

# Movie Rateapp: Online Reviews Mining & Predicting Sales Performance

Ruman Bagwan<sup>1</sup> Aditya Bankar<sup>2</sup> Vikram Bankar<sup>3</sup> Chetan Gorane<sup>4</sup> Rahul Samant<sup>5</sup>

<sup>1,2,3,4,5</sup>Department of Information Technology  
<sup>1,2,3,4,5</sup>NBN Sinhgad School of Engineering 410041

**Abstract**— Nowadays posting online review has become a way of expressing our feeling. With the help of reviews the products sales can be affected hugely. The products Economic Value can get fluctuated. In this paper, we are going to focus on Movie Domain and predict the sales of particular movie. For this purpose we are proposing Autoregressive Sentiment-Aware model for sales prediction (ARSA). An Autoregressive Sentiment and Quality Aware model (ARSQA) assist us to further improve the accuracy of prediction by considering the quality factor. Rigorous Experiment conducted on a large movie data set confirms the effectiveness of the proposed approach.

**Key words:** Review mining, Quality analysis, prediction

## I. INTRODUCTION

Posting reviews online has become an increasingly popular way for people to express out there opinion or feeling with other users toward products and services. There is some dedicated review websites that allows people to give there reviews. Reviews can be in the form of blogs posted, social networking sites. Online reviews present a good amount of information about the product and services. And if used properly can be very useful for vendors to improve their business. As a result, review mining has got a good deal of attention.

Recent studies have concentrated on the profit values of reviews, studying the relationship between the sales performance of products and their reviews since what the public think of a product by expressing it in reviews can give you the sells, of a product, understanding the opinions and sentiments expressed in the relevant reviews is of high importance, and can give a very good of the product's future sales performance.

Developing models and algorithms that give a actionable knowledge extracted from reviews. Such models and algorithms can be used to effectively predict the future sales of products, which can in turn provide useful knowledge to a vendor.

Online setting where anybody can post anything, reviews can vary from person to person in terms of quality. Examples of "bad" reviews include very short insulting comments with no substance like "This book sucks", or long and tedious reviews. Reviews poorly written, reviews containing no real meaning, or even spam reviews, may actually negatively affect the accuracy of the prediction, if they are not properly taken care of. Prediction of product sales is a highly domain-driven task, for which a deep understanding of various factors involved is essential.

## II. RELATED WORK

### A. Ranking System:

An important solution to the information overload problem where people find it more and more difficult to identify the useful information effectively .there are three directions 1) content-based filtering: It rely on content descriptions of

behavioral user data, 2) collaborative filtering: takes the rating data as input that constitute the user model, 3) hybrid methods: that combines both content information and collaborative filtering to solve the cold-start problems.

## III. EXISTING APPROACHES

In the past few years, domain driven data mining has emerged as an important new paradigm for knowledge discovery. Motivated by the significant gap between the academic goals of many current KDD methods and the real-life business goals, Domain driven data mining advocates the shift from data-centered hidden pattern mining to domain-driven actionable knowledge discovery (AKD). Mining opinions and sentiments from blogs, which is necessary for predicting future product sales, presents unique challenges that cannot be easily addressed by conventional text mining methods. With the rapid growth of online reviews, review mining has attracted a great deal of attention. Early work in this area was primarily focused on determining the semantic orientation of reviews. Among them, some of the studies attempt to learn a positive/negative classifier at the document level. The existing approach deals with the three streams of forecasting models of motion picture revenues, diffusion models of new product adoption, and methodologies for measuring word-of-mouth. There are large on-line collections of such reviews, and because reviewers often summarize their overall sentiment with a machine-extractable rating indicator, such as a number of stars; Word-of-mouth communicates previous moviegoers' assessment of a movie's quality and might encourage or discourage some prospective moviegoers from watching the movie.

### A. Drawbacks:

- It is mainly focused on the determining the semantic orientation of reviews rather than extract the subjective portion of text, and then feed them into the sentiment classifier.
- Quality-wise, not all reviews are created equal
- Strong correlation between the volume of reviews and sales spikes, using the volume or the link structures do not provide satisfactory prediction performance.
- It is not able to perform visual comparison of consumer opinions.

## IV. PROPOSED METHODOLOGY

### A. ARSA: A sentiment-Aware Model:

The model to provide product sales predictions based on the sentiment information taken from reviews. As our methodologies proposes to focus on the predicting box office revenues. Model aims two different factors that can affect the box office revenues. One factor is the box office revenue of current day. As it state the performance in

upcoming days. Second factor is the users sentiment's about the movies.

#### 1) The Autoregressive Model:

The model that captures the first factor as described in above and avoids the second factor into the model. The relationship between the current and preceding days of the box office revenue can be modeled in autoregressive process. Let us denote the box office revenue of the movie of interest at day  $t$  by  $x_t$  ( $t = 1, 2, \dots, N$ , where  $t = 1$  corresponds to the release date and  $t = N$  corresponds to the last date we are interested in), and we use  $\{x_t\} (t = 1, \dots, N)$  to denote the time series  $x_1, x_2, \dots, x_N$ . 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:

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

Once the model is learned from data present it can predict for next. Note AR model is appropriate for time series that are stationary. As the box office revenues always peak on weekends and low on weekdays. To model order in proper time series some preprocessing steps are required. By computing the difference between the box office revenues of a particular day and seasonality factor due to different days of week.

As AR model not only for the movies domain it can be use in many different product domains (such as electronics and music CD's) and used in predicting sales performance.

#### 2) Helping the ARSA Model:

The ARSA model involves the learning the set of parameters obtain from training data that consist of box office revenues and from the reviews data.

For a particular movie  $m (m=1, 2, \dots, M)$ , where  $m$  is the total number of movies in the training data, and a given date  $t$ , let us add the subscript  $m$  to  $y_t$  and  $w_{t-i, j}$  in to be more precise.

Let

$$\alpha_{m, t} = (\gamma_{m, t-1}, \dots, \gamma_{m, t-p}, w_{m, t, 1}, \dots, w_{m, t-q}, k).$$

#### B. The Quality Factors In Prediction:

The quality factor means the accuracy. Reviews vary a great deal in quality. Quality of review can lead in greater prediction of accuracy. In this we predict the quality of reviews using the extension of ARSA, which is called the ARSQA model.

##### 1) Modeling Quality:

We will develop the model on the bases of writhing style on previous study which shows the syntactical features like speech inductor of the reviews .we will tag the review on the bases of following:

- 1) Qualifiers: one that modifies, reduce, tempers, or restrains (e.g., quite, rather, enough).
- 2) Modal auxiliaries: a type of verbs used to indicate modality. They give additional information about the function of the main verb that follows it (e.g., can, should, will).
- 3) Nominal pronouns (e.g., everybody, nothing).
- 4) Comparative and superlative adverbs: indicators of comparison (e.g., harder, faster, most prominent).

- 5) Comparative and superlative adjectives: indicators of comparisons, again (e.g., bigger, chief, top, largest).
- 6) Proper nouns: reference to a specific item, which will begin with a capital letter no matter where it occurs in a sentence (e.g., Caribbean, Snoopy).
- 7) Interjections/exclamations: strong signs of opinion (e.g., ouch, well).
- 8) wh-determiners (e.g., what, which), wh-pronouns (who, whose, which), and wh-adverbs: wh-words that signify either questions or other interesting linguistic constructs such as relative clauses. (e.g., how, where, when).

In addition, we even compare the movie review from the IMDB web site. To ensure the robustness of the predictive model, we only consider the reviews that have received at least 10 votes. And the purpose of training and testing the reviews with usefulness is used.

#### V. FUTURE WORK

For future work, like to explore its role in clustering and classification of reviews based on their sentiments. It would also be interesting to explore the use of S-PLSA as a tool to help track and monitor the changes and trends in sentiments expressed online. Also note that the ARSA and ARSQA models are general frameworks for sales performance prediction, and would certainly benefit from the development of more sophisticated models for sentiment analysis and quality prediction.

#### VI. CONCLUSION

As the general Comments & Blogs has stated Uniqueness in the general public Sentiments and the business intelligence. in this paper As we driving the sentiments out of Comments and Blogs which helps to derive the power of predictivity and predicting sales performance of Movie. ARSA is a model to predict Sales Performance of Box office .the Quality model which is ARSQA a general framework for quality sales performance predictions, which benefits in the development of more sophisticated model for Sentiment Analysis and Quality Prediction.

#### REFERENCES

- [1] "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 2012)
- [2] Predicting Missing Items in Shopping Carts(Ieee Transactions On Knowledge And Data Engineering, Vol. 21, No. 7, July 2009)
- [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] A. Ghose and P. G. Ipeirotis, "Designing novel review ranking systems: predicting the usefulness and impact of reviews," in ICEC, 2007, pp. 303-310.
- [5] C. Dellarocas, X. M. Zhang, and N. F. Awad, "Exploring the value of online product ratings in revenue forecasting: The case of motion pictures," Journal of Interactive Marketing, vol. 21, no. 4, pp. 23-45, 2007.