ONLINE REVIEW MINING FOR FORECASTING SALES

Nihalahmad R. Shikalgar [1], Deepak Badgujar [2]

[1] *Department of Computer, PVPIT,Maharashtra,India,nihalcse@gmail.com*
[2] *Assistant Professor, Department of Computer, PVPIT, Maharashtra, India, deepakbadgujar@gmail.com*

Abstract

The growing popularity of online product review forums invites people to express opinions and sentiments toward the products .It gives the knowledge about the product as well as sentiment of people towards the product. These online reviews are very important for forecasting the sales performance of product. In this paper, we discuss the online review mining techniques in movie domain. Sentiment PLSA which is responsible for finding hidden sentiment factors in the reviews and ARSA model used to predict sales performance. An Autoregressive Sentiment and Quality Aware model (ARSQA) also in consideration for to build the quality for predicting sales performance. We propose clustering and classification based algorithm for sentiment analysis.

Index Terms: *Online Review mining, Text mining,reviews, S-PLSA, ARSA, Clustering, Classification.*

----------***----------

1. INTRODUCTION

The growing pervasiveness of the Internet has changed the way that people shop for goods, watch the movie. The growing popularity of online product review forums invites the development of models and metrics that allow firms to harness these new sources of information for decision support. Whereas in a blog (blogspot.com), IMDB (www.imdb.com) websites visitors can usually evaluate movie review before watching it. Online people increasingly rely on alternative sources of information such as "word of mouth" in general, and user-generated movie reviews in particular. In fact, some researchers have established that user-generated movie information on the Internet attracts more interest than vendor information among consumers. In contrast to movie descriptions provided by vendors, consumer reviews are, by construction, more user oriented. In a review, customers describe his/her sentiment in terms of different usage scenarios and evaluate it from the user's perspective. Despite the subjectivity of people evaluations in the reviews, such evaluations are often considered more credible and trustworthy by people than traditional sources of information (Bickart and Schindler 2001).

The hypothesis that movie reviews affect movies box office collection has received strong support in prior empirical studies. However, these studies have only used the numeric review ratings (e.g., the number of stars) and the volume of reviews in their empirical analysis, without formally incorporating the information contained in the text of the reviews. To the best of our knowledge, only a handful of empirical studies have formally tested whether the textual information embedded in online user-generated content can have an economic impact. Ghose et al. (2007) estimate the impact of buyer textual feedback on price premiums charged by sellers in online second-hand markets. Eliashberg et al. (2007) combine natural language- processing techniques and statistical learning methods to forecast the return on investment for a movie, using shallow textual features from movie scripts. Netzer et al. (2011) combine text mining and semantic network analysis to understand the brand associative network and the implied market structure. Decker and Trusov (2010) use text mining to estimate the relative effect of product attributes and brand names on the overall evaluation of the products. But none of these studies focus on estimating the impact of user-generated movie reviews in influencing there box office collection beyond the effect of numeric review ratings, which is one of the key research objectives of this paper.

There is a potential issue with using only numeric ratings as being representative of the information contained in movie reviews. By compressing a complex review to a single number, we implicitly assume that the product quality is one-dimensional, whereas economic theory tells us that movie have multiple attributes and different attributes can have different levels of importance to people. Thus, unless the person reading a review has exactly the same preferences as the person who wrote the review, a single number, like an average movie rating, might not be sufficient for the reader to extract all information relevant to the watching decision.

Recent studies have been focused on the reviews for finding the relationship between the sales performance of the products and reviews. The actionable knowledge developed by using the average of the number of the quality reviews presented and also the number of the people rated the reviews in the blogs and mdb websites. The actionable knowledge is the last part and which can be developed by the base models and algorithms which is used to effectively predict the sales performance and which can be shared to all the peoples across the world. Predicting sales performance is completely a domain driven task, it gives us to analyze the public sentiments, past sales performance and box office revenues. Such that the actionable knowledge can be

developed by the sentiments and the quality reviews also plays as an important role here.

Some wrong reviews also affect the prediction but only we are taking the whole reviews So it will not be impact the sales of the movie.

2. RELATED WORK

2.1 Online Review Mining

Review Mining is one of the growing mining sectors. It is very predictive for analysis review. Many online blog and social networking sites are available, where many people are expressing their review with respect to product and movie. If we considering these reviews then it is very helpful to increase sales performance.

2.2 Domain Driven Task

Domain-driven data mining [5] generally targets actionable knowledge discovery in complex domain problems. It aims first to utilize and mine many aspects of intelligence for example, in-depth data, domain expertise, and real-time human involvement as well as process, environment, and social intelligence. Domain-driven data mining works to expose next generation methodologies for actionable knowledge discovery, identifying how KDD can better contribute to critical domain problems in theory and practice. It uncovers domain-driven techniques to help KDD strengthen business intelligence in complex enterprise applications.

Domain Driven task [5] is categorized into three level. These three are Human Intelligence, Domain intelligence, Network intelligence.

2.3 Human Intelligence

In this domain driven task, we considered number of blog. These blogs are considered with opinions of user .Opinions of user or information of other thing is posted on blog. So blogs are the way to express opinion of people. Let's consider any upcoming movie is there or movie already released, in that case number of blog are generated to express thought about the movie. So this is one way to get the people sentiment analysis.

2.4 Domain Intelligence

In this domain driven task, we considered data from various websites e.g. IMDB sites or else we have to create movie database. In the IMDB website, we get movie rating and recommendation as well as revenue data collected on box office. So this is very important factor to consider the popularity of any kind of movie.

2.5 Knowledge discovery from database

It is a part of KDD [6].Knowledge discovery from database. We discover knowledgeable data by mining raw data.raw data is considered as input to the system. By using this data we apply data mining algorithm so that knowledge is discovered from it. This knowledge is of the pattern or a collective analysis report. This is very important for business to increase profit of organization. Here actionable knowledge [6] is considered to the knowledge discovery. AKD is an iterative optimization process toward the actionable pattern, considering surrounding business environment and Problem states.

2.6 Clustering and Classification

Clustering can be considered the most important unsupervised learning problem, so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

So our problem solution is not using kind of supervised learning. By using clustering, the records of the movies data is collected simultaneously and made available for analysis.

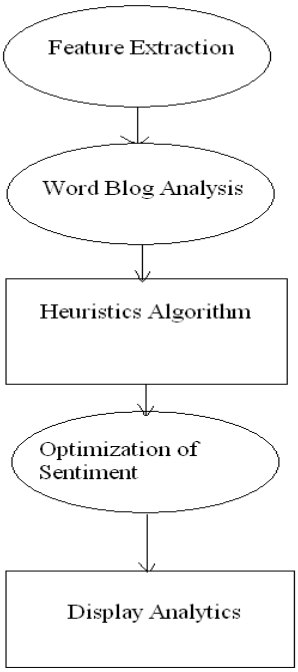

Fig -1: Overall Process

3. S PLSA

Sentiment Probabilistic Latent Semantic Analysis [7] in which a review can be considered as being generated under the influence of a number of hidden sentiment factors. For the said purpose numbers of blogs are available. Separating number of blogs from other blog is very tedious task.

Categorize each one blog and consider it only those blog which is relating with movie review for subjecting movie.

If we considering word review pair then each word expressing the positive and negative comments. Initial task is to separate the sentiment in different categories like positive, negative, average. Number of other word is present in to the sentiment .those unnecessary words are increasing calculation. Optimization is done to illuminate unnecessary calculation. The people can observe their opinions and increase their feel to watch the movie if the reviews are good.

If the movie is bad then the reviews are bad then it will become serious impact for the movie. Reviews posted in online are so important so that it is directly affecting the movie's sales and the bad reviews about the movie failed to get the right place in the box office in the movie database site.

Movie Sentiments are as follow:

Movie Name: SSTG
Comments:
Positive Comments:

I love action of only sunny as it suits him anytime. A mast entertainer movie
Beautiful movie with nice storyline.
Very Good Movie
Excellent movie. Must see
Must watch...better than stupid ramlila or Chennai express

Negative Comments:

Not a good one....just time pass
Only bang bang nothing else.
Time pass movie.
1 time watch, average movie

CONCLUSIONS

The increasing use of online reviews as a way of conveying views and comments has proved a unique way to find sales performance and derive business intelligence. In this paper, we have studied the problem of predicting sales performance using sentiment information mined from reviews. The outcome of this generates knowledge from mined data that can be useful for forecasting sales.S-PLSA is useful for analysis of sentiment that help us to classify different categorization of sentiments in blogs. By using ARSA, we can easily predict sales performance.

REFERENCES

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

[2]. A. Ghose and P.G. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews," Proc. Ninth Int'l Conf. Electronic Commerce (ICEC), pp. 303-310, 2007.

[3]. Y. Liu, X. Huang, A. An, and X. Yu, "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 607-614, 2007.

[4]. Y. Liu, X. Yu, X. Huang, and A. An, "Blog Data Mining: The Predictive Power of Sentiments," Data Mining for Business Applications, pp. 183-195, Springer, 2009

[5]. L. Cao, C. Zhang, Q. Yang, D. Bell, M. Vlachos, B. Taneri, E. Keogh, P.S. Yu, N. Zhong, M.Z. Ashrafi, D. Taniar, E. Dubossarsky, and W. Graco, "Domain-Driven, Actionable Knowledge Discovery," IEEE Intelligent Systems, vol. 22, no. 4, pp. 78-88, July/Aug. 2007.

[6]. L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang, and E.K. Park,"Flexible Frameworks for Actionable Knowledge Discovery,"IEEE Trans. Knowledge and Data Eng., vol. 22, no. 9, pp. 1299-1312, Sept. 2009.

[7]. XiaohuiYu , Jimmy Xiangji Huang ,"Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain" IEEE Trans. Knowledge and Data Eng., Vol. 24, No. 4, April 2012.

BIOGRAPHIES

Mr.Nihalahmad Shikalgar is an PG student of Computer Department,His reaserch area is Data Mining.Pursuing ME Degree.

Mr.Deepak Badgujar is an Assistant Professor in PVPIT,Pune.His reaserch area is in Data mining.He is completed his MTech.