International Journal for Housing Science and Its Applications
Publish August 3, 2025. Volume 46, Issue 3 Pages 615-626

https://doi.org/10.70517/ijhsa46345

Text Analysis and Risk Identification of Financial Reporting Based on Natural Language Processing Algorithms in the Era of Intelligence

Xiangling Wang[1,*]
[1] Department of Economics and Trade, Yongcheng Vocational College, Yongcheng, Henan, 476600, China
Corresponding authors: (e-mail: gwangxiangling@163.com).

Abstract As an important data of annual operation and production overview and financial situation of listed companies, the analysis of their text sentiment has an important application value in financial risk identification. In this paper, the financial reports of listed companies are taken as the research object, and TF-IDF is used to extract the structured, data-oriented and visualized information in the text. Then, using N-Gram model, the text information is processed by word vector. Subsequently, the improved sentiment co-occurrence algorithm is used to extract and expand the general sentiment lexicon to construct the financial report sentiment lexicon. Meanwhile, the SEN-TF-IDF algorithm is introduced to build the annual report sentiment dataset. The construction and improvement of the financial report sentiment dictionary is completed through the extraction of financial report text information and the learning of word vectorized representation. Comparing with the general sentiment dictionary, the financial report sentiment dictionary has the highest F1 value of 0.872 under the research threshold of 0.6, which demonstrates its superiority in analyzing and mining the sentiment tendency in the field of financial reporting.

Index Terms financial reporting, sentiment dictionary, risk identification, SEN-TF-IDF

I. Introduction

With the rise of the knowledge economy and the accelerated development of global economic integration, the annual report system, as an important public system, has become a means of communication between enterprises and external enterprises [1], [2]. Among them, the enterprise financial report is an important document to carry out the company's financial status, business performance and future development plan for the first time, carrying rich information about the company [3]. By exploring the information disclosure mechanism of corporate financial statements, it enriches the relevant theories, provides a deeper understanding of the core competitiveness and development dynamics of enterprises, and provides investors with more comprehensive references for investment decisions [4]-[6]. At the same time, it also facilitates regulators, auditors, analysts and investors to read and analyze and evaluate the content of financial statements, so as to better promote the formulation and implementation of China's corporate financial standards [7]-[9].

In the current capital market, all publicly disclosed data of listed companies is one of the most important information for management, investors and the Securities and Futures Commission (SFC) and other organizations, which contains both numerical and textual information [10], [11]. Currently, the wide application of data makes companies rely more on numerical information for risk disclosure. With the arrival of the era of big data, the complexity of operations becoming higher and the changing needs of investors, the traditional financial reporting analysis methods have been difficult to cope with the changing market environment and business risks [12], [13]. Faced with such an era, there is more and more supplementary information in financial reports, more textual information loaded in the note section on top of the numerical information, and more and more enthusiasm for risk disclosure by managers with the help of text [14]-[16]. As a result, investors' decisions not only rely on the numerical information displayed on the surface of financial statements, but also deeply interpret the potential real meaning behind the text [17], [18]. At the same time, the improvement of data processing capabilities implies the improvement of financial analysis thresholds and the expansion of the scope, which provides the possibility of processing textual information at the technical level [19], [20].

A large number of scholars have explored the impact of risk disclosure on corporate development and regulation from financial reporting, with a view to providing specific framework ideas and improving the feasibility of text-level risk disclosure research. Lewis, C. and Young, S. showed that text automation analysis methods are well suited to the process of analyzing corporate reports with progressively more textual content, and that the application of natural language processing (NLP) to corporate financial reporting that has a significant impact on financial

reporting regulation by continuously improving the effectiveness of NLP processing [21]. Wujec, M. carries out sentiment analysis on text documents containing corporate financial information and proposes a sentiment analysis method that does not require manual labeling of data and subjective assessment to obtain accurate predictions related to the direction of market reactions and business risks of companies [22]. Chan, S. W. and Chong, M. W. proposed a grammar-based Sentiment Analysis Engine (SAE) which showed high effectiveness in extracting value expressions from unstructured text and predicting financial market trends [23]. Pejić Bach, M. et al. investigated the application of text mining techniques based on big data analytics in the field of financial information analysis, which can effectively integrate semi-structured and unstructured data information in the financial sector of the enterprise and provide support for decision-making on enterprise development [24]. Wei, L. et al. utilized a semi-supervised text mining algorithm for risk disclosure of textual information in corporate financial statements to analyze in detail the trends of corporate risk factors over time and rank them in order of importance [25]. Hsu, M. F. et al. introduced text mining and statistical methods to construct an extraction model of business risk factors for corporate financial statements and input the results into a prediction model based on support vector machines, which contributed greatly to the development of business performance and financial reporting regulation [26].

This paper firstly describes the preprocessing steps of the text information of financial annual report, selects TF-IDF and N-Gram model for the extraction of keywords and features of text information as well as the vectorized representation of text words, and carries out the data preparation for the financial report-specific sentiment dictionary. Then the expansion of the general sentiment dictionary based on the improved sentiment co-occurrence algorithm is elaborated to construct the financial report-specific sentiment dictionary. The SEN-TF-IDF algorithm is introduced to calculate the sentiment tendency of financial annual reports, and the financial annual report sentiment dataset is established. Finally, the effectiveness of the sentiment dictionary for financial annual reports is examined by analyzing the financial annual reports of a group, comparing similar algorithms and applying experiments.

II. Text processing and word vectorization

II. A. Text information pre-processing

The detailed processing steps for the text information are as follows:

(1) Extract the MD&A text content in the annual reports of the target samples. Firstly, crawl and download the samples and the relevant annual reports corresponding to the paired samples from the web. Second, select the annual reports before updating, exclude the annual report summaries, and retain only the annual reports published for the first time in the year. Finally, we extracted the sections of the annual reports involving “Board of Directors' Report”, “Discussion and Analysis of Business Situation”, and “Management's Discussion and Analysis”. The detailed procedure for extracting the text of MD&A from the annual reports of listed companies is as follows:

1) Convert the text format of the annual report. In order to facilitate the extraction of the text of the annual report, the need to PDF format text into txt form of text, and only the first 100 pages of the annual report to extract the text content.

2) Regular matching of the target text content. The use of regular expressions and manual screening methods to extract the annual report in the MD&A part of the relevant text. For the use of regular expressions caused by the “wrong” extraction of the text judgment is as follows: first of all, record the location of each paragraph of the MD&A text in the annual report, in the text of the annual report of the more forward marking the position of the manual review. Secondly, the ratio of the extracted text length to the whole text of the annual report is calculated, and the text with higher ratio is checked manually. The combination of regular extraction and manual checking is used to extract the MD&A text content of the sample, and finally get the “clean” MD&A text information to be processed.

3) The MD&A text information extracted in step (1) is cleaned, mainly dealing with the header and footer of the text and form data, and deleting the content of the “annual report”. For the footer content, delete only a number or contain “xx page” content. For the table data, if the number accounts for a high proportion of the contents of a line, it means that the line is a line in the table, but also delete the contents of these lines.

4) Segmentation of MD&A text content. Step (2) after cleaning the text can not be used directly to analyze, need to carry out the word processing to build text indicators, this paper uses Python comes with the jieba library to the text information for word processing. Because the annual report contains many financial accounting field proper nouns, and at present there is no more authoritative, complete financial accounting field-specific lexicon. The use of generalized lexical dictionaries will lead to the original is a word but due to the wrong cut into more than one word, which in turn affects the number of real contextual emotional words in the text, resulting in the composition of unreasonable text emotional indicators. In view of this, this paper expands the names of listed companies, proper nouns in the field of financial accounting, place names and other proper nouns into the generalized participle lexicon to improve the accuracy of MD&A text slicing.

The text preprocessing process is shown in Figure 1.

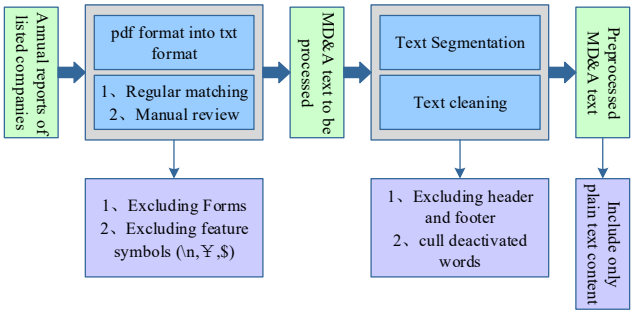

Figure 1: Text preprocessing process

II. B.TF-IDF

Word Frequency-Inverse Document Frequency (TF-IDF) is a common weighting technique used in data mining for information critical information retrieval, which is usually used to assess the importance of a word in a text set for the whole corpus or the importance of a document in a document set.TF is used to represent word frequency, and IDF is used to represent the inverse text frequency index of a word.In the dataset, the importance of the word token is usually is proportional to the number of times it appears in a document, and will be inversely proportional to the frequency of its appearance in the corpus, i.e., if a word token appears more often in a document and not in other documents, it means that the word is important for the classification of that document, and if the number of times it appears in other documents is also more, it means that the word is not very distinguishing, and the IDF is used to reduce the weight of the word .

Word frequency refers to the frequency of a given word token in the document, that is, the number of times the word appears in the document / the total number of words in the document, the TF value is usually normalized to prevent it from bias longer text, regardless of the importance of the word, to the same word in the longer text will usually have a higher probability of occurrence than in the shorter documents. However, it should be noted that some generic words such as deactivated words are not very useful for classifying text, and it is usually the case that words that occur less frequently in the dataset but more frequently in a particular document tend to accurately express the theme of the text, and thus the use of TF alone is not appropriate. The design of weights must satisfy: the stronger the ability of a word to predict the theme, the greater the weight, and vice versa, the smaller the weight, in all the statistical articles, some words appear in only a few of them, then such a word on the theme of the article is very useful, the weight of these words should be designed to be larger, the inverse document frequency IDF is accomplishing such a job as in equation (1):

$$IDF = \log\left(\frac{\text{Total number of documents in corpus plus}}{\text{Number of documents containing the lemma } \omega + 1}\right) \tag{1}$$

The high frequency of a word token in a document and the low frequency of that token in the whole dataset documents can produce a highly weighted TF-IDF, therefore, the TF-IDF tends to tend to filter out the common words and keep the important words as in equation (2):

$$TF - IDF = TF * IDF \tag{2}$$

II. C.N-Gram Model

The basic idea of N-Gram, as a classical statistically based language modeling algorithm, is to statistically predict the contents of a document according to a sliding window of length N to form a sequence of byte fragments of length N, where each byte fragment is called a gram.

In the statistical process, the frequency of occurrence of all grams is counted. The grant title filters the text by a fixed window size according to a set hyperparameter N, forming a list of grams to represent the vector feature space of that text, and each gram item in the list is a feature vector dimension.The N-Gram is based on the assumption that in a given gram, the occurrence factor of the first N word is only related to the preceding $N-1$ words and not with any other word, and the probability of the entire sentence combination is the product of the probabilities of the individual word occurrences. The probability of occurrence of each gram can be computed in the whole corpus by the list of grams. And then based on the given textual information, the next most likely word occurrence is predicted. $N=1$ is called uni-gram, which indicates that the occurrence of the next word does not depend on any of the previous words. $N=2$ is called bi-gram, which indicates that the next word depends only on one of the immediately preceding words, and so on, and the complete sentence is obtained as in equation (3):

$$P(\omega_1,\omega_2,\ldots,\omega_n)=P(\omega_1)\cdot P(\omega_2\mid\omega_1)\cdots\cdots P(\omega_n\mid\omega_1,\omega_2,\ldots,\omega_{n-1}) \tag{3}$$

III. Improvements to the Sentiment Dictionary for Financial Reporting

After the preprocessing of the annual financial report text in Chapter 2 and the vectorized representation of the text, research data that can be input into the sentiment dictionary for analysis are obtained. In preparation for this research, this chapter expands the generic sentiment dictionary using an improved sentiment co-occurrence algorithm to construct a sentiment dictionary for the financial reporting domain, and utilizes the SEN-TF-IDF algorithm to perform sentiment annotation in the annual financial report sentiment data.

III. A. Improved sentiment co-occurrence algorithm dictionary expansion

In this paper, we use an improved sentiment co-occurrence algorithm to expand the sentiment lexicon. Point Mutual Information PMI algorithm is an important concept in information theory, this method measures the degree of association between two words, two words $word_1$ and $word_2$ in information theory PMI value is calculated as equation (4):

$$PMI(word_1,word_2)=\log_2\left[\frac{p(word_1,word_2)}{p(word_1,word_2)}\right] \tag{4}$$

where $word_1$ and $word_2$ represent two words, and $p(word_1)$ and $p(word_2)$ represent the probabilities of occurrence of $word_1$ and $word_2$ occurrence probabilities in the target text. $p(word_1,word_2)$ represents the probability of $word_1$ and word_2 appearing together. The probability values are derived by counting the number of times $word_1$ and $word_2$ appear together in the selected text range. The ratio of $p(word_1,word_2)$ to $p(word_1)p(word_2)$ is the $word_1$ and $word_2$ word The statistical measure of independence, $PMI(word_1,word_2)$, results from taking the logarithm of their ratio. The value of $PMI(word_1,word_2)$ represents the probability of the two words occurring together in the selected range, which represents the degree of co-occurrence of $word_1$ and $word_2$, and the smaller the value is, the greater the independence of the two words. The smaller the value, the greater the independence of the two words, the smaller the degree of association, the less frequent the co-occurrence, and conversely, the larger the value, the less independent the two words are, the greater the degree of association, the more frequent the co-occurrence.

The main idea of the above algorithm is based on word frequency and does not consider the relative distance between words. In order to improve the above defects, this algorithm is improved in this paper by introducing the concept of word spacing for sentiment word expansion. When two words co-occur, they may be adjacent to each other or they may be far away from each other. It is assumed that if the relative distance between two co-occurring words is considered to be close, the stronger the correlation will be, and vice versa, the weaker it will be. The number of words spaced between two words is used to indicate the relative distance between the two words, if its relative distance has more than one value, the value with the smallest absolute value is taken as its spacing distance. The formula for calculating the relative distance between words is shown in equation (5):

$$d=\min\left|d_{word_1}-d_{word_2}\right| \tag{5}$$

where d_{word_1} denotes the length (number) of words in the corpus from the beginning of the corpus to the position of the earlier of the two words, and d_{word_2} denotes the length (number) of words in the corpus from the beginning of the corpus to the position of the later of the two words. Then the updated PMI is calculated as in equation (6):

$$PMI(word_1,word_2)=\log_2\left[\frac{d\cdot p(word_1,word_2)}{N\cdot p(word_1)p(word_2)}\right] \tag{6}$$

where N denotes the total number of times all words in the corpus are used.

In view of this, in order to determine the sentiment orientation (SO) of the words, several existing positive sentiment words and negative sentiment words in the sentiment dictionary are selected as seed words, which are applied to the point mutual information algorithm, and the formula for the sentiment orientation is shown in Equation (7):

$$SO(\alpha)=PMI(\alpha,\alpha^+)-PMI(\alpha,\alpha^-) \tag{7}$$

where α is the word to determine the emotional tendency, α^+ and α^- represent the seed words with positive and negative emotions, respectively, if the emotional tendency SO value is greater than the threshold, it means that the word to be determined is more relevant to the positive word, and its probability of being a positive word, and vice versa for the probability of being a negative word is greater, so as to determine the emotional polarity of the word.

The basic idea of the improved sentiment co-occurrence algorithm is: firstly, a group of words with positive sentiment tendency and a group of words with negative sentiment tendency are selected as the base words, and the two groups of words are represented by posword and negword respectively. The emotion tendency of the emotion words in the selected two groups of words should be representative and the emotion tendency of the words in the target domain is relatively stable. The difference obtained by subtracting the point-to-point mutual information of word and posword from the point-to-point mutual information of word and negword to be determined and comparing it with 0, the emotional tendency of the target word can be judged accordingly. The formula is shown in equation (8):

$$SO-PMI(word)=\sum_{posword\in poset}PMI(word,posword)-\sum_{negword\in negset}PMI(word,negword) \tag{8}$$

Using this algorithm in the expansion of the emotion vocabulary, the text of the annual report after text pre-processing to obtain a number of words, these words and the fusion of the basic emotion dictionary of the vocabulary for comparison, the same words will be screened out to generate the pre-expansion lexicon, in the lexicon to select a number of representative of the negative and positive vocabulary as a benchmark word. The vocabulary obtained after text preprocessing and the selected benchmark words are used to calculate the SO-PMI values in turn, and a number of words whose absolute value of SO-PMI is higher than the set threshold are incorporated into the emotion dictionary according to the emotional tendency, so that the purpose of expanding the emotion dictionary can be achieved.

III. B. Sentiment labeling of datasets based on SEN-TF-IDF weighting

In this paper, the weight of sentiment positive words is set to 1 and the weight of sentiment negative words is set to -1. Since the text in annual reports is more formal compared to ordinary spoken language, the weight of degree adverbs is not considered in the sentiment analysis.

To calculate the sentiment value of a single annual report, the first step is to compare the annual report sentiment dictionary to extract the sentiment words contained in the annual report, and then multiply the sentiment weights of the sentiment words with the TF-IDF weights of the words to get the SEN-TF-IDF weights of each word, and then the sentiment value of the annual report can be obtained by summing up the SEN-TF-IDF weights of all the sentiment words of the annual report as in Equation (9):

$$emo_{word}=SEN-TF-IDF=w_i\times q_i \tag{9}$$

where w_i 1 represents the TF-IDF weights of sentiment words and q_i 2 represents the sentiment weights of sentiment words. The sentiment value of the annual report text is the sum of the sentiment weights of all its sentiment words, calculated as in equation (10):

$$emo_{essay}=sum(emo_{word}) \tag{10}$$

When $emo_{essay}\geq 0$, i.e., the annual report text sentiment value is greater than or equal to 0, it is judged to be a positive sentiment. When $emo_{essay}<0$, i.e., the annual report text sentiment value is less than 0, it is judged as negative sentiment.

Based on the above rules, the steps of the algorithm to calculate the annual report sentiment value are as follows:

Step 1: Take all the annual report texts as a training sample, compare the sentiment dictionary, extract all the sentiment words in it, and calculate the weight w_i of each sentiment word using the TF-IDF algorithm.

Step 2: Compare the annual report sentiment dictionary to extract the sentiment words contained in the annual report, for the sentiment words in the annual report, it is recorded as the set $Word\{word_1,word_2,word_3,\ldots,word_i\}$.

Step 3: The words in the set $Word$ are weighted and summed to obtain the sentiment value of the annual report.

Step 4: Label the annual report sentiment, calculate the annual report with a sentiment value greater than or equal to 0 to label it as positive, and with a sentiment value less than 0 to label it as negative. Annual reports with positive sentiment are labeled with 1 and annual reports with negative sentiment are labeled with 0.

Based on the above methodology, the annual report sentiment dataset was derived. The total number of data is 26,610, of which 15,665 are positive and 10,945 are negative. The annotation process of annual report sentiment dataset is shown in Figure 2.

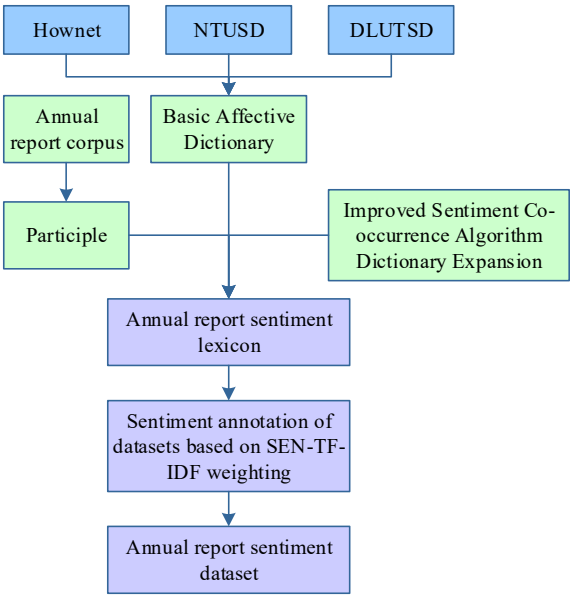

Figure 2: Annual report emotional data set annotation process

IV. Test and Analysis of the Sentiment Lexicon for Financial Reporting

This chapter unfolds the examination of the sentiment lexicon of financial annual reports in the form of analyzing the notes to the financial reports, comparing generalized algorithms, and applying an assessment.

IV. A. Analysis of notes to the financial report

In this section, the financial reports of A-share listed companies are selected as samples, and the financial reports are textualized and entered into the financial report database through manual processing. The sample industry in this section is airlines, and the sample source for a single report is the 2020 annual financial report of China E Airlines, and the industry analysis is the 2020 annual financial statements of 15 companies in China's aviation industry.

IV. A. 1) Formation of High Frequency Words after Segmentation

Segmentation is the first step of text mining analysis, and it is also a key step to decompose unstructured text data into vocabulary that can be used for statistics and analysis. Before word separation, the first step is to set the word list, that is, the content to be analyzed in accordance with certain standards for word separation, financial reports are different from general text, with a very large number of accounting terminology, and according to different reporting subjects, some of the proper nouns inside the special definition. Set up a customized glossary of accounting terms, on the one hand, including the accounting term “accounting phasing”, “financial instruments”, etc., on the other hand, including the company's proper nouns, i.e., the company's full name, the company's abbreviation, etc., but also including “2017”, etc., as a whole. The term “fiscal year” has a special meaning as a whole.

In the case of a sentence such as “Statement of Compliance with Accounting Standards”, for example, the result is “Statement of Compliance with Accounting Standards” after the word separation process. The analyzed text is presented in the form of individual words and phrases. A word frequency table for financial reporting, i.e., a count of the number of occurrences of different words, can be obtained by counting the words. Due to space constraints, this paper extracts the top 30 words with the highest number of occurrences in the word frequency table, and compiles the high-frequency vocabulary of financial reporting disclosure texts in Table 1. It was observed that the keyword "Group" appeared most frequently. The second is "Assets", "E-Airlines", "Recognition", "Fair Value", "Amount", "Measurement", etc. The four assumptions of accounting are the premise of accounting, which is well reflected in the word frequency statistics, the accounting subject "the group" is the word that appears the most, followed by the company abbreviation "E Airlines", and the frequency of keywords related to "amount" and monetary measurement is also in the forefront, and accounting "confirmation" and "measurement" are the methods of financial information recording. Word frequency statistics and analysis is the simplest step of text mining, which can directly reflect the key information of the mined text content and the frequency of key information appearing,

and its disadvantage is that it can only observe the independent key information, but not the connection between them.

Table 1: Note high-frequency words

Serial number	Keyword	Frequency
1	This group	270
2	Property	168
3	E Airlines	167
4	Affirm	153
5	Fair value	150
6	Amount of money	133
7	Metering	128
8	Project	114
9	Financial assets	113
10	Subsidiary corporation	111
11	Reckon in	105
12	This company	102
13	RMB	99
14	Shanghai	86
15	Aviation	85
16	Current profit and loss	85
17	Income	85
18	Provision	84
19	Aircraft	83
20	Investment	82
21	This year	81
22	Liabilities	79
23	According to	74
24	Deal with	73
25	Consolidated financial statemen	71
26	On the basis of	70
27	Book value	69
28	Cost	69
29	Reserve for bad-debt	68
30	Service	68

IV. A. 2) Co-occurrence analysis

Co-occurrence refers to the phenomenon that different keywords obtained from word splitting in the mined text appear together, and co-occurrence analysis can be realized to compare a set of keywords directly related, i.e., whether there is a connection between two keywords and which two keywords are related to each other. Whether there is a connection can be measured by whether there is a co-occurrence frequency, and the magnitude of the degree of association can be measured by the magnitude of the co-occurrence frequency. The text after word separation is imported into the ROST CM software, which can construct a word list of co-occurrence matrix for all the associated words, and only some keywords are extracted in this paper, presenting the 11×11 co-occurrence matrix of keywords of financial report disclosure information in Fig. 3. The keywords extracted are (W1) Group, (W2) assets, (W3) Air E, (W4) recognition, (W5) fair value, (W6) amount, (W7) measurement, (W8) items, (W9) financial assets, (W10) subsidiaries, and (W11) crediting.

It was observed that the keywords “assets” and “recognition” appeared together 44 times, indicating that the disclosure focused on the recognition of assets. The keyword “subsidiaries” appears only in conjunction with “Air E”, indicating that the disclosed information on subsidiaries is centered around the main body of the company. Among the 11 keywords, the two keywords “recognized” and “credited” are the most frequent words co-occurring with other keywords, each co-occurring with seven words, indicating that the disclosed items and their amounts revolve around how the amounts are recognized and what items are credited.

Co-occurrence analysis is two-dimensional compared to high-frequency word analysis, i.e., high-frequency word analysis can only observe the number of occurrences of the key items of the financial report disclosure, whereas co-occurrence analysis can show the link between the two key words of the disclosure. A major disadvantage of co-occurrence analysis is that it can only analyze the direct linkage of a set of keywords, i.e., it can only compare the direct linkage between two keywords and cannot see the linkage between multiple keywords as well as the indirect linkage between keywords.

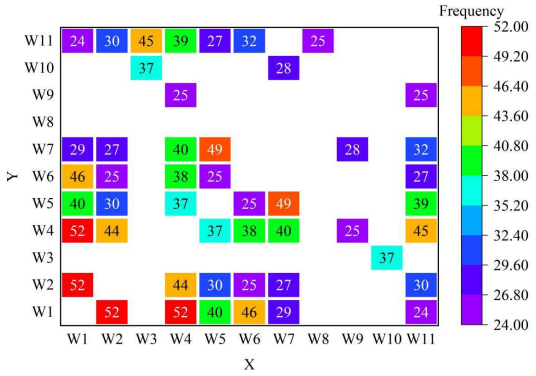

Figure 3: Co-occurrence matrix of auxiliary high-frequency words

IV. A. 3) Analysis of industry statements

The 12×12 co-occurrence matrix is shown in Figure 4. Compared with the high-frequency words, the new words (W12) Company, (W12) Unit, (W14) Aviation, (W15) Year, (W16) Balance are added to the industry report. It can be found that “amount” reflecting the presentation figures of the financial report is still the word with a high frequency of co-occurrence with the keywords, indicating that the financial report of the aviation industry as a whole to the statement of the amount of each item presented in the statement of the reasonable recognition of the key elements of disclosure. "Assets" are the key items presented in the financial report, and the way in which the superficial assets such as "recognition" and "accounting" are recorded in the financial report.

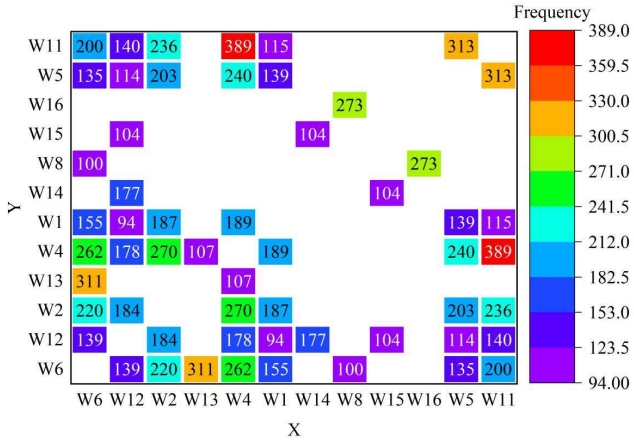

Figure 4: High frequency word co-occurrence matrix of industry report notes

IV. B. Performance evaluation results and analysis

In order to verify the classification effect of the sentiment dictionary in this paper, 300 annual reports of listed companies are randomly selected, and the sentiment tendency value of MD&A part is calculated. The analysis method of annual report sentiment dictionary constructed in this paper (T4), as well as the analysis method of basic L&M sentiment dictionary (T5), convolutional neural network-based sentiment dictionary (T3), deep learning-based sentiment dictionary (T2), and semantic rule-based sentiment dictionary (T1) are compared to evaluate the effectiveness of the sentiment dictionary. The comparative analysis can verify the effectiveness of the sentiment dictionary analysis method (T5) in this paper, and can verify the effectiveness of using this sentiment dictionary (T5) for financial risk identification.

The sentiment value of the whole annual report is calculated, and in order to find a threshold value applicable to differentiate the sentiment value of positive and negative annual reports, the classification effect for the text of the annual report is calculated for different thresholds, and the changes in the F1 value of the evaluation indexes are shown in Fig. 5.

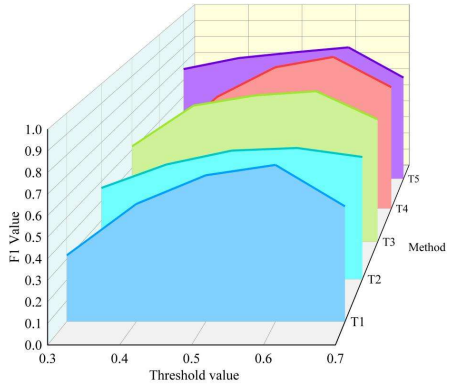

Figure 5: The evaluation index changes under different thresholds

The X coordinate in the figure represents the threshold value, the Y coordinate represents the evaluation index F1 value, and the Z coordinate represents the 5 sentiment dictionary methods. It can be found that the five methods gradually become better at categorizing as the threshold value gradually increases. When the threshold value is lower than 0.5, the classification effect of emotion dictionary based on convolutional neural network (T3) is better and the F1 value is 0.740, but when the threshold value is greater than 0.5, the classification of specialized emotion dictionary constructed based on this paper (T4) is better. When the threshold value is selected as 0.6, the F1 value of all five methods reaches the highest value. Therefore, the evaluation threshold of the emotion dictionary research method is determined to be 0.6, and the F1 values of the six methods under this threshold are shown in Figure 6.

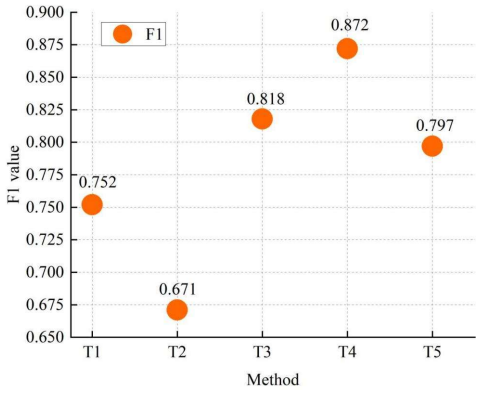

Figure 6: F1 results for 5 methods at the 0.6 threshold

Where the horizontal coordinate represents the research method and the vertical coordinate represents the evaluation index F1 value. As can be seen from Figure 6, based on the dedicated sentiment dictionary constructed in this paper (T4) financial risk identification is improved on the indicators, the classification results of the F1 value is the highest than the rest of the four sentiment dictionary methods, 0.872. It shows that the sentiment dictionary established in this paper can improve the effect of sentiment analysis of the text of the annual report of listed companies, which is more suitable for the identification of financial risk. Compared with the selection of existing publicly available sentiment dictionaries in other scholars' studies, this paper uses the method of constructing a sentiment dictionary that combines the textual information of the annual report to obtain a more accurate representation of the textual sentiment, which effectively improves the prediction ability of the financial risk company.

IV. C. Application experiments and analysis

IV. C. 1) Analysis of results of indicator selection

The LDA method was used to obtain the financial data information of Group D. The results of determining the set of financial risk information evaluation indexes realized by analyzing it are shown in Table 2. 16 financial risk information evaluation indexes were obtained by statistically and analytically determining the financial risk influencing factors of the group companies.

Table 2: Financial risk information evaluation index selection results

Index	Measured variable
Asset-liability ratio	X1
Liquidity ratio	X2
Quick ratio	X3
Cash ratio	X4
Return on assets	X5
Net assets income rate	X6
Operating margin	X7
Ratio of profits to cost	X8
Retained earnings to assets ratio	X9
Total assets turnover	X10
Current asset turnover	X11
Inventory turnover ratio	X12
Turnover of account receivable	X13
Growth rate of total assets	X14
Capital accumulation rate	X15
Increase rate of business revenue	X16

IV. C. 2) Assessment of financial risk information

The financial risk information of ten subsidiaries (D1, D2, D3, D4, D5, D6, D7, D8, D9, D10) of Group D in 2020-2022 is identified by utilizing the financial reporting sentiment dictionary. The identification performance of the financial reporting sentiment dictionary is verified by analyzing the financial risk information assessment results, and the experimental results of the financial risk information assessment are shown in Fig. 7. where the risk scoring interval is [-1,8], and the lower the scoring value, the higher the risk, and the risk level is classified as ultra-high risk (R1), high-risk (R2), medium-risk (R3), and low-risk (R4).

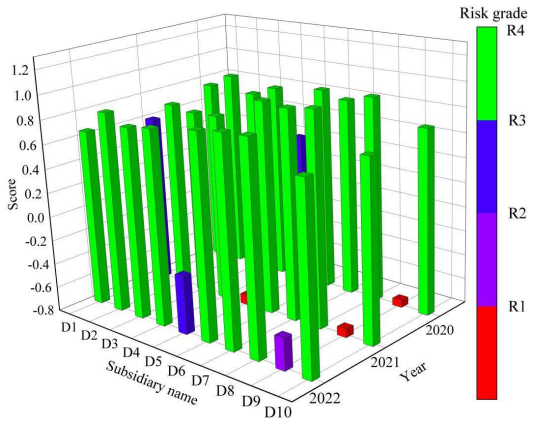

Figure 7: Financial risk information assessment results

Observing Figure 7, the financial risk information of 10 subsidiaries of Group D in the past 3 years was identified by using the sentiment dictionary algorithm in this paper, and the financial risk information assessment result of D5 company in 2020 was 0.44 at "medium risk", and the financial risk of D9 company was high, at -0.74, reaching the "ultra-high risk" level. In 2021, the financial risk of D5 companies intensified, with an assessment of -0.72, reaching the "high risk" level, while the financial risk of D9 companies was reduced by one level. D1 companies have hidden financial risks, and their scoring results have decreased, from 0.75 to 0.66, approaching the critical value of

medium and low risk, and it is necessary to strengthen the awareness of risk prevention. In 2022, the financial status of D1 and D5 companies continued to deteriorate, rising to the "medium risk" and "ultra-high risk" levels, respectively, while the financial risk information assessment scores of D9 companies increased, but they still did not get out of the "high risk" level, and the financial status of the remaining subsidiaries was better and was at the low risk level. Experimental results show that the sentiment dictionary in this paper can realize the automatic identification of financial risk information.

V. Conclusion

This paper designs an exclusive sentiment dictionary for financial annual reports based on an improved sentiment co-occurrence algorithm. The sentiment dictionary algorithm is based on natural language processing technology, and by extracting and mining the textual vocabulary features of the company's annual financial reports, it analyzes the sentiment tendency embedded in them, so as to effectively identify the financial risk information of listed companies.

The highest F1 value of the financial annual report sentiment dictionary under the evaluation threshold is 0.872, which is far more than similar general sentiment dictionary algorithms. In the analysis of a listed company's annual reports for three consecutive years, it is able to accurately identify the changes in the financial risks of its subsidiaries in different years.

References

[1] Easton, P. (2016). Financial reporting: An enterprise operations perspective. Journal of Financial Reporting, 1(1), 143-151.

[2] Kuprina, N., & Volodina, O. (2019). Features of the analysis of financial results of the enterprise activities in modern conditions. Food Industry Economics, 11(3).

[3] Alduais, F. (2024). Textual analysis of the annual report and corporate performance: evidence from China. Journal of Financial Reporting and Accounting, 22(5), 1221-1252.

[4] Subramanian, D., Bhattachrajya, D., Torrado, R. R., Kephart, J., Chenthamarakshan, V., & Rios, J. (2017, December). A cognitive assistant for risk identification and modeling. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 1570-1579). IEEE.

[5] Li, C., Liu, Q., & Huang, L. (2021). Credit risk management of scientific and technological enterprises based on text mining. Enterprise Information Systems, 15(6), 851-867.

[6] Gardi, B., Hamza, P. A., Sabir, B. Y., Aziz, H. M., Sorguli, S., Abdullah, N. N., & Al-Kake, F. R. A. (2021). Investigating the effects of financial accounting reports on managerial decision making in small and medium-sized enterprises. Turkish Journal of Computer and Mathematics Education, 12(10), 2134-2142.

[7] Alfiah, A., Bakri, A. A., Fatimah, F., Syahdan, R., & Rusman, H. (2023). Capability to manage financial reports for MSMEs utilizing accounting information systems. Jurnal Ekonomi, 12(02), 1356-1363.

[8] Renaldo, N., & Putri, N. Y. (2021). ACCOUNTING INFORMATION SYSTEMS INCREASE MSMEs PERFORMANCE. Journal of Applied Business and Technology, 2(3), 261-270.

[9] Wei, R., & Yao, S. (2021). Enterprise financial risk identification and information security management and control in big data environment. Mobile Information Systems, 2021(1), 7188327.

[10] Wei, L., Deng, Y., Huang, J., Han, C., & Jing, Z. (2022). Identification and analysis of financial technology risk factors based on textual risk disclosures. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 590-612.

[11] Wang, Y., Li, G., Li, J., & Zhu, X. (2018). Comprehensive identification of operational risk factors based on textual risk disclosures. Procedia computer science, 139, 136-143.

[12] Smailović, J., Žnidaršič, M., Valentinčič, A., Lončarski, I., Pahor, M., Martins, P. T., & Pollak, S. (2017). Automatic analysis of annual financial reports: A case study. Computación y Sistemas, 21(4), 809-818.

[13] Zhang, Z., Luo, M., Hu, Z., & Niu, H. (2022). Textual emotional tone and financial crisis identification in Chinese companies: A multi-source data analysis based on machine learning. Applied Sciences, 12(13), 6662.

[14] Ding, P., Zhuoqian, L., & Yuan, D. (2019, December). Text Analysis of Enterprise Financial Report Based on Semantic Perception. In Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City (pp. 59-64).

[15] Liu, W., & Le, W. (2023, November). Enterprise Digital Transformation Risk Identification and Dynamic Assessment Model under Big Data Perspective. In Proceedings of the 2023 3rd International Conference on Big Data, Artificial Intelligence and Risk Management (pp. 714-719).

[16] Cohen, J., Krishnamoorthy, G., & Wright, A. (2017). Enterprise risk management and the financial reporting process: The experiences of audit committee members, CFO s, and external auditors. Contemporary Accounting Research, 34(2), 1178-1209.

[17] Madu, M., & Hassan, S. U. (2021). Enterprise risk management and financial reporting quality: Evidence from listed Nigerian non-financial firms. Journal of Risk and Financial Studies, 2(1), 43-70.

[18] Gao, S., Hsu, H. T., & Liu, F. C. (2025). Enterprise Risk Management, Financial Reporting and Firm Operations. Risks, 13(3), 48.

[19] Zhu, X., Wang, Y., & Li, J. (2022). What drives reputational risk? Evidence from textual risk disclosures in financial statements. Humanities and Social Sciences Communications, 9(1), 1-15.

[20] Alduais, F., Ali Almasria, N., Samara, A., & Masadeh, A. (2022). Conciseness, financial disclosure, and market reaction: A textual analysis of annual reports in listed Chinese companies. International Journal of Financial Studies, 10(4), 104.

[21] Lewis, C., & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. Accounting and Business Research, 49(5), 587-615.

[22] Wujec, M. (2021). Analysis of the financial information contained in the texts of current reports: A deep learning approach. Journal of Risk and Financial Management, 14(12), 582.

[23] Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. Decision Support Systems, 94, 53-64.

[24] Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. Sustainability, 11(5), 1277.
[25] Wei, L., Li, G., Zhu, X., & Li, J. (2019). Discovering bank risk factors from financial statements based on a new semi‐supervised text mining algorithm. Accounting & Finance, 59(3), 1519-1552.
[26] Hsu, M. F., Chang, C., & Zeng, J. H. (2022). Automated text mining process for corporate risk analysis and management. Risk Management, 24(4), 386-419.