

# Sentiment Analysis on Power Rationing Micro Blog Comments Based on SnowNLP-SVM-LDA Model

Bingqian Zhou, Yuhan Zhu\*, Xiabing Mao

College of Maritime Economics and Management, Dalian Maritime University, Dalian, Liaoning, 116000, China

\*Corresponding author. Email: zhuyuhan0418@163.com

**Abstract.** In the context of energy conservation and emission reduction, many places across the country have introduced "power rationing" policies. In order to understand the public's emotional attitude and emotional characteristics of the policy, help government departments to grasp the trend of public opinion and optimize strategic decision-making, this paper proposes a public opinion analysis method based on the SnowNLP-SVM-LDA model. This paper takes micro blog comments on power rationing as the research object, and obtains a total of 5242 micro blog comments through Python collection and data cleaning. Firstly, the word cloud map is drawn for the comment text after Jieba word segmentation, and then the SnowNLP-SVM model is used to classify the sentiment of micro blog comments, and then LDA topic mining is carried out for positive and negative comments respectively. Research shows that most micro blog netizens (69.34% of the total sample) hold a negative attitude towards power cuts. This paper identifies three major topics of national policy, support for the government, and power supply security in positive emotions, and three major topics of people's livelihood issues, rising coal prices, and government work in negative emotions. Through the analysis of this paper, it provides data support and theoretical support for government departments to guide the trend of public opinion and optimize work decisions.

**Keywords:** power rationing; sentiment analysis; SnowNLP; SVM; LDA.

## 1. Introduction

In the strategic context of "2030 to achieve carbon peaks, 2060 to realize carbon neutrality", The state strengthens the control of the energy consumption dual control policy, and strengthens the control of high energy consumption and high emission projects. Therefore, various provinces have introduced various power rationing measures to respond to the national call. "Power Rationing" has become one of the most popular topics at the end of 2021. The reason is that the issue of "Power Rationing" has brought certain effects to the industrial sector and people's lives. With the arrival of the big data era, people are free to express their views and remarks in various network social platforms. Therefore, the Internet has become the main channel for people to obtain information. Due to the characteristics of real-time interaction and diversity, micro blog has become an important channel for netizens to communicate and express their emotions. Since these comments contain a lot of emotional information, it has become a research focus of many scholars. The use of various text excavation and emotional analysis techniques for extraction, treatment and analysis of the subjective online comments with emotional color, can better understand the emotional tendency of the whole social group, and analyze key factors affecting the emotional attitudes. Therefore, this paper takes the comments related to power rationing in micro blog as the research object to conduct sentiment analysis research, in order to help government departments control public opinion and optimize decision-making.

## 2. Research methods

At present, there are two main methods for sentiment analysis. The first is sentiment analysis based on sentiment dictionary, which records sentiment words and phrases, as well as their related sentiment tendencies and other information, which are used to calculate the sentiment of each document. The traditional emotional dictionary is: Harvard University's GI (General Inquirer) English emotional

dictionary, Chinese emotional words of Dalian University of Technology and so on. The advantage of this method is that the accuracy is high, but it is very dependent on the reasonable construction of the emotional dictionary.

The second type of method is based on machine learning emotional analysis, which is to identify the test data using training sets labeling emotional category. First extract the feature vector in the text, then the support vector machine, logistic regression, Naive Bayes and other algorithms are used to train the classification model, specifying data samples of unknown categories by classification model. Liu L[1] et al. performed sentiment analysis on micro blog user comments through SVM, Naive Bayes and Random Forest. Xue [2] et al. used the LDA topic model to better realize the identification of topics related to Novel coronavirus pneumonia in 22 million Twitter messages.

Therefore, this paper will comprehensively use two sentiment analysis methods to perform sentiment classification and topic identification on micro blog comments about power rationing. Firstly, SnowNLP based on sentiment dictionary is used to calculate the sentiment score of the micro blog corpus, and the comments with high positive and negative polarity are extracted to train the SVM model, and then use the trained SVM model to perform sentiment classification on micro blog comments. Finally, LDA topic mining is carried out on the classified comments to explore the focus of netizens under different emotional attitudes.

### 3. Theoretical basis

#### 3.1. SnowNLP

The SnowNLP library is a class library written in Python that specializes in processing Chinese text, with its own corpus and sentiment dictionary. SnowNLP supports a variety of Chinese text processing operations including: Chinese word segmentation, part-of-speech tagging, sentiment analysis, text classification, converting to pinyin, converting traditional to simplified, extracting text keywords, extracting text summaries and calculating text similarity, etc. The function is very comprehensive.

#### 3.2. SVM(Support Vector Machine)

The method adopts a supervised learning approach to model the binary classification problem. Usually used to solve two types of problems: One is linearly separable, and the samples are separated by a hyperplane, and an optimal hyperplane is found to classify them; the other category is linearly inseparable problems, which are usually transformed into linearly separable using kernel function (Polynomial kernel, Radial basis function kernel, Laplacian kernel, and *Sigmoid* kernel). In the case of linear separability, if the given sample set has  $m$  samples, that is:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}, y_i \in \{-1, +1\}$$

Where  $x_i$  represents the  $i$ th sample in the feature space, and  $y_i$  represents its classification result. The core of the support vector machine is to find an optimal classification surface  $w * x + b = 0$  in the feature space of the sample set  $T$ . The classification surface should make the classification spacing maximum, that is, the number of samples ( support vector) to the plane is the highest. The process of solving the optimal classification is essentially solving a Quadratic convex programming problem, that is,  $w$  and  $b$  that need to be found, but also need to maximize the classification interval.

#### 3.3. LDA(Latent Dirichlet Allocation) model

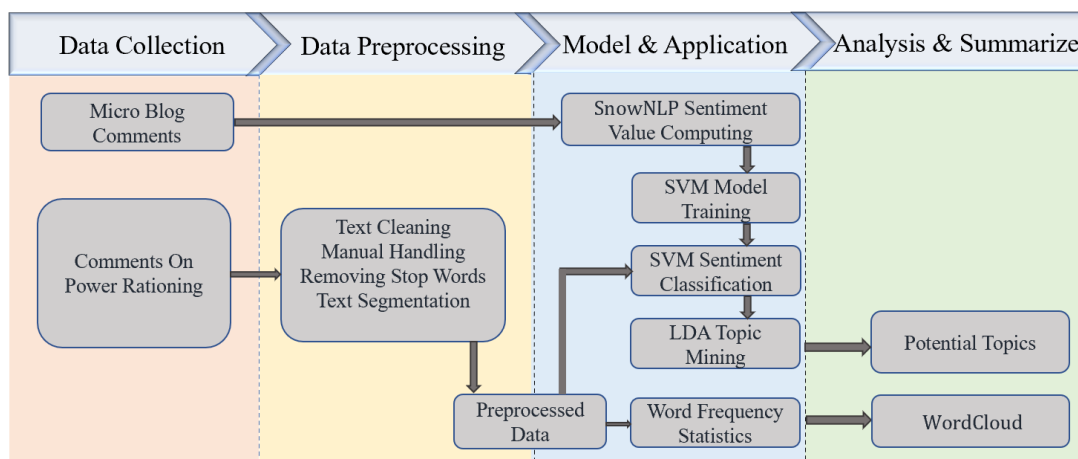
LDA model (Latent Dirichlet Allocation, LDA) is a topic model algorithm based on probability model proposed by Blei et al. LDA is an unsupervised machine learning text mining technique that can be used to identify potentially hidden topic information in large-scale document sets or corpora [5]. The model believes that each word of each document is "selected with a certain probability of a certain topic, and selected a certain word from this topic with a certain probability". The LDA model,

also known as the Three-layer Bayesian probability model, contains a three-layer structure of document, topic, and word, which can effectively model text. Through the LDA topic model, the potential topics in the dataset can be mined, and then the focus of the dataset and its related feature words can be analyzed.

## 4. Model and Data

### 4.1. Overall Framework

This paper first uses python to crawl the relevant comment data in micro blog, and then preprocesses the comment data, including text cleaning, manual processing, removing stop words, text segmentation, etc. Then draw a word cloud map for the comments after text segmentation processing, and get a preliminary understanding of the public's focus. At the same time, in order to make the SVM model training more accurate and improve the classification accuracy of the model, this paper uses SnowNLP to calculate the sentiment value of the collected more than 50,000 micro blog comment corpus, and selects the micro blog comment data with high positive and negative polarity as the SVM model training dataset. Then, use the trained SVM model to perform sentiment classification on the crawled micro blog comments to obtain sentiment classification results with higher accuracy. Finally, LDA topic mining was performed on positive and negative comments to obtain the emotional characteristics of netizens with different attitudes.



**Figure 1** Sentiment Analysis Flow Chart of "Power Rationing" Comments

### 4.2. Data Collection and Processing

#### 4.2.1. Data Sources

This paper uses the "micro blog" community to obtain the relevant opinions and attitudes of the public on the topic of "power curtailment", and uses the python crawler tool to crawl the comment data under the relevant blog posts published by news media such as "People's Daily" and "CCTV". A total of 5242 micro blog comments were collected.

#### 4.2.2. Data Preprocessing

In order to facilitate subsequent sentiment classification and topic mining, the target text is preprocessed. The first is to manually remove meaningless comments such as text, pictures and links. Then, use regular expressions to remove meaningless symbols, spaces and blank lines in the comments, and convert them into text containing only text. There are a total of 4951 comments. Then, combine the stop words list of Harbin Institute of Technology and Baidu stop word list to build the final stop word list, use a more comprehensive stop word list to remove stop words, and use the jieba component in python to segment the target text.



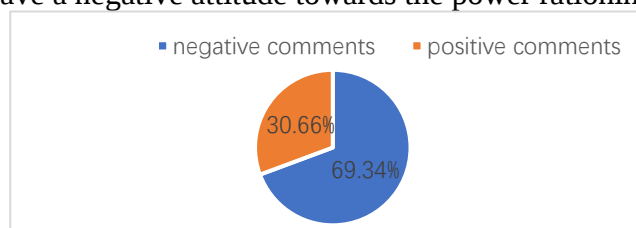
gamma=0.1) is used for training on reduced dimensionality data. The final accuracy of the model is 0.87. Therefore, the model has high accuracy and can be used for sentiment classification of micro blog comment text.



**Figure 3** SVM Model Construction Flow Chart

#### (5)Sentiment Classification

Using the trained SVM model to perform sentiment classification on the preprocessed 4951 micro blog comment texts, 1518 positive comments and 3433 negative comments were obtained. Since the negative comments accounted for 69.34% of the total comments, it can be preliminarily understood that most of the public have a negative attitude towards the power rationing.



**Figure 4** Sentiment Classification Results

### 5.3. Topic Mining Based on the LDA Model

Although the word cloud map can be used to preliminarily judge the public's concerns about power rationing, it cannot identify emotional tendencies. The sentiment classification of the comment text through the SVM model can intuitively understand the public's emotional tendencies, but cannot obtain the reasons why users hold positive or negative attitudes. Therefore, in order to help relevant departments optimize decision-making in a targeted manner, this paper uses the LDA topic model to further mine potential topics in positive and negative reviews, and identify key factors that affect public sentiment and attitudes.

#### 5.3.1. LDA Topic Mining

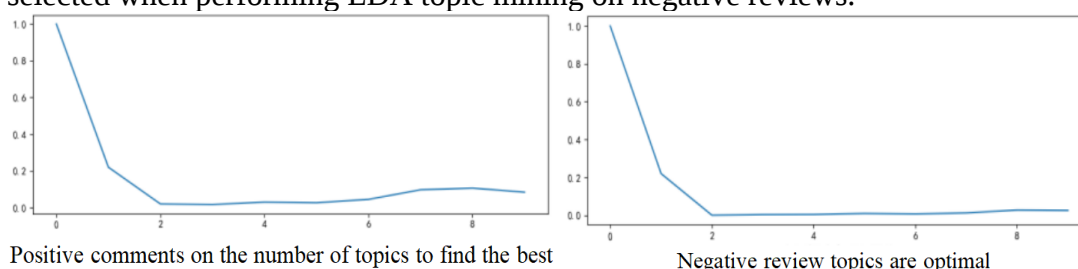
##### (1)Build Dictionaries and Corpora

Perform word segmentation processing on the classified positive and negative comments, and build a dictionary and corpus.

##### (2)Find the Optimal Number of Topics

A similarity-based adaptive optimal LDA model is selected to determine the optimal number of topics. This method can find the optimal topic structure with relatively few iterations without the need to manually debug the number of topics. The final average cosine similarity between topics is shown in Figure 5.

It can be seen from Figure 5 that when the number of topics is 3, the average cosine similarity between the positive comment topics is optimal. Therefore, the number of topics can be selected as 3 when performing LDA topic mining on positive reviews. When the number of topics of negative comments is 2 or 3, the average cosine similarity between topics is optimal. Therefore, 3 topics can also be selected when performing LDA topic mining on negative reviews.



**Figure 5** Average Cosine Similarity Analysis Between Topics

### (3)Build the LDA topic model

According to the optimal number of topics obtained in the previous section, use the gensim module in python to construct LDA topic models for positive and negative reviews respectively. Then, set the number of topics to 3 to conduct LDA topic analysis on positive and negative comments respectively, and obtain the 10 most likely words and their corresponding probabilities under each topic. The final LDA topic analysis results are shown in Table 1 and Table 2.

**Table 1** Active Text Topic Mining Results Statistics Table

| Topic   | Key Words                                                                                                             |
|---------|-----------------------------------------------------------------------------------------------------------------------|
| Topic 0 | develop,low carbon,China, policy,power consumption, local, electricity,switch off, resident, nationwide               |
| Topic 1 | cooperate, policy, gratitude, influence, government, civil, power rationing, support, people, power grid              |
| Topic 2 | country, electricity, people's livelihood, electricity price, true, guarantee, industrial, limited, south, switch off |

**Table 2** Negative Text Topic Mining Results Statistics

| Topic   | Key Words                                                                                                                        |
|---------|----------------------------------------------------------------------------------------------------------------------------------|
| Topic 0 | northeast, electricity, elevator, water cut off, heating, notice, coal, air conditioner,comments, location                       |
| Topic 1 | coal price, power outage, power rationing, price increase, notice, reason, resident, stop, advance, Three northeastern provinces |
| Topic 2 | notice, switch off, advance, power off, engage, home, people's livelihood, night, comment, nationwide,                           |

### 5.3.2. Potential topic analysis

Table 1 shows the mining results of active text topics mined by the LDA topic model. Combined with the statistical text, it can be found that: Topic 0 is mainly about the discussion of the national power curtailment policy. Some netizens discussed the reasons for the country's introduction of the power curtailment policy and the transformation of national industrial structure. Topic 1 is mainly about the support and cooperation of the people for the government. Some people think that power cuts may have a certain impact on production and life, but considering the shortage of resources and sustainable development, the people's attitude towards the government is support, cooperation and gratitude. Topic 2 mainly discusses the national power supply guarantee. Some netizens are more concerned about the issue of basic living security after power cuts, and hope that the state will guarantee the basic electricity demand for people's livelihood, hold the safety bottom line of people's livelihood development, and avoid power cuts as much as possible.

Table 2 shows the mining results of negative text topics mined by the LDA topic model. Combined with the statistical text, it can be found that: Topic 0 is mainly about the discussion about the inconvenience brought by power cuts to people's lives. Some people worry that power outages will bring many troubles to life, including elevators, water outages, heating and air conditioning and many other impacts. Topic 1 is mainly about the discussion of rising coal prices. Some netizens believe that the main reason for this round of power outages and blackouts is the imbalance between supply and demand. On the one hand, due to the shortage of coal in the country and the inversion of coal-electricity prices, many provinces have experienced a shortage of power supply; on the other hand, the demand for electricity has soared. Topic 2 is mainly about the discussion on the adjustment of the government's work plan. Some residents believe that the government did not give a comprehensive notice before the power cut, and the residents did not make adequate preparations before the power cut, which brought great inconvenience to their lives.

## 6. Summary and Outlook

With the advent of the era of big data, people freely express their opinions and remarks on various online social platforms, so the Internet has also become the main channel for people to obtain information. However, some netizens spread false information online, causing adverse effects to the society, so public opinion governance is crucial. Not only that, the remarks published by netizens on the Internet usually contain a lot of information, and the emotional tendencies contained in these information have important analytical value. Therefore, understanding and grasping the emotional attitudes of netizens will help managers guide the development of network public opinion and adopt targeted policies in a timely manner.

Based on the method of sentiment analysis, this paper takes the power rationing policy under the low-carbon background as an example to analyze the emotional tendencies and potential topics in the network public opinion. This paper firstly draws a word cloud map on the preprocessed comment text to gain a preliminary understanding of the focus of public attention, and then uses the SnowNLP-SVM model to classify the sentiment of the micro blog comment text. The results show that most netizens have a negative attitude towards the power rationing policy. Then, LDA topic mining is carried out on the comment texts with different emotional attitudes. Three major themes of national policy, support for the government, and power supply security are identified in positive comments, and three major themes of people's livelihood, rising coal prices, and government work are identified in negative comments. The following policy recommendations are made on the negative themes:

(1) Guide local governments to scientifically and rationally formulate an orderly energy consumption plan, improve the dynamic adjustment mechanism, and ensure that the public is informed, the contract is agreed, and the plan is executable and operable. Adhere to the bottom line of ensuring energy consumption for people's livelihood, and resolutely avoid situations that restrict energy consumption by residents.

(2) The rise in coal prices under the power rationing policy is mainly caused by the excessive reduction of production capacity in the last round of supply-side reforms and the restriction of imported coal. Therefore, the reform of the market power field can be gradually carried out from the perspective of separation and independent operation of power supply, power grid and distribution plants.

(3) Establish a complete power service system to ensure that residents can know the time of the power rationing implementation plan at the first time, and make full preparations. At the same time, the regional government strengthens multi-level supervision from top to bottom, and controls the implementation of power rationing work at every level.

## References

- [1] LIU L R, FENG S, WANG D L, et al. An Empirical Study on Chinese Microblog Stance Detection Using Supervised and Semi-supervised Machine Learning Methods [C]//Natural Language Understanding and Intelligent Applications. Kunming: Springer, 2016: 753-765.
- [2] XUE J, CHEN J X, HU R, et al. Twitter discussions and concerns about COVID-19 pandemic: Twitter data analysis using a machine learning approach [J/ OL].arXi: 2005.12830 [cs.SI].[2012.11.16].2020.<https://arxiv.org/abs/2005.12830v2>.
- [3] Zhang Jie. Research on the development of forest health tourism in Guangdong Province [D]. Guangzhou: Guangzhou University of Traditional Chinese Medicine, 2019:1-7.
- [4] Li Hang. Statistical Learning Methods (2nd Edition) [M]. Beijing: Tsinghua University Press, 2019: 111-153.
- [5] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, (3): 993-1022.