Sentiment analysis measure of topic mapping in LDA-LSTM combination model

Yi-hang Lv¹, Qin Liu²*, Chang-bin Jiang³

¹School of computer, Wuhan University of Technology, China
²School of Entrepreneurship, Wuhan University of Technology, China
³ School of Management, Wuhan University of Technology, China

¹2830019035@qq.com
²* Corresponding author: Qin Liu, email: isunnygirl@qq.com
³175521@qq.com

Abstract. New energy market competition incentive, how to improve the product accurately, so as to attract consumers, will become the key point for enterprises to occupy the market. From the perspective of improving products, this paper constructs a review information dimension mapping theme emotion analysis model. Firstly, LDA model is used to extract the topic of comment information. Secondly, topic mapping is carried out on the comment information to get the correlation degree between each comment information and each dimension topic. Finally, the LSTM model is constructed for sentiment analysis, and the emotional intensity of consumers in each dimension theme is measured.

1. Introduction
In recent years, with the popularity of the Internet and the rapid development of online shopping, more and more consumers choose to share their shopping experience on social networks. As an open and accessible information resource, a large number of online reviews contain value information about the various dimensions of product quality. Classifying and mining these information helps enterprises to understand the needs and preferences of consumers.

At present, the processing of comment information is a main direction of consumer research. Liu et al. Use a large number of user behavior data such as browsing, clicking and purchasing to predict consumers' purchasing situation through support vector machine, and get more correct results. Yang SA ran used xgboost and lightgbm algorithm to model and predict customers' purchase behavior. Chen Junhua constructs a customer preference recommendation algorithm based on hidden multi-dimensional using customer purchase behavior data. Silahtaroglu et al. use decision tree and neural network to predict whether customers will buy the goods in their shopping cart based on the historical shopping behavior data and demographic data. The existing analysis methods of comment information mainly focus on whether to buy or not, but ignore how to improve products to guide and promote consumers to buy, and the existing research basis is mainly based on rational analysis and lack of emotional measurement. Therefore, from the perspective of product improvement, this paper constructs a comment information dimension mapping theme emotion analysis model.
Emotion is the internal subjective experience and special reflection form of human beings to the real world. When consumers comment on products, the measure of satisfaction is mostly reflected in the intensity of emotion. Therefore, it is necessary to analyze the sentiment of comment information to obtain the demand orientation of consumers. Enterprises need to make use of consumer demand as a guide to adjust their products, so as to improve consumer satisfaction and build product awareness.

At present, the new energy vehicle market is in a state to be developed. At present, the emerging new energy vehicles have not been widely recognized by the public, and have broad market prospects. The user's hobby is ever-changing, how to stand out in many new energy vehicle products is very important.

Users' liking degree and preference dimension of products are often expressed by the emotion in the comment information. Therefore, this paper proposes LSTM emotion analysis model based on topic classification algorithm. Firstly, the car review data is classified into various dimension topics to improve the accuracy of the prediction module, and then the emotional analysis of the content of the dimension is carried out to realize the measurement of consumer emotion of specific topics in a small range of car review information.

2. Topic classification of comment information

The first paragraph after a heading is not indented (Bodytext style).

2.1. Data preprocessing

Text preprocessing is necessary for text data. Because most of the text data obtained is too complicated and has no practical value, it will affect the effectiveness of data output. If these comments data are introduced into the subsequent word segmentation and emotion classification model, it will have a great impact on the analysis results, and the quality of the analysis results will be inevitably damaged.

The main purpose of data preprocessing is to remove the useless information in the text, make the data clearer and improve the usefulness of comments, which is easy to analyze.

The data preprocessing in this paper is mainly divided into three steps, which are text de-duplication, text de-noising and text short sentence deletion.

Text de-duplication is to delete the repeated comments in the comment information and reduce the interference of redundant information. In most comment sites, there is an automatic comment system, which lacks practical value and research value. It needs to be screened and deleted to prevent interference to the comment processing.

Text denoising is to delete some special characters such as expression and web address in the comment information, which is lack of usability for text research and will affect the following word segmentation.

The deletion of short text sentences refers to the deletion of data with too few words in the comment information. The number of words in online reviews is an important factor affecting the degree of usefulness. The more word data, the more product information it conveys. From a psychological point of view, comments with a large number of words have a profound understanding of the product. From an analytical point of view, the less the number of words, the less the value of the information contained in the review information, and it is likely to be randomly generated by consumers, which has no feedback significance and research value, such as "very satisfied".

2.2. Text segmentation

Word segmentation is to divide a complete sentence into several words. In English, the space between words is a natural separator. For Chinese, there are three commonly used classification algorithms, which can be based on understanding, statistics or string matching, as well as three or a mixture of two algorithms. This paper uses Python's open source Jieba library for accurate word segmentation of the comment set. The algorithm is simple to use, high efficiency, and the basic vocabulary is up to 350000.
It can also customize the dictionary, add a user-defined stop word list, delete some useless modal particles and determiners, and increase the accuracy of word segmentation.

2.3. Topic classification based on LDA Algorithm

LDA is an unsupervised machine learning method, which has a certain effectiveness in identifying the topic information hidden in large-scale document sets and corpora. It adopts a bag of words model, which regards every document as a word frequency vector, so as to transform text information into digital information which is easy to analyze. The core idea is a probability model, each document represents a probability distribution of some topics, and each topic represents a probability distribution of many words. Therefore, the purpose of LDA is to identify topics, that is, to change the document vocabulary matrix into document topic matrix distribution and topic vocabulary matrix distribution. The core formula is as follows;

$$P(w | d) = P(w | t) * P(t | d)$$

Based on the above data, LDA model is used for topic classification. The whole data is divided into 6 topics, and the top 10 keywords are extracted from each topic.

| Theme 1 | Theme 2 | Theme 3 | Theme 4 | Theme 5 | Theme 6 |
|---------|---------|---------|---------|---------|---------|
| cost    | appearance | power | space | steering | interior |
| performance | Design | acceleration | seat | handling | decoration |
| Price | valence | Design | acceleration | seat | handling | plastic |
| cheap | preferential | good-looking | speed up | comfortable | precision | style |
| oil | consumption | modelling | modeling | trunk | flexibility | work |
| maintain | atmosphere | high speed | ride | chassis | Design |
| choice | body | Electric vehicle | air | sensitive | Materials |
| configuration | The | overtake | cosy | feel | used |
| Endurance | line | beautiful | Enough | front row | lithe | material |
| | | | fast | spacious | direction | feel |

In theme 1, the keywords such as "cost performance", "price", "cheap", "preferential" and "configuration" reflect the price advantage of new energy vehicles. New energy vehicles have the advantages of low cost, including government subsidies, environmental support and low fuel consumption. At present, due to the rise in oil prices, the factors that lead to the inability to buy a car gradually turn into the current situation that the car can not afford. Therefore, in the long run, new energy vehicles have greater cost performance.

In Theme 2, the keywords such as "appearance", "design", "fashion", "good-looking" and "modeling" reflect the appearance characteristics of new energy vehicles. From the keyword analysis, it can be seen that users generally have a high evaluation on the appearance of new energy vehicles. Good visual effects will always bring users enjoyment in the visual dimension. Therefore, new energy vehicles have achieved good results in appearance design, which has been praised by most users.

In theme 3, the keywords such as "power", "acceleration", "start", "speed up" and "mode" reflect the power situation of new energy vehicles. It can be seen that users also have higher requirements on performance when they purchase cars. Simple good visual effect is unable to persuade users to make purchase decisions, but also to meet certain requirements in terms of performance.
In theme 4, the keywords such as "space", "seat", "back row", "comfortable" and "trunk" reflect the user experience of new energy vehicles in space dimension. Now in this era, the car seems to have become an essential part of the travel process in life. Not only on the way to work, but also on the way to school or self driving. Users stay in the car longer and longer, so the design requirements of car space are higher and higher. If there is no complete planning for the car space, it is difficult to meet the living needs of users.

In theme 5, keywords such as "steering wheel", "handling", "braking", "precision" and "flexibility" reflect the feeling of handling and user feedback of new energy vehicles. Vehicle handling refers to the difference between the results achieved by the operation of the direction, brake, throttle and various technical and condition configurations of the vehicle and the expected purpose and requirements of the driver when the vehicle's own problems or external factors occur in driving. This is the handling performance. In the process of using the car, good handling performance can bring users a good experience. Therefore, to improve the controllability of new energy vehicles can persuade users from the technical level and urge users to make purchase decisions.

In theme 6, keywords such as "interior decoration", "plastic", "peculiar smell", "style" and "material" reflect the interior decoration and design style of new energy vehicles. Automotive interior system is an important part of the car body, and the design workload of interior system accounts for more than 60% of the car styling design workload, far more than the car shape, is one of the most important parts of the car body.

Excellent car interior can bring users more advanced and convenient in car experience. Therefore, it is necessary to improve the rationality of the car interior and build a life enjoyment close to the user at the design level. So as to make new energy vehicles unique in similar products.

To sum up, the LDA model is used to summarize the text into six themes: price advantage, appearance characteristics, power, space dimension, handling feeling and interior design. In the third part, we will build an emotion analysis model to measure the emotion of each topic.

3. Construction of emotion analysis model

Consumers express their feelings for products and services through comments, so we need to build an appropriate model to reflect consumers' emotional feedback.

3.1. Comment topic mapping

If the emotional analysis of a comment directly has certain blindness, it may ignore the important information. Therefore, this paper first maps the comment information to six topics, and then carries on the emotion analysis, which improves the guidance of the comment information and the accuracy of the comment information. First, we get the vector vector of each sentence, and then calculate the vector of each comment word. Cosine similarity: the cosine value of the angle between two vectors in vector space is used to measure the difference between two individuals. The closer the value is to 1, the closer the angle is to 0°, that is, the more similar the two vectors are.

\[ \text{similarity} = \cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} \]  

Finally, Levin steinby is used to map each comment to each topic.

3.2. Sentiment analysis

With the development of machine learning, the efficiency and accuracy of neural network are more and more recognized and applied to all areas of life. RNN neural network is widely used in natural language processing, but the original RNN will encounter a big problem, which is called the vanishing gradient problem for RNNs. With the deepening of the network level, the network can not carry out the subsequent training, resulting in gradient elimination and gradient explosion problems. In order to solve this problem, LSTM network, a variant of RNN network, is produced. LSTM network has "memory", because there are connections between different "time points", rather than feedforward or
feedback at a single time point. LSTM is good at dealing with multiple variables, which makes it helpful to solve the problem of time series prediction.

There are three stages in LSTM

- Forgetting stage. This stage is mainly to selectively forget the input from the previous node. Simply remember that "it's important not to forget."
- Choose the memory stages. This stage selectively "remembers" the input of this stage. The main purpose is to select and memorize the input data. Write down what is important and what is not.
- Output phase. This phase will determine which outputs will be taken as the current state.

In this paper, LSTM network is used to construct emotion analysis model. There are 8673 comments in the data set of this training. The emotion of the text is divided into two categories: "positive" and "negative".

Then we need to do a simple analysis of the corpus, the statistical chart of sentence length and frequency in the data set (Figure 1), and the cumulative distribution function of comment sentence length in the data set (Figure 2). It can be seen that the sentence length of most samples is between 1-200, and the cumulative frequency of sentence length is 0.90, so the length is about 167.

![Figure 1. Statistical chart of sentence length and frequency](image1.png)

![Figure 2. Cumulative distribution function of sentence length](image2.png)

The parameters of LSTM model are constructed, and the feedforward network layer with 100 neurons is defined. The dropout layer is also introduced in the network, and the parameters are defined as 0.2. Some neurons are ignored randomly to prevent over fitting. The activation function is defined as softmax function:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{i=1}^{K} e^{z_i}}$$  \hspace{1cm} (3)
Where $\sigma(z)_j$ is the $j$ active output and $e^j_j$ is the $j$ element. The loss function is set as the cross entropy loss function:

$$
L = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]
$$

Where $y$ is the real sample label and $\hat{y}$ is the prediction sample label.

LSTM network structure is as follows:

The data set is divided into training set and test set according to 9:1, and trained for 5 times. The accuracy of the model in the training set is more than 95%, and the accuracy in the test set is more than 90%. The effect is quite good.

4. Experimental
Crawling the comment data of aika auto.com in 2020, the distribution is brought into LSTM sentiment analysis model for sentiment division, topic mapping and relevance analysis.

Draw the comment information, and take the average correlation degree of each dimension of each month as the dimension feedback of that month, as shown in the figure:

![Figure 4. Positive comment mapping](image-url)
Figure 5. Negative comment mapping

It can be seen that in the positive comments, there is a high degree of correlation between appearance characteristics and power, indicating that consumers have a strong sense of identity in these two aspects, but there is a low degree of correlation between space dimension and interior design, indicating that new energy vehicles still need to make improvements in this aspect, and there is still a large space for development.

For consumers, there are two dimensions that are less related to interior design, which means that the two dimensions of interior design information and operation information are less related.

To sum up, the appearance characteristics and power of these two dimensions have been more mature development, with the difficulty of further development and creative development. There is still a lot of room for improvement in the two dimensions of space dimension and interior design, which can be improved to win more consumers’ favor.

5. Summary
Under the condition of big data, the emotional dimension mapping analysis of comment information can effectively grasp the pain points of consumers and promote enterprises to make industrial decisions. In this paper, LDA model is used to extract the dimensions of the comment data, and then LSTM sentiment analysis model is constructed. After topic mapping, sentiment analysis is carried out to measure the consumer sentiment of a specific topic in a small range of comment information.

Acknowledgment
This work is supported by the National Social Science Foundation of China(No. 19BSH105), the Fundamental Research Funds for the Central Universities(WUT: 2020VI028), and National innovation and entrepreneurship training program for college students S202010497038 and 202010497018.

References
[1] Liu X. Li J. Using support vector machine for online purchase predication [C] / 2016 International Conference on Logistics, Informatics and Service Sciences( LISS) IEEE, 2016: 1－6
[2] Yang SA ran. Prediction model of online customers' repurchase behavior based on machine learning theory [D]. Tianjin: Tianjin University, 2018
[3] Chen Junhua. *Modeling and prediction of customer purchase behavior* [D]. Chengdu: University of Electronic Science and technology, 2018

[4] Silahtarogu G., Donertasli H. *Analysis and Prediction of E- Customers' Behavior by Mining Clickstream Data* [C]. IEEE International Conference on Big Data. Santa Clara, CA, USA:IEEE, 2015.