# The research of natural language processing (NLP) technology based on statistical machine learning on the investment decision of port and shipping enterprises

Xu Bin*[a], Gong Xiaoxing[a], Xue Xuebo[a], Liu Chao[a]

[a]School of Transportation Engineering, Dalian Maritime University, Dalian, Liaoning, 116026

* Corresponding author: xzyx_bin@163.com

**ABSTRACT.**This paper takes China's listed port and shipping companies as the research object, and uses 426 Chinese annual reports from 2001 to 2021 as a text corpus to quantify the emotional tone of the annual report texts of port and shipping companies based on natural language processing (NLP) of statistical machine learning. Study the relationship between the emotional tone index of the annual report text of port and shipping listed companies and the financial performance of the company. The innovations and characteristics of this paper are as follows: Firstly, construct a dictionary of emotional intonation in the field of port and shipping, and use this as a basis to construct an index of emotional intonation in annual reports of listed companies in port and shipping. The second is to establish a two-stage research method to study the relationship between the emotional tone of the annual report text of port and shipping listed companies and the financial performance of the company. In the first stage, the improved LASSO model is used to screen the control variables. In the second stage, a two-step regression based on the Fama-MacBeth model is used to explore the relationship between text tone and financial performance. The study found that the emotional tone index of annual reports constructed based on statistical machine learning NLP technology is significantly negatively correlated with the financial performance of port and shipping companies, which shows that the more positive(negative) the tone of the annual report of listed companies in port and shipping companies, the worse the financial performance of the company(good). The research results provide an analysis tool for the application of NLP technology in the field of port and shipping. It also provides decision support for the management and investment of port and shipping enterprises.

**Keywords:** transportation economics; NLP; domain sentiment lexicon; LASSO; Fama-MacBeth; annual report tone; financial performance

## 1 INTRODUCTION

In the context of the widespread application of NLP, business managers and investors have begun to pay attention to the important role of emotional tone contained in corpo-


A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_40

rate disclosure texts. This means that the study of unstructured text information in corporate annual reports can provide a new perspective for the strategy formulation of port and shipping companies and the decision-making of investors[1]. As an important information disclosure document, the annual report of the port and shipping company not only transmits data information to the outside world, but also serves as an external expression of the emotional and psychological activities of corporate managers. In addition, the emotional tone of the annual reports of port and shipping enterprises has a certain correlation with many economic indicators such as corporate financial performance. Therefore, this study aims to use natural language processing technology to analyze the text information of the annual reports of Chinese port and shipping companies to quantify the emotional tone indicators, and further explore the relationship between the emotional indicators and corporate financial performance. Specifically, this study is dedicated to studying whether the tone of voice in the annual reports of port and shipping companies can reflect accurate information on the level of financial performance of the company.

Existing research on the relationship between tone of text in annual reports of listed companies and corporate performance mainly focuses on two aspects. The first is the quantitative research of text intonation. The premise of quantifying text tone is to construct a sentiment dictionary. Loughran and McDonald[2] used annual reports, news reports, etc. as corpora, manually classified emotional words based on the Harvard Psychosociological Dictionary, and constructed a business sentiment dictionary to quantify text emotional intonation. However, from a linguistic point of view, Chinese is a high-context language, and the emotional expression of English words cannot be fully applied to the emotional expression of Chinese texts. Xie Deren and Lin Le[3] based on the English emotional dictionary combined with Chinese word habits to manually select words suitable for expression in the Chinese context to build an emotional dictionary, and verified that the tone of the text can provide incremental information about the company's future performance. However, the construction of the above-mentioned dictionaries is constructed by artificial discrimination methods, which cannot avoid the deviation caused by personal classification and screening of emotional words. Existing sentiment dictionaries for Chinese text sentiment analysis, such as Hownet Emotion dictionary and Dalian University of Technology sentiment dictionary and other general dictionaries. Such dictionaries are not very applicable to text sentiment analysis in professional industries. It is found that the constructed domain dictionary is conducive to improving the performance of prediction and correlation analysis compared with the basic sentiment dictionary. Yao et al.[1] established a corpus based on the comments posted on financial forums, combined with the existing general dictionary to construct a Chinese sentiment dictionary in the financial field, and found that the sentiment indicators of netizens' speeches in financial forums could predict the risk of stock crash. Liu Wenlong et al.[4] based on the cluster analysis of the components of the dictionary in the field of landscape architecture, can help accurately extract the sentiment analysis of Internet reviews. Jiang Cuiqing et al.[5] found unknown sentiment words from the user-generated content of social media, constructed a domain sentiment dictionary, and applied it to sentiment analysis of car reviews.

To sum up, there is no Chinese sentiment dictionary built for the domestic port and shipping field in the existing research. In fact, the operation and investment activities of port and shipping enterprises have strong information dependence. The annual report of port and shipping enterprises is an important way of information transmission. How to quantify the emotional tone of annual report text according to the characteristics of port and shipping field and apply it to the actual investment decision is one of the objectives of this paper.

The second is the research method of text emotional tone analysis. Liu Yishuang et al[6] used Logistic regression, decision tree and support vector machine to verify the change of financial risk warning ability after adding text emotional tone. The newly added index of textual emotional tone provides information that cannot be reflected by quantitative financial data. Bian Hailong et al.[7] used LIBSVM model to construct financial prediction model and mixed prediction model adding sentiment value of Web financial information. It is concluded that the introduction of sentiment indicators for Web financial information improves the forecasting performance. Calomiris et al.[8] studied the impact of textual emotional tone on the risk and return of financial market through elastic regression model. Bai et al.[9] studied the influence of sentiment index on freight rate index based on vector autoregression (VAR) model.

Most of the above emotional intonation research models are constructed using literature analysis to determine the control variables. Because there are many kinds of indicators related to finance. The method of subjective determination of control variables is not rigorous enough. It is more ideal to use objective screening methods. Moreover, Logistic regression and decision tree models have limitations in dealing with samples with heterogeneous characteristics. VAR model is easy to produce "small sample bias problem" when dealing with small sample data. The Fama-MacBeth regression model provides an efficient method for estimating standard errors that can take into account the effects of heterogeneous features and panel data structure. This helps to more accurately assess confidence intervals and significance of regression coefficients.

To sum up, this paper conducts extended research in the following aspects: Firstly, construct the annual report sentiment index of port and shipping listed companies. Innovatively established a corpus of annual reports of listed port and shipping companies, and constructed a Chinese sentiment dictionary suitable for the field of port and shipping through the NLP technology of statistical machine learning, and then constructed the annual report sentiment index of listed port and shipping companies based on the dictionary. Second, the EBIT after excluding earnings management is used as the financial performance index, which effectively eliminates the "noise" of earnings management. Third, establish a two-stage research method to study the relationship between the emotional tone of annual reports of port and shipping listed companies and corporate financial performance. In the first stage, the improved LASSO model is used to screen the control variables. The LASSO model is improved by eliminating the individual effect and time effect in the data, and the control variables are determined by combining subjective and objective methods. In the second stage, a two-step regression based on the Fama-MacBeth model is used to explore the relationship between text tone and financial performance, and to solve the heteroscedasticity problem in the regression model while considering cross-individual effect and time effect. Fourth, the research conclusions are of great significance to the operation and investment decisions of port and shipping companies, that is, to provide text emotional tone analysis tools in the field of port and shipping companies and to provide important theoretical references for rationally analyzing the disclosure texts of port and shipping companies to make scientific decisions.

## 2 RESEARCH DESIGN

Firstly, construct a dictionary of emotional intonation of annual reports of listed companies in the field of port and shipping, and build an emotional tone index based on this. Secondly, construct financial performance indicators that eliminate the "noise" of earnings management, and initially select control variables. Then, based on the improved LASSO model screening to determine the control variable indicators. Finally, the Fama-Macbeth regression model is constructed to explore the relationship between the emotional tone of annual reports of port and shipping listed companies and corporate financial performance.

### 2.1 Construction of port and shipping motional dictionary and emotional intonation index

The Construction of the Sentiment Dictionary of Annual Reports of Port and Shipping Enterprises.

The digital information in the annual report of the port and shipping enterprises can no longer bring the advantage of information asymmetry to the port and shipping enterprises or investors. However, the rich text content has become a new focus under the background of the application of big data technology and text sentiment analysis technology. In the process of researching the emotional tone of the annual report, it is found that the annual report contains a large number of cryptic emotional words, as shown in the following example, where the words in bold are the words expressing cryptic emotional intonation.

Example: "During the reporting period, in the face of the extremely severe external situation, the company made every effort to stabilize operations, tap potentials, and increase efficiency. Based on the overall situation of ensuring supply, the company gave full play to the advantages of professional ship operation and management, and successfully completed the task of... (Source: Ningbo Shipping 2022 Annual Report)".

In this paper, based on the annual report corpus of port and shipping companies, the emotional dictionary of annual reports in the field of port and shipping is constructed. The construction idea is shown in Figure 1.

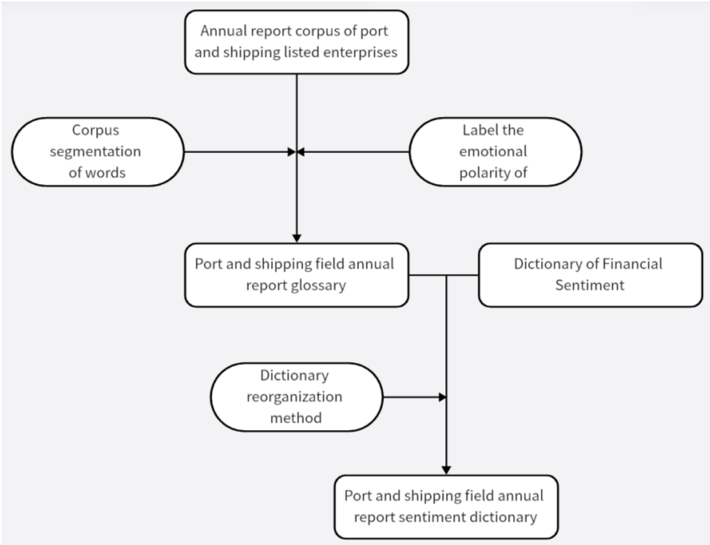


**Fig. 1.** Conceptual map of port and shipping sentiment dictionary construction.

The specific research steps are as follows:

1. Obtain annual reports from the official websites of shipping companies and Sina Finance and Economics websites, convert the file format into .txt files, and establish a corpus of annual reports of listed companies in port and shipping companies.
2. The corpus is cleaned, and the natural language processing technology based on statistical machine learning (jieba module in Python) segments continuous words into words, and removes useless words and symbols.
3. Classify and mark the emotional polarity according to the emotions expressed by the words, and build the port and shipping annual report word library.
4. Using the dictionary reorganization method, the financial sentiment dictionary is integrated, and the results of words segmented from the annual reports of port and shipping companies are added to deduplicate to obtain the sentiment dictionary of the annual report of the port and shipping industry.

This paper collects 426 annual reports of A-share port and shipping companies from 2001 to 2021 as a dictionary construction corpus (source: Sina Finance Network), and constructs an emotional dictionary of annual reports in the field of ports and shipping. The sentiment dictionary has a total of 2901 words, including 2016 positive sentiment words and 885 negative sentiment words. Examples are shown in Table 1.

**Table 1.** Example of port and shipping sentiment dictionary.

| positive words | | | negative words | | |
|---|---|---|---|---|---|
| navigator | Stabilize | on time | hijack | embargo | anchor |
| unimpeded | build | durable | collision | out of control | lower price |

Construction of annual report emotional tone indicators.

Based on the emotional intonation dictionary of the annual reports of port and shipping listed companies, the emotional intonation of the text is quantified by referring to the calculation method defined by Davis et al.[10] as shown in formula (1), formula (2) and formula (3).

$$S_{(i,t)} = P_{S(i,t)} - N_{S(i,t)} \tag{1}$$

$$P_{S(i,t)} = \sum\nolimits_p P_{(i,t)} / T_{(i,t)} \tag{2}$$

$$N_{S(i,t)} = \sum\nolimits_n N_{(i,t)} / T_{(i,t)} \tag{3}$$

In formula (1), $i$ is the individual variable of the enterprise; $t$ is the time variable of the year; $S_{(i,t)}$ is the sentiment index of the annual report of the company $i$ in $t$; $P_{S\,(i,t)}$ is the positive Sentiment index, $N_{S\,(i,t)}$ is the negative sentiment index of company $i$'s annual report for year $t$. In formula (2), $p$ is the word individual with positive emotion; $\Sigma_p P_{(i,t)}$ is the sum of the occurrence times of all positive words in the annual report of company $i$ in year $t$; $T_{(i,t)}$ is the emotion in the annual report of company $i$ in year $t$ sum of words. In formula (3), $n$ is the word individual with negative emotion; $\Sigma_n N_{(i,t)}$ is the sum of the times of all negative words appearing in the annual report of company $i$ in year $t$.

## 2.2 Determination of financial performance indicators and primary selection of control variables

Construction of financial performance indicators.

Financial performance is an important indicator that reflects the level of a company's financial operations. Most of the existing research measures the financial performance of enterprises through indicators such as return on assets (ROA) and earnings per share(EPS). However, the above indicators do not remove the "noise" of earnings management. This paper refers to the revised accrual profit calculation model of Cornett et al.[11] and selects the EBIT margin of total assets after removing the "noise" of earnings management as the financial performance indicator. The calculation steps are as follows:

The first step is to estimate the revised accrual profit rate, as shown in formula (4):

$$\frac{TA_{i,t}}{Assets_{i,t-1}} = \alpha_0 \frac{1}{Assets_{i,t-1}} + \beta_1 \frac{\Delta Sales_{i,t}}{Assets_{i,t-1}} + \beta_2 \frac{PPE_{i,t}}{Assets_{i,t-1}} \tag{4}$$

In formula (4), $TA_{i,t}$ represents the accrued profit of enterprise $i$ in year $t$, which is the difference between net profit and cash flow of operating activities. $Assets_{i,t-1}$ represents the total assets of enterprise $i$ in year $t$-$1$. $\Delta Sales_{i,t}$ represents the change in annual sales, which is the difference between the annual sales of company $i$ in year $t$ and the previous year. $PPE_{i,t}$ represents the total fixed assets of company $i$ in year $t$. $\alpha_0$ represents regression intercept. $\beta$ is the regression coefficient.

The second step is to calculate the difference between the actual accrual rate of profit and the estimated contingency rate of profit to obtain the manipulated accrual rate of profit, as shown in formula (5):

$$DA = \frac{TA_{i,t}}{Assets_{i,t-1}} - \left( \alpha_0 \frac{1}{Assets_{i,t-1}} + \beta_1 \frac{Sales_{i,t}}{Assets_{i,t-1}} + \beta_2 \frac{PPE_{i,t}}{Assets_{i,t-1}} \right) \tag{5}$$

In formula (5), $DA$ represents the manipulated accrual rate. $\hat{\alpha}_0$ represents the estimated value of the regression intercept. $\hat{\beta}$ represents the estimated value of the regression coefficient.

The third step is to calculate the EBIT margin of total assets after removing the "noise" of earnings management, as the financial performance indicator of the enterprise, as shown in formula (6):

$$UnEBIT_{i,t} = \frac{EBIT_{i,t}}{Assets_{i,t}} - DA \tag{6}$$

In formula (6), $UnEBIT_{i,t}$ represents the profit margin before interest and taxes of total assets of enterprise $i$ after excluding the "noise" of earnings management in year $t$, and represents the financial performance index of this paper, which will be represented by $U_{i,t}$ in the following. $EBIT_{i,t}$ is the EBIT profit of enterprise $i$ in year $t$.

The primary control variables were selected by literature analysis.

Based on existing literature research, corporate financial performance is an indicator that reflects whether corporate strategy and its implementation and execution are contributing to the final business performance. The selection of control variables in this paper starts from four aspects: profitability, solvency, capital structure, and governance capacity. The specific description and calculation methods of control variables are shown in Table 2.

**Table 2.** Control variable names and their explanations.

| Index classification | Indicator name | Variable symbol |
|---|---|---|
| Profitability | ROE (weighted) | $W_R$ |
| | Cash recovery rate of total assets | $O_C$ |
| | Net profit from operating activities | $O_P$ |
| | EPS growth rate | $E_P$ |
| | Growth rate of total assets | $T_Y$ |
| Ability to pay | Cash flow non - current Debt Ratio | $N_C$ |
| | Current ratio | $C_U$ |
| | Cash ratio | $C_R$ |
| | Overspeed ratio | $S_U$ |
| Capital Structure | Long-term asset fit Ratio | $L_F$ |
| | Fixed assets ratio | $F_A$ |
| | Total working capital | $W_C$ |
| | Total assets of port and shipping equipment | $T_P$ |
| Governance capacity | Number of managers receiving compensation | $N_P$ |
| | Total salary | $T_O$ |
| | Equity Concentration I | $O_{CI}$ |

### 2.3 Selection of control variables based on the improved LASSO method

Although the control variable framework constructed by the literature analysis method used above includes the basic financial situation indicators and governance capacity indicators of port and shipping companies, it still needs to be further streamlined and optimized. This paper improves the LASSO model and further selects the control variables. The steps are as follows:

Firstly, construct the following regression model, as shown in formula (7):

$$y_{it} = \alpha_0 + \beta_1 x_{1it} + \cdots + \beta_n x_{nit} + \mu_i + v_t + \varepsilon \tag{7}$$

In formula (7), $y_{it}$ represents the dependent variable of the model, here refers to the financial performance of company *i* in year *t*. $x_{nit}$ represents the independent variable of the model, here refers to the *nth* candidate control variable of company i in year *t*. $\alpha_0$ represents the intercept item. $\beta_n$ represents the regression coefficient of the *nth* candidate control variable ($n=1, 2...N$). $\mu_i$ represents the individual effect of the enterprise. $v_t$ represents the time effect of the enterprise. $\varepsilon$ represents the random error item.

Then, eliminate the individual effect ($\mu_i$) and time effect ($v_i$) in the model, as shown in formula (8):

$$Y_m = \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{8}$$

Detailed variant methods and explanations are shown in formulas (9)-(13):

$$Y_m = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y} \tag{9}$$

$$X_n = x_{nit} - \bar{x}_{ni} - \bar{x}_{nt} + \bar{x}_n \tag{10}$$

$$\bar{y}_i = \alpha_0 + \beta_1 \bar{x}_{1i} + \cdots + \beta_n \bar{x}_{ni} + \mu_i + \bar{v} + \varepsilon \tag{11}$$

$$\bar{y}_t = \alpha_0 + \beta_1 \bar{x}_{1t} + \cdots + \beta_n \bar{x}_{nt} + \bar{\mu} + v_t + \varepsilon \tag{12}$$

$$\bar{y} = \alpha_0 + \beta_1 \bar{x}_1 + \cdots + \beta_n \bar{x}_n + \bar{\mu} + \bar{v} + \varepsilon \tag{13}$$

In formula (8), $Y_m$ represents the dependent variable that eliminates the individual effect and time effect, and this paper represents the corresponding financial performance of the enterprise; $m=1,2...M$ represents the financial performance of each enterprise that eliminates the individual effect and time effect. $X_n$ represents the candidate control variable to eliminate the individual effect and time effect; where $n=1,2...N$ represents the number of variable dimensions. In formula (9), $\bar{y}_i$ represents the individual average value of the financial performance of enterprise *i* within the statistical time range. $\bar{y}_t$ represents the time average value of the financial performance of all enterprises in year *t*, that is, the average value on the time section. $\bar{y}$ represents the average value of all enterprises in the statistical time range the mean value of the financial performance without considering the individual effect and time effect. In formula (10), $\bar{x}_{ni}$

represents the individual mean value of the *nth* candidate control variable of the *ith* enterprise within the statistical time range. $\bar{x}_{nt}$ indicates the time mean value of the *nth* candidate control variable for all enterprises in year *t,* that is, the mean value on the time section. $\bar{x}_n$ indicates the indifferent mean value of the *nth* candidate control variable for all enterprises within the statistical time *range.* In formula (11), $\bar{v}$ represents the mean value of the time effect without considering individual differences. In formula (12), $\bar{\mu}$ represents the mean value of the individual effect without considering the influence of time.

Next, add the L1 norm penalty term to get the minimization formula:

$$\sum_{m=1}^{M}(Y_m-\beta_0-\sum_{n=1}^{N}\beta_n X_{mn})^2+\lambda\sum_{n=1}^{N}|\beta_n|=RSS+\lambda\sum_{n=1}^{N}|\beta_n| \tag{14}$$

In formula (14), RSS is the residual sum of squares. $\lambda$ is the adjustment parameter.

Finally, transform the regression model of formula (14) into the objective function of solving the minimum $\beta$ as shown in formulas (15) and (16):

$$LASSO(\beta)=\arg\min_{\beta}\left\{\sum_{m=1}^{M}(Y_m-\beta_0-\sum_{n=1}^{N}\beta_n X_{mn})^2+\lambda\sum_{n=1}^{N}|\beta_n|\right\} \tag{15}$$

$$\text{s.t.}\ \sum_{n=1}^{N}|\beta_n|\leq c \tag{16}$$

Among them, Equation (15) represents the transformation of the model into $\beta$ an objective function for finding the minimum regression coefficient. Equation (16) is the constraint condition of the objective function, and c≥0.

As the value of the adjustment parameter $\lambda$ increases, the compression strength of the variable regression coefficients is continuously enhanced, so that the regression coefficients of some control variables to be selected are compressed to 0 one by one, thereby obtaining an optimized set of control variables.

### 2.4 Research Model on the Correlation Between Annual Report Emotional Tone and Corporate Financial Performance

This paper studies the relationship between the financial performance and the emotional tone of annual reports of port and shipping listed companies. The reasons for using the Fama-MacBeth regression model are as follows: Firstly, it is beneficial to control the individual effect and time effect in the panel data, and control the unobserved individual characteristics and invariant time characteristics on the regression results by introducing fixed effects of individuals and time, which can eliminate individual and temporal heterogeneity. Second, heteroskedasticity in regression models can be addressed by estimating in two stages, first estimating the effect of individual effect through a fixed-effects model, and then estimating the effect of time effect in a second stage. The principle of the Fama-MacBeth model is as follows:

Firstly, the first step of time series regression is performed to obtain the regression coefficients of the independent variable and the control variable. The regression model is shown in formula (17):

$$y_{it} = \beta_{i0} + \beta_{i1}x_{1it} + \ldots + \beta_{in}x_{nit} + \alpha_i + \varepsilon_{it} \tag{17}$$

In formula (17), $y_{it}$ represents the dependent variable of the first step of the time series regression of the Fama-MacBeth model, where it represents the financial performance of company *i* in year *t*. *i* represents the individual sample number of the enterprise. *t* represents time sample serial number. $x_{nit}$ indicates the independent variable of the first step of the time series regression of the Fama-Macbeth model, where it indicates the *nth* variable of company *i* in year *t*; $n=1,2...N$ indicates the serial number of the variable here. $\beta_{i0}$ represents the intercept item. $\beta_{in}$ represents the variable regression coefficient obtained from the first step of the time series regression of the Fama-MacBeth model, here represents the regression coefficient of the *nth* variable of enterprise *i*; $\alpha_i$ represents the individual effect error of enterprise *i* difference; $\varepsilon_{it}$ represents the random disturbance term.

$\beta_{in}$ obtained by formula (17) is used as the independent variable of the second step of cross-sectional regression, and cross-sectional regression is performed at each time point *t*. The regression model is as follows:

$$y_{it} = \gamma_{11}\beta_{i1} + \ldots + \gamma_{tn}\beta_{in} + \alpha_{it} \tag{18}$$

In formula (18), $\gamma_{tn}$ represents the regression coefficient of the variable $\beta_{in}$. $\alpha_{it}$ represents the error term.

Finally, an independent cross-section regression is conducted in each t, and the parameters obtained from these T cross-section regressions are taken as the estimation of the regression. Such as formulas (19), (20) shown:

$$\gamma = \frac{1}{T}\sum_{t=1}^{T}\gamma_t \tag{19}$$

$$\alpha_i = \frac{1}{T}\sum_{t=1}^{T}\alpha_{it} \tag{20}$$

In the Fama-Macbeth cross-section regression, the sample points of T periods are processed independently to obtain T sample estimates of $\lambda$ and $\alpha$. We can then find the standard errors of $\lambda$ and $\alpha$.

# 3 EMPIRICAL ANALYSIS

## 3.1 Data collection and processing

This chapter uses the annual report data and financial ratio data of 19 listed port and shipping companies from 2008 to 2021 for empirical analysis (from: RESSET database). In order to avoid the impact of the magnitude difference of data in dimension and value, this paper standardizes the data. In addition, there are a small number of missing values in the collected data, and the interpolation method is used in this paper to fill in the missing data.

### 3.2 Screening results of control variables based on the improved LASSO model

In this paper, the LASSO model regression uses the coordinate descent method to obtain the regression coefficients. Adjust the parameter settings $\gamma = e^{x}(x \in [-10, +\infty))$, and the regression results of the control variables to be selected are shown in Figure 2.

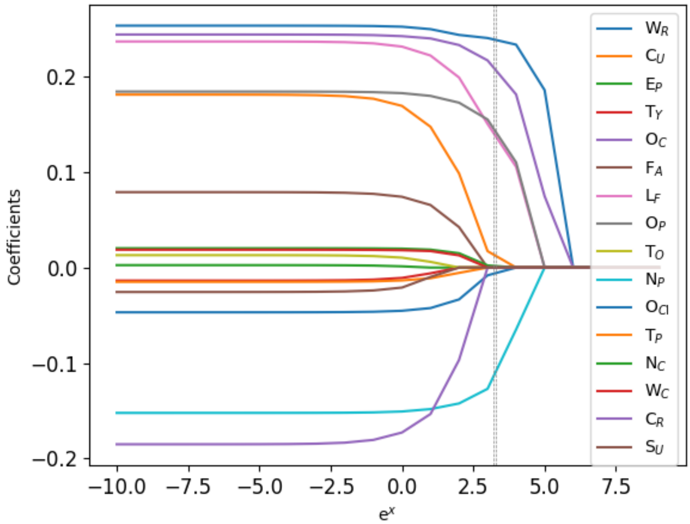


**Fig. 2.** Regression coefficient curve of control variables.

Figure 2 shows the gradual compression of the regression coefficients of the control variables to be selected with the change of the regulatory parameters. There are two dashed lines in the figure, the first represents the value of $\lambda$ when the mean squared error is the smallest, and the other is the value of $\lambda$ when the mean squared error is one standard error away. When the value of the adjustment parameter $\lambda$ is close to the minimum value of the mean square error, the selected control variables are also consistent with the four basic financial aspects of port and shipping enterprises, and the results of selected control variables are shown in Table 3.

**Table 3.** Improved LASSO screening control variable results.

| Index classification | Indicator name | Variable symbol |
|---|---|---|
| Profitability | ROE (weighted) | $W_R$ |

| | | |
|---|---|---|
| | Cash recovery rate of total assets | $O_C$ |
| | Net profit from operating activities | $O_P$ |
| Ability to pay | Cash flow non-current debt ratio | $N_C$ |
| | Current ratio | $C_U$ |
| Capital Structure | Long-term asset fit ratio | $L_F$ |
| Governance capacity | Number of managers receiving compensation | $N_P$ |
| | Equity concentration I | $O_{CI}$ |

The control variables obtained by using the improved LASSO model not only conform to the four aspects that reflect the basic financial situation of port and shipping companies, but also objectively reduce the number of control variables, thereby optimizing the combination of control variables.

### 3.3 Research results based on the Fama-MacBeth

Table 4 shows the regression results of the annual report emotional tone index on the corporate financial performance based on the sentiment dictionary of the annual report of port and shipping listed companies. The results show that there is a significant negative correlation between the emotional tone index of the annual report of port and shipping listed companies and the financial performance of the company at the level of 5%, indicating that the more positive (negative) the emotional tone of the annual report of the listed port and shipping listed companies, the worse the real financial performance of the company (Good), a positive and optimistic annual report emotional tone will reduce the financial performance of port and shipping companies.

Based on the framing effect, unusually positive or negative emotional tone in annual reports may be an illusion of rise and fall created by corporate managers to maintain an image conducive to corporate development. In the context of the current port and shipping market, contemporary investors not only pay attention to the hard data indicators in the disclosure of annual reports of enterprises, but also gradually understand and judge the incremental information in the disclosure of annual reports and other texts with the development of text analysis technology. The two sides form a game situation of "rendering" and "parsing", resulting in serious information asymmetry.

**Table 4.** Fama-MacBeth regression results.

| | $U_{i,t}$ | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| $S_{(i,t)}$ | -0.160*** | -0.173*** | -0.102*** | -0.063** |
| | (-2.97) | (-2.76) | (-3.83) | (-2.14) |
| $O_C$ | 0.246* | 0.284* | 0.159* | 0.190* |
| | (1.76) | (1.82) | (1.87) | (1.92) |
| $W_R$ | 0.250*** | 0.197*** | 0.172** | 0.116 |
| | (2.65) | (2.63) | (1.99) | (1.63) |
| $N_P$ | -0.139** | -0.146** | -0.096** | -0.120*** |

| | | | | |
|---|---|---|---|---|
| | (-2.26) | (-2.42) | (-2.23) | (-2.70) |
| $O_{CI}$ | -0.122 | -0.139 | -0.084 | -0.131** |
| | (-1.18) | (-1.54) | (-1.30) | (-2.30) |
| $L_F$ | 0.117 | 0.151 | 0.142** | 0.095 |
| | (1.24) | (1.57) | (2.01) | (1.25) |
| $O_P$ | | -0.008 | 0.004 | 0.029 |
| | | (-0.12) | (0.07) | (0.43) |
| $N_C$ | | | 0.234 | 0.238 |
| | | | (1.41) | (1.44) |
| $C_U$ | | | | 0.137*** |
| | | | | (4.25) |
| Intercept | 0.315*** | 0.354*** | 0.281*** | 0.273*** |
| | (2.85) | (3.22) | (2.73) | (2.75) |
| Adj. $R^2$ | 57.82% | 58.38% | 65.84% | 65.07% |

## 3.4 Stability test

In order to better verify the stability of the model, this paper uses ROA (return on assets) and EPS (earnings per share) as proxy variables of corporate financial performance for regression analysis. The results are shown in Table 5. When ROA is the proxy variable of financial performance, $S_{(i,t)}$ and financial performance are highly significant at the 1% level. When EPS is the proxy variable of financial performance, $S_{(i,t)}$ and financial performance are still highly significant at the 10% level. It is verified that the research model in this paper has high robustness.

**Table 5.** Robustness test results.

| | ROA | | EPS | |
|---|---|---|---|---|
| | (1) | (2) | (1) | (2) |
| $S_{(i,t)}$ | -0.127*** | -0.091*** | -0.059 | -0.098* |
| | (-3.93) | (-2.63) | (-1.51) | (-1.73) |
| $O_C$ | -0.145*** | -0.127*** | 0.149** | 0.179** |
| | (-3.27) | (-3.34) | (2.45) | (2.41) |
| $W_R$ | 0.797*** | 0.742*** | 0.719*** | 0.712*** |
| | (7.23) | (7.84) | (7.03) | (6.94) |
| $N_P$ | -0.028 | -0.044 | -0.149*** | -0.144*** |
| | (-0.70) | (-1.17) | (-3.88) | (-3.88) |
| $O_{CI}$ | -0.004 | -0.064* | 0.020 | 0.011 |
| | (-0.10) | (-1.93) | (0.60) | (0.26) |
| $L_F$ | 0.123 | 0.076 | -0.043 | 0.018 |
| | (1.50) | (0.96) | (-0.53) | (0.38) |
| $O_P$ | 0.052 | 0.089 | 0.158* | 0.140** |
| | (0.70) | (1.25) | (1.84) | (2.13) |
| $N_C$ | 0.033 | 0.018 | -0.329** | -0.354** |
| | (0.37) | (0.21) | (-2.20) | (-2.26) |
| $C_U$ | | 0.170*** | | -0.050 |
| | | (3.71) | | (-1.13) |
| Intercept | 0.162** | 0.156** | 0.119 | 0.157* |

| | | | | |
|---|---|---|---|---|
| | (2.47) | (2.49) | (1.26) | (1.66) |
| Adj. $R^2$ | 80.26% | 83.02% | 69.15% | 67.40% |

# 4 CONCLUSION

This paper focuses on A-share listed port and shipping enterprises. Based on Chinese word segmentation system and text sentiment classification algorithm, this paper constructs a Chinese sentiment intonation dictionary suitable for port and shipping field. Based on the dictionary, this paper constructs the emotional tone index of the annual report of port and shipping listed enterprises. Then a two-stage research method is used to study the relationship between the emotional tone of the annual report text and the financial performance of the port and shipping listed enterprises. The results show that there is a significant correlation between the emotional tone of annual reports and the financial performance of listed companies. From the perspective of port and shipping enterprises, the editors of annual reports may modify the tone of the text disclosed in annual reports to a certain extent on the premise of not violating laws and regulations, so as to achieve the expected rendering effect. From the perspective of investors, some investors have combined text sentiment analysis methods to quantify indicators, thus affecting investment decisions. The two sides form a game situation of "rendering" and "parsing", which intensifies the degree of information asymmetry.

# ACKNOWLEDGMENT

Fund project: National Natural Science Foundation of China (71974023).

# REFERENCES

1. Yao, J. Q., Feng, X., Wang, Z. J., Ji, R. R., Zhang, W. (2021) Tone, Sentiment and Market Impacts:The Construction of Chinese Sentiment Dictionary in Finance. Journal of Management Science, 24: 26-46. DOI:10.19920/j.cnki.jmsc.2021.05.002
2. Loughran, T, Mcdonald, B. (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66: 35-65. DOI:10.1111/j.1540-6261.2010.01625.x
3. Xie, D. R., Lin L. (2015) Do management tones help to forecast firms' future performance: A textual analysis based on annual earnings communication conferences of listed companies in China. Accounting Research, 2: 20-27. DOI:10.3969/j.issn.1003-2886.2015.02.003
4. Liu, W. L., Huang, W. (2021) Sentiment clustering of landscape constituents of Liuyuan Garden based on domain dictionary. Science Technology and Engineering, 21: 3174-3179. DOI:10.3969/j.issn.1671-1815.2021.08.028
5. Jiang, C. Q., Guo, Y. B., Liu, R. (2019) Constructing a domain sentiment lexicon Based on Chinese social media text. Data Analysis and Knowledge Discovery, 3: 98-107. DOI:CNKI:SUN:XDTQ.0.2019-02-012
6. Liu, Y. S., Chen, Y. Y. (2018) Tone at the top and credit risk warning for listed companies: textual analysis of company annual reports. Financial Economics Research, 33: 46-54. DOI:CNKI:SUN:JIRO.0.2018-04-005

7. Bian, H. R., Wan, C. X., Liu, D. X., Jiang, T. J. (2013) Studyon financial crisis prediction model with Web financial information for listed companies. Computer Science, 40: 295-315. DOI:10.3969/j.issn.1002-137X.2013.11.060
8. Calomiris, C. W., Mamaysky H. (2018) How news and its context drive risk and returns around the world. Journal of Financial Economics, 133: 299-336. DOI:10.2139/ssrn.2944826
9. Bai, X. W., Lam, J., Jakher, A. (2021) Shipping sentiment and the dry bulk shipping freight market: New evidence from newspaper coverage. Transportation Research Part E: Logistics and Transportation Review, 155. DOI：10.1016/j.tre.2021.102490
10. Davis, A. K., Ge, W., Matsumoto, D., Zhang, J. L. (2015) The effect of manager-specific optimism on the tone of earnings conference calls. Review of Accounting Studies, 20:639-673. DOI:10.1007/s11142-014-9309-4
11. Cornett, M. M., Marcus, A. J., Tehranian, H. (2008) Corporate governance and pay-for-performance: The impact of earnings management. Journal of Financial Economics ,87: 357-373. DOI：10.1016/j.jfineco.2007.03.003