IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 5, 2019, accepted May 19, 2019, date of publication May 29, 2019, date of current version July 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919734

Online Sales Prediction: An Analysis With Dependency SCOR-Topic Sentiment Model

LIJUAN HUANG[1], ZIXIN DOU[1], YONGJUN HU[1], AND RAOYI HUANG[2]
[1]School of Management, Guangzhou University, Guangzhou 510275, China
[2]Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Zixin Dou (13826064897@163.com)

This work was supported in part by the 2015 National Social Science Fund of China under Grant 15BGL201, in part by the 2018 National Social Science Fund of China under Grant 18BGL236, in part by the ‘‘13th Five-Year Plan’’ Thinktank Project of Social Sciences of Guangzhou under Grant 2018BGZZK23, and in part by the Research Plan for Graduate Joint-Training Programs of Guangzhou University.

ABSTRACT This study aims to find a robust method to improve the accuracy of online sales prediction. Based on the groundings of existing literature, the authors proposed a Dependency SCOR-topic Sentiment (DSTS) model to analyze the online textual reviews and predict sales performance. The authors took the online sales data of tea as empirical evidence to test the proposed model by integrating the auto-regressive review information model into the DSTS model. The findings include: 1) the effect of the distribution of SCOR-topic from reviews on sales prediction; 2) the effect of review text sentiment on sales prediction increases as the specific topic probability dominates; and 3) the effect of review text sentiment on sales prediction increases as the rest topic probability evenly distributes. These findings demonstrate that the DSTS model is more precise than alternative methods in online sales prediction. This study not only contributes to the literature by pointing out how the distribution of sentiment topic impacts on sales prediction but also has practical implications for the e-commerce practitioners to manage the inventory better and advertise by this prediction method.

INDEX TERMS Sentiment analysis, SCOR-topic distribution, sales prediction.

I. INTRODUCTION

Online reviews play essential roles in shaping customers' awareness and perceptions about products [1]. They have become popular with the development of the Internet. Many e-commerce websites such as JD have established online review systems to encourage consumers to post product reviews [2], and this enables companies to predict sales performance before making operation decisions. Therefore, online reviews are considered as the main driver for product sales prediction.

A considerable number of researches have studied the relationship between online reviews and product sales [3], [4]. Although most pieces of evidence suggest that online reviews have an impact on predicting sales, these findings are not always consistent. Ye *et al.* [5] and Segal *et al.* [6] found that the volume and the star ratings of reviews have a positive effect on product sales. However, Hu *et al.* [7] pointed out, the content of review information is more important than the simple statistics such as ratings and volume. Accordingly, a growing number of researchers studies the sentiment embedded in the reviews to forecast product sales. Liu and Zhang [8] found that the sentiment index extracted from online reviews was applied in many fields, particularly in box office [9], automotive industry [10] and stock market [11]. It has been reported that combining the sentiment of reviews has the potential to predict sales more precisely [12]. These indicate that sentiments of reviews are capable of helping businesses with forecasting sales performance. However, most of the existing studies have not further examined the distribution of sentiment from online reviews. This paper applies a DSTS model to examine the sentiment's distribution from online reviews. Some research has studied the relationship between textual topics and product sales. For example, Yu *et al.* [13] found that sentiment topics have a substantial effect on box office forecasting. However, they did not consider the relationship between the specific topics o review's contents and sales. Therefore, it is challenging to find which

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

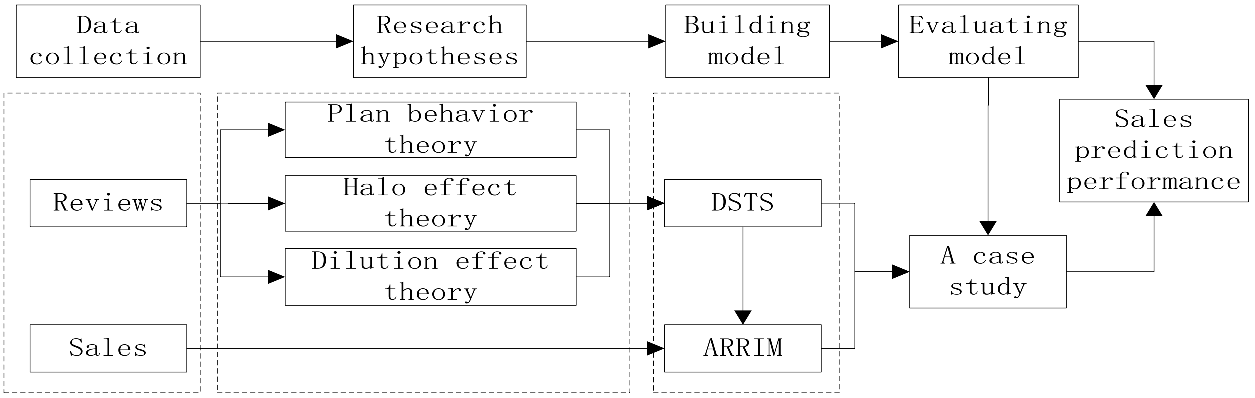


FIGURE 1. Research framework.

kind of the topic's distribution of online reviews can be evaluated meaningfully by companies. Some studies explored the relationship between specific topics and sales performance. For example, Park and Ha [3] thought those specific topics accounted for the upward or downward trends of sales performance. The phenomena above indicate that specific topics are possibly attributable to the trends. However, it did not further investigate the joint effect of the relationship between the specific topic's distribution and product sales prediction.

While previous studies have recognized the importance of sentiment topics of reviews, few studies examine the relationship between the topic's distribution of reviews and sales prediction. This study takes efforts to see how the topic's distribution are related to sales prediction. The main objective of this study is to investigate how the distribution of the sentiment topic influences sales prediction.

These findings imply theoretical bases. They explain the reasons why the sentiment information embedded in the reviews can help businesses forecast more accurately. Because the existing studies have not investigated the effects between the topic's distribution and sales prediction, this is the first study to investigate how the distribution of topics optimizes product sales forecasting.

These results also contribute to the practical application. This research finds which kind of the distribution of the topic of online reviews can be evaluated as useful or helpful by companies. It may help related parties know that the review text sentiments and their relationship with product sales prediction. For example, related parties could improve their products and services to drive customers to publish sentiment words, which is more closely related to sales.

The rest of the paper is organized as follows. The next section presents the methodology, including data collection, research hypotheses, and building model. Then the empirical analyses are present, including procedures and results. Finally, the conclusion is presented.

II. METHODOLOGY

The purpose of this paper is to investigate the relationship between the topic's distribution and sales prediction. The research framework is illustrated in Figure 1:

- Data collection: two types of data are collected, i.e., historical sales, online reviews (involving number and content of reviews).
- Research hypotheses: Based on three theories (i.e., the theory of plan behavior, the theory of halo effect, and theory of dilution effect), hypotheses are developed to examine how textual topic distribution influences sales prediction.
- Model development: A sentiment prediction method is presented. It not only analyzes the review data, calculates, maps the sentiment factors to SCOR-topic, allowing users to simplify the opinions, but also it makes use of the characteristics of reviews, including the volume and value information to predict sales performance.

A. DATA COLLECTION

In this study, two types of data, namely historical sales and online reviews (involving number and content of reviews), were collected. For a specific online shop, reviews are collected from one day before the sales to 22 days later. A group of 16,155 online reviews of 5 different online tea shops was collected from JD.com through the crawler program developed by the authors. Meanwhile, the historical sales data of the 5 shops are also collected from the same website.

B. RESEARCH HYPOTHESES

Consumers at times may find it difficult to make purchase decisions based on the online reviews when they are aimless. In this case, they may read the review texts carefully for each product. If consumers have a positive attitude towards the quality of the product, they are more likely to choose a product which has the quality content of reviews. Such phenomena could be explained by the planned behavior. The planned behavior indicates that consumer's attitude is one of the determinant factors influencing purchasing decisions. To better find out the consumer's attitudes hidden in the review, the specific topics of review contents need to be distinguished.

The Supply-Chain Operations Reference (SCOR) model is our first choice. It could allow checking each whole topic, based on the attitudes of all aspects of the reviews. We map the relevant sentiment words based on their attributes to construct SCOR-topics. Thus, the following hypotheses are proposed:

H1: The distribution of SCOR-topics from online textual reviews positively influences on sales prediction.

Consumers may miss helpful reviews when encountering too many review texts. Such phenomena could be explained by the halo effect [14]. The halo effect indicates that global evaluation of things can be affected by the evaluation of the attributes, even when there is sufficient information to make independent assessment [14]. Consumers may feel more comfortable and secure about sentiment information when they observe specific topics of review. Therefore, the following hypotheses are proposed:

H2: The moderating effect of the review text sentiment on sales prediction will increase as the specific topic probability domination.

Though most of the consumers may immediately be aware of their emotional information they need based on the observation of the comments of a particular topic, consumers may doubt the credibility of information when reading many review texts that they need. However, Lee *et al.* [12] found that consumers may increase the credibility of the particular information when observing evenly distributed review text sentiment of the irrelevant information. Such phenomena could be explained by the dilution effect. The dilution effect referring to irrelevant information could weaken the effect on individual stereotypes [15]. That is, after offering a large number of specific commodity information to the customer, meanwhile, through providing a small amount of irrelevant information for him, it can reduce the stereotype, which will positively influence product sales. Hence, the following hypotheses are proposed:

H3: The moderating effect of the review text sentiment on sales prediction will increase as the rest topic probability evenly distributed, under the circumstance of the specific topic probability domination.

C. MODEL DEVELOPMENT

1) DSTS MODEL

In this study, appraisal words are exploited to compose the feature vectors for reviews, which are then used to infer the hidden sentiment factors. It is supposed that a set of reviews $D = \{d_l, d_2, \ldots, d_N\}$ are given, and a set of words (appraisal words) from a vocabulary $W = \{w_1, w_2, \ldots, w_M\}$. The website review entry is associated with some unobserved hidden factors, $K = \{k_I, k_2, \ldots, k_K\}$. Then, the distribution of sentiment of each review is calculated by using equation (1):

$$P(W|K) = \frac{n_k^w + \beta_w}{\sum_{w=1}^{W} n_k^w + \beta_w}, \tag{1}$$

where n_k^w represents the frequency of sentiment words in topics, β_w is the parameter of the Dirichlet prior on the per-topic wth-word distribution.

To this end, this paper puts forward the DSTS, which is used to integrate and analyze the emotional information variables. This model maps emotional words into the plan (P), demand (D), quality (Q), and logistic (L) topics, and works out the cumulative sum of all the words appearing on the topic distribution under each topic. People who reach a consensus on the most important demand-side believe that this model can search the special sentiment topics effectively, which can influence the public's purchasing decisions and optimize future product sales prediction.

After mapping, the word group W is divided into four groups, $PW = \{pw_1, pw_2, \ldots, pw_{M1}\}$ represent the plan word group, $DW = \{dw_1, dw_2, \ldots, dw_M 2\}$ represent the demand word group, $QW = \{qw_1, qw_2, \ldots, qw_{M3}\}$ is the quality word group and $LW = \{lw_1, lw_2, \ldots, lw_{M4}\}$ is the logistic word group.

As discussed above, sales might be greatly influenced by people's opinions at the same time. This paper uses a different approach to the pre-processing sentiment index. Let K be the number of topics for a given tea on day t, and $\varepsilon_{t,k,w}$ be the probability of the w-th sentiment word conditional on the kth sentiment factor on day t, i.e, $\varepsilon_{t,k,w} = P(W = w|t, k)$. Then, the average probability of factor $K = k$ at a time t is defined as the equation (2):

$$\omega_{t,w} = \frac{1}{K}\sum_{k=1}^{K} \varepsilon_{t,k,w} = \frac{1}{K}\sum_{k=1}^{K} P(W = w|t, k) \tag{2}$$

where $W = \{pw, dw, qw, lw\}$. Then the equation can be rewritten as the equation (3):

$$\begin{aligned}\omega_{t,w} = &\frac{1}{K}\sum_{k=1}^{K} P(PW = pw|t, k)\\ &+ \frac{1}{K}\sum_{k=1}^{K} P(DW = dw|t, k)\\ &+ \frac{1}{K}\sum_{k=1}^{K} P(QW = qw|t, k)\\ &+ \frac{1}{K}\sum_{k=1}^{K} P(LW = lw|t, k).\end{aligned} \tag{3}$$

Then sum up the plan words, the demand words, the quality words, and the logistic words to generate an independent SCOR-topic. The formula is equation (4):

$$\omega_{t,w} = Plan_t + Demand_t + Quality_t + Logistic_t. \tag{4}$$

Intuitively, $\omega_{t,w}$ represents the average fraction of the sentiment "mass" that can be attributed to the hidden sentiment topic K. $\omega_{t,w}$ represents the average fraction of the sentiment "mass" that can be attributed to the sentiment word w. $Plan_t, Demand_t, Quality_t, Logistic_t$ represent the sum of the sentiment SCOR topic that can be attributed to the sentiment word *PW, DW, QW, LW*, respectively. Finally, there are 11 words mapped into P topic, 14 words are mapped into D topic, 86 words are mapped into Q topic, 5 words are mapped into L topic.

2) ARRIM

A prediction model is proposed. The equation (5) as follows:

$$Sales_{i,t} = \lambda Sales_{i,t-1} + \delta Number_{i,t-1} + \sum_{w=1}^{W} \beta_w \omega_{i,t-1,w}. \tag{5}$$

After mapping all sentiment words to the SCOR-topic, Equation (5) can be converted as equation (6):

$$\begin{aligned} &Sales_{i,t} \\ &= \lambda Sales_{i,t-1} + \delta Number_{i,t-1} + \beta_1 Plan_{i,t-1} \\ &\quad + \beta_2 Demand_{i,t-1} + \beta_3 Quality_{i,t-1} + \beta_4 Logistic_{i,t-1} \end{aligned} \tag{6}$$

The dependent variable $Sales_{i,t}$ denotes the value of sales of tea i on day t. Its lagged variable is defined as $Sales_{i,t-1}$. $Number_{i,t-1}$ denotes the number of reviews of tea i on day *t-1*. $Plan_{t-1}, Demand_{t-1}, Quality_{t-1}, Logistic_{t-1}$ are the key variables that denote the SCOR-topics of the review text sentiment of tea *i* on day *t-1*. Where W and K are user-chosen parameters, while λ, δ and β_w are parameters whose values are to be estimated using the training data.

Let $b_{i,t} = (Sales_{i,t-1}, Number_{i,t-1}, Plan_{i,t-1}, Demand_{i,t-1}, Quality_{i,t-1}, Logistic_{i,t-1})^{\mathrm{T}}$. Then, equation (6) can be rewritten as the equation (7):

$$b_{i,t}^{\mathrm{T}} H = Sales_{i,t}. \tag{7}$$

Let B be a matrix composed of all $b_{i,t}$ vectors corresponding to each tea shop and each day, i.e., $B = (b_{1,1}, b_{1,2}, ..)^T$. Similarly, let C be a vector composed of all possible, i.e., $C = (Sales_{1,1}, Sales_{1,2}, ..)^T$. Then, based on the training and testing data, this paper seeks to find a solution C for the equation $BH \approx C$.

III. EMPIRICAL ANALYSIS

A. PROCEDURE

In each stage of the experiment, the procedures are as follows:

- 80% of the tea sales dataset are randomly chosen for training, and the rest for testing; the historical sales data and online reviews data (involving number and content of reviews) are correspondingly partitioned into training and testing data sets.
- Using the training reviews content, a DSTS Model is trained. For each review, the sentiments are summarized by using a vector of the posterior probabilities of the sentiment index.
- Feed the probability vectors obtained in the DSTS Model, along with the sales of the previous days and the number of reviews of the previous days, into the ARRIM model.
- Comparing the prediction performance of different value of topic in the ARRIM model, forecasting performance is evaluated by using specific measures.

TABLE 1. Comparison with other methods.

Model	TEA 1	TEA2	TEA3	TEA 4	TEA 5	MEAN
DSTSM	1.43%	3.18%	1.00%	1.09%	1.76%	1.69%
LDA	1.80%	3.70%	0.99%	0.86%	1.80%	1.83%
PLSA	1.96%	3.28%	1.08%	0.98%	1.86%	1.83%

TABLE 2. Mean value of different prediction model.

VARIABLE	TEA 1	TEA2	TEA3	TEA 4	TEA 5
Sales	1382	537	834	726	462
Num	151	98.10	115.24	96.10	74.24
P	0.0610	0.0791	0.0776	0.0759	0.0845
D	0.0201	0.0038	0.0367	0.0703	0.0425
Q	0.8893	0.8905	0.8458	0.8231	0.8228
L	0.0296	0.0266	0.0399	0.0307	0.0502
R1	0.2405	0.2330	0.2260	0.2585	0.2610
R2	0.2612	0.2434	0.2375	0.2384	0.2453
R3	0.2435	0.2333	0.2717	0.2599	0.2432
R4	0.2547	0.2903	0.2648	0.2432	0.2504

In this study, the Mean Absolute Percentage Error (MAPE) is utilized to measure the prediction accuracy as the equation (8):

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|\mathrm{Pred}_i - True_i|}{True_i} \tag{8}$$

where n is the total number of predictions made on the testing data, *Pred* is the predicted value, and *True* represents the actual value of the sales.

B. RESULTS

1) EVALUATING MODEL

This section aims to prove the effectiveness of SCOR-topics as the input vectors, and we experimentally compare ARRIM model with those using other topic distributions. For all the model, the number of the topic is set to 4.

As shown in TABLE 1, the DSTS model is generally superior to the LDA model and the PLSA model. That is, the prediction model is more accurate for SCOR-topics, which supports hypothesis *H1*.

2) A CASE STUDY

Descriptive statistics of the sample is presented as TABLE 2 and TABLE 3, where P is the plan topic, D is the demand topic, Q is the quality topic, L is the logistic topic and *R1-R4* are random topics. tea1-tea5 are different shop respectively. Variable sales and review number are non-negative and integer. Their standard deviation is less than their mean. It is unnecessary to pre-process the data. Also, from the sentiment topics distribution, it is found that how people express emotions when purchasing goods. From TABLE 2 and TABLE 3, it showed that SCOR-topics distribution was decentralized, where the quality theme was dominant, and the remaining three topics (plan, demand, logistic) were inconsistent, which is supportive to our hypothesis *H2*.

Additionally, as shown in TABLE 1, TABLE 2 and TABLE 3, as the distribution of quality topics dominates,

TABLE 3. STD value of different prediction model.

VARIABLE	TEA 1	TEA2	TEA3	TEA 4	TEA 5
Sales	28.07	66.34	57.91	44.28	38.56
Num	64.84	99.23	16.69	17.92	12.46
P	0.0138	0.0549	0.0249	0.0249	0.0237
D	0.0122	0.0060	0.0150	0.0162	0.0195
Q	0.0294	0.0515	0.0374	0.0364	0.0436
L	0.0150	0.0264	0.0182	0.0156	0.0250
R1	0.0528	0.0685	0.0433	0.0456	0.0578
R2	0.0455	0.0698	0.0374	0.0484	0.0544
R3	0.0575	0.0786	0.0428	0.0364	0.0516
R4	0.0323	0.0821	0.0415	0.0411	0.0538

TABLE 4. The correlations of SCOR topic.

CLUSTER	NUM	P	D	Q	L
Tea 1	0.079	-0.329	-0.138	0.009	0.399
	0.733	0.145	0.55	0.97	0.073
Tea 2	.900**	-0.296	738**	0.022	0.405
	0	0.193	0	0.924	0.069
Tea 3	-0.413	0.136	0.339	-0.292	0.135
	0.062	0.557	0.133	0.198	0.56
Tea 4	-.503*	0.367	0.384	-.507*	0.2
	0.02	0.101	0.085	0.019	0.386
Tea 5	-.564**	-0.192	0.364	-0.182	0.214
	0.008	0.404	0.105	0.43	0.351

*****.Correlation is significant at 0.01 level (2-tailed), *. Correlation is 0.05.***

TABLE 5. The extrema of SCOR topic.

ERROR VALUE	P	D	Q	L
Mean	0.0235	0.0502	0.0677	0.0236
Std	0.0411	0.0136	0.0221	0.0115

TABLE 6. The error of an independent method for sales prediction.

METHOD	TEA 1	TEA2	TEA3	TEA 4	TEA 5	MEAN
M1	1.43%	3.18%	1.00%	1.09%	1.76%	1.69%
M2	1.31%	3.03%	0.96%	1.67%	1.57%	1.71%
M3	1.49%	3.17%	0.97%	0.87%	1.65%	1.63%
Original	1.43%	3.18%	1.00%	1.09%	1.76%	1.69%

the effect of commenting on the textual mood on sales prediction optimizes, which is consistent with hypothesis *H2*. This finding suggests that when people have the same preference for the same tea, it could be difficult for them to realize which emotions can help them to make a decision effectively if the sentiment topics are distributed evenly. When the sentiment topics distribution gap is large, it is possible to find the emotional information needed immediately when observing specific review information.

At the same time, the influence of the review data and the different SCOR-topic on sales volume is examined, which aims to find out the factor which is most related to the sales and consistent with the sales volume's changing pattern.

From TABLE 4, it is concluded that the only *L* topic is positively correlated to tea sales among all the correlation coefficient in the ARRIM model and SCOR-topics.

To further know the reason for the correlation between the topic distribution and sales prediction, extreme condition (maximum-minimum value) is performed from four SCOR topics' means and standard deviations in TABLE 2 and TABLE 3, to evaluate the dispersion of relevant data in different tea stores. In general, the difference between any two units' standard deviation does not exceed the range. The larger the range is, the larger the degree of dispersion will be [16].

As shown in TABLE 5, the L topic is positively related to sales, because its Mean dispersion is smaller than other topics. However, although the P topic's mean range is the smallest, its standard deviation range is more significant than its mean range, which is not consistent with the range operation requirements. Furthermore, it could be concluded that when the topic distribution's range is small, it is positively correlated with sales.

In a research context, sales prediction might be dominated by a specific topic, as the specific topic is exposed to the review websites. However, there is a relationship between the distribution of topic and the review text sentiment. In order to measure the moderating effect of the specific topic distribution of the review text sentiment, a forecasting model is applied to moderate each topic distribution value to examine their increasing effect on sales prediction. The following are the strategies taken.

Method1 (*M1*): Raise the probability distribution of *P* topic, add the corresponding shops' *P* topic of probability distribution mean to the daily *P* topic distribution, which accordingly reduces the same number of *Q* topic.

Method2 (*M2*): Raise the probability distribution of *D* topic, add the corresponding shops' *D* topic of probability distribution mean to the daily *D* topic distribution, which accordingly reduces the same number of *Q* topic.

Method3 (*M3*): Raise the probability distribution of *L* topic, add the corresponding shops' *L* topic of probability distribution mean to the daily *L* topic distribution, which accordingly reduces the same number of *Q* topic.

As can be seen in TABLE 6, the impact of the *L* topic on sales is generally superior to other topics, which are in line with the correlation test results in TABLE 4. Furthermore, from TABLE 6 it is found that, although the result of *M3* is better than those of *M1* and *M2* in total error, a portion of the prediction errors in 5 shops are lower than those of *M1* and *M2*. Therefore, we need to consider combining strategies to verify which SCOR-topic commonly have a more significant impact on sales.

Method4 (*M4*): Raise the probability distribution of *P* and *D* topic, add the corresponding shops' probability distribution mean of these topics to the daily *P* and *D* topic distribution, which accordingly reduces the same number of *Q* topic.

Method5 (*M5*): Raise the probability distribution of *D* and *L* topic, add the corresponding shops' probability distribution mean of these topics to the daily *D* and *L* topic distribution, which accordingly reduces the same number of *Q* topic.

TABLE 7. The errors of the hybrid method for sales prediction.

Method	Tea 1	Tea2	Tea3	Tea 4	Tea 5	Mean
M4	1.30%	2.97%	0.95%	1.65%	1.57%	1.69%
M5	1.37%	2.97%	0.93%	1.50%	1.47%	1.65%
M6	1.48%	3.13%	0.95%	2.71%	1.64%	1.98%
M7	62.94%	2.89%	0.93%	1.58%	1.53%	13.98%
Original	1.43%	3.18%	1.00%	1.09%	1.76%	1.69%

TABLE 8. The mean of SCOR-topic of different methods.

Method	Variable	Tea 1	Tea2	Tea3	Tea 4	Tea 5	Mean
M1	P	0.1220	0.1582	0.1552	0.1518	0.1690	0.1512
	D	0.0201	0.0038	0.0367	0.0703	0.0425	0.0347
	Q	0.8283	0.8114	0.7683	0.7471	0.7383	0.7787
	L	0.0296	0.0266	0.0399	0.0307	0.0502	0.0354
M2	P	0.0610	0.0791	0.0776	0.0759	0.0845	0.0756
	D	0.0402	0.0077	0.0733	0.1407	0.0850	0.0694
	Q	0.8692	0.8867	0.8092	0.7527	0.7803	0.8196
	L	0.0296	0.0266	0.0399	0.0307	0.0502	0.0354
M3	P	0.0610	0.0791	0.0776	0.0759	0.0845	0.0756
	D	0.0201	0.0038	0.0367	0.0703	0.0425	0.0347
	Q	0.8597	0.8639	0.8060	0.7923	0.7726	0.8189
	L	0.0592	0.0531	0.0798	0.0614	0.1004	0.0708
M4	P	0.1220	0.1582	0.1552	0.1518	0.1690	0.1512
	D	0.0402	0.0077	0.0733	0.1407	0.0850	0.0694
	Q	0.8082	0.8076	0.7316	0.6768	0.6958	0.7440
	L	0.0296	0.0266	0.0399	0.0307	0.0502	0.0354
M5	P	0.0610	0.0791	0.0776	0.0759	0.0845	0.0756
	D	0.0402	0.0077	0.0733	0.1407	0.0850	0.0694
	Q	0.8396	0.8601	0.7693	0.7220	0.7301	0.7842
	L	0.0592	0.0531	0.0798	0.0614	0.1004	0.0708
M6	P	0.1220	0.1582	0.1552	0.1518	0.1690	0.1512
	D	0.0201	0.0038	0.0367	0.0703	0.0425	0.0347
	Q	0.7988	0.7849	0.7284	0.7164	0.6881	0.7433
	L	0.0592	0.0531	0.0798	0.0614	0.1004	0.0708
M7	P	0.1220	0.1582	0.1552	0.1518	0.1690	0.1512
	D	0.0402	0.0077	0.0733	0.1407	0.0850	0.0694
	Q	0.7786	0.7810	0.6917	0.6461	0.6456	0.7086
	L	0.0592	0.0531	0.0798	0.0614	0.1004	0.0708

Method6 (*M6*): Raise the probability distribution of *P* and *L* topic, add the corresponding shops' probability distribution mean of these topics to the daily *P* and *L* topic distribution, which accordingly reduces the same number of *Q* topic.

Method7 (*M7*): Raise the probability distribution of *P*, *D* and *L* topic, add the corresponding shops' probability distribution mean of these topics to the daily *P*, *D* and *L* topic distribution, which accordingly reduces the same number of *Q* topic.

As can be seen in TABLE 7, the impact of the probability distribution of the promotion *R* and *L* topic on sales (*M5*) is generally superior to other portfolio strategies. Although its effect predicted is slightly lower than that of *M4* in tea1 case sample and that of *M7* in tea2 case sample, it is found that the rest of error result of M5 are better than those of *M4* and *M7*. As shown in TABLE 6 and TABLE 7, *M5* is generally superior to other combination strategies. Nevertheless, it is still slightly lower than the results of *M3*, which ranks in second place. At the same time, it is found that four of five shops' sales prediction in *M5* are better than that of *M3*. As a result, M5 is adopted. At the same time, these results match with our hypotheses to some extent. That is, the SCOR-topics play a leading role in the sales prediction under the background of the quality topic dominant.

To better discover the mutual effects between topics, we further analyze every SCOR-topic's distribution in different strategies. As shown in TABLE 8, sales prediction error in *M5* is low, because the rest of the three topics' distributions are even under the quality topic dominant. In other words, on the premise of the *Q* topic dominant, when the distribution of the *P*, *D, L* topic is the same, the effect of the text of the reviews on the sales prediction is optimized. The above-concluded phenomena are in favor of the hypothesis *H3*. It reveals that because people have the same preferences for the same tea, most consumers may immediately be aware of their emotional information they need based on the observation of the specific topic of the comments. However, when consumers encounter more commentary texts than they need, they may doubt the credibility of the information. When consumers observe that the text of the review emotion is evenly distributed across the remaining irrelevant information (e.g., *P*, *D, L*), the stereotype can be reduced, which increases the creditability of the quality topic information from other sides. As a result, the reviews, which are evenly distributed among the rest of unrelated emotional topics, have a positive impact on product sales prediction.

IV. CONCLUSION

While existing studies have been conducted to examine the importance of sentiment topics of reviews, the number of studies is still small on the relationship between the distribution of the topics of reviews and sales prediction. This study fills some gap in the literature to find out how the distribution of sentiment topic is related to sales prediction.

In this paper, we conduct a case study in the tea industry and focus on the problem of the topic's distribution of reviews for predicting product sales performance. The paper firstly developed a DSTS Model to extract the sentiment SCOR-topic information from online textual reviews; Also, we prove that the prediction accuracy is improved through integrating the ARRIM into DSTS model. These findings could explain why the sentiment information embedded in the reviews can help the business to make more accurate forecasting.

Furthermore, the experiment results show that the DSTS Model is more precise than other methods, and the topic's distribution has three positive moderating impacts on product sales: (1) The distribution of SCOR-topic from online textual reviews positively influences on sales prediction. (2) The moderating effect of the review text sentiment on sales prediction increases as the specific topic probability domination. (3) The moderating effect of the review text sentiment on sales prediction increases as the rest topic prob-ability evenly distributed, under the circumstance of the specific topic probability domination. These results also show that quality topic information is more likely to attract consumers to buy products, and the distribution of the rest irrelevant topics have particular impacts on sales. It indicates that consumers are more likely to trust review information when they equally encounter specific topic and the unrelated information review texts. Through applying the proposed method, companies will be able to better harness the predictive power of reviews and manage their business more effectively.

Like other studies, this study has limitations. This analysis is limited to online users who leave reviews at a Chinese review website. Hence, this analysis focuses on review texts written in Chinese. It would be interesting if future research expands the study to a global context.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable remarks and comments.

REFERENCES

[1] S. Xiao, C.-P. Wei, and M. Dong, "Crowd intelligence: Analyzing online product reviews for preference measurement," *Inf. Manage.*, vol. 53, no. 2, pp. 169–182, 2016.

[2] K. Zhao, A. C. Stylianou, and Y. Zheng, "Sources and impacts of social influence from online anonymous user reviews," *Inf. Manage.*, vol. 55, no. 1, pp. 16–30, 2018.

[3] K. Park and S. H. Ha, "Mining user-generated opinions online with LDA model to discover service complaints," *Int. Inf. Inst. (Tokyo). Inf.*, vol. 21, no. 3, pp. 875–884, Mar. 2018.

[4] Z. Li and A. Shimizu, "Impact of online customer reviews on sales outcomes: An empirical study based on prospect theory," *Rev. Socionetw. Strategies*, vol. 12, no. 2, pp. 135–151, 2018.

[5] Q. Ye, R. Law, and B. Gu, "The impact of online user reviews on hotel room sales," *Int. J. Hospitality Manage.*, vol. 28, no. 1, pp. 180–182, 2009.

[6] J. Segal, M. Sacopulos, V. Sheets, I. Thurston, K. Brooks, and R. Puccia, "Online doctor reviews: Do they track surgeon volume, a proxy for quality of care?" *J. Med. Internet Res.*, vol. 14, no. 2, p. e50, 2012.

[7] N. Hu, L. Liu, and J. J. Zhang, "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects," *Inf. Technol. Manage.*, vol. 9, no. 3, pp. 201–214, 2008.

[8] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. Boston, MA, USA: Springer, 2012, pp. 415–463.

[9] M. Hur, P. Kang, and S. Cho, "Box-office forecasting based on sentiments of movie reviews and Independent subspace method," *Inf. Sci.*, vol. 372, pp. 608–624, Dec. 2016.

[10] Z.-P. Fan, Y.-J. Che, and Z.-Y. Chen, "Product sales forecasting using online reviews and historical sales data: A method combining the bass model and sentiment analysis," *J. Bus. Res.*, vol. 74, pp. 90–100, May 2017.

[11] R. Batra and S. M. Daudpota, "Integrating StockTwits with sentiment analysis for better prediction of stock price movement," in *Proc. IEEE Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Mar. 2018, pp. 1–5.

[12] J. H. Lee, S. H. Jung, and J. Park, "The role of entropy of review text sentiments on online WOM and movie box office sales," *Electron. Commerce Res. Appl.*, vol. 22, pp. 42–52, Mar./Apr. 2017.

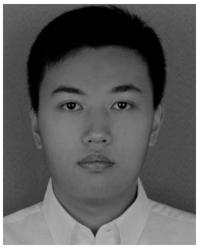
[13] X. Yu, Y. Liu, X. Huang, and A. An, "Mining online reviews for predicting sales performance: A case study in the movie domain," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 720–734, 2012.

[14] R. E. Nisbett and T. D. Wilson, "The halo effect: Evidence for unconscious alteration of judgments," *J. Personality Social Psychol.*, vol. 35, no. 4, p. 250, 1977.

[15] R. E. Nisbett, H. Zukier, and R. E. Lemley, "The dilution effect: Nondiagnostic information weakens the implications of diagnostic information," *Cogn. Psychol.*, vol. 13, no. 2, pp. 248–277, 1981.

[16] G. B. Thomas, Jr., M. D. Weir, J. Hass, C. Heil, and A. Behn, *Thomas' Calculus: Early Transcendentals*. Boston, MA, USA: Pearson, 2016.

LIJUAN HUANG received the M.A. and Ph.D. degrees from Nanchang University, in 1991 and 2006, respectively. She is currently pursuing the Ph.D. degree with the Jiangxi University of Finance and Economics. She is currently the Director of the Academy of E-Commerce Research, Guangzhou University. Her current research interests include e-commerce and logistics, and supply chain management.



ZIXIN DOU received the B.S. degree in mathematics and applied mathematics from Guangzhou University, Guangdong, China, in 2016, where he is currently pursuing the M.S. degree in technology economy and management. From 2017 to 2018, he was a Visiting Research Student with The University of Sydney, Australia. His current research interest includes e-commerce and sentiment analysis.

YONGJUN HU received the B.S. degree from Shandong University, in 2000, and the M.A. degree in electronic and communication engineering and the Ph.D. degree in management science and engineering from Sun Yat-sen University, in 2005 and 2013, respectively. From 2016 to 2018, he was a Visiting Research Fellow with The University of Sydney, Australia. He is currently the Director of the Institute of Business Intelligence and Data Science, Guangzhou University. His current research interest includes machine learning and sentiment analysis.

RAOYI HUANG is currently pursuing the bachelor's degree with The Hong Kong Polytechnic University. She has a strong interest in e-commerce, data analytics, modeling, and statistics. She has been to USA, U.K., and Australia for further study.