Predicting Sentiment Analysis from Online Reviews

S.Balaji., M.E.,

Assistant Professor,

Department of Information Technology,

V.R.S College of Engineering and Technology

Arasur, Tamilnadu, India.

---***---

Abstract - ***Abstract:-Posting online reviews has become an increasingly popular way for people to share with other users their opinions and sentiments toward products and services. Both the sentiments expressed in the reviews and the quality of the reviews has a significant impact on the future sales performance of products in question. To tackle the problem of mining reviews for predicting product sales performance, propose Sentiment PLSA (S-PLSA). Based on S-PLSA, ARSA, an autoregressive sentiment-aware model for sales prediction and then seek to further improve the accuracy of prediction by considering the quality factor, with a focus on predicting the quality of a review in the absence of user-supplied indicators, and present ARSQA, an autoregressive sentiment and quality aware model, to utilize sentiments and quality for predicting product sales performance.***

Key Words: *Mining online reviews, Sentiment-Probabilistic latent semantic analysis (SPLSA), Autoregressive Sentiment Quality Aware model (ARSQA).*

1. INTRODUCTION

Online review has become a common practice for e-commerce websites to provide the venues and facilities for people to publish their reviews, with a prominent example being Amazon (www. amazon.com). Reviews are also prevalent in blog posts, social networking websites as well as dedicated review websites such as Epinions (www.epinions.com). Those online reviews present a wealth of information on the products and services, and if properly utilized, can provide vendors highly valuable network intelligence and social intelligence to facilitate the improvement of their business. As a result, review mining has recently received a great deal of attention. A growing number of recent studies have focused on the economic values of reviews, exploring the relationship between the sales performance of products and their reviews [1], [2], [3], [4]. Prior studies on the predictive power of reviews have used the volume of reviews or link structures to predict the trend of product sales [1], [8], failing to consider the effect of the sentiments present in the blogs. It has been reported [1], [8] 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. Indeed, as we will illustrate with an example, the sentiments expressed in the reviews are more predictive than volumes. In addition, another important aspect that has been largely overlooked by those prior studies is the effect of the reviews quality on their predictive power. Quality wise, not all reviews are created equal. Especially in an online setting where anybody can post virtually anything, the quality of reviews can vary to a great extent. We believe that prediction of product sales is a highly domain-driven task, for which a deep understanding of various factors involved is essential. In this paper, using the movie as a domain, we investigate the various issues encountered in modeling reviews, producing box office predictions, and deriving actionable knowledge. To this end, we identify three factors that play important roles in predicting the box office revenues in the movie domain, namely, public sentiments, past box-office performance, and review quality, propose a framework for box-office prediction with all those factors incorporated. We start with modeling sentiments in reviews, which presents unique challenges that cannot be easily addressed by conventional text mining methods. In order to model the multifaceted nature of sentiments, we view the sentiments embedded in reviews as an outcome of the joint contribution of a number of hidden factors, and propose a novel approach to sentiment mining based on Probabilistic Latent Semantic Analysis (PLSA), which we call Sentiment PLSA (S-PLSA). Different from the traditional PLSA [6], S-PLSA focuses on sentiments rather than topics. In S-PLSA, appraisal words are exploited to compose the feature vectors for reviews, which are then used to infer the hidden sentiment factors. The second factor we consider past box office performance of the same movie. Combining this AR model with sentiment information mined from the reviews, we propose a new model for product sales prediction called the Autoregressive Sentiment Aware(ARSA) model. Extensive experiments show that the ARSA model provides superior predication performance compared to using the AR model alone, confirming our expectation that sentiments play an important role in predicting future box-office performance. The quality factor is then incorporated into the ARSA model, resulting in an Autoregressive Sentiment and Quality Aware (ARSQA) model for sales prediction.

2.RELATED WORK

D.Gruhl,R,Guha,gives The predictive power of Online Chatter.[1].A. Ghose and P. G. Ipeirotis, gives a novel review ranking systems for predicting the usefulness and impact of reviews [2]. Y. Liu, X. Huang, A. An, and X. Yu, gives a Autoregressive sentiment-aware model for predicting sales performance using blogs.[3].And also gives Blog data mining that gives the predictive power of sentiments, in Data Mining for Business Applications[4].B. Pang and L. Lee, gives A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[6]. Some surveys of [8][9] gave a detail review of existing methods of mining on line reviews.N.Z. Foutz and W. Jank, “Pre-Release Demand Forecasting for Motion Pictures Using Functional Shape Analysis of Virtual Stock Markets[12].T. Hofmann gives a Probabilistic Latent Semantic Analysis[14].C. Whitelaw, N. Garg, and S. Argamon using Appraisal Group for Sentiment Analysis[10].

3. METHODOLOGY OF WORK

3.1 S-PLSA for Sentiment Mining

Mining opinions and sentiments present unique challenges that cannot be handled easily by traditional text mining algorithms. This is mainly because the opinions and sentiments, which are usually written in natural languages, are often expressed in subtle and complex ways. Moreover, sentiments are often multifaceted, and can differ from one another in a variety of ways, including polarity, orientation, graduation, and so on. Therefore, it would be too simplistic to just classify the sentiments expressed in a review as either positive or negative. For the purpose of sales prediction, a model that can extract the sentiments in a more accurate way is needed. In its traditional form, PLSA [6] assumes that there are a set of hidden semantic factors or aspects in the documents, and models the relationship among these factors, documents, and words under a probabilistic framework. With its high flexibility and solid statistical foundations, PLSA has been widely used in many areas, including information retrieval, Web usage mining, and collaborative filtering. Nonetheless, to the best of our knowledge, we are the first to model sentiments and opinions as a mixture of hidden factors and use PLSA for sentiment mining.

We now formally present S-PLSA. Suppose we are given a set of reviews $B=\{b_1 .....,b_N\}$and a set of words (appraisal words) from a vocabulary=$(w_1....,w_M)$. The review data can be described as a N ×M matrix $D=(c(b_i,w_j))_{i,j}$ where $c(b_i,w_j)$ is the number of times w_j appears in review b_i. Each row in D is then a frequency vector that corresponds to a review S-PLSA is a latent variable model for co-occurrence data ((b,w) pairs) that associates with each(w,b) observation an unobserved hidden variable from the set of hidden sentiment factors, $Z=\{z_1...,z_k\}$..Just like in PLSA where hidden factors correspond to the “topics” of the documents, in S-PLSA those factors may correspond to the sentiments embodied in the reviews (e.g., joy, surprise, disgust, etc.). Such sentiments are not directly observable in the reviews; rather, they are expressed through the use of combinations of appraisal words. Hence, we use hidden factors to model sentiments.

For a word-review pair (w,b), S-PLSA models the co occurrence probability as a mixture of conditionally independent multinomial distributions

$$\Pr(b,w)=\sum_{z\in Z}\Pr(z)\Pr(w|z)\Pr(b|z)$$

where we consider both review b and word d to be generated from the latent factor z in similar ways, using the conditional probabilities Pr(b|z) and Pr(w|z), respectively. The assumption made here is that b and w are independent given the choice of the latent factor.

To explain the observed (b,w) pairs, we need to estimate the model parameters Pr(z),Pr(b|z) and Pr(w|z). To this end, we seek to maximize the following likelihood function:

$$L(B,W)=\sum_{b\in B}\sum_{w\in W}c(b,w)\log\Pr(b,w)$$

Where c(b,w) represents the number of occurrences of a pair (b,w) in the data.

A widely used method to perform maximum likelihood parameter estimation for models involving latent variables (such as our S-PLSA model) is the Expectation-Maximization (EM) algorithm [32], which involves an iterative process with two alternating steps.

1. An Expectation step (E-step), where posterior probabilities for the latent variables (in our case, the variable z) are computed, based on the current estimates of the parameters.

2. A Maximization step (M-step), where estimates for the parameters are updated to maximize the complete data likelihood.

In our model, with the parameters Pr(z),Pr(w|z), and Pr(b|z) suitably initialized, we can show that the algorithm requires alternating between the following two steps:

In E-step, we compute

$$\Pr\left(z|b,w\right)=\frac{\Pr(z)\Pr(b|z)\Pr\left(w\mid z)\right)}{\sum_{z'\in z}\Pr(z')\Pr(b\mid z')\Pr(w\mid z')}$$

In M-step, we update the model parameters with

$$\Pr\left(w|z\right)=\frac{\sum_{b\in B} c(b,w)\Pr(z\mid b,w)}{\sum_{b\in B}\sum w'\in Wc(b,w')\Pr(z\mid b,w')}$$

$$\Pr\left(b|z\right)=\frac{\sum_{w\in} c(b,w)\Pr(z\mid b,w)}{\sum_{b\in B}\sum_{w'\in} Wc(b',w)\Pr(z\mid b',w)}$$

and

$$\Pr\left(z\right)=\frac{\sum_{b\in B}\sum_{w\in W} c(b,w)\Pr(z\mid b,w)}{\sum_{b\in B}\sum_{w'\in} Wc(b,w)}$$

It can be shown that each iteration above monotonically increases the complete data likelihood, and the algorithm converges when a local optimal solution is achieved. Once the parameter estimation for the model is completed, we can compute the posterior probability Pr(z|b) using the Bayes rule

$$\Pr(z\mid b)=\frac{\Pr(b\mid z)\Pr(z)}{\sum_{z\in Z}\Pr(b\mid z)\Pr(z)}$$

Intuitively, Pr(z|b) represents how much a hidden sentiment factor z(ϵZ)"contributes" to the review b. Therefore, the set of probabilities $\left\{\Pr(z\mid b)z\in Z\right\}$ can be considered as a succinct summarization of b in terms of sentiments.

3.2 ARSA for BoxOffice Prediction

The temporal relationship between the box office revenues of the preceding days and the current day can be well modeled by an autoregressive process. Let us denote the box office revenue of the movie of interest at day t by x_t (t = 1, 2, . . .,N, where t = 1 corresponds to the release date and t = N corresponds to the last date we are interested in), and we use {xt}(t = 1, . . .,N) to denote the time series x_1, x_2, . . . , x_N. Our goal is to obtain an AR process that can model the time series $\{x_t\}$. A basic (but not quite appropriate, as discussed below) AR process of order p is as follows:

$$X_t=\sum_{i=1}^{p}\phi_i X_{t-i}+\in_t$$

where ϕ1,ϕ2,. . . , ϕp are the parameters of the model, and $\in_t$ is an error term (white noise with zero mean).

After the preprocessing step, a new AR model can be formed on the resulting time series $\{y_t\}$

$$y_t=\sum_{i=1}^{p}\phi_i y_{t-i}+\in_t$$

It is worth noting that although the AR model developed here is specific for movies, the same methodologies can be applied in other contexts. For example, trends and seasonality are present in the sales performance of many different products (such as electronics and music CDs). Therefore, the preprocessing steps described above to remove them can be adapted and used in the predicting the sales performance.

Our new model, which we call the ARSA model, can be formulated as follows:

$$y_t=\sum_{i=1}^{p}\phi_i y_{t-i}+\sum_{i=1}^{q}\sum_{k=1}^{K}\rho_i,k\omega t_{-i,k}+\in_t$$

where p, q, and K are user-chosen parameters, while ϕ_i and $\rho_{i,k}$ are parameters whose values are to be estimated using the training data. Parameter q specifies the sentiment information from how many preceding days is taken into account, and K indicates the number of hidden sentiment factors used by S-PLSA to represent the sentiment information. In summary, the ARSA model mainly comprises two components. The first component, which corresponds to the first term in the right hand side of (2), reflects the influence of past box office revenues. The second component, which corresponds to the second term, represents the effect of the sentiments as reflected from the reviews.

3.3 ARSQA for Quality Factor Prediction

Recall that $\{y_t\}$ denotes the time series representing the sales figures after proper treatment as described in the preceding section. Let v_t be the number of reviews posted at day t. Also, recall that $\psi_{t,j,k}$ is the inferred probability of the kth sentiment factor in the jth review at time t, which we assume can be obtained based on S-PLSA. Denote by $\mu_{t,j}$ the quality of the jth review (either readily available or predicted by some model) on day t. Then, the prediction model can be formulated as follows:

$$y_t=\sum_{i=1}^{p}\phi_i y_{t-i}+\sum_{i=1}^{q}\frac{1}{v_{t-i}}\sum_{j=1}^{vt-i}\mu_{t-i,j}\sum_{k=1}^{K}\rho_{i,k}\psi_{t-i,j,k}+\in_{t,}$$

where p, q, and K are user-defined parameters, ϵ_t is an error term (white noise with zero mean), and ϕ_i,$\mu_{t-i,j}$, and $\rho_{i,k}$ are parameters to be estimated from the training data p and q specify how far the model "looks back" into the history, whereas K specifies how many sentiment factors we would like to consider. What differentiates ARSQA from ARSA is that, the sentiment factors are weighted by the quality of the reviews, which reflects the fact that reviews of Different levels of quality have different degrees of influence on the prediction. With the sentiment and quality factors already known (in the case of available quality ratings) or predicted, parameter estimation (forϕ_i,$\mu_{t-i,j}$, and $\rho_{i,k}$) can be done using least squares regression in a fashion similar to that for ARSA.

4. EXPRIMENTAL WORK

In this section, we report the results obtained from a set of experiments conducted on a movie data set in order to validate the effectiveness of the proposed model, and compare it against alternative methods.

The movie data we used in the experiments consists of three components. The first component is a set of blog documents on movies of interest collected from the Web, the second component contains the corresponding daily box office revenue data for these movies, and the third component consists of movie reviews and their helpfulness scores that are obtained from the IMDB websites.

In each run of the experiment, the following procedure was followed:

1. We randomly choose half of the movies for training, and the other half for testing the blog entries and box office revenue data are correspondingly partitioned into training and testing data sets.
2. Using the training blog entries, we train an S-PLSA model. For each blog entry b, the sentiments toward a movie are summarized using a vector of the posterior probabilities of the hidden sentiment factors, Pr(z|b)
3. We feed the probability vectors obtained in Step 2, along with the box revenues of the preceding days, into the ARSA model, and obtain estimates of the parameters.
4. We evaluate the prediction performance of the ARSA model by experimenting it with the testing data set.

In addition, to evaluate the effectiveness of our quality aware model, we collected movie reviews that were published on the IMDB website.

In this paper, we use the Mean Absolute Percentage Error (MAPE) [19] to measure the prediction accuracy

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\Pr ed_i - True_i|}{True_i}$$

where n is the total amount of predictions made on the testing data, Pred_i is the predicted value, and True_i represents the true value of the box office revenue. In statistics, MAPE is a measure of accuracy in a fitted time series value, specifically trending. The difference between actual value and the forecast value is divided by the actual value. The absolute value of this calculation is summed up for each fitted or forecast point and divided again by the number of fitted points. This makes it a percentage error so we can compare the error of fitted time series.

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,” in ICEC, 2007, pp. 303–310.
3. Y. Liu, X. Huang, A. An, and X. Yu, “ARSA: a sentiment-aware model for predicting sales performance using blogs,” in SIGIR, 2007, pp. 607–614.
4. Y. Liu, X. Yu, X. Huang, and A. An, “Blog data mining: The predictive power of sentiments,” in Data Mining for Business Applications. Springer, 2009, pp. 183–195.
5. L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang, and E. Park, “Flexible frameworks for actionable knowledge discovery,” IEEE Transactions on Knowledge and Data Engineering, vol. 99, no. preprints, 2009.
6. B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in *ACL*, 2004, pp. 271–278.

BIOGRAPHIES

Mr.S.Balaji is working as a assistant professor in V.R.S College of Engineering and Technology. He completed his undergraduate degree in Sri Venkateswara College of Engineering and postgraduate in Prathyusha Institute of Technology and Management.