Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 509–517

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Information Technology and Quantitative Management

Building public market opinion indices for electricity consumption prediction of Chinese nonferrous metal industry

Wei Shang[a,b], Lei Cao[a,b], Feng Li[c], Xuemei Jiang[d], Xuerong Li[a, *]

[a]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences,Beijing 100190,China
[c.] Beijing Wuzi University, Beijing 101149, China
[d] School of Economics, Capital University of Economics and Business, Beijing 100390, China

Abstract

Nonferrous metal is an essential natural resource for national economy and industry production. Understanding the supply, demand and price trend of nonferrous metal market is important to economic prosperity analysis and might contribute to electricity power consumption forecast. This research builds related market opinion indices to understand the supply, demand, and price changes in nonferrous metal market. We propose a self-learning graph neural network model-based sentiment index network method for value mining of sentiment-related features, addressing the issue of sentiment loss caused by discontinuities in domain-specific news. Based on the structured indicators and the integration of Internet sentiment indices, we establish two single-item models: Granger Causality-based ARDL Model (G-ARDL) and the Factor Analysis-based Transformer Model (F-Transformer). Finally, we construct a Bagging random sampling integration prediction framework to obtain the weight coefficients for single model predictions. We merge the prediction results in different single models with weighted fusion, resulting in smaller prediction errors compared to single models and equal-weight integration models.

Keywords: Nonferrous metal industry; Electricity consumption; Transformer model; Integration forecasting framework

1. Introduction

Nonferrous metal is an essential natural resource for national economy and industry production. As one of the most energy consuming industry, electricity consumption of nonferrous metal smelting and pressing industry

* Corresponding author. Tel.: +10-13401088139.
E-mail address: lixuerong@amss.ac.cn

10.1016/j.procs.2023.08.008

ranked second among all the industries, just behind the power and heat supply industry in the year 2020, according to Chinese Statistics Bureau. Understanding the supply, demand and price trend of nonferrous metal market is important to economic prosperity analysis and might contribute to electricity power consumption forecast.

However, the prediction of electricity power consumption of nonferrous metal industry (hereinafter referred to as EPC-NM) is a difficult task, and there are few researches in related fields. EPC-NM involves various factors of supply and demand, which are difficult to distinguish and may add uncertainty to the prediction results. As a complementary information source, the internet data provides various information for market analysts and decision makers. Recent researches have been trying to collect news, forum articles, user comments from the Internet and use text mining techniques to extract meaningful information for business decision making [1-3]. Instead of reading all the information by human experts, computer algorithms process all the text data and convert it into relations or signals to support decisions.

This research aims at using a text mining approach to extract meaningful market opinion signals from industry Internet website and build related market opinion indices to understand the supply, demand, and price changes in nonferrous metal market. The construction of these market opinion indices is following a formal and rigorous approach, and intents to provide a way of identify the market trends represented by the media and participants' opinions. We propose a self-learning graph neural network model-based sentiment index network method for value mining of sentiment-related features, addressing the issue of sentiment loss caused by discontinuities in domain-specific news, making it suitable for EPC-NM forecasting.

Based on the analysis of structured indicators and the integration of Internet sentiment indices, we establish two single-item models: Granger Causality-Based ARDL Model (G-ARDL) and the Factor Analysis-Based Transformer Model (F-Transformer). Finally, we construct a Bagging random sampling ensemble prediction framework to obtain the ensemble weight coefficients for individual predictions. We merge the model prediction results in different subsets with weighted fusion, resulting in a non-ferrous metal industry electricity consumption prediction with smaller prediction errors compared to single-item models and equal-weight ensemble models.

The remaining parts of this paper are arranged as follows: Section 2 reviews relevant literature on sentiment indices construction and energy forecasting based on text mining; Section 3 introduces the proposed framework of EPC-NM prediction; Section 4 shows the data collection and empirical results; Section 5 is the conclusion and discussion.

2. Literature review

2.1. Sentiment indices construction

Some studies use time series data of macroeconomic indicators and capital markets to generate sentiment indices, such as Internet Finance Investor Sentiment Index (IFIS) [4], Consumer Sentiment Index (CSI) [5], etc. However, such methods of ignoring Internet text public opinion have unstable results. Methods such as dictionaries or machine learning can identify textual sentiment and generate sentiment indices, which directly and effectively aggregate public opinion. Hausler et al. [6] build sentiment index of stock market and real estate market based on support vector institution, and explore the relationship between sentiment index and total return of two markets by using vector autoregressive forecasting framework. Xu et al. [7] use the grounded theory and semi-automated method to construct the microblog emotion dictionary, and derive eight microblog emotion categories using the bottom-up method. The empirical results showed that the emotion recognition effect of the constructed dictionary is better than that of the five basic dictionaries. Ardia et al. [8] set up a text public opinion index aiming at the American industrial production index based on the American newspaper news data in the LexisNexis database, and find that the time dimension, subject distribution and calculation method of emotion of the public opinion index have different degrees of influence on the emotion index, and the emotion index can improve the prediction accuracy of economic growth. Different from Internet text data, Internet search data is also used in the construction of public opinion index, for example, Ito et al. [9] construct a sentiment index by using Google Trends to capture market sentiment in Japan and the United States.

2.2. Energy forecasting based on text mining

Considering that the Internet public opinion information can expand the characteristic dimension of energy prediction and better grasp the trend of energy market, Internet information is especially widely used in energy price prediction. Zhao et al. [10] propose a mixed oil price prediction model based on text mining. VADER method is used to divide online text into four kinds of emotions: compound, positive, negative and neutral, which are respectively added into the international oil price prediction model, and explore the influence of different emotional categories on oil price prediction. The results show that the stability and accuracy of the model are improved after the addition of emotion, and the improvement effect is more significant in the period of strong emotion. Cai et al. [11] use the BERT-BiLSTM combination forecasting method to analyze the emotional direction of investor and consumer comments. The research shows that the combination forecasting is more accurate than the single model and can better predict the emotional tendency of Internet users to energy events, which provides certain data support for grasping energy market trends. Loureiro et al. [12] analyze Twitter messages about climate change and energy-related issues in Spain and the United Kingdom. They use NCR dictionary to classify positive and negative emotions in tweets, investigate people's preferences for energy policies, and find that people obviously support renewable energy policies to reduce emissions and have a positive attitude towards the development prospects of renewable energy. Liu et al. [13] collect ten years of clean energy-related headlines to construct investor sentiment index, and establish a time-varying parameter vector autoregressive model to study the volatility correlation between investor sentiment and clean energy stocks, and find that the spillover effect between investor sentiment and clean energy stocks is small.

In summary, when quantifying unstructured news data, most studies directly use general sentiment dictionaries without building sentiment dictionaries for specific research objects. At the same time, they fail to consider the complex correlation between public opinion in related fields when processing news in specific fields. These problems will ultimately reduce the accuracy of quantifying emotions. In this study, we comprehensively consider the indicators related to electricity consumption in the nonferrous metal industry and unstructured influencing factors such as Internet news. In the selection of electricity consumption prediction model, many studies use a single prediction model or a simple combined model, but the performance of each model is uneven, and the accuracy of different models in different time periods is inconsistent. The integration model can eliminate this problem and improve the accuracy and stability of prediction results.

3. Methodology

This section introduces the proposed framework of EPC-NM prediction. Firstly, public opinion index construction method of non-ferrous metal industry used in this paper are introduced, then we present two prediction model of EPC-NM, namely G-ARDL model and F-Transformer model.

3.1. Public opinion index construction of nonferrous metal industry

Based on the Internet news data of the nonferrous metal industry, this paper constructs a public opinion dictionary in related fields. TextRank and TF-IDF algorithms are used to select trend words as seed words. On the basis of the extension of these seed words and their synonyms, the correlation calculation results of Word2Vec and emotion oriented mutual information (So-PMI) are integrated to determine the attribution of emotion words in the public opinion dictionary, and the label propagation algorithm is used to attach the opinion value of emotion words to the public opinion dictionary. A degree dictionary and a negative dictionary are added to the public opinion dictionary to describe the intensity level of semantics, and subject matching is carried out at the sentence level to improve the accuracy of quantified text. The news data under the same topic is subdivided, and the dictionary is used to quantify the news text of each sub-item, so as to construct the daily public opinion value of each sub-item.

To solve the problem of missing news data, a dynamic graph self-learning graph neural network (AGNN) is proposed to depict the correlation between public opinion indexes, so as to realize the formation of public opinion index network and achieve data completion of missing index data nodes. Graph neural network (GNN) is a deep learning method to learn graph structure data. GNN is used to supplement public opinion data, which can make full use of the characteristic information of observed values of other nodes in the network at different times [14]. Based on the distribution of electricity consumption in the nonferrous metal industry, electrolytic aluminium accounts for the highest proportion of electricity consumption, reaching 80%, followed by copper and zinc. We take the price, supply and demand of major nonferrous metals (aluminium, copper and zinc) as the theme sub-item, and regard each sub-item as the node, and the public opinion value of the sub-item as the characteristic sequence of the node. The designed graph neural network model is used to learn the existing hidden graph structure, discover the hidden correlation between nodes, and capture the spatial-temporal dependence between nodes. Finally, we construct three public opinion indices in the field of nonferrous metal industry: Price Public Opinion Index ($DPOI_{price}$), Supply Public Opinion Index ($DPOI_{supply}$) and Demand Public Opinion Index ($DPOI_{demand}$). Fig. 1 shows the construction framework.

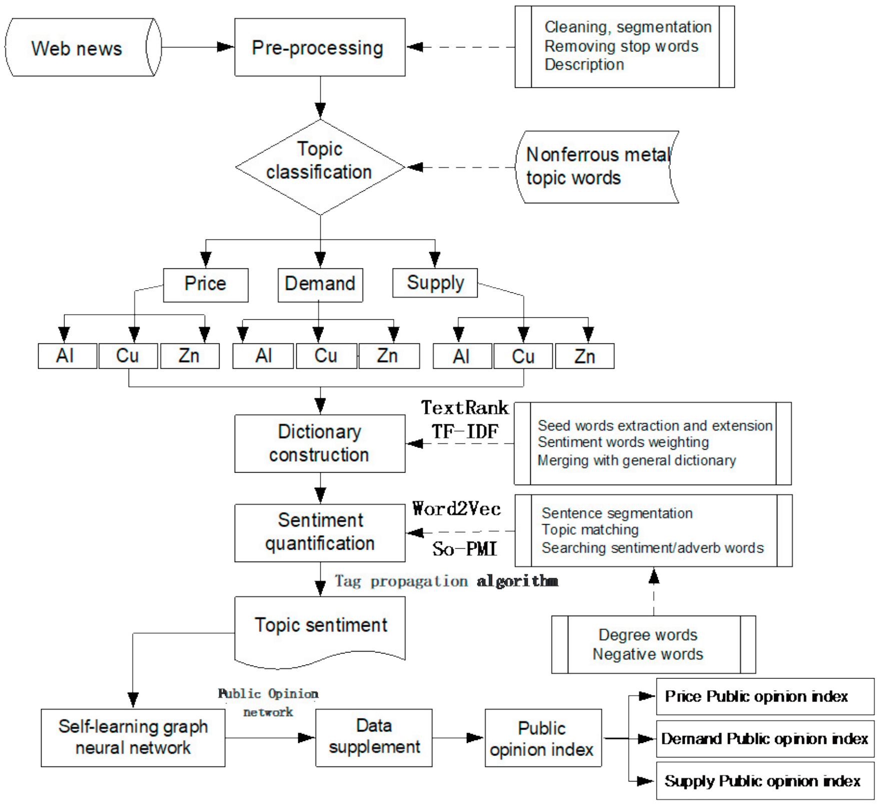

Fig. 1. Public opinion index construction

3.2. G-ARDL model of EPC-NM prediction

Based on the results of Granger causality test [15], this paper adds influence indicators with Granger causality to EPC-NM into the autoregressive model, and constructs an ARDL model based on Granger causality (G-ARDL model). In order to select reasonable exogenous variables to be added to the ARDL model, the optimal subset regression method is used for variable screening. Since the ARDL model is a time series model and allows the variable lag term to participate in the regression process, the 1-4 order lags of all relevant indicators including the EPC-NM is generated.

After optimal subset regression on the training set, according to AIC, BIC, Cp criteria and goodness of fit R^2, the first-order lag ($D_{Power}(-1)$) of EPC-NM, first order lag of Demand Public Opinion Index ($DPOI_{demand}(-1)$), first order lag and second order lag of Price Public Opinion Index ($DPOI_{price}(-1)$ and $DPOI_{price}(-2)$), second order lag of money supply M2 (M2(-2)), the fourth-order lag of aluminium production (Al(-4)), the fourth-order lag of copper production (Cu(-4)), the fourth-order lag of tin production (Sn(-4)) and the third-order lag of ten nonferrous metals production (Non(-3)) The following ARDL models are generated:

$$D_{Power} = \alpha D_{Power}(-1) + \beta_1 M2(-2) + \beta_2 Al(-4) + \beta_3 Cu(-4) + \beta_4 Sn(-4) + \beta_5 Non(-3) + \beta_6 DPOI_{demand}(-1) + \beta_7\, DPOI_{price}(-1) + \beta_8 DPOI_{price}(-2) + \sigma \tag{1}$$

3.3. F-Transformer model of EPC-NM prediction

In order to better enhance the explanatory power of variables in the model, we have grouped the collected variables according to the categories of production, supply-demand, macroeconomics and product gross profit, and conducted factor analysis on the grouped data to deduce appropriate factors for each group. Based on the factor analysis under each category, we get a total of 6 factors. Then we combine the 6 factor scores and the public opinion index data to form a new data set for the construction of EPC-NM forecast model.

In this paper, *PytorchL* is used to build Transformer model based on factor analysis. Transformer is a model for sequence-to-sequence tasks, which was first proposed in the paper [16] in 2017. Unlike recursive neural network or convolutional neural network, Transformer is the first representation that fully relies on self-attention to compute input and output. When defining the neural network, the Module class of *torch.nn* is used and the initialization (_int_) and forward propagation (forward method) are re-implemented. Specifically, the required network layer is added to the (_int_) method, and then define the forward propagation between layers to achieve the function of input to output. MSE is taken as loss function in transformer framework to carry out gradient descent training. In order to facilitate comparison, F-Transformer model takes nonferrous metal electricity consumption, 6 composed factors, Demand Public Opinion Index and Price Public Opinion Index as feature input, *input_steps*=6, *output_steps*=1, and uses relevant indicators of the past 6 months to predict the electricity consumption of the next 1 month. After the training set training, in the model structure, we take 8 head attention layers in the Encoder (*nhead*=8); *epochs* is set to 500 and *batch_size* to 16. In addition, the super-parameter settings of transformer are: dropout probability is 0.1 to prevent overfitting; The initial learning rate is 0.001.

4. Empirical results

4.1. Data collection and preprocessing

This study adopts monthly data from January 2015 to August 2021, and relevant indicators are derived from Wind database. Due to the influence of holidays and other factors, there are some missing values in the original data, and there are both daily and monthly data. Therefore, linear interpolation is used to complete the missing data, and monthly mean value is used to re-sampler the daily data into monthly data.

The Internet text data is mainly collected from the website of China Nonferrous Metal Daily (https://www.cnmn.com.cn/) from January 2015 to November 2021, including two columns of nonferrous metal futures and spot, and 56,184 texts are obtained after manually selecting and removing news apparently unrelated to the subject. The emotions of 808 pieces of news are manually marked. The examples are shown in Table 1.

Table 1. Examples of text news

Date	Title	Pageviews	Content	Column
2019/07	Consumption off-season demand to light; aluminium prices have downward pressure	3988	Starting in June, alumina prices began to fall, reducing the cost of electrolytic aluminium production and weakening support...	Futures
2021/06	ICSG: World refined copper production rose 3.2% year-on-year in the first half of 2021	1081	China's refined copper output rose 6% in the January-June period from a year earlier according to preliminary official data…	Spots

After the original text data is obtained by crawler software, text preprocessing operations should be carried out to clean redundant and chaotic text data and eliminate useless text and symbols. In the follow-up empirical analysis, the index data from January 2015 to December 2020 and the quantified public opinion index data are mainly used as training samples to predict EPC-NM from January to August 2021.

4.2. ARDL model based on Granger Causality (G-ARDL model)

In this section, the public opinion index and other relevant indicators are used to construct four G-ARDL models about EPC-NM. These models are established respectively for dynamic out-of-sample prediction. Model 1: G-ARDL model without considering public opinion index; Model 2: G-ARDL model considering Demand Public Opinion Index; Model 3: G-ARDL model considering Price Public Opinion Index; Model 4: G-ARDL model considering both Demand and Price Public Opinion Index. The prediction evaluation of the model is based on five criteria, namely mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute relative error (MAPE) and symmetric mean absolute percentage error (SMAPE). The prediction results were shown in Table 2.

Table 2. Prediction results of G-ARDL models

	MSE	RMSE	MAE	MAPE	SMAPE
Model 1	0.0207	0.1442	0.1368	1.9828	2.0041
Model 2	0.0167	0.1293	0.1152	1.6724	1.6893
Model 3	0.0117	0.1084	0.1001	1.4574	1.4699
Model 4	0.0085	0.0923	0.0785	1.1448	1.1538

The prediction results show that adding the public opinion index in the field of non-ferrous metals can improve the prediction accuracy, indicating that the public opinion index synthesized with the subject of price and demand can capture more information, so as to better predict EPC-NM. The prediction results verify the validity of the public opinion index constructed. The proposed prediction models of EPC-NM can improve the prediction accuracy of electricity consumption and contribute to the research of the national non-ferrous metal market.

4.3. Transformer model based on factor analysis (F-Transformer model)

In order to further demonstrate that the public opinion index and group factor analysis method established in this paper can effectively improve the forecasting performance of electricity consumption, F-Transformer models with and without considering public opinion index are compared, and grouped factor analysis method and non-grouped factor analysis method are compared respectively. Finally, four F-Transformer models are constructed, namely Model 1: Transformer model without considering public opinion index and ungrouped factor analysis;

Model 2: Transformer model considering public opinion index and ungrouped factor analysis; Model 3: Transformer model without considering public opinion index and grouping factor analysis; Model 4: Transformer model considering public opinion index and grouping factor analysis. Five criteria of MSE, RMSE, MAE, MAPE and SMAPE are used to evaluate four models. The results were shown in Table 3.

Table 3. Prediction results of F-Transformer models

	MSE	RMSE	MAE	MAPE	SMAPE
Model 1	0.0071	0.0841	0.063	0.926	0.928
Model 2	0.0070	0.0840	0.067	0.991	0.993
Model 3	0.0051	0.0712	0.050	0.738	0.739
Model 4	0.0015	0.023	0.020	0.305	0.305

According to Table 3, comparing the results of Model 1 and Model 4, it can be seen that the prediction errors of the prediction model that considers both public opinion index and grouped factor analysis are reduced. Compared with the results of Model 1 and Model 2, the public opinion index can improve the prediction accuracy of EPC-NM, indicating that the public opinion index contains information that is not available in other related variables. According to the comparison results between model 2 and Model 3, the grouped factor analysis method has better performance than the non-grouped factor analysis method, which indicates that the group factor analysis method is effective and reasonable.

4.4. Integrated prediction based on Bagging method

In this paper, a Bagging integrated prediction framework is constructed to obtain the integrated weight coefficient of single prediction models. With the monthly prediction error values of the two models as input, 10000 times of random sampling (with put back) are carried out, and the model prediction results in different subsets are combined in a way of weighting, thus improving the stability and generalization ability of the model. Based on the relevant indicator data from January 2015 to December 2020, we first forecast EPC-NM from January 2021 to August 2021 by using the above G-ARDL and F-Transformer models, and extract the forecast results for m times. For each sampling, The MAPE values of the two prediction methods are calculated, and the minimum MAPE is used to select the best model. The weight coefficient is increased 1/m times for each time the model is selected.

In this study, 10,000 sampling times are selected, and after calculation by bagging integration method, the weight coefficient of G-ARDL model is 0.1238 and that of F-Transformer model is 0.8762. It can be seen that the weight of Transformer model based on factor analysis is larger. After 100 times of bagging random protection sampling, its coefficient remains above 0.84, which also proves the effectiveness of F-Transformer model integrated into Internet public opinion index in this paper. However, the coefficient of the ARDL model based on Granger causality is lower, which is also consistent with the evaluation results given by the single model.

In order to verify the effectiveness of bagging random sampling integration adopted in this paper, five criteria of MSE, RMSE, MAE, MAPE and SMAPE are taken to evaluate prediction results. Compared with the predicted results of equal-weighted integration method, G-ARDL model and F-Transformer model, the results are shown in Table 4.

Table 4. Prediction results of Bagging models

Date	True value	Bagging integration	Equal-weighted integration	G-ARDL	F-Transformer
		Predicted value	**Predicted value**	**Predicted value**	**Predicted value**
Jan-21	6.7607277	6.7707287	6.8264529	6.8884938	6.7644120

Feb-21	6.6970772	6.7155281	6.7512998	6.799836	6.7027629
Mar-21	6.7520789	6.7879842	6.8262296	6.8781228	6.7743363
Apr-21	6.7418850	6.7432918	6.8019195	6.8543314	6.7295075
May-21	6.7571312	6.8026291	6.8267672	6.8595189	6.7940154
Jun-21	6.7512081	6.7719404	6.8058855	6.8519439	6.7598270
Jul-21	6.742676	6.7429295	6.7325397	6.7455793	6.7295000
Aug-21	6.7531543	6.7549473	6.7507135	6.7314002	6.7700266
MSE		0.001	0.003	0.009	0.002
RMSE		0.023	0.055	0.092	0.023
MAE		0.018	0.049	0.079	0.020
RMSE		0.265	0.718	1.154	0.305
SMAPE		0.266	0.721	1.145	0.305

Table 4 shows that the prediction performance of Bagging integration method with put back sampling is better than that of equal-weighted integration model. Compared with the two single prediction models, the Bagging integration method has better fitting effect and higher accuracy, indicating that the integration model adopted in this paper can break through the limitation that a single model can only capture partial features of sample data.

5. Conclusions and discussions

This paper studies the forecast of electricity consumption in the nonferrous metal industry. Aiming at the extraction of public opinion trend in the nonferrous metal market in the Internet news, a public opinion network construction method based on dynamic graph self-learning graph neural network is proposed. On this basis, the demand public opinion index, price public opinion index and supply public opinion index in the field of nonferrous metal are constructed. Then, two single models are established: ARDL model based on Granger causality and Transformer model based on group factor analysis to forecast the electricity consumption of nonferrous metal industry. Finally, the Bagging sampling integration framework is used to calculate the weight coefficients of the two models. The integrated prediction results are calculated according to the weights, and the performances of the single models, equal-weighted integration and Bagging integration model are compared. The results show that the prediction error of the integrated framework adopted in this paper is smaller and the prediction results are more accurate. It provides some insights for other high energy consumption industries to carry on electricity prediction analysis.

The public opinion network framework built in this paper is based on dynamic graph self-learning graph neural network, which is a general and systematic method to extract market trend views from unstructured data. All the information used to build a public opinion network comes from industry-specific news data and a small amount of given label data, and no human experience is required to intervene in the model quantification process. The overall process of public opinion network construction is fully applicable to the quantification of public opinion in other industries, and the synthesized public opinion index can be applied to the monitoring of public opinion and the analysis of the economy in the industry.

The research method of this paper is still to be improved. For example, the public opinion index generated by dynamic graph self-learning graph neural network mainly uses long text data of news, which may be insufficient in processing short text (such as news headlines). This paper does not consider the authenticity of information for the time being. Future research plans to consider authenticity or credibility, so as to build a public opinion index with more dimensions. In terms of model selection, in the future, we can also try to take the randomness of model parameters into consideration, so as to balance the advantages and disadvantages of various models more effectively and obtain more stable prediction results.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 72273137).

References

[1] Ardia D, Bluteau K, Boudt K. Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. International Journal of Forecasting 2019;35(4):1370-1386.
[2] Baker SR, Bloom N, Davis SJ. Measuring economic policy uncertainty. The Quarterly Journal of Economics 2016;131(4):1593-1636.
[3] Alola AA, Uzuner G. The housing market and agricultural land dynamics: Appraising with Economic Policy Uncertainty Index. International Journal of Finance and Economics 2020;25(2):274-285.
[4] Adam S. Forecasting returns with fundamentals-removed investor sentiment. International Journal of Financial Studies 2015;3(3):319-341
[5] Tsalikis J, Seaton B. Business Ethics Index: Measuring consumer sentiments toward business ethical practices. Journal of Business Ethics 2006;4:64.
[6] Hausler J, Ruscheinsky J, Lang M. News-based sentiment analysis in real estate: a machine learning approach. Journal of Property Research 2018;35(4):344-71.
[7] Xu L, Li L, Jiang Z, et al. A novel emotion lexicon for chinese emotional expression analysis on weibo: using grounded theory and semi-automatic methods. IEEE Access 2020;9:92757-92768.
[8] Ardia D, Bluteau K, Boudt K. Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. International Journal of Forecasting 2019;35(4):1370-1386.
[9] Ito T, Masuda M, Naito A, et al. Application of Google Trends ‐ based sentiment index in exchange rate prediction. Journal of Forecasting 2021;40(7):1154-1178.
[10] Zhao LT, Zeng GR, Wang WJ, et al. Forecasting oil price using web-based sentiment analysis. Energies 2019;12(22):4291.
[11] Cai R, Qin B, Chen Y, et al. Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM. IEEE Access 2020;8:171408-171415.
[12] Loureiro ML, Alló M. Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain. Energy Policy 2020;143:111490.
[13] Liu T, Hamori S. Does investor sentiment affect clean energy stock? Evidence from TVP-VAR-based connectedness approach. Energies 2021;14(12):1-21.
[14] Cai J, Cand EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization 2010;20(4):1956-1982.
[15] Granger C, Wiener J. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 1969;37:424-438.
[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). New York: Curran Associates Inc.; 2017, p. 6000–6010.