Effective Approach for Data Utilization in the Online Marketplace

Tianrui Lin[1], Siwei Yan[1], Yuxin Wang[2]

[1]No.2 High School of East China Normal University, Shanghai;

[2]Shanghai Weiyu International School, Shanghai.

Abstract

In the big data era, it is important for customer-based companies to collect the data of customer and make good use of them. In this paper, we are asked to help Sunshine Company to analyze the data. Firstly, we process data for further analysis, including data screening, data imputation and string segmentation. In data screening, useless or abnormal data for product measures is removed. To deal with the missing data used in time series model, the parabolic interpolation method is employed in this paper. The most essential part of data processing is the conversion of text-based review, so we firstly segment the text and then covert each word to vector space for the next evaluation model. Secondly, term frequency–inverse document Frequency (TF-IDF) method is applied to evaluate the weight of text-based reviews to build data measures model on the basis of rating and reviews. Considering the importance of star rating and reviews, we create weighted scoring model to measure the data. In this way, each data can be measured quantitatively. Next, we build two kinds of time series prediction model to predict the product's reputation, which are long short-term memory recurrent neural network (LSTM RNN) and autoregressive integrated moving average (ARIMA) model. We calculate the reputation by monthly star rating and convert these data into time series. Then the processed data is utilized to create the prediction model and the error analysis results indicate LSTM RNN model has higher accuracy than that of ARIMA model. So time series prediction model using LSTM RNN is employed to predict the product's reputation. Afterwards, the product's potential evaluation model is established with the combination of text-based review and star rating based on time series model. In this model, we firstly calculate the comprehensive monthly scores of products based on proposed weighted scoring model and then predict the product's average scores for next 6 month. To evaluate whether the product is potentially successful or not, we calculate the average scores of products positioned at top 10% as a successful metric, but that positioned at last 10% as a failing metric, which are 1.31 and -0.46 respectively. If the predicted score is higher than 1.31, the product is potentially successful; if lower than -0.46, it's likely to be unsuccessful; if fall in [-0.46, 1.31] interval, it is potentially mediocre. And we also find the common features of products at top 10%: nice figure and good durability, which can be reference to new products' design features. In terms of causality between specific stars and reviews, we do Spearman correlation analysis using time series data, then we find the lower stars rating level the product had in a period of time, the more negative reviews will appear then. And to figure out the relevance between descriptors in reviews and star rating levels, we calculate weighting of every descriptor in reviews based on TF-IDF to find the three most weighted descriptors in each star rating level. To verify the strong association between the descriptors and star rating, we also calculate the proportion of different rating level in each descriptor contrarily, and then we find many relevance, for example, once positive descriptors such as "Great", "love", "nice" and "ok" occur in reviews, high rating levels will be obtained then. Finally, our models perform well in sensitivity analysis, and we write a letter to Sunshine Company summarizing all our findings and suggestions on data using.

Keywords

TF-IDF; Time Series Prediction; LSTM RNN; ARIAMA; Spearman Correlation Analysis.

1. Introduction

1.1 Background

Amazon provides customers with an opportunity to rate and review purchases, and then collect these data for further analysis. The data will be used to gain insights into the markets in which they participate, the timing of that participation, and the potential success of product design feature choices. It is quite important for the marketplace to make good use of data for further development.

1.2 Problem restatement and analysis

The questions and corresponding analysis can be summarized as follows:

(1) **Question 1**

Identify data measures based on ratings and reviews.

Analysis: At first, we should process the data and find a proper way to convert the text-based review, and then build a comprehensive measure models

(2) **Question 2**

Identify and discuss time-based measures and patterns within each data set that might suggest that a product's reputation is increasing or decreasing in the online marketplace.

Analysis: Time series analysis will be suitable for this question

(3) **Question 2**

Identify and discuss time-based measures and patterns within each data set that might suggest that a product's reputation is increasing or decreasing in the online marketplace.

Analysis: Time series analysis will be suitable to analyze the product's reputation. And we can compare different kinds of time series prediction model and select the better one.

(4) **Question 3**

Determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product

Analysis: The measures of product's potential should consider reviews and star rating based on time series, and we can combine the model established in question 1 and question 2.

(5) **Question 4**

Do specific star ratings incite more reviews? For example, are customers more likely to write some type of review after seeing a series of low star ratings?

Analysis: First to do is code the data as time series, and then Spearman correlation analysis can be applied to find the relevance.

(6) **Question 5**

Are specific quality descriptors of text-based reviews such as 'enthusiastic', 'disappointed', and others, strongly associated with rating levels?

Analysis: We can evaluate the text-based descriptors by weight, and find the three most weighted descriptors in each rating level.

2. Basic assumption and Symbols

2.1 Assumption

(1) According to Amazon.com, one to two stars indicate negative ratings, three stars are neutral, and four to five stars are positive ratings

(2) The reviews and star rating from the customers that are invited to become Amazon Vine Voices are more reliable than normal reviews

(3) The data provided is reliable and can reflect the real comments of customers

(4) The reviews of customers who are neither invited to become Amazon Vine Voices nor buy the products at Amazon have less reference value.

2.2 Symbols

Symbols	Definition
N	the number of documents in the corpus
$F_{k,j}$	the frequency of all terms t_k that occur in document d_j
S_1	Score of reviews
S_2	Score of star ratings
α	weight of reviews
β	weight of star rating,
y	the real output in time series prediction model
p	the predicted output in time series prediction model

More specific definitions of symbols in models are illustrated in corresponding section

3. Data processing

3.1 Data screening

The three data sets provided by Sunshine Company contain tens of thousands of data and it's quite large. Therefore, to save computational time, we should do data screening at first according to the usefulness of the information. Considering the situation in real-world online shopping, if a customer does not buy the product but make review and rating on this product, he or she may be hired to make false evaluations on specific products, therefore, these kind of reviews and ratings are biased and thus useless. However, some good customers are invited to become Amazon Vine Voices, they are free to use the products to make real and valuable evaluations on products, so their evaluations should be more expensive and important for reference. Then we delete some useless data base on the metric shown in Figure 5.1. Note that: we code N as 0 and Y as 1.

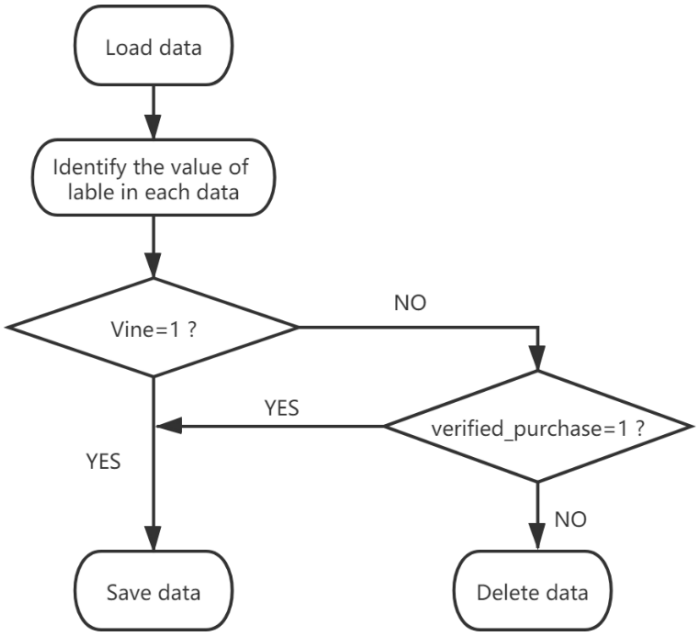

Figure 3.1: Metric to delete useless data

3.2 Missing data imputation

In the database, missing data is quite common that may cause bias on the analysis results to some extent. Therefore, we should do data imputation to eliminate the uncertain bias as much as possible. Considering the task 2, we are asked to identify and discuss time-based measures and patterns within

each data set that might suggest that a product's reputation is increasing or decreasing in the online marketplace. However, the data provided is not continuous in date, which will influence our further analysis. Therefore, we should take some methods to fill up the missing data. And there are plenty of methods for data imputation, such as Stochastic regression, mean substitute, multiple imputation and parabolic interpolation method so on.

In the next section, we will use the parabolic interpolation method to fill up the missing data when evaluating the product's reputation.

3.3 Convert the text to vector space

For further analysis of reviews, the texts of reviews are supposed to be quantizable [1]. Therefore, we parse the text and form a vector. Every review is regarded as a document while a word in the review is a term. All control characters, spaces between words, dots, commas, and similar characters are removed in the parsing process. And then the text can be segmented and converted to vector. An example from the data file "**hair_dryer. tsv"** is shown as follows to illustrates the procedure.

Example:

Text: "Petite, but Powerful Performance."

Preprocessing: Petite but Powerful Performance

Base vector: mainVec = [Petite but Powerful Performance]

mainVec [0] =Petite

mainVec [1] =but

mainVec [2] =Powerful

mainVec [3] =Performance

In addition, the structure of the main vector object is shown in Figure 5.2

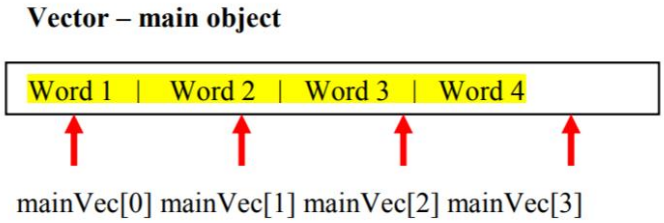

Figure 3.2 Graphical presentation of vector

4. Data measures model based on rating and reviews

4.1 Evaluate text-based review by TF-IDF method

After data processing, every review has been transferred into vector space, each document is represented by a vector in a *n*-dimensional space, where each dimension corresponds to a term from the overall vocabulary of a given document collection, therefore, we obtained Vector Space Model (VSM). Then we should weight the terms and measure the feature vector similarity. TF-IDF (Term Frequency-Inverse Document Frequency) method [2], a widely used weighting method, is employed in this study. Note that: in our model, every review is considered as a document. There are three basic assumptions of TF-IDF method [3]:

(1) **IDF assumption:** rare terms are not less relevant than frequent terms

(2) **TF assumption:** multiple occurrences of a term in a document are not less relevant than single occurrences

(3) **Normalization assumption:** long documents are not preferred to short documents

The TF-IDF function can be expressed as follows:

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) \cdot log \frac{N}{n_k} \quad (4.1)$$

Where *N* is the number of documents in the corpus, and n_k the number of documents in the collection in which the term *tk* occurs at least once.

$TF(t_k, d_j)$ can be calculated by Equation (4.2):

$$TF(t_k, d_j) = \frac{F_{k,j}}{\max f_{z,j}} \quad (4.2)$$

where $F_{k,j}$ is the frequency of all terms t_k that occur in document d_j. To ensure the weights to fall in the [0,1] interval and the documents to be represented by equal-length vectors. The weight obtained by Equation (4.1) can be normalized by cosine normalization:

$$w_{k,j} = \frac{TF-IDF(t_k,d_j)}{\sqrt{\sum_{s=1}^{|T|} TF-IDF(t_S,d_j)^2}} \quad (4.3)$$

Then we divide the database into 5 collections according to the star rating, the structure can be described as Figure 4.1:

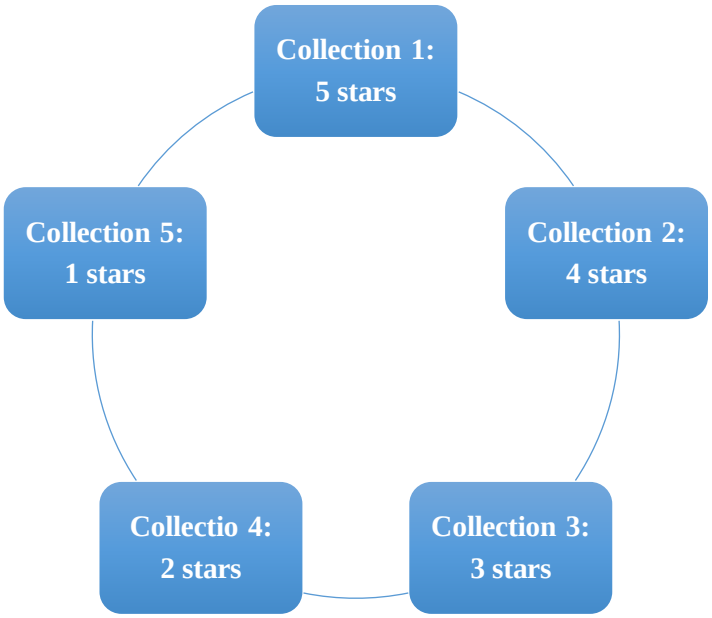

Figure 4.1: Classification of initial data

Next, we use the TF-IDF method to calculate the weight of each review (we define the reviews are formed by both headline and body), and then define their final review scores by the Equation (4.4):

$$S_1 = Star \times w_{k,j}$$

Where *Star* means star rating corresponding to the reviews. According to Amazon.com, one to two stars indicate negative ratings, three stars are neutral, and four to five stars are positive ratings. Therefore, we define the *Star* can be calculated by

$$Star = \begin{cases} 2, \ if \ the \ \text{corresponding rating is } 5 \\ 1, \ if \ the \ \text{corresponding rating is } 4 \\ 0, \ if \ the \ \text{corresponding rating is } 3 \\ -1, \ if \ the \ \text{corresponding rating is } 2 \\ -2, \ if \ the \ \text{corresponding rating is } 1 \end{cases} \quad (4.4)$$

For example, “Works great”, a review from **“hair_dryer.tsv”** ,calculated by TF-IDF method, its weight is 0.19, and its corresponding rating is 5, therefore, its final score is $2 \times 0.19 = 0.38$. Using the proposed method, all of reviews in the database are transferred into scores.

4.2 Comprehensive measures model based on ratings and quantitative reviews

To measure a data of a product, we take both the ratings and reviews into consideration, because they are most informative for Sunshine Company to track. The assessment procedure can be summarized as follows:

Step 1: Extract the information of rating and review

Step 2: Transfer the text of review into vector space by the proposed method in section 3.2

Step 3: Calculate the score of review S_1 by the proposed method in section 4.1

Step 4: Calculate the score of star rating S_2 based on Equation (4.4)

Step 5: Obtain the final score of data base on Equation (4.5)

$$Score = v(\alpha S_1 + \beta S_2) \tag{4.5}$$

Where α and β is the weight of reviews and star rating, and they are defined to be 0.4 and 0.6 respectively in our model, because we think direct ratings are more valuable than reviews. And v represents an adjustment coefficient, which is used to enhance the value of reviews and rating from customers invited to become Amazon Vine Voice, their reviews and rating will be more reliable and meaningful. Note that: if the reviews and ratings come from customers invited to become Amazon Vine Voice, v is 1.2, otherwise v is 1.

Then the data measures model based on rating and reviews has been built, we can measure all the data via MATLAB software for the next analysis.

5. Product's reputation times series prediction model

As we all know, a product's reputation may increase or decrease in the online marketplace during a period. To analysis the rules of reputation's change according to time-based data, two kinds of time series prediction models are employed in this section for comparison.

5.1 Long short-term memory recurrent neural network (LSTM RNN) model

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture, which is widely used in time series prediction model. The LSTM RNN model performs well when the objective functions are non-linear and complicated. A LSTM cell is composed by an input gate, an input modulation gate, an output gate and a forget gate [4], the structure of LSTM cell is shown in Figure 5.1.

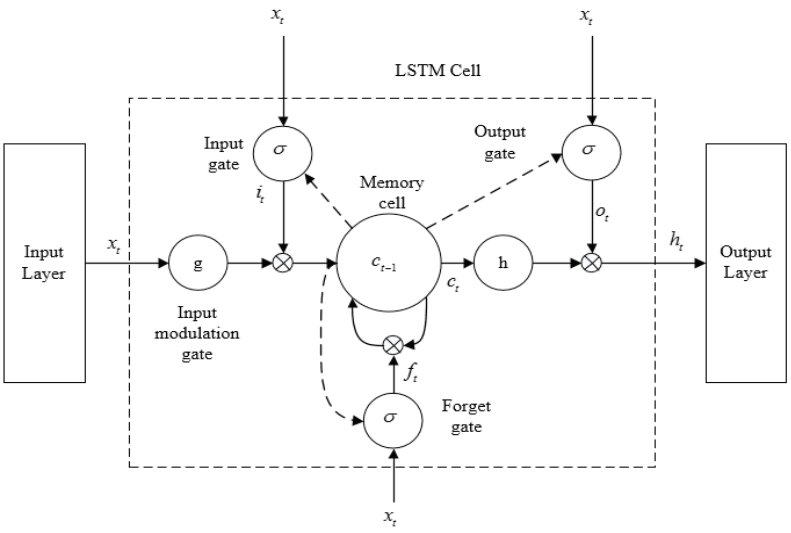

Fig 5.1. Structure of LSTM RNN cells

Define the input time series as $X = (x_1, x_2, \dots, x_n)$, hidden state of memory cells as $H = (h_1, h_2, \dots, h_n)$, output time series as $Y = (y_1, y_2, \dots, y_n)$. LSTM RNN run the computation as below:

$$h_1 = H(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \quad (5.1)$$

$$p_t = W_{hy}h_{t-1} + b_y \quad (5.2)$$

Where *W* is the weight matrices and *b* are bias vectors. The hidden state of memory cells is computed as follows [5]:

$$i_t = \sigma(W_{ix}x_t + W_{hh}h_{t-1} + W_{ic}c_{t-1} + b_i) \quad (5.3)$$

$$f_t = \sigma\left(W_{fx}x_t + W_{hh}h_{t-1} + W_{fc}c_{t-1} + b_f\right) \quad (5.4)$$

$$c_t = f_t{}^* c_{t-1} + i_t{}^* g(W_{cx}x_t + W_{hh}h_{t-1} + W_{cc}c_{t-1} + b_c) \quad (5.5)$$

$$\sigma_t = \sigma(W_{\sigma x}x_t + W_{hh}h_{t-1} + W_{oc}c_{t-1} + b_o) \quad (5.6)$$

$$h_t = o_t{}^* h(c_t) \quad (5.7)$$

Where g and h are extends of stand sigmoid function with the range changing to [-2,2] and [-1,1.],* stands for the scalar product of two matrices or vectors, and the standard sigmoid function defined in Equation (5.8) is denoted as σ.

$$\sigma(x) = \frac{1}{1+e^x} \quad (5.8)$$

As for objective function, the square loss function as shown in Equation (5.9) is used.

$$e = \sum_{i=1}^{n}(y_t - p_t)^2 \quad (5.9)$$

Where *y* and *p* are the real output and predicted output respectively.

5.2 Autoregressive integrated moving average (ARIMA) model

ARIMA model is also a popular algorithm used in time series analysis, it performs well when the objective function and constraints are linear. In an ARIMA (p, d, q) model, the predicted value of a variable is assumed to be a linear function of several past observations and random errors [6-7]. Equation (5.10) can illustrate how ARIMA model generates the time series with the mean μ

$$\emptyset(B)\nabla^d(y_t - \mu) = \theta(B)a_t \quad (5.10)$$

Where y_t is the actual value and a_t is the random error at time period *t*. a_t is assumed to be independently and identically distributed with a mean of zero and a constant variance of σ^2;

$\emptyset(B) = 1 - \sum_{i=1}^{p}\varphi_i B^i$, $\theta(B) = 1 - \sum_{j=1}^{1}\theta_j B^j$ are polynomials in *B* of degree *p* and *q*, φ_i and θ_j are l parameters in ARIMA model ; $\nabla = (1 - B)$, *B* is the backward shift operator, *p* and *q* are integers and often referred to as orders of the model, and *d* is an integer and usually referred to as the order of differencing.

5.3 Application of two models in product's reputation prediction

We define the product's reputation can be measured by the average rating of a month because it's the most intuitional way to judge a product. However, the data provided is not continuous in time, therefore, the parabolic interpolation method is applied to do missing data imputation. The computational procedure of product's reputation prediction model based on time series prediction model can be expressed as follows:

Step 1: Data processing

Extract time and rating data from the database and do data imputation to fill up the missing data by parabolic interpolation method. Then calculate the average rating of product in each month and give the processed data a serial number starting from 1 according to the time series. Note that: the unit of time series model is a month in this model. And the range of rating is [1,5] that is quite narrow, therefore, we needn't do normalization processing.

Step 2: Construct time series prediction model by LSTM RNN and ARIMA

Load the processed data, and then take 80% data as training points while 20% as test points. The predicted value depends on the data of last three months. Then apply two algorithms respectively in MATLAB, set proper parameters for models.

Step 3: Error and sensitivity analysis of two model

Test the accuracy of the model and do the sensitivity analysis of two model

Step 4: Comparison between two models

Compare the accuracy of two models and select the better one for product's reputation prediction.

5.4 Error analysis and comparison of two models

To test the accuracy of our model, we use both mean square error (MSE) and mean average error (MAE) as measures, they are defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{5.11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{5.12}$$

Where $\widehat{y_i}$ is the predicted value and y_i is the real value.

We take the data of a product with the product parent ID "732252283" as an example to compare LSMT RNN and ARIMA. The time of rating data is from January, 2012 to September, 2015, at a total of 45 sample points (missing two months data, we fill up by parabolic interpolation method). Then calculating in MATLAB, we obtained the results shown in Table 5.1 and the Figure 5.2 is the comparison among the predicted value of two models and the real value.

Table 5.1 Error analysis

	LSTM RNN	ARIMA
MSE	0.2421	0.4412
MAE	0.4121	0.9412

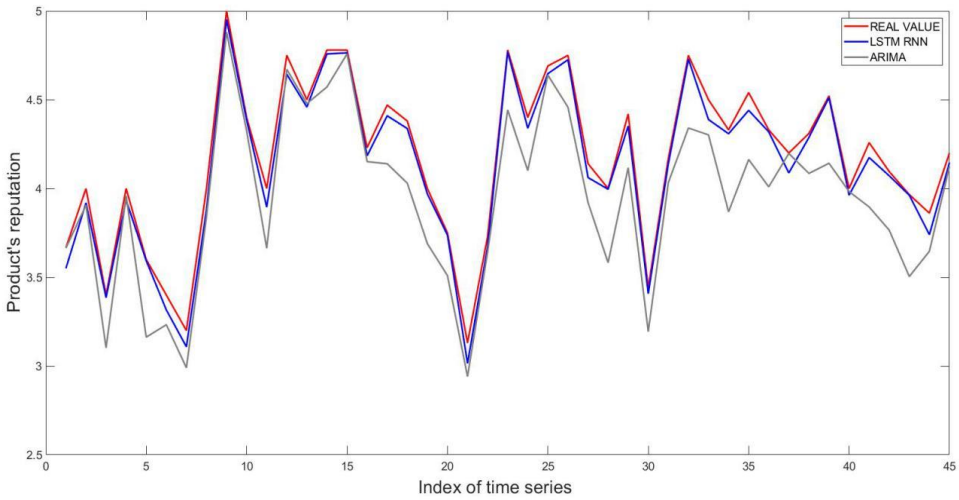

Figure 5.2 Comparison among real value, LSMT RNN and ARIMA over selected 45 time series

It can be concluded from the results that time series model using LSMT RNN algorithm performs better than that of ARIMA in accuracy. Therefore, the time series model using LSMT RNN algorithm is built to predict the product's reputation.

6. Comprehensive measure model to evaluate the product's potential

As mentioned above, we have built data measures model based on rating and reviews and time series prediction model to predict product's reputation. In this section, we will combine the two models to construct a comprehensive measure model to evaluate the product's potential, which will be useful for Sunshine Company to judge a product.

6.1 How to evaluate a product is successful or not

To evaluate a product comprehensively, we should consider both of the reviews and the rating, however, after analyse the data provided we find a product may have high rating and good reviews in a certain period of time but may also obtain negative reviews and low rating sometimes. Therefore, we should evaluate a product in long-term development rather than a specific period. The evaluation model built in section can be used to calculate the comprehensive scores in a specific period, and then the scores can be taken as input data to predict the performance of products in the future.

6.2 Product's potential prediction based on time series model

The time series prediction model using LSTM RNN can be used to predict the future of products based on their last data. In this model, the scores of products in last are set as input data while the future scores of products are set as the output data. The following is the computational procedure:

(1) **Step1: Data processing and preparation**

Transfer the text-based reviews into vector and calculate their weighting by TF-IDF method, then evaluate the monthly average scores of each product. Note that: the unit of time series model is month in this study.

(2) **Step2: Construct time series model using LSTM RNN**

Input: the monthly average scores of a product, 80% of samples are used for training while 20% of samples are used for test.

Set parameters of LSTM RNN: initial hidden layer=10, feedback delays = 1:3

(3) **Step3: Error analysis of model**

Evaluate the accuracy of models by MSE and MAE, if the accuracy is low, the parameters of model will be adjusted until high accuracy models are obtained.

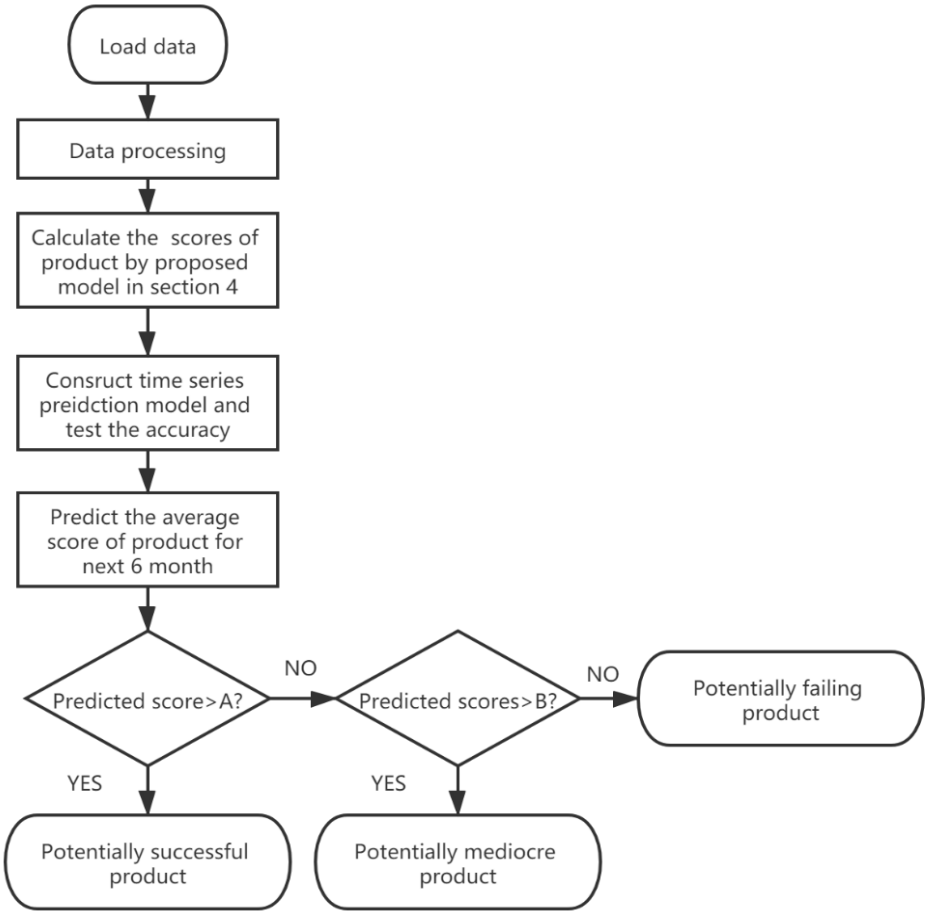

Figure 6.1: Process of comprehensive evaluation model for product's potential

(4) **Step4: Product's potential evaluation**

Predict the scores of products for the next 6 months, and then evaluate the potential of product via comparison with the metric score *A*, which is the average scores of products positioned at top 10% ranked by scores in database. If the scores of products are higher than the threshold, we can say they are potentially successful products, otherwise we continue to compare the scores with *B*, which is the average scores of products positioned at last 10% in database, if the scores of products are lower than the threshold, they are defined as potentially failing products. In addition, the products that scores between *A* and *B* are regarded as potentially mediocre products.

The process is illustrated in Figure 6.1.

6.3 Model Application and the features of the most potential products

Using the proposed model, we evaluate all the products in the database, and then we obtain the value of *A* is 1.31 and B is -0.46. Therefore, the potential of every product's can be evaluated by the proposed model. The products positioned at top 10% and last 10% ranked by scores in the database are listed in the appendix.

After analysis the reviews of the products at top 10%, we obtain the conclusion that they have a common feature: nice figure and good durability, which can be a valuable reference for Sunshine company to make strategies for new products

7. Relevance analysis between star rating levels and text-based review

7.1 Analysis of causality between specific stars and reviews based on the time series model

According to the proposed model in section 4, we convert the data into the time series, which will be easy to analyze the causality between specific stars and reviews. To find the impact the low star rating may cause, we select the data with low star rating and analyze the time-neighboring data, meanwhile the record the corresponding helpful votes. We define and calculate some parameters for next correlation analysis, they are the number of helpful votes in a review (defined as N_{help}), the number of negative reviews (quality of reviews are evaluated by proposed model in section 4.2) and good reviews of time-neighboring data (defined as N_1 and N_2). Then we calculate these parameters to do Spearman correlation analysis [8] in SPSS, the results are shown in Table 7.1:

Table 7.1. Spearman correlation coefficient

	Star_rating	N_help	N1	N2
Star_rating	1	0.4321545	-0.9142121	-0.1754521
N_help	0.4321545	1	-0.1754521	0.2124511
N1	-0.9142121	-0.1754521	1	-0.184271573
N2	0.8941211	0.2124511	-0.184271573	1

The results indicate that the star rating level has a strong association with the number of reviews and rating of time-neighboring data, and we can find the correlation coefficient between N_1 and star rating

level is negative, which indicates the lower stars rating level the product had in a period, the more numbers of negative reviews will be incited then. In a similar way, the higher stars rating level the product had in a period, the more numbers of good reviews will be incited then.

7.2 Analysis of descriptors of text-based reviews and star rating levels

To analyze if the specific quality descriptors of text-based reviews are strongly associated with rating level, we can use the TF-IDF method mentioned above to find the three most weighted descriptors in each rating levels. According to Amazon's rating rules, we divide the rating into three levels as shown in Figure 7.1

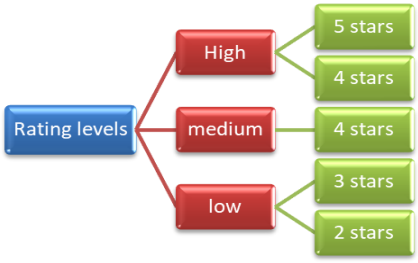

Figure 7.1. Illustration of rating levels

In each rating level, we calculate the weighted of every descriptor in reviews based on TF-IDF to find the three most weighted descriptors, the results are shown in Table 7.2

Table 7.2 Top 5 word in each rating level

	High	Medium	Low
1	Great	ok	Bad
2	Love	nice	Junk
3	Perfect	Little	disappointed

To identify if these kinds of words are strongly associated with rating levels, we calculated the proportion of different rating level in each descriptor in return in the database and the results are shown in Figure 7.2

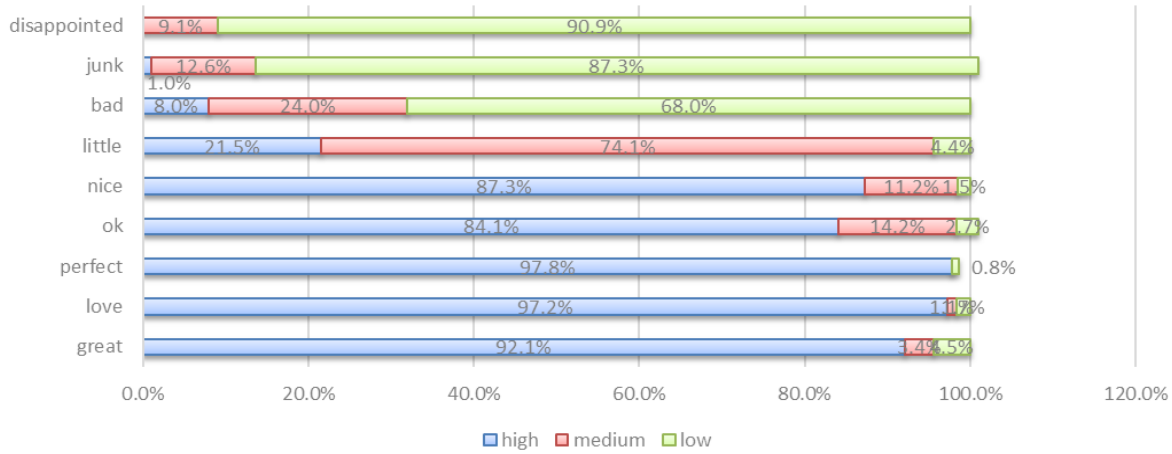

Figure 7.2 Proportion of different rating level in each descriptor

What can be concluded from the above results is that the rating levels do associate with specific quality descriptors in reviews and the relevance can be summarized as follows:

(1) The descriptors such as "Great", "love", "perfect", "nice" and "ok" usually indicate high rating levels (more than two stars rating).

(2) The descriptors such as “little” most likely obtain a medium rating level.

(3) As we known, descriptors such as “Bad”, “junk” and “disappointed” are often used for something bad. And according to the statistical analysis, of course, once they appear in the review, low rating levels will follow

8. Sensitivity analysis of models

To test the sensitivity of time series prediction model using LSTM RNN, we selected another product’s data, and change the parameters of LSTM RNN model, the hidden layer changed from 10 to 30, the results show that the model still has great accuracy, which indicate the model is good at sensitivity. Figure 8.1 is the structure of LSTM RNN model. Figure 8.2 and Figure 8.3 illustrate the error analysis of model.

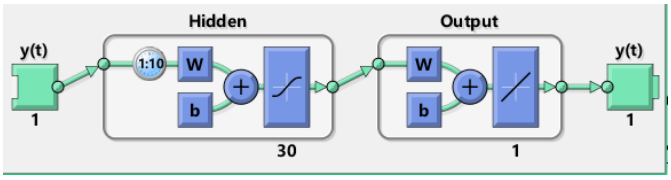

Figure 8.1 The structure of LSTM NN model

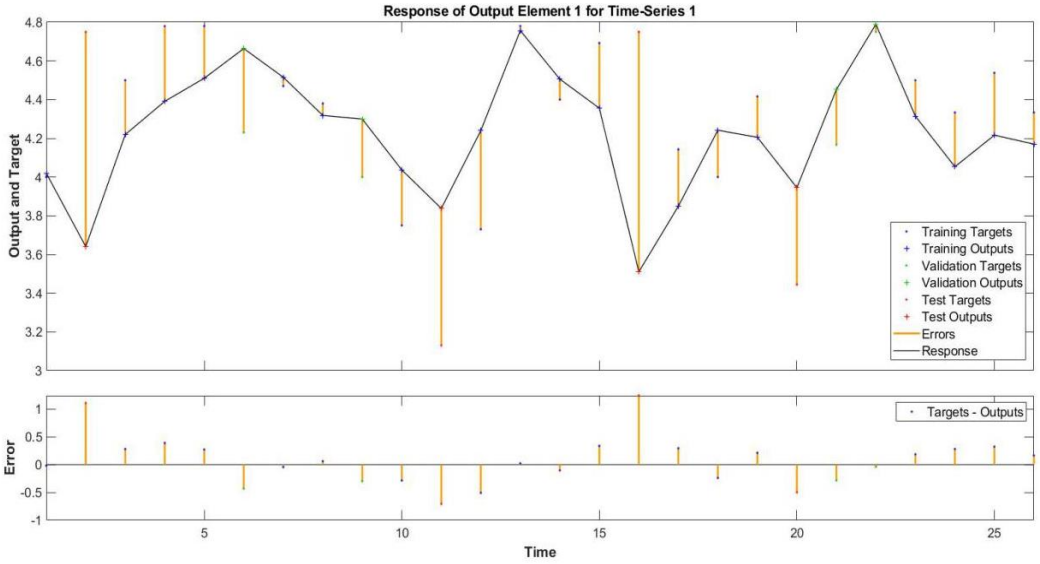

Figure 8.2 Response of time series model

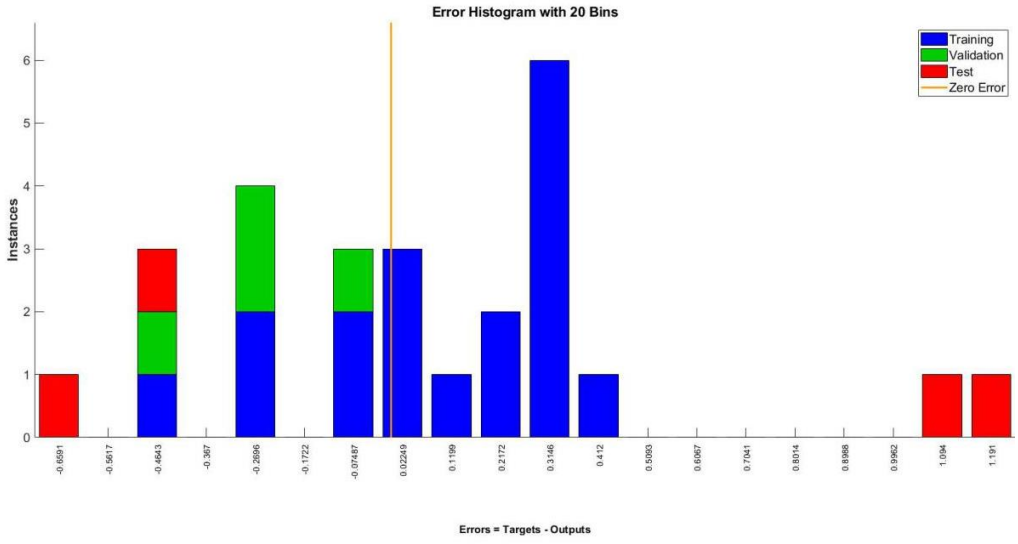

Figure 8.3 Error histogram of time series model

Therefore, the product's reputation can be predicted using the proposed model in this section, which will be helpful for Sunshine Company to make strategies for new products.

9. Conclusion and evaluation of models

9.1 Conclusions of each question

(1) **For question 1**: The text-based reviews are evaluated by TF-IDF model, and weighted scoring model based on reviews and star rating is established to measure the data.

(2) **For question 2:** We convert the data to time series, then analyze and predict the product's reputation by LSTM RNN model. The model shows high accuracy.

(3) **For question 3:** Time series prediction model using LSTM RNN is established to evaluate the potential of products. We select the products at top 10% as a metric to judge if a product is successful. And analyze the features of products at top 10% by the reviews in database, we find the similarity among them: nice figure and practicability, which is worth of reference for Sunshine company to design new products

(4) **For question 4:** The lower stars rating level the product had in a period, the more numbers of negative reviews will be incited then. In a similar way, the higher stars rating level the product had in a period, the more numbers of good reviews will be incited then.

(5) **For question 5:** The descriptors such as "Great", "love", "perfect", "nice" and "ok" usually indicate high rating levels (more than two stars rating); the descriptors such as "Great", "love", "perfect", "nice" and "ok" usually indicate high rating levels (more than two stars rating). As we known, descriptors such as "Bad", "junk" and "disappointed" are often used for something bad. And according to the statistical analysis, of course, once they appear in the review, low rating levels will follow

9.2 Strength and Weakness

Strength

(1) In data processing, we use the parabolic interpolation method to fill up the missing data. And covert the text to vector space for further evaluation.

(2) To evaluate text-based reviews, TF-IDF method is employed in this paper, which is popular in the field of recommender system

(3) Two kinds of time series prediction model are compared in order to select the one with higher accuracy for product's reputation prediction. And the LSTM RNN model proves to be more reliable.

(4) Evaluate the potential of products by comprehensive time-based measures, which is reasonable and practical.

Weakness

(1) Emotion-based measures of text can be used to obtain a better evaluation.

(2) More time series models can be employed for comparison.

Acknowledgements

These authors are contributed equally to this work.

References

[1] Trstenjak, B, Mikac, S., & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. (Vol.69, pp.1356-1364). Elsevier Ltd.

[2] Wu, Ho Chung, Luk, Robert Wing Pong, Wong, Kam Fai, & Kwok, Kui Lam. Interpreting tf-idf term weights as making relevance decisions. Acm Transactions on Information Systems, 26(3), 1-37.

[3] Yacine Rezgui. (2007). Text-based domain ontology building using tf-idf and metric clusters techniques. Knowledge Engineering Review, 22(4), 379-403.

[4] Guo, T., Xu, Z., Yao, X., Chen, H., Aberer, K., & Funaya, K. (2016). Robust Online Time Series Prediction with Recurrent Neural Networks. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). doi:10.1109/dsaa.2016.92.

[5] Wen, Tsung-Hsien, Gasic, Milica, Mrksic, Nikola, Su, Pei-Hao, Vandyke, David, & Young, Steve. (2015). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. Computer Science.

[6] Paul Newbold. (1983). Arima model building and the time series analysis approach to forecasting. Journal of Forecasting, 2(1), 23-35.

[7] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, & Charles Korede Ayo. (2014). Stock price prediction using the ARIMA model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.

[8] Chengwei Xiao, Jiaqi Ye, Rui Máximo Esteves, & Chunming Rong. (2016). Using spearman's correlation coefficients for exploratory data analysis on big dataset. Concurrency & Computation Practice & Experience, 28(14), 3866-3878.