second input will be the box office revenue of the movie in rupees. It will be taken from the websites, according to the week wise collection after the release of the movie.
- Both the inputs are given to ARSA and ANFIS model for sales prediction.
- The output for the pair of above input will be the final category of the movie, defined in different categories of the movie starting from Disaster to All Time Blockbuster.

3.1 Review Mining

With the rapid growth of posting comments and reviewsrelated to the product, review mining has involved a great deal of consideration. Prior studies in this area were mainly focused on deriving the sentiment reviews [2].

The movie review data set was obtained from the publicly accessible IMDB Website. Specifically, collected the reviews for movies released in the bollywood. Snapshot [17] of English Vinglish movie review is shown the Figure 2

Figure 2. Snap Shot of a Movie Review

3.1.1 S- PLSA Model

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, does not provide a comprehensive understanding of the sentiments reflected in blog reviews [7].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. To evaluate hidden factors which are present in reviews posted by customers, a new approach is used to review mining based on Probabilistic Latent Semantic Analysis (PLSA), which is called as Sentiment PLSA.

It would be too simplistic to just classify the sentiments expressed in a review as either positive or negative. Moreover, mining opinions and sentiments present unique challenges that cannot be addressed easily with traditional text mining

algorithms, due to the fact that opinions and sentiments, which are usually written in natural languages, are often expressed in subtle and complex ways. All these concerns call for a model that can extract the sentiments in a more accurate manner. To this end, Liu et al. [3] [9] propose the S-PLSA model, in which a review can be considered as being generated under the influence of a number of hidden sentiment factors [8]. Inspired the PLSA model [4], [5] the use of hidden factors in S-PLSA provides the model the ability to accommodate the intricate nature of sentiments, with each hidden factor focusing on one specific aspect [10].

3.2 Box Office Revenue

Box office revenue is taken as the second input in this paper. Along with the sentiment rating, box office revenue site collected from boxofficeindia.com and koimoi.com websites are used for predicting movie performance. A product can attract a lot of attention (thus a large number of blog mentions) due to various reasons, such as aggressive marketing, unique features, or being controversial. This may boost the product's performance for a short period of time. But as time goes by, it is the quality of the product and how people feel about it that dominates. For that reason instead of taking the first day or first week box office collection, average collection is considered.

3.3 Train the Arsa Model (Existing Model)

For this Existing regression model, the author focused on extracting sentiment information from the public reviews incorporated in the online blogs. From the above cited work, reviews are used for predicting product sales performance. Based on study of the compound nature of sentiments in the reviews, they proposed S-PLSA, in which a entries in the blog is analyzed as a document created by a number of unknown sentiment factors. As a result of training the model, sentiment's summary is obtained from the blogs. That sentiment's summary which is retrieved from the PLSA method is given as the input to the ARSA model, for prediction. Wide ranges of testing were conducted on a movie data set. Then Comparison is done by including sentiments and in the absence of sentiments.
In Summary,

- First, Input and Output factors are decided for prediction using ARSA model.Here in this paper, Ratings of the movie and Box office revenue are the input factors of the movie and output factor is the Category of the movie.
- Ratings can be collected either from S-PLSA model or IMDB website.
- Box office revenue generation is easily available on the internet and revenue information is collected from those websites.
- Training the ARSA model based on the Input factors and output category is calculated.
- Now the result is used to analayzed to calculate prediction accuracy and the error rate.
- For comparison, same procedure is followed using sentiment information and the absence of sentiment information.

4 Proposed Model

The architecture and learning procedure underlying ANFIS is propsed, which is fuzzy inference system, implemented in the frame work of adaptive networks. The proposed ANFIS can construct an Input-output mapping based on human knowledge. Basic aspect of this model is that standard techniques are not existed for converting human knowledge or experience into the rule base and database of a fuzzy inference systemand there is a need for efficient methods for tuning the membership functions to reduce the output error measure or increase the performance index. In this perception, the objective of this paper is to propose a new architecture design known as Adaptive Network based Fuzzy Inference System (ANFIS), which is used as a source for the construction of a set of fuzzy if-then rules with proper membership functions to create the set of input-output pairs. Sugeno model is shown in the Figure 3.

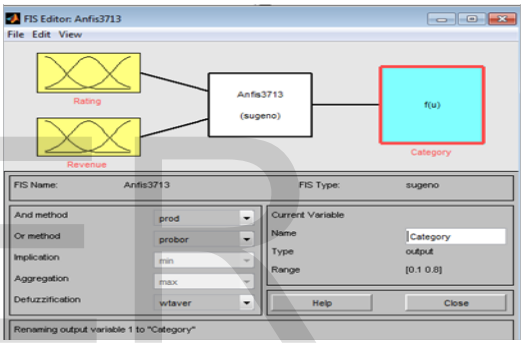

Figure 3. Two Input-Output Sugeno Model

4.1 Input- Output Factors in Prediction

The movie data which is proposed to conduct experiments consists of two components. The first component is a set of blog documents on movies which are of interest, collected from the Web, and the second component contains the corresponding daily box office revenue data for these movies. Predicting sales performance is done with the help of different factor like, past box office performance, box office collection and main important factor is online reviews which are present on different movie websites. After collecting the reviews/comments from different web sites/blogs/discussion forums [16], [17], [18], [19], it will be analyzed by the sentiment analyzer tool [19] so that proper rating is retrieved by considering the sentiment factor present in the online reviews. Here we will obtain the overall probabilistic sentiment rating of the movie based on the comments/reviews through the analyzers then and the box-office revenue will be the inputs for the proposed system. Once we obtain the overall ratings of the movies, and the box-office collection then these two components will act as the input for the proposed learning model and the predicted output will be the categorization of the movie in the predefined linguistic type. Table 1 show the input and output factor for the prediction. Ratings and Revenue is consid-

ered as an input and output will describe the category of the movie.

TABLE 1
INPUT-OUTPUT FACTORS FOR PREDICTION

RATING	REVENUE	CATEGORY
Poor	Small	Disaster
Mediocre	Average	Flop
Average	Medium	Below average
Decent	Good	Average
Good	Very Good	Hit
Very Good	Better	Super hit
Excellent	Best	Block buster
Superb	Excellent	All time block buster

Title	Release Date (mm/dd/yyyy)	Actual Category for our reference	Rating	Revenue in Cr.	RATING x1	REVENUE x2	Proposed Category Y
Kahaani	3/8/2012	Super hit	8.2	59.26	0.8	0.4	0.6
Barfi!	9/13/2012	Super hit	8.3	120	0.8	0.8	0.8
Gangs of Wasseypur	5/22/2012	Average	8.2	27.8	0.8	0.2	0.4
English Vinglish	9/14/2012	Hit	7.9	85	0.7	0.6	0.5
OMG: Oh My God!	9/28/2012	Super hit	7.9	81.5	0.7	0.6	0.5
Vicky Donor	4/20/2012	Super hit	7.9	40.01	0.7	0.3	0.4
Ishaqzaade	5/11/2012	Hit	6.6	62.2	0.6	0.5	0.5
Agneepath	1/25/2012	Super hit	7	123.05	0.7	0.8	0.7
Shanghai	6/7/2012	Losing	7.3	22.05	0.7	0.2	0.3
Talaash	11/29/2012	Hit	7.3	93	0.7	0.7	0.7
Jab Tak Hai Jaan	11/13/2012	Super hit	7.2	120.65	0.7	0.8	0.7
Luv Shuv Tey Chicken Khurana	11/2/2012	Losing	6.7	8	0.6	0.1	0.2
Chakravyuh	10/11/2012	Losing	6.8	16	0.6	0.2	0.3
Dabangg 2	12/20/2012	Super hit	5.2	158.5	0.5	0.8	0.6
Ferrari Ki Sawaari	6/14/2012	Average	6.5	29.3	0.6	0.2	0.3
Rowdy Rathore	5/31/2012	Super hit	5.5	131	0.5	0.8	0.6
Ek Tha Tiger	8/15/2012	Super hit	5.2	198	0.5	0.8	0.6
Bol Bachchan	7/6/2012	Super hit	5.5	102	0.5	0.7	0.6
Tere Naal Love Ho Gaya	2/23/2012	plus	5.5	22	0.5	0.2	0.3
Cocktail	7/13/2012	Hit	5.8	76	0.5	0.6	0.4
Agent Vinod	3/22/2012	Average	5.1	44.06	0.5	0.3	0.3
Housefull 2	4/5/2012	Super hit	5	114	0.5	0.8	0.6
Teri Meri Kahaani	6/21/2012	Flop	4.9	27.5	0.5	0.2	0.2
London Paris New York	3/1/2012	Flop	5.7	7.02	0.5	0.1	0.1
Heroine	9/20/2012	plus	4.8	44.25	0.4	0.3	0.3
Son of Sardaar	11/13/2012	Hit	4.2	105.43	0.4	0.8	0.5

Figure 4. Snapshot of Input - Output data taken for the proposed system

Based on the Input – output factors shown in the table 1, Input and output data are collected for the bollywood movies from online to implement the proposed prediction model.Snapshot of input-output data taken for impementation is shown in the Figure 4.

4.2 Fuzzy Inference System

Adaptive Neuro fuzzy inference system (ANFIS) is a kind of neural network that is based on Takagi–Sugeno fuzzy inference system [6]. Ever since it integrates commonly neural networks and fuzzy logic principles, it has feasible to capture the benefits of together in a single framework. Its inference system corresponds to a set of fuzzy IF–THEN rules that have learning capability to approximate nonlinear functions. [13] Hence, ANFIS is considered to be universal approximator. ANFIS 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. The rule based system for the current prediction model is shown in the Figure 5.

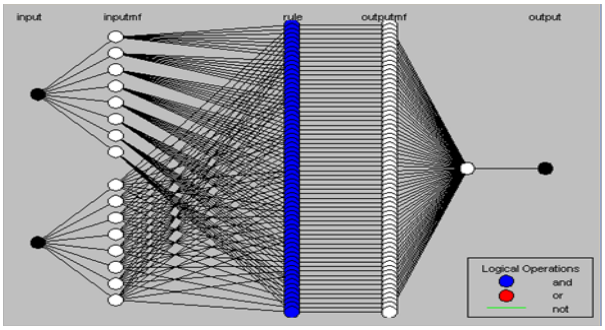

Figure 5. Anfis Model Architecture – Rule based System

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. This method plays a particularly important role in the induction of rules from observations within fuzzy logic. The convention of artificial intelligence has been applied broadly in a large amount 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.

4.3 Training the ANFIS Model

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. Box office revenue will be collected from the IMDB website [16] and different blog entries to find the probability of the sentiments.

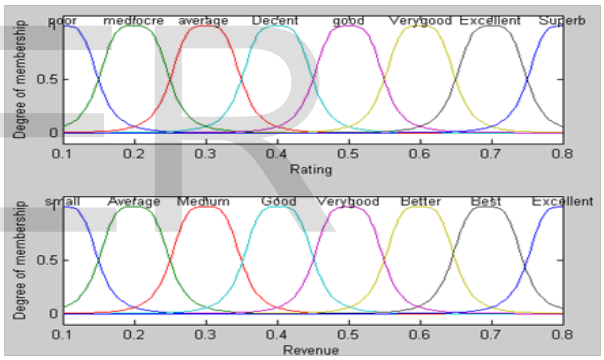

Figure 6. Memebership Functions for Rating and Revenue

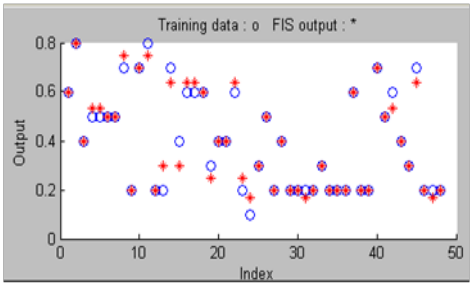

Figure 7. Training Error

The two input memebership functions of the proposed model and training error is shown in the Figure 6 and Figure 7 respecdtively.On the basis of these two inputs, prediction of upcoming movie will be done through ARSA model. The mean absolute percentage errors (MAPE) and Mean Squared Error (MSE) will be used to measure the prediction accuracy.

which can be represented as

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}((Y_t - F_t)/Y_t)*100$$

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

For the above learning model many possible training samples can be taken from the movies which are released from 2010 - 2013.

5 Results

The extensive use of online reviews and comments as a way of conveying opinions and views has provided a distinctive opportunity to identify with the user's opinions and obtain business intelligence. In this paper, the predictive power of comments and sentiments incorporated in the reviews are investtigated using the movie domain as a case study, and implemented the problem of predicting sales performance using sentiment information mined from reviews. As mentioned in this paper, we are predicting the performance of a particular movie by considering two important input factors like, sentiment information incorporated in online reviews and box office revenue of that particular movie and from different websites.

TABLE 2
MEASUREMENT OF PREDICTION ACCURACY USING MAPE & MSE

Method	ARSA	ANFIS
MAPE	0.2747	0.0160
MSE	3.8548e-005	3.5025e-005

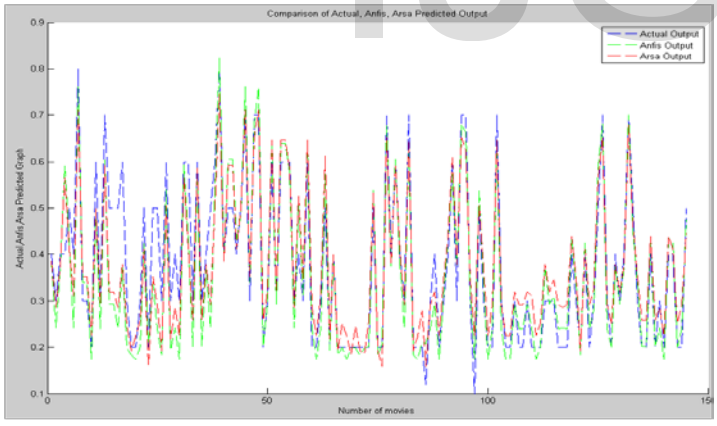

Figure 8. Comparison of Predicted Vs. Actual output

Figure. 8. Shows the precited output of the movies which is released in the bollywood and the comparison of the proposed oupt, ARSA output and ANFIS output is also shown. From the Table 2, it is clearly shown that Anfis is giving the best prediction performance when compared to ARSA error accuracy. From the prediction accuracy, it is easy to state that propsed model can work well as a very good prediction system for deciding product performances for future movies.

6 Conclusions and Future Work

In this paper, the model of predicting sales performance of a movie is implemented using sentiment information mined from reviews and box office revenue. This model is explored as a domain-driven task, and managed to produce human intelligence (identifying imperative characteristics of online reviews related to movie), domain intelligence (the knowledge of movie and box office revenues), and network intelligence (online reviews posted by public users). The result of the proposed model leads to an actionable knowledge that can readily used by decision makers to decide the awards for the movies and can make the critical business decisions better and it yields to significant competitive advantages in the entertainment industry.

It is worth noting that, only two input factors are used for prediction in this paper. For future work, this present system can be enhanced by considering few more inputs like budget of the movie (promotional budget & production budget), CBFC rating, movie genre, targeted audience of the movie to improve the accuracy and quality of the prediction. It would also be interesting to enhance the systemto keep a track and monitor the current trends and changes in sentiments posted in the blogs.

References

[1] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The Predictive Power of Online Chatter," Proc. *11th ACM SIGKDD Int'lConf. Knowledge Discovery in Data Mining (KDD)*, pp. 78-87, 2005.

[2] Rubicon Consulting, "Online Communities and Their Impact on Business: Ignore at Your Peril," 25 Mar. 2009; http://rubiconconsulting.com/downloads/whitepapers/Rubiconwebommunity

[3] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information Diffusion through Blogspace," *Proc. 13th Int'l Conf. World Wide Web (WWW)*, pp. 491-501, 2004

[4] Xiaohui Yu, Yang Liu, "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, APRIL

[5] Y. Liu, X. Huang, A. An, and X. Yu, —ARSA: A Sen-timent-Aware Model for Predicting Sales Performance Us-ing Blogs,‖ *Proc. 30th Ann. Int'l ACM SIGIR Conf. Re-search and Development in Information Retrieval (SIGIR)*, pp. 607-614, 2007

[6] Jyh-Shing roger Jang, "ANFIS : Adaptive-Network-Based fuzzy Inference system", IEEE Transactions onSytems, Man and cybernetics, pp.No 665-685, VOL 23, No.3,May/June 1993

[7] T. Hofmann, "Probabilistic Latent Semantic Analysis," Proc. Uncertainty in Artificial Intelligence (UAI), 1999.

[8] Thomas Hofmann and Jan Puzicha. Latent class models for collaborative filtering. In IJCAI, pages 688–693, 1999.

[9] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou,"Low-Quality Product Review Detection in Opinion Summarization," Proc. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP), pp. 334-342, 2007.

[10] C. Elkan, Method and System for Selecting Documentsby Measuring Document Quality. US patent 7,200,606, Washington, D.C.: Patent and Trademark Office, Apr.2007.

[11] Li Zhuang "Movie Review Mining and Summarization",Microsoft Re-

search Asia Department of Computer Science and Technology, Tsinghua University Beijing
[12] A. Ghose and P.G. Ipeirotis, "Designing Novel ReviewRanking Systems: Predicting the Usefulness and Impact of Reviews," Proc. inth Int'l Conf. ElectronicCommerce (ICEC), pp. 303-310, 2007.
[13] P.D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," Proc. 40th Ann. Meeting on Assoc. forComputational Linguistics (ACL), pp. 417-424, 2001.
[14] J. Kamps and M. Marx, "Words with Attitude," Proc.First Int'l Conf. Global WordNet, pp. 332-341, 2002.
[15] Nedjah, Nadia, ed. Studies in Fuzziness and Soft Computing. Germany: Springer Verlag. pp. 53–83.ISBN 3-540-25322-X.
[16] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," Proc. 14th Int'l Conf. World Wide Web (WWW), pp. 342-351, 2005.
[17] http://www.rottentomatoes.com/m/english_vinglish/
[18] http://www.mouthshut.com/Hindi-Movies/3-Idiotsreviews-925106887
[19] http://www.imdb.com/title/tt1187043/reviews
[20] http://www.cs.bham.ac.uk/~axk/Assign1.doc
[21] http://sentiment.brandlisten.com/analyse
[22] http://socialmediatoday.com/sector45/1433331/why-online-reviews-matter
[23] http://theonlinedepartment.com/8-reasons-why-online-reviews-are-important-to-your-business/
[24] http://knowledge.brandify.com/why-online-reviews-are-important-for-your-business/

- *Mary Margarat Valentine is currently pursuing masters degree program in Computer engineering inMumbai University,India E-mail: e.marymargret@gmail.com*
- *Mrs.Veena Kulkarni is currently working as a Assistant Professor in Thakur-College of Engineering & Technology, Mumbai*
- *Dr.R.R.Sedamakr is working as a Professor, Dean in academics, HOD-CMPN in ThakurCollege of Engineering & Technology, Mumbai*