Research on Text Emotion Analysis and Product Performance based on NLP and VAR Model

Zhongji Wei*, Baiyang Han, Enxian Zhou
School of Economics and Management, Tsinghua University, Beijing, 100084
*Corresponding author email: zhongjiwei@tsinghua.edu.cn

Abstract. Customers’ evaluations, in terms of text or ratings, are important sources of information for online shops to decide products and marketing strategies. In this paper, we discuss methods to conclude and predict merchants’ performance and characteristics of customers’ behavior of target products based on Amazon dataset. First, we design a Natural Language Processing (NLP) model, Amazon-BERT, to predict text to rating based on transfer learning. We improve the model accuracy by 20% compared to the original pre-trained BERT, and we use the predicted rating in our following steps of modeling. Second, we use PCA to extract 5 principle components, which estimate review’s popularity, product’s reputation and so on. And then, for pacifier, we find that vector auto regression (VAR) model fits better, we apply it to the regression and doing a Granger causality test. Finally, we looked at the impulse response and variance decomposition. All 3 factors included do have a small impact.

1. Introduction
Customer feedback has always been an important indicator for suppliers to decide what and how many products to supply. Specifically, for online business, customers’ ratings and reviews provide not only a kind of feedback, but also an atmosphere of public opinion which has a big impact on sales (Wang, 2015). Now, Sunshine Company is going to take advantage of the information on Amazon reviews and ratings over a period of time regarding to the three products they intend to sell. In this paper, we discuss methods to conclude and predict merchants’ performance and characteristics of customers’ behavior of target products based on this dataset.

2. Text-Based Comments Overview with BERT Model
There are altogether 18939 reviews and ratings for baby pacifiers available at hand. It can be easily obtained that all the reviews are from the US. The rough analysis above indicates that the variables of marketplace and product category could be omitted in the following modeling process.

Deep learning is a class of machine learning algorithms mainly based on artificial neural networks (ANN). The structures of the networks are usually nonlinear, multiple-layered and fully connected, the refore have a better performance when dealing with large dataset if the network is trained well. We can think of a deep neural network N as a complex system, that is:

$$N: X \rightarrow Y , \text{ where } X \text{ is the input and } Y \text{ is the output}$$ (1)
There are many applications of deep learning algorithm such as computer vision (CV) and natural language processing (NLP). Text classification is one of the most common NLP tasks and can be used in many real-life applications such as label prediction and sentiment analysis. We can represent the text classification task in the following math representation,

\[
\text{find the Best } N: X_i \rightarrow \hat{Y}_i, \text{ such that } \hat{Y}_i \approx Y_i
\]  

(2)

\(X_i\): Input text, \(\hat{Y}_i\): predicted label, and \(Y_i\): real label for the i-th data.

We run a binary text classification task on Amazon Datasets using BERT model: input a comment, is the user’s attitude positive or negative, satisfied or unsatisfied? If the star rating is greater than three stars, we considered the user satisfied \(Y_i = 1\); and if the star rating is less than three stars, we considered the user unsatisfied \(Y_i = 0\); if the star rating is exactly three stars, we randomly picked a label for the output label.

For data preprocessing, we merge the ‘review headline’ and ‘review body’ together as the input text \(X_i\), and mapped the ‘star rating’ to a real label \(Y_i \in \{0, 1\}\). Then we run the pre-trained BERT model on the input text. Since the last layer of text classification task is a sigmoid function, the output of the network \(B\) is actually a possibility \(p \in (0, 1)\), which can represent the possibility of the input text being “positive”. Therefore, if \(p > 0.5\) we considered the predicted label satisfied \(\hat{Y}_i = 1\); if \(p < 0.5\) we considered the predicted label unsatisfied \(\hat{Y}_i = 0\); similarly if \(p = 0.5\) we randomly picked a label for the predicted label. After predicting, we can evaluate the model’s performance by its accuracy: the ratio of the predicted labels to the total labels such that \(\hat{Y}_i = Y_i\).

To improve accuracy, we do transfer learning: We add a linear neural network layer after the pre-trained BERT output layer, and do training using stochastic gradient descend to make it fit on our input dataset (we trained three different models to fit the three products). The new model is called the Amazon-BERT model, symbols \(B'\). The mathematical formula can be written into \(B' = L(B)\), where \(L\) is a linear layer. The accuracy of the two models is listed in the table below:

| Table 1. Accuracy of the two models on Amazon Datasets. |
|-------------------------------------------------------|
| Pacifier                                             |
| Pre-trained BERT \(B\) (without training)             | 58.42%   |
| Amazon-BERT \(B'\) (After training)                   | 85.04%   |

After training, our Model outperform the pre-trained BERT model and have an average accuracy of 75%. We considered our Amazon-BERT model reliable. Therefore, we add a new factor to every comment: the output probability \(p\) predicted from BERT model and consider it a vital factor.

3. Evidence from PCA

3.1. Obtaining Potential Variables from the Initial Datasets

Now there are six potential criteria for Sunshine Company to track. We would like to process the PCA measure to find out the most informative combinations among them. To begin with, however, we need to test the validity of PCA. Kaiser (Kaiser, Rice, 1974) pointed out that if the overall kmo value was below 0.5, PCA was not allowed to carry out. This case (kmo value < 0.5) illustrates the inadequate information the present variables provide, which means that other potential criteria need to be dug for. Unfortunately, the six initial variables (star rating, helpful votes, total votes, vine, verified purchase and \(p\)) report a 0.4942 kmo value based on the dataset of pacifiers, while the test on the other two products present a kmo value very close to 0.5.
Previous studies provide other potentially useful variables to be considered. The most important ones are the length and the helpfulness rate of the reviews. The former one indicates the writers’ attitude while the latter one shows the viewers’ attitude. So, three more variables are added into consideration: length of the review headline, length of the review body and helpful rate of the review (defined as follows):

\[
\text{helpful rate} = \begin{cases} 
\frac{\text{helpfulness votes}}{\text{total votes}} & \text{total votes} > 0 \\
0 & \text{total votes} = 0
\end{cases}
\]  

This time, the nine variables of the three products respectively survive the KMO test, reporting significant difference from 0.5. The notations of the standardized variables are shown in the notation part.

3.2. PCA of Pacifiers: Results and Explanation

![Scree plot of eigenvalues after pca](image.png)

**Figure 1.** Pacifiers: evidence for choosing the principle component.

**Table 2. Results of PCA of Pacifiers.**

| Component | Cumulative |
|-----------|------------|
| Com1      | 0.2704     |
| Com2      | 0.4366     |
| Com3      | 0.5669     |
| Com4      | 0.6683     |
| Com5      | 0.7646     |

It can be observed from the above figure that the first five component takes up more than 75% of the cumulative proportion of variance and they all report eigenvalues larger than or close to 1. So, it is reasonable to choose the first five components as the principle ones. They can be described using the nine initial variables as follows:

\[
(Com_1, Com_2, Com_3, Com_4, Com_5) = A_p \times I
\]  

It can be easily obtained that vh and vt contributes to Com1 most, which indicates the review popularity; rr contributes to Com2 most, along with rs, vt, vh, lh and lb, so Com2 is a good indicator of the total reputation of pacifiers; v contributes to Com3 most, which illustrates influencers’ power; lh contri
butes to Com4 most, so it mainly describes review quality; rh contributes to Com5 most, so we can just consider Com5 as the helpfulness rate.

4. Vector Autoregression Model

4.1. Model Establishment

During the process of modeling, we find that another model, vector autoregression (VAR), can be more informative when there exist strong interactions between variables. In our case, the past has a strong impact on the future and the causal relationship is difficult to determine. In a word, variables have dynamic interactions, so we apply VAR to the problem.

Generally, a VAR model is as follow (Michael, Abrigo & Love, 2015):

\[
Y_t = C_i + \Phi_1 Y_{i,t-1} + \Phi_2 Y_{i,t-2} + \ldots + \Phi_p Y_{i,t-p} + \varepsilon_t, \quad i = 1, 2, \ldots, N; t = 1, 2, \ldots, T;
\]

\[
Y_t = \begin{bmatrix}
Y_{1t} \\
Y_{2t} \\
\vdots \\
Y_{kt}
\end{bmatrix},
Y_{i,t-1} = \begin{bmatrix}
Y_{1,t-1} \\
Y_{2,t-1} \\
\vdots \\
Y_{k,t-1}
\end{bmatrix},
Y_{i,t-p} = \begin{bmatrix}
Y_{1,t-p} \\
Y_{2,t-p} \\
\vdots \\
Y_{k,t-p}
\end{bmatrix},
C_i = \begin{bmatrix}
c_{1i} \\
c_{2i} \\
\vdots \\
c_{ki}
\end{bmatrix},
\varepsilon_t = \begin{bmatrix}
\varepsilon_{1t} \\
\varepsilon_{2t} \\
\vdots \\
\varepsilon_{kt}
\end{bmatrix},
\Phi_1 = \begin{bmatrix}
\phi_{11}^{(1)} & \phi_{12}^{(1)} & \ldots & \phi_{1k}^{(1)} \\
\phi_{21}^{(1)} & \phi_{22}^{(1)} & \ldots & \phi_{2k}^{(1)} \\
\vdots & \vdots & \ddots & \vdots \\
\phi_{k1}^{(1)} & \phi_{k2}^{(1)} & \ldots & \phi_{kk}^{(1)}
\end{bmatrix},
\Phi_p = \begin{bmatrix}
\phi_{11}^{(p)} & \phi_{12}^{(p)} & \ldots & \phi_{1k}^{(p)} \\
\phi_{21}^{(p)} & \phi_{22}^{(p)} & \ldots & \phi_{2k}^{(p)} \\
\vdots & \vdots & \ddots & \vdots \\
\phi_{k1}^{(p)} & \phi_{k2}^{(p)} & \ldots & \phi_{kk}^{(p)}
\end{bmatrix}.
\]

(5)

We select the optimal lag order first, see Table 3:

| Lag | AIC   | BIC   | HQIC  |
|-----|-------|-------|-------|
| 1   | 2.559 | 3.951*| 3.112 |
| 2   | 2.381 | 4.005 | 3.030 |
| 3   | 2.229 | 4.137 | 2.995 |
| 4   | 2.208*| 4.470 | 3.120*|

* means the lowest which indicates the best choice in terms of this criterion.

In order to avoid high precision and dimensional disasters, we choose the first order lag. The result is as followed Table 4:

|       | com1          | com2          | com5          |
|-------|---------------|---------------|---------------|
| L.com1| -0.402*** (0.156) | 0.378*** (0.146) | 0.378*** (0.155) |
| L.com2| -0.379*** (0.126) | 0.299*** (0.121) | 0.322*** (0.132) |
| L.com5| -0.489*** (0.115) | 0.254* (0.110)  | 0.558*** (0.110) |

*, ** and *** means significant at 10%, 5% and 1% level. In parentheses is the standard error.
4.2. *Granger Causality Test*

For the same reason, we do a Granger causality test for the model.

**Table 5.** The result of Granger causality test.

| Variable | Excluded | Chi-square |
|----------|----------|------------|
| com1     | com2     | 9.002(1) ※ ※ ※ |
| com1     | com5     | 18.204(1) ※ ※ ※ |
| com1     | All      | 18.262(2) ※ ※ ※ |
| com2     | com1     | 6.720(1) ※ ※ ※ |
| com2     | com5     | 5.429(1) ※ ※ |
| com2     | All      | 7.124(2) ※ ※ |
| com3     | com2     | 6.668(2) ※ ※ |
| com3     | com5     | 5.992(1) ※ ※ |

※, ※※ and ※※※ means significant at 10%, 5% and 1% level. In parentheses is the degree of freedom.

The results of Table 5 tell us that all 3 variables are each other’s Granger cause, which indicates that their causality cannot be excluded.

4.3. *Sensitivity and Practical Analysis*

Since the models are of the linear forms, the sensitivity of each variable can be easily obtained by simply viewing the coefficient. The review popularity contributes to the reputation of the pacifiers most. This time, it influences the reputation in the same direction. Again, the results suggest that tracking the reviews popularity is important for the pacifiers.

Success, review popularity and helpfulness rate, all these three in the past have a positive impact on the future success. And what deserves attention is that even though the past is influential, the magnitude is not large and it diminishes quickly. This is because as a baby-used product, their customers, parents, cannot feel and try it directly, and product updates relatively slow. Thus, it is a weakly mouth-word-based product. Once used, it keeps its position for a long time. This can also be seen from its high average rating. Therefore, keeping the quality and making sure there aren’t big mistakes are enough.

5. **Impact of Star Rating on Reviews**

In this section, we’re going to discuss the impact of star rating on reviews, namely, whether one’s opinion on something will be influenced by others.

For the same reason, we still use dynamic panel model to regress. For simplification, we will give the Arellano-Bond adjusted form and its Granger causality test directly. Not only the star rating but also the rating and number of reviews will have impacts, so we use the latter two as independent variables. The unit root test shows that all are stationary series. The results for the three products are presented respectively as follow Table 6.
Table 6. Hair dryers: Result and Z-par value.

| Dependent Independent Variable’s Coefficient | r_r | review_num | L. r_r | L. r_s | L. review_num |
|---------------------------------------------|-----|------------|--------|--------|--------------|
| r_r                                         | 0.068*** | 0.001 | 0.124* | -0.013 | 0.001* |
|                                             | (0.009)  | (0.001) | (0.067) | (0.010) | (0.001) |

※, ※※ and ※※※ means significant at 10%, 5% and 1% level. In parentheses is the standard error

| Independent Variable’s Z-par Value (1st order lag) |
|--------------------------------------------------|
| r_r review_num                                  |
| 0.120 | 1.535 |

※, ※※ and ※※※ means significant at 10%, 5% and 1% level.

The result shows that for hair dryers, review number and rating are not a Granger cause of review rating, thus not the real cause of the changes in review rating though they have linear correlation.

Table 7. Microwave ovens: Result and Z-par value.

| Dependent Independent Variable’s Coefficient | r_s | review_num | L. r_r | L. r_s | L. review_num |
|---------------------------------------------|-----|------------|--------|--------|--------------|
| r_r                                         | 0.005 | 0.003※ | -0.493※※※ | 0.002 | -0.003※ |
|                                             | (0.005) | (0.001) | (0.124) | (0.006) | (0.001) |

※, ※※ and ※※※ means significant at 10%, 5% and 1% level. In parentheses is the standard error

| Independent Variable’s Z-par Value (1st order lag) |
|--------------------------------------------------|
| r_r review_num                                  |
| 1.447 | 3.618※※※ |

※, ※※ and ※※※ means significant at 10%, 5% and 1% level.

The result of Table 7 shows that for microwave ovens, review rating is not but review number is a Granger cause of review rating, thus review number may be the real cause of review rating.

Table 8. Pacifiers: Result and Z-par value.

| Dependent Independent Variable’s Coefficient | r_s | review_num | L. r_r | L. r_s | L. review_num |
|---------------------------------------------|-----|------------|--------|--------|--------------|
| r_r                                         | 0.037※※※ | 0.001※※ | 0.308※※※ | 0.019※※※ | 0.001※※※ |
|                                             | (0.004) | (0.000) | (0.048) | (0.004) | (0.000) |

※, ※※ and ※※※ means significant at 10%, 5% and 1% level. In parentheses is the standard error

| Independent Variable’s Z-par Value (1st order lag) |
|--------------------------------------------------|
| r_r review_num                                  |
| 2.978※※※ | 8.243※※※ | 2.404※※ |

※, ※※ and ※※※ means significant at 10%, 5% and 1% level.

The result of Table 8 shows that all three above are granger cause of review rating.
6. Conclusion
In this paper, we use Amazon data to build a product performance prediction model based on NLP and Vector Autoregression. The NLP helps mine the information in the text while having some drawbacks to describe the change and relationship on time series statistically. The data at first was not so informative with some unconcentrated information like some products appeared only once. And the information of the original variables are inadequate. We use PCA and rearrange it into panel data to make it informative. After analysis, we come to the conclusions that the pacifiers have a relatively stable, persistent and weak word-of-mouth. The past reviews and performances have positive impact on present reviews and performances.

References
[1] Chen, Qiang. Advanced Econometrics with Stata Applications (2nd edition). Beijing. 2014.
[2] Devlin, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805. 2018.
[3] Ding, Mingzhou, Chen, Yonghong, Bressler, Steven L. Granger Causality: Basic Theory and Application to Neuroscience. Quantitative Biology: 826-831. 2016.
[4] Jörg Breitung, Samarjit Das. Panel Unit Root Tests Under Cross-sectional Dependence. Statistica Neerlandica, 59(4): 414-433. 2005.
[5] Kaiser, H. F., & Rice, J. Little Jiffy, Mark IV. Educational and Psychological Measurement, 34 (1), 111-117.
[6] Liu, Xuetian. Feature Analysis and Detection of Review Spam Based on Web Quality Feature Model. Southwest Jiaotong University of computer science and technology. 2017.