ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20230629000774
Original scientific paper

Construction of an Inquiry Letter Sentiment Dictionary Using *SO-PMI* and Word2Vec for Sentiment Analysis

Wei WANG, Guiying WEI, Sen WU*, Huixia HE

Abstract: Sentiment analysis of financial texts provides valuable insights into market dynamics, but relies on domain-specific dictionaries. Inquiry letters as particular financial texts lack tailored sentiment dictionaries, limiting research in the important domain. This study develops an Inquiry Letter Sentiment Dictionary (ILSD) using innovative integration of semantic orientation pointwise mutual information (*SO-PMI*), word2vec, and manual screening. The ILSD leverages co-occurrence and contextual information to expand coverage and sentiment word capture capabilities beyond existing general and financial dictionaries. We collected 1754 inquiry letters from the Shanghai Stock Exchange (SSE). The experiment results indicate that the ILSD performs better than other sentiment dictionaries according to coverage, accuracy (*ACC*), *F*1 weighted score ($F1_{Weighted}$), Matthews correlation (*MCC*), and geometric mean (*G*-mean), which proves the effectiveness of the ILSD in practice.

Keywords: inquiry letter; sentiment classification; sentiment dictionary; *SO-PMI*; word2vec

1 INTRODUCTION

Under the background of continuous strengthening of front-line supervision in the securities market, inquiry letters have gradually become an important tool for non-punitive supervision of exchanges [1]. When a listed company is considered to have deficiencies in information disclosure, or has doubts about major matters such as the company's merger and reorganization or other matters in the company's business activities, the exchange will send an inquiry letter to the company, requiring the listed company to reply and disclose the relevant information of the received letter and reply letter in a timely manner. Naturally, the inquiry letter also contains sentiment information. By analyzing the sentiment tendency of the inquiry letter, we can understand the regulatory attitude and intention, and thus gain insight into the potential risks of the listed company. This will provide a reference for market participants such as regulators, listed companies and investors to make decisions. In recent years, with the development of text mining technology, financial text sentiment has attracted great attention from academia and industry [2]. This is because the sentiment in financial texts is the inner sentiment expression of managers or market participants towards relevant scene events, which is a supplement to potential risk information to some extent [3]. Therefore, some researchers extract financial text sentiment indicators and apply them to stock price trend prediction, financial distress warnings, and so on. Better prediction results were obtained compared with traditional methods [4, 5]. Generally speaking, financial text sentiment analysis mainly includes the sentiment dictionary method and the machine learning method [6]. The sentiment analysis method based on machine learning requires existing sentiment dictionary and a large number of corpora labels for training, which is difficult for financial texts. On the contrary, the method based on the sentiment dictionary only needs to calculate the sentiment score of the text through the sentiment dictionary to complete the sentiment analysis of the text. In short, the method of the sentiment dictionary is simple, efficient, and does not rely on corpus labels [7]. At present, there are mainly two research ideas on sentiment dictionary construction. (1) Methods based on knowledge bases. It is often used to construct a general sentiment dictionary (GSD), which is constructed by experts using the existing knowledge base and manually screening according to semantic understanding. Although the general sentiment dictionaries have great versatility, its accuracy is relatively low in the financial field because it cannot effectively cover the professional financial words. In addition, the general sentiment dictionaries have the disadvantage of not updating in time. (2) Methods based on corpus bases. The domain sentiment dictionary is mainly constructed by the co-occurrence (or similarity) of seed sentiment words and words in the corpus, such as Financial Sentiment Dictionary (FSD). The sentiment words in the domain sentiment dictionary can be updated with the corpus. The domain sentiment dictionary can improve the accuracy of sentiment analysis in the field, but it depends on the corpus. The inquiry letter belongs to a typical type of official document text under the financial text. There are some differences between the inquiry letter and the general financial text. Specifically, the inquiry letter has the characteristics of typical official document text: more professional terms, fewer sentiment words, and relatively long text content. It can be seen that the inquiry letter has a certain field specificity, which causes many difficulties in the extraction of sentiment words. Some words are not even sentiment words in many domains, but they are sentiment words with an obvious sentiment tendency in the inquiry letters domain. For example, words like excess profit, new materials, and shortage reflect positive sentiment tendency. But words like decline and risk reflect negative sentiment tendency. We will not be able to achieve if we use the GSD or the FSD directly to measure the sentiment of the inquiry letters. In other words, the sentiment dictionary is sensitive to the sentiment words in the professional field, and only the construction of the sentiment dictionary in the specific domain can improve accuracy [8, 9]. As far as we are aware, the current research lacks the domain sentiment dictionary tailored to inquiry letters. Therefore, constructing an inquiry letter sentiment dictionary (ILSD) is significant and is an urgent problem to be solved in the research on the sentiment of inquiry letters. This paper attempts to fill this research gap and provide support for further research on the sentiment analysis of inquiry letters. We developed the ILSD using innovative integration of semantic orientation pointwise mutual information (*SO-PMI*) [10, 11], word2vec [12], and manual screening. We construct ILSD by integrating methods based on both knowledge bases and corpus bases. The purpose is to make ILSD have better accuracy, comprehensiveness and timeliness, so that it can deal with various sentiment analysis tasks in the field of inquiry letters. The accuracy of the proposed ILSD is compared and evaluated by several experiments and classification evaluation performance metrics. This study mainly answers the question of whether the ILSD improves the sentiment word recognition ability of inquiry letters. The experimental results show that the ILSD is evaluated on inquiry letter data and has been shown to improve sentiment analysis performance compared with the benchmark sentiment dictionaries. The ILSD's effectiveness and excellent adaptability of sentiment word recognition in inquiry letters are fully demonstrated. The contributions and innovations of this paper are mainly reflected in the methodological innovation and the unique data sample. (1) We present the ILSD that has been constructed using *SO-PMI*, word2vec, and manual screening. Furthermore, this paper's results may offer a fresh viewpoint for inquiry letters research in academia. (2) We found that the ILSD had superior domain sentiment word coverage after reconstructing the GSD, the FSD, and the Basic Inquiry Letter Sentiment Dictionary (BILSD). (3) The coverage and the four performance evaluation metrics were carried out to confirm the ILSD's validity. The experimental results indicated that the ILSD outperformed the benchmark sentiment dictionaries in terms of its capacity to extract and capture sentiment words. It is helpful to provide some reference for research in other related fields. The remainder of this paper is organized as follows: Section 2 introduces the literature review. Section 3 introduces the methods and framework of the study. Section 4 presents the results of this study. Section 5 presents the conclusions and future perspectives of the study.

2 LITERATURE REVIEW

2.1 Literature Review of Sentiment Dictionaries

In the sentiment analysis of Chinese text, words are the smallest units of analysis, and Chinese words contain three sentiment situations: positive, negative, and neutral [13]. The sentiment dictionary can quickly extract and identify the sentiment words of Chinese text, which is an indispensable tool for sentiment analysis [14]. The GSD is commonly used in current Chinese text sentiment analysis. Three representative dictionaries of the GSD are: HowNet Sentiment Dictionary (HowNetSD) [15], National Taiwan University Sentiment Dictionary (NTUSD) [16], and Tsinghua University Sentiment Dictionary (TUSD) [17]. These three sentiment dictionaries contain positive and negative words, which are mainly obtained from literary works, media news, and other texts. The sentiment analysis based on the sentiment dictionary mainly relies on the sentiment words in the text to measure sentiment. However, the GSD above have some shortcomings, such as a lack of field sentiment words and polysemy sentiment words. For example, the application effect of the GSD in the field of finance does not work well. The relevant scholars have constructed some financial sentiment dictionaries. Loughran and McDonald found that about 70% of the negative words in the GSD was not applicable to the analysis of financial documents, so they extracted high-frequency words from the annual reports of listed companies and constructed the Loughran and McDonald Financial Sentiment Dictionary (LMFSD) through manual screening [18]. The LMFSD has been widely used in the research on sentiment analysis of relevant English financial texts [19]. The LMFSD has also been widely used in the sentiment measurement of Chinese financial texts after being translated by Chinese scholars. However, it has been found that some sentiment words in LMFSD are different from Chinese sentiment words in practice. Therefore, when using LMFSD to conduct sentiment analysis on Chinese financial texts, some researchers will conduct manual screening and translation considering the Chinese context to improve the effect of sentiment analysis [20]. Considering the specificity of terms in the field of finance and economics in China, some researchers have constructed the Financial Sentiment Dictionaries by restructuring the finance and economics words in the GSD and the LMFSD (translation) that are in line with Chinese situations. In the current stage, more and more financial sentiment dictionaries have been constructed, most of which are based on the existing sentiment dictionaries and construction techniques.

Table 1 Description of common sentiment dictionary information

Categories	Name	Time	Method	Resources	Number of negative words	Number of positive words
General Sentiment Dictionary (GSD)	NTUSD	2006	Manual label	News, web blog articles	8276	2812
	TUSD	2007	Dictionary reconstruction	Online reviews	4469	5567
	HowNetSD	2008	Measure semantic similarity	CNKI	4370	4566
Financial Sentiment Dictionary (FSD)	LMFSD	2011	Manual screening	10-K, MD&A	2080	1076
	BCFSD	2018	Word2vec	Prospectus, annual report	1488	1109
	YCFSD	2021	Dictionary reconstruction	Annual report	1633	3592
	JCFSD	2021	Dictionary reconstruction, LM translation, word2vec	The Infobank database	5890	3338

The Bian Shi Bo Chinese Financial Sentiment Dictionary (BCFSD), which was based on the prospectus and annual reports, was constructed by using word2vec. [21]. Jiang Fuwei Chinese Financial Sentiment Dictionary (JCFSD) was constructed based on the Infobank database, the annual reports of listed companies [22]. The sentiment dictionary has been successfully applied to the calculation of sentiment index of Chinese financial media. Yao

Jiaquan constructed Chinese Financial Sentiment Dictionary (YCFSD) by using dictionary restructuring and machine learning methods [23]. The annual report intonation index of listed companies is constructed by using the YCFSD. Furthermore, some technical indexes of stock prices were predicted based on the annual report tone index. The experimental results show that the YCFSD outperforms other sentiment dictionaries in terms of accuracy. The mentioned GSD and FSD have been widely used in the field of sentiment analysis. GSD was the basic tool of sentiment analysis, while FSD was constructed for the specific needs of the financial field, which has higher accuracy for financial market sentiment analysis. The GSD, the FSD mentioned above are summarized in Tab. 1 below. As can be seen from Tab. 1, GSD is commonly constructed by manual screening method, while FSD is considered to be constructed or expanded by calculating word similarity method (such as word2vec).

2.2 Literature Review of Sentiment Dictionary Construction Methods Based on Corpus Bases

There are many methods to construct sentiment dictionary based on corpus. However, researchers often use *SO-PMI* of word co-occurrence method or word2vec of word similarity method to construct sentiment dictionary. The literature related to sentiment dictionary construction using word2vec and *SO-PMI* will be reviewed in this section, respectively. Word sentiment polarity can be effectively determined using *SO-PMI* [24]. Turney et al. used Point Mutual Information (*PMI*) to classify a target word as positive or negative based on its correlation with the seed words. Some researchers have applied *SO-PMI* to the construction tasks of various sentiment dictionaries, and have achieved satisfactory accuracy [25]. Zhao et al. used *SO-PMI* to discriminate the sentiment polarity of subjective evaluations in TV programs [26]. Yang et al. constructed a Chinese sentiment dictionary suitable for the domain of hotel reviews based on *SO-PMI* [27]. Liu et al. used *SO-PMI* to judge the sentiment words not included in the current microblog and achieved good performance [28]. In a word, *SO-PMI* is a method to construct sentiment dictionaries via word co-occurrence. However, it relies on the quality of seed words and does not consider contextual semantics. Since word2vec was proposed by Mikolov in 2013, it can capture the semantic relationships of the context and has been used in tasks such as text sentiment analysis. [29]. Researchers chose word2vec to train models and generate useful word vector representations for text sentiment analysis tasks [30]. The word2vec gives the generated vector rich meaning information, which provides convenience for calculating the similarity between words [31-33]. The similarity between sentiment words can be quantitatively measured by cosine similarity [34]. The word2vec extracts word vectors based on the context information of words in the text, and the generated word vectors carry contextual semantic information [35]. Some researchers have applied word2vec to constructing sentiment dictionaries [36, 37]. Li et al. built a tourism-specific sentiment lexicon via word2vec [38]. Based on the word vector trained by the word2vec model, Yuan et al. first judged the sentiment type of the candidate word by calculating the similarity between the seed word and the candidate word. Then the sentiment lexicon was constructed [39]. Li et al. used word2vec model to expand sentiment words to construct a Chinese financial domain sentiment lexicon (CFDSL). The experiments show that sentiment features sentiment derived from CFDS exhibit superior performance in FDP when compared to other sentiment dictionaries [40]. In summary, the domain-specific sentiment dictionary constructed by word2vec has better performance than the existing general sentiment dictionary.

3 RESEARCH METHODS

3.1 Sentiment Dictionary Construction Process

3.1.1 Collection and Processing of Inquiry Letter Data

The inquiry letters in this paper were obtained from the information disclosure section of the official websites of Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE). Python was used to crawl the data of inquiry letters from December 1, 2014 to March 10, 2023. These data inquiry letters main attributes include: company code, company abbreviation, letter date, letter categories, letter content and other attributes.

Table 2 A basic description of the crawled inquiry letter example data

Number	Code	Name	Time	Categories	Part of content
1	603393	Xintai Gas	2023/3/3	Comment letter of material assets reorganization plan review	After reviewing the report on material asset purchase and related party transactions (draft) submitted by your company, the following problems need further explanation and explanation from your company...
2	603655	Langbo Technologies	2023/3/1	Inquiry letter	On March 1, 2023, your company disclosed that it intends to change the control of the company by transferring shares through the agreement of the controlling East...
3	000931	Centergate Technologies	2023/3/8	Concern letter	Your company has repeatedly disclosed the announcement that part of the shares of the controlling shareholders and their concerted actions are frozen by the judiciary and waiting to be frozen...
4	002072	Kai Rui De	2023/3/2	Annual report inquiry letter	The following matters are noted: 1. Gross margin. During the reporting period, your company's coal trade revenue and sales gross margin were 359 million yuan and 3.11% respectively, up 206.48% and down 15.08% year-on-year respectively...
5	603393	Inventronics	2023/3/9	Non-licensed reorganization Inquiry letter	Please further verify and explain the following questions: 1. According to the reply letter, Seller 1 signed the Trademark License Agreement and Osram Brand License Agreement with the German target company, agreeing to transfer 226 trademarks...

The crawled inquiry letter example data are shown in Tab. 2 below: We first apply textual preliminary processing to the obtained inquiry letters. The text preliminary processing mainly includes deleting missing values, word segmentation, removing stop words, and labelling the text sentiment.

(1) Deleting missing values.

In the obtained text of the inquiry letters, after screening and viewing, some missing values were found in the specific content data of the letter of inquiry, and these missing values were deleted.

(2) Word segmentation.

There are no spaces between words in Chinese sentences. It will bring a huge workload to digitize sentences directly, and it is unsuitable for analysis, so it is necessary to segment sentences. In this study, the inquiry letters were segmented into words set using Jieba, and the precise word segmentation mode was adopted.

(3) Removing stop words.

Many auxiliary semantic expression words in Chinese sentences have no specific meaning themselves. However, their existence greatly increases the sentence length and increases the amount of calculation in the subsequent sentiment analysis. Therefore, we expand the stop word list provided by Jiang in the financial field according to the actual situation. Finally, we will remove the stop words from the word set.

(4) Labelling the texts sentiment.

The inquiry letters lack a defined sentiment, making it challenging to manually identify it. Therefore, the cognitive structure of emotions model (i.e., the OCC model) was used to label the inquiry letter sentiment labels [41]. The basic principle of OOC is to judge sentiment through the cognitive evaluation process. According to the OCC model framework, investors' reactions after reading the inquiry letter are judged. If the inquiry letter is believed to be positive, the buy decision will be executed with probability. When buying investor power outperforms selling investor power, the company's stock price will rise in the next period of time. The cumulative excess return rate of the companies listed three days after receiving the inquiry letters were chosen as the basis for determining the positive and negative sentiment of the inquiry letters, with reference to the financial text labelling approach proposed by Engelberg [42]. The specific labelling methods are as follows: the cumulative rate of return is positive, and the sentiment of the inquiry letter is marked as positive. The cumulative return rate is negative, and the sentiment of the inquiry letter is marked as negative. This method can effectively avoid the bias caused by manual judgment of text information and the method framework is shown in Fig. 1.

3.1.2 Constructing the Basic Inquiry Letter Sentiment Dictionary (BILSD)

Given that China National Knowledge Infrastructure (CNKI) is the country's largest academic database, a significant amount of research literature on inquiry letters can be found there. These works of literature contain a large number of keywords that the authors believe best reflect the intrinsic content. Sentiment words in the inquiry letter domain may also be achieved by screening and labelling these keywords. Firstly, we chose the CNKI literature that contains the keyword "inquiry letter." Subsequently, the inquiry letter literature's keywords were acquired through iterative manual screening. We built the Inquiry Letter Literary Keyword Sentiment Dictionary (ILKSD), which offers advantages for both professional and academic studies. There are 119 positive words and 194 negative words in the ILKSD. However, some sentiment words rely on the personal experience of the literature authors and are not included in the text of the inquiry letters.

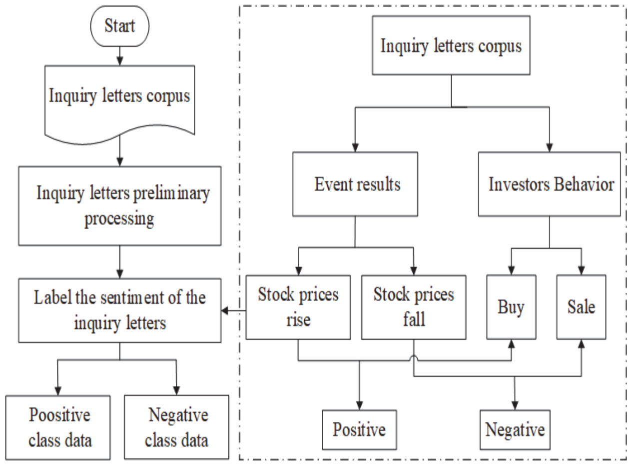

Figure 1 Sentiment labels for inquiry letters based on OCC theory

As we all know, Term Frequency-Inverse Document Frequency (TF-IDF), which is frequently employed in text sentiment analysis, may also be used to assess the importance of words within a corpus [43]. First, the inquiry letter's text data were used as the corpus. Then we calculated each word's TF-IDF value inside the corpus. Finally, the TF-IDF Inquiry Letter Sentiment Dictionary (TILSD) was constructed, which includes 212 positive words and 220 negative words. The sentiment words of the TILSD are highly correlated with inquiry letters. However, there are also disadvantages, such as a few words and the importance of words being only measured by word frequency without considering each word's context. For the convenience of the following, the ILKSD and the TILSD were utilized as the components of the Basic Inquiry Letter Sentiment Dictionary (BILSD).

3.1.3 Reconstructing the Sentiment Dictionary

The reconstruction in these sentiment dictionaries mainly refers to the fusion of the existing GSD, the FSD, and the BILSD. The GSD collected in this paper includes: HowNetSD, NTUSD and TUSD. The FSD collected in this paper includes: LMFSD, JCFSD, YCFSD. The BILSD collected in this paper includes: the TFILSD and the ILKSD. The main steps of reconstructing these sentiment dictionaries are as follows: First, the positive and negative word sets from all the sentiment dictionaries mentioned above were pulled out, respectively. Then, the duplicate sentiment words were removed. Furthermore, we manually determine a sentiment word's tendency when it appears in both positive and negative sentiment word sets simultaneously. Finally, the reconstruction process of these sentiment dictionaries was completed.

3.1.4 Constructing the *SO-PMI* Inquiry Letter Sentiment Dictionary (*SO-PMI* ILSD)

The basic assumption of *SO-PMI* is that words have a certain part of sentiment tendency, which can be generally divided into positive words and negative. The *SO-PMI* consists of two parts: Pointwise Mutual Information (*PMI*) and Semantic Orientation Pointwise Mutual Information (*SO-PMI*). *PMI* is used to calculate the probability of a word and a benchmark sentiment word occurrence in the corpus. The formula is shown in Eq. (1):

$$PMI=\log_2\left(\frac{P(\text{word1, word2})}{P(\text{word1})P(\text{word2})}\right) \tag{1}$$

where, P(word1, word2) is the probability that word1 and word2 appear in the corpus at the same time, P(word1) and P(word2) are the probabilities that word1 and word2 appear in the corpus alone, respectively. If $PMI > 0$, there is a semantic correlation between the two words, and the larger the *PMI* value is, the stronger the correlation is. If $PMI = 0$, it means that the two words have a weak semantic correlation, which is neither related nor mutually exclusive. If $PMI < 0$, the semantic correlation between the two words is the weakest or even irrelevant, and there is a mutually exclusive relationship. The *SO-PMI* is an algorithm for calculating the sentiment tendency of words. The overall idea of the *SO-PMI* algorithm is as follows: First, a set of positive seed words and a set of negative seed words are selected as the benchmark sentiment words, respectively. Then, we need to calculate the *SO-PMI* value, which is expressed as the value of the *PMI* value of the candidate sentiment word and the positive seed words minus the *PMI* value of the candidate sentiment word and the negative seed words. Finally, the *SO-PMI* value is used to judge the sentiment tendency of the candidate sentiment word. The formula is shown in Eq. (2):

$$\begin{aligned} & SO\text{-}PMI(\text{word})= \\ & =\sum_{i=1}^{num(pos)} PMI\left(\text{word, pos}_i\right)-\sum_{i=1}^{num(neg)} PMI\left(\text{word, neg}_i\right) \end{aligned} \tag{2}$$

where, pos_i represents the i-th positive word in the set of positive seed words, neg_i represents the ith negative word in the negative seed words. If the *SO-PMI*(word) > 0, the sentiment tendency of word is positive. If the *SO-PMI*(word) = 0, the sentiment tendency of the word is neutral. If the *SO-PMI*(word) < 0, the sentiment tendency of the word is negative. The construction process of the domain sentiment dictionary based on the *SO-PMI* is shown in Fig. 2.

In this paper, the *SO-PMI* is used to calculate the *SO-PMI* values of all candidate sentiment words from inquiry letters corpus. Then, the effective positive and negative sentiment words were extracted after setting the threshold and removing the duplication to construct the *SO-PMI* inquiry letter sentiment dictionary (*SO-PMI* ILSD). For example, the "scientific and reasonable" process of using *SO-PMI* to calculate sentiment words polarity is as follows. Firstly, the PMI^+ value between the word and the positive seed words set is 33.4, and then the PMI^- value between the word and the negative seed words set is 21.09. Finally, the PMI^+ value minus the PMI^- value is 12.31, which is the "scientific and reasonable" *SO-PMI* value. Due to the *SO-PMI* value > 0, the word was initially added to the list of positive words for *SO-PMI* ILSD.

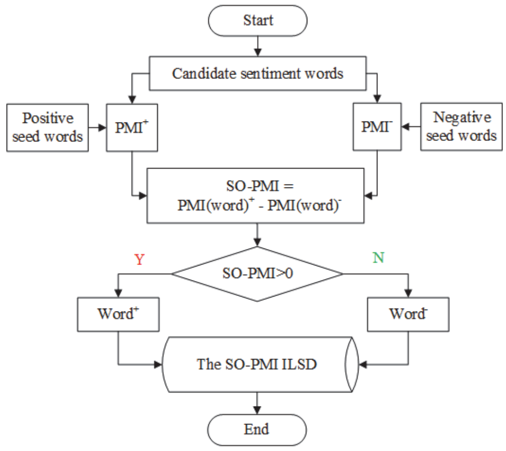

Figure 2 The construction process of *SO-PMI* ILSD

3.1.5 Constructing the Word2vec Inquiry Letter Sentiment Dictionary （Word2vec ILSD）

The word2vec generates word vectors based on the context of words. The word information relationship is transformed into a vector representation, and the similarity formula can be further used to calculate the similarity of two words. The word2vec has two main training models: CBOW and Skip-gram [44]. We use the corpus of ["assets", "raise money", "complete set", "funding", "plan"] as an example to describe the CBOW and Skip-gram model process, as shown in Fig. 3 and Fig. 4. The CBOW model inputs the word vector of the surrounding words and outputs the word vector of the central target word. The CBOW needs to minimize the formula is shown in Eq. (3):

$$L=-\frac{1}{T}\sum_{t=1}^{T}\log P\left(w_t \mid w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}\right) \tag{3}$$

The kip-gram model inputs the word vector of the central target word words and outputs the word vector of the surrounding words. The Skip-gram needs to minimize the formula is shown in Eq. (4):

$$L=-\frac{1}{T}\sum_{t=1}^{T}\log P\left(w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k} \mid w_t\right) \tag{4}$$

where, T is the number of words in the corpus, w_t is the corresponding word at t time steps, and k is the window length.

In general, this study used the Skip-gram model because it offers better predicted accuracy for low-frequency words than the CBOW model. The specific steps of the Word2vec Inquiry Letter Sentiment Dictionary (Word2vec ILSD） construction are as follows:

(1) Obtain the seed set. The GSD, the FSD, and the BILSD are merged to remove duplication. Then, the intersection of the merged sentiment dictionary above and the inquiry

letter corpus after word segmentation is used as the seed set.
(2) Inquiry letter corpus vectorization and seed set vectorization. The word2vec of genism package is used to train the inquiry letter corpus in this paper. The specific model parameter setting results are shown as Tab. 3.

Table 3 The specific model parameter settings for word2vec

Parameter	Description	Value
Size	Dimensions of word vectors,denote as n	$n = 100$
Window	The maximum distance between the central word and the surrounding words, denote as k	$k = 2$
Sg	Set the model type: CBOW ($sg = 0$) or Skip-gram ($sg = 1$).	$sg = 1$
Others	Default	Default

(3) Obtain candidate sentiment words. We use the cosine similarity between the seed sentiment word and the corpus word vector to obtain the candidate sentiment words. The cosine similarity of $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ word vectors calculation formula is shown in Eq. (5):

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{5}$$

where, $\cos(\theta)$ represents the similarity between vectors X and Y, and $n = 100$ in this paper, $\cos(\theta) \in [-1, 1]$. The $\cos(\theta)$ closer to 1 indicates a stronger similarity.

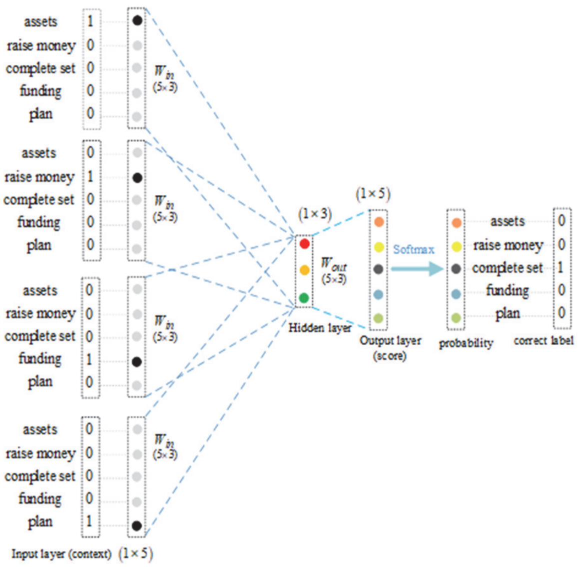

Figure 3 An example of the CBOW model architecture ($k = 2$)

We set the model output parameter topn = 5, which means the top 5 words with the highest cosine similarity between the output word and the seed word.
(4) Selection of candidate words. If the candidate words are in the seed set, we will remove them. If the candidate words are not in the seed set, they will be stored in the word2vec ILSD. Repeat steps (3) to (4) until all the seed sets of words have been traversed.
(5) Obtain the word2vec ILSD. The construction process of the word2vec ILSD is shown in Fig. 5.

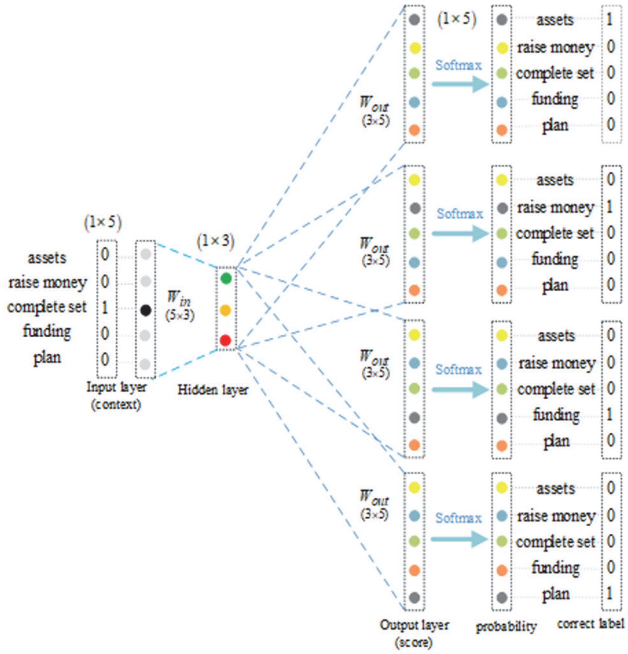

Figure 4 An example of the Skip-gram model architecture ($k = 2$)

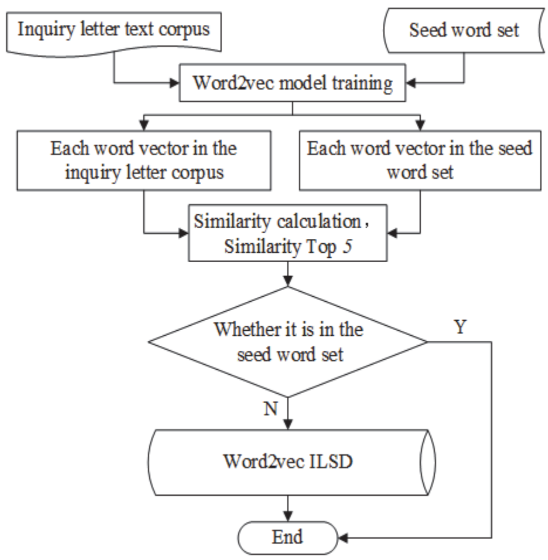

Figure 5 The construction process of word2vec ILSD

3.1.6 The Overall Construction Framework of Inquiry Letter Sentiment Dictionary (ILSD)

The construction of the ILSD in this paper includes five parts: text processing, sentiment dictionary reconstruction, seed word selection, sentiment dictionary expansion, and ILSD generation. The specific steps of ILSD construction are as follows:
Step 1: Text processing. The data in this paper mainly come from the main board of the information disclosure board of the SSEand the main board inquiry letters of the information disclosure board of the SZSE and the Growth Enterprise Market (GEM). First, we delete the collected data with missing text content. Then, the Jieba module is used for preliminary word segmentation, and the stop words are removed via the stop words list. Finally, the OCC model is used to label the sentiment of the inquiry letters.
Step 2: Sentiment dictionary reconstruction. We merge and deduplicate the existing GSD, FSD, and the basic ILSD.

Step 3: Seed words selection. The intersection words of the reconstruction sentiment dictionary and the processed inquiry letter corpus are obtained. And then, we manually screen the intersection words to obtain the seed words needed for subsequent sentiment dictionary expansion.
Step 4: Sentiment dictionary expansion. The *SO-PMI* and word2vec are used to obtain the *SO-PMI* ILSD and the Word2vec ILSD, respectively.
Step 5: The ILSD generation. We first construct a draft ILSD by merging the *SO-PMI* ILSD, the word2vec ILSD, and seed sentiment dictionary. Then, the intersection words of positive words and negative words in the draft ILSD were deleted, respectively. Finally, the ILSD is constructed via manually reviewing intersection words and adding them to the corresponding positive words or negative words list. The ILSD overall construction framework is shown in Fig. 6.

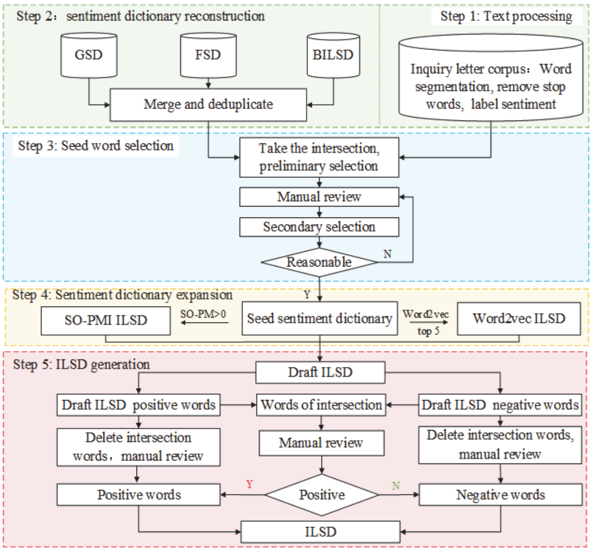

Figure 6 The Inquiry ILSD overall construction framework

3.2 Inquiry Letter Sentiment Dictionary (ILSD) Introduction

Compared with other sentiment dictionaries, the ILSD constructed in this paper has many advantages. First, according to the dictionary structure, the ILSD proposed in this paper can be updated at any time according to the corpus, which makes the dictionary strong in timeliness. In addition, the ILSD not only draws from many people's strengths but also considers the domain characteristics of inquiry letters. Finally, the ILSD can more objectively represent the sentiment of the inquiry letters.

Table 4 The sizes of ILSD

Name	Number of negative words	Number of positive words
ILSD	1311	1511

Table 5 Example of the ILSD negative words

Negative words				
Liabilities	maturity	decline	inventory	render
Explanation	involve in	volatility	situation	check
Provision	alteration	submit	suspicious	risk
Verification	disclosure	inquiry	limitation	loss
Termination	bad debt	check	lower than	loan
Impairment of value	resulting in	price decline	increasing loss	fixed assets

According to Tab. 4, the ILSD has a moderate number of positive and negative words. Both positive and negative affective words are strongly associated with the inquiry letters, as shown by Tab. 5 and 6.

Table 6 Example of the ILSD positive words

Positive words				
Top 100	moderation	credit	heavy	favor
Shortage	refinement	adhering	success	offering
Overcome	robustness	authority	revenue	revenue
Speed	doubling	get high	shaping	create
Get through	key items proportion	remarkable effect	excess profit	raise salary
Virtuous cycle	high starting point	new materials	ease up	explosive

3.3 Evaluation Metrics

(1) Coverage evaluation metrics.

We verify ILSD's ability to extract sentiment words from inquiry letters using coverage testing. The calculation formula is shown in Eq. (6):

$$coverage = \frac{count_{hit}}{count_{all}} \times 100\% \tag{6}$$

where, $count_{hit}$ is the number of sentiment words matched in the inquiry letters, and $count_{all}$ is the total number of words in the inquiry letters.
(2) Performance evaluation metrics.

The accuracy (short as $ACC) \in [0, 1]$ is the percentage of correctly predicted samples in all samples, which is an overall evaluation metrics of the sentiment classification performance [45]. However, the ACC fails to accurately evaluate the imbalanced class samples. The $F1$ weighted score (short as $F1_{Weighted}) \in [0, 1]$ is a comprehensive evaluation metrics that integrates false positive, false negative and sample classes weights [46]. The matthews correlation (short as $MCC) \in [-1, 1]$ provides more information and true score than $F1_{Weighted}$ when evaluating balanced or imbalanced binary classes [47, 48]. The geometric mean (short as G-mean)$\in [0, 1]$ also takes account of both the true positive and the true negative, which can be used as a comprehensive metrics of sentiment classification performance [49]. Therefore, we will use the four metrics as the sentiment classification performance of different sentiment dictionaries for inquiry letters. The four performance evaluation metrics mentioned above all rely on the classification results, of which four possible results are presented in the confusion matrix in Tab. 7.

Table 7 The confusion matrix for classification results

	Predict negative (0)	Predict positive (1)
Actual negative (0)	*TN*	*FP*
Actual positive (1)	*FN*	*TP*

In Tab. 7, the true negative (*TN*) is denoted as the actual negative class and predicted negative class. The false negative (*FN*) is denoted as the actual negative class and predicted positive class. The false positive (*FP*) is denoted as the actual negative class and predicted positive class. The true positive (*TP*) is denoted as the actual positive class and predicted positive class. These evaluation metrics are used to comprehensively evaluate the results, and the specific calculation formulas are shown in Eq. (7) to Eq. (10).

$$ACC = \frac{S_1 + S_3}{S} \tag{7}$$

$$F1_{Weighted} = \frac{S_{12}}{S} F1_N + \frac{S_{34}}{S} F1_P \tag{8}$$

$$MCC = \frac{S_1 \times S_3 - S_2 \times S_4}{\sqrt{S_{23} \times S_{34} \times S_{12} \times S_{14}}} \tag{9}$$

$$G\text{-mean} = \sqrt{\frac{S_3}{S_{34}} \times \frac{S_1}{S_{12}}} \tag{10}$$

where S_1, S_2, S_3, S_4 denote the TN, FP, TP, and FN, respectively. $S = TP + FP + FN + TN$, S denote the total number of the samples. $S_{12} = TN + FP$, S_{12} denotes the number of samples by actual negative class. $S_{14} = TN + FN$, S_{14} denotes the number of samples by predict negative class. $S_{23} = FP + TP$, S_{23} denotes the number of samples by predict positive class. $S_{34} = TP + FN$, S_{34} denotes the number of samples by actual positive class. The $F1_N = \frac{2TN}{2TN + FP + FN}$, $F1_P = \frac{2TP}{2TP + FP + FN}$ denote $F1$ score of the negative and positive classes, respectively.

4 RESULTS AND DISCUSSION

4.1 Data Collection and Processing

The experiment in this section aims to verify the validity of the LSD proposed in this paper. We collected the inquiry letters issued under the information disclosure section of the SSE from December 26, 2014, to March 10, 2023. It provides an important data for verifying the validity of the ILSD. In the process of text data processing, the missing data was first removed. Then, the jieba was used to implement the word segmentation process on the inquiry letters. Finally, the stop words were removed via the stop word list. See Tab. 8 for an example of the processed inquiry letters.

Table 8 Example of processed inquiry letters

Number	Code	Name	Time	Class	Inqul,iry letter segmentation
1	600531	Yuguang Gold Lead	2016/02/23	0	Annual report, post-audit, inquiry, profit and loss, net profit, negative, weak, loss low level, ...
2	600136	Daobo shares	2016/03/01	1	Gross margin, development, openness, competitiveness, core, disclosure, content, Advantage, copyright, extensive, ...
3	600764	CEC CoreCast	2016/03/03	0	Intense, fail, intensive, competitive, drive, product, update, iteration, segmentation, field, ...
4	600785	Xinhua Commercial	2016/03/07	1	Merger, increase, investment, real estate, transaction, situation including, transaction, subject matter, transaction, ...
5	600987	Hangmin shares	2016/03/07	1	Growth, percentage points, Industry, environment, operation, situation, supplement, disclosure, gross margin, increase, ...

4.2 Experiment Results and Discussion

To illustrate the sentiment word extraction performance and classification performance of the presented ILSD, the evaluation metrics in the previous section were used to evaluate the experimental results.

4.2.1 Analysis of Sentiment Dictionary Coverage

The coverage results of sentiment words in the ILSD and benchmark sentiment dictionaries for the SSE inquiry letters are shown in Fig. 7.

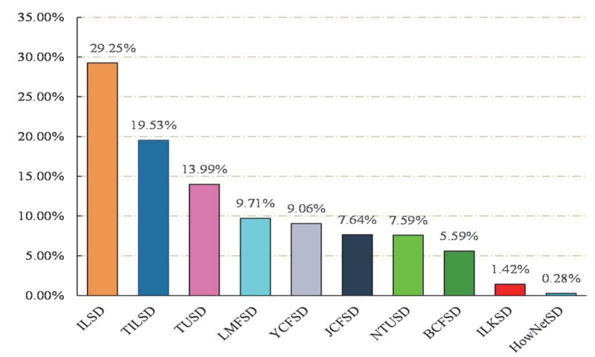

Figure 7 Comparison of sentiment dictionaries coverage results

As shown in Fig. 8 and Tab. 9 the coverage of ILSD is 29.25%, which is the highest among all sentiment dictionaries. The coverage of the ILSD is superior to the other sentiment dictionaries by 9.72% to 28.97%. Preliminarily, it shows that the sentiment dictionary of this paper has a better capturing ability for the sentiment words of inquiry letters. To further verify that the sentiment dictionary has the advantage of excellent coverage in a specific domain, we research from the perspective of the ILSD, the BILSD, the FSD and the GSD.

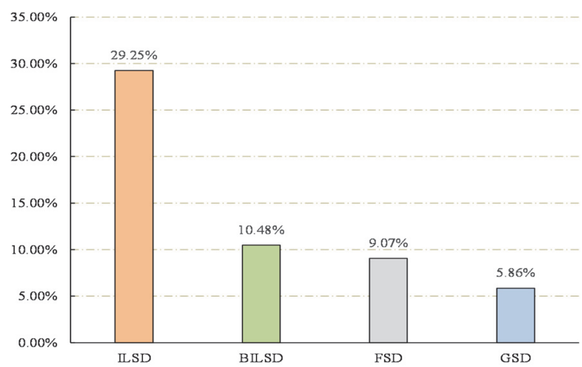

Figure 8 Comparison of domain sentiment dictionaries average coverage results

Fig. 8 and Tab. 9 show the average coverage of ILSD > BILSD > FSD > GSD. The results show that the domain sentiment dictionary is superior to the other sentiment dictionaries in its application field. It reflects that domain sentiment dictionary can improve the utilization of text, increase the coverage of sentiment dictionary to text, and more effectively extract the sentiment words of the domain texts. Therefore, it is necessary to construct a domain sentiment dictionary for inquiry letters.

Table 9 Comparison of domain sentiment dictionaries average coverage results

		Coverage	Average coverage
ILSD	ILSD	29.25%	29.25%
BILSD	TFILSD	19.53%	10.48%
	ILKSD	1.42%	
FSD	LMFSD	13.99%	9.07%
	JCFSD	7.64%	
	YCFSD	9.06%	
	BCFSD	5.59%	
GSD	HowNetSD	0.28%	5.86%
	NTUSD	7.59%	
	TUSD	9.71%	

4.2.2 Analysis of Classification Results of ILSD

The sentiment classification results of ILSD for inquiry letters are presented in the confusion matrix, as shown in Tab. 10.

Table 10 Confusion matrix from ILSD sentiment classification results

		Predict		Total
		negative (0)	positive (1)	
Actual	negative (0)	467	364	831
	positive (1)	401	522	923
Total		868	886	1754

The $S_1 = TN = 467$, $S_2 = TP = 364$, $S_3 = TP = 522$, $S_4 = FN = 401$, $S = TP + FP + FN + TN = 1754$. What's more, $S_{12} = TN + FP = 831$, $S_{14} = TN + FN = 868$, $S_{23} = \mathrm{FP} + TP = 886$ and $S_{34} = \mathrm{TP} + FN = 923$ are listed in Tab. 9. The $F1_N$ and $F1_P$ can be calculated as:

$$F1_N = \frac{2TN}{2TN + FP + FN} = \frac{2 \times 467}{2 \times 467 + 364 + 401} = 0.5479,$$

$$F1_P = \frac{2TP}{2TP + FP + FN} = \frac{2 \times 522}{2 \times 522 + 364 + 401} = 0.5771.$$

According to those results above, ACC, $F1_{Weighted}$, MCC and G-mean can be calculated as:

$$F1_{Weighted} = \frac{S_{12}}{S} F1_N + \frac{S_{34}}{S} F1_P = \\ = \frac{831}{1754} \times 0.5479 + \frac{923}{1754} \times 0.5771 = 0.5641$$

$$MCC = \frac{S_1 \times S_3 - S_2 \times S_4}{\sqrt{S_{23} \times S_{34} \times S_{12} \times S_{14}}} = \frac{467 \times 522 - 364 \times 401}{\sqrt{886 \times 923 \times 831 \times 868}} = \\ = 0.1274$$

$$G\text{-mean} = \sqrt{\frac{S_3}{S_{34}} \times \frac{S_1}{S_{12}}} = \sqrt{\frac{522}{923} \times \frac{467}{831}} = 0.5638$$

From the above calculation results, the proposed ILSD provides a better comprehensive performance evaluation metrics for the sentiment classification of inquiry letters. The ACC result of 0.5639 indicates that more than 50% samples are correctly classified. Moreover, the $F1_{Weighted}$, MCC and G-mean results of 0.5641, 0.1274 and 0.5638, respectively. The results of the three comprehensive performance evaluation metrics show that the ILSD is effective for the sentiment classification of inquiry letters.

4.2.3 Analysis of Classification Results Comparison with Other Sentiment Dictionaries

To further evaluate ILSD's performance, we conducted a comparison of its performance with the results of other benchmark sentiment dictionaries. Those sentiment dictionaries are HowNetSD, TUSD, NTUSD, BCFSD, JCFSD, YCFSD, LMFSD, ILKSD, and TILSD. The performance of ILSD and other benchmark sentiment dictionaries is shown in Tab. 11 and Fig. 9 on the four metrics. Overall, the ILSD has the best performance in all four performance evaluation metrics (ACC, $F1_{Weighted}$, MCC and G-mean) compared with other benchmark sentiment dictionaries in this paper. Specifically, the ACC of the ILSD is 0.5639, which is better than other benchmark sentiment dictionaries by 2.17% to 6.55%. The $F1_{Weighted}$ of the ILSD is 0.5641, which is better than other benchmark sentiment dictionaries by 2.64% to 16.33%. The MCC of the ILSD is 0.1274, which is better than other benchmark sentiment dictionaries by 5.20% to 16.73%. The G-mean of the ILSD is 0.5638, which is better than other benchmark sentiment dictionaries by 3.60% to 33.86%. Therefore, ILSD generally outperforms other benchmark sentiment dictionaries in the inquiry letters sentiment analysis. As shown in Tab.11 and Fig. 9, the classification results of most benchmark sentiment dictionaries in this paper (excluding LMFSD) are biased towards the positive class, resulting in higher false positive rates than ILSD. This indicates that most sentiment dictionaries in this paper exhibit weaker ability in classifying negative categories compared to ILSD. The ILSD lies in its accurate extraction of positive and negative sentiment words specific to inquiry letters, thus reducing the likelihood of negative comments being wrongly judged as positive. Compared to other sentiment dictionaries, the accuracy of ILSD in classifying both positive and negative sentiment has been enhanced.

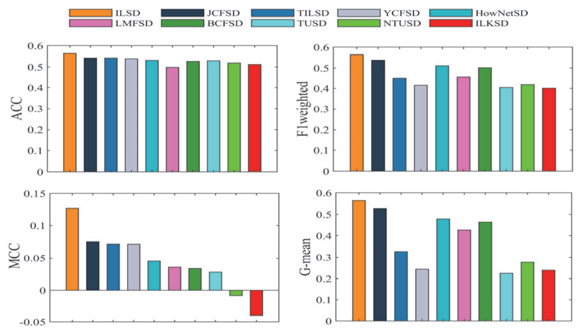

Figure 9 Comparison classification performance results of each sentiment dictionary

However, 43.31% of the samples remain misclassified, indicating that there is still room for improvement in accuracy in future iterations. In a word, the experimental results show that the ILSD has relatively accurate prediction ability for both positive and negative sentiment classes of inquiry letters compared with other benchmark sentiment dictionaries. The experimental results illustrate the effectiveness of the proposed ILSD in the sentiment classification of inquiry letters and the ILSD's greater relevant domain adaptability in this field.

Table 9 The sentiment dictionary classification performance results

Sentiment dictionary		Predict negative (0)	Predict positive (1)	*ACC*	$F1_{Weighted}$	*MCC*	*G*-mean
ILSD	Actual negative (0)	467	364	0.5639	0.5641	0.1274	0.5638
	Actual positive (1)	401	522				
JCFSD	Actual negative (0)	364	467	0.5422	0.5377	0.0754	0.5278
	Actual positive (1)	336	587				
TILSD	Actual negative (0)	96	735	0.5421	0.4495	0.0718	0.3271
	Actual positive (1)	68	855				
YCFSD	Actual negative (0)	51	780	0.5387	0.4149	0.0717	0.2439
	Actual positive (1)	29	894				
HowNetSD	Actual negative (0)	263	568	0.5313	0.5108	0.0452	0.4789
	Actual positive (1)	254	669				
LMFSD	Actual negative (0)	668	163	0.4984	0.4551	0.0357	0.4257
	Actual positive (1)	715	208				
BCFSD	Actual negative (0)	241	590	0.5267	0.5014	0.0336	0.4632
	Actual positive (1)	240	683				
TUSD	Actual negative (0)	44	787	0.5296	0.4045	0.0279	0.2252
	Actual positive (1)	38	885				
NTUSD	Actual negative (0)	70	761	0.5194	0.4180	–0.0082	0.2770
	Actual positive (1)	82	841				
ILKSD	Actual negative (0)	52	779	0.5120	0.4008	–0.0399	0.2393
	Actual positive (1)	77	846				

5 CONCLUSIONS

In this study, we pioneered the construction of a specialized ILSD to enable more effective sentiment analysis of regulatory communications. The key innovation lies in integrating *SO-PMI*, word2vec embeddings, and manual screening to develop a sentiment dictionary tailored to inquiry letters. This approach captures co-occurrence relationships and word contexts lacking in existing general and financial sentiment dictionaries. Compared with the baseline sentiment dictionaries, although the computational complexity increases, our proposed approach reduces the time required for constructing the sentiment dictionary and improves its accuracy. Rigorous empirical analysis of SSE 1754 actual inquiry letters demonstrates the ILSD's superior coverage (29.25%) and performance on four evaluation metrics compared to baseline sentiment dictionaries. The results validate the importance of domain-specific dictionaries for financial text sentiment mining. The proposed technical paradigm provides a dynamic framework to continuously expand the ILSD by incorporating new textual data. This study makes key contributions to constructing the first known sentiment dictionary for inquiry letters, advancing knowledge on financial text sentiment analysis. What is more, the ILSD unlocks new research opportunities in leveraging regulatory communications for stock price trend prediction, risk monitoring, and investment decision support. However, this study only verifies the validity of the ILSD through the SSE inquiry letter sentiment classification experiment. In future research, we can verify the effectiveness of ILSD in different data sources and domain applications. Overall, this study pioneers a specialized sentiment dictionary to enable more nuanced modelling of dynamic financial market sentiment through inquiry letter mining.

Acknowledgements

This work was supported by the National Social Science Fund of China 22 & ZD153.

6 REFERENCES

[1] Lu, J. & Qiu, Y. (2023). Does non-punitive regulation diminish stock price crash risk? *Journal of Banking & Finance, 148*, 106731. https://doi.org/10.1016/j.jbankfin.2022.106731

[2] Renault, T. (2020). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance, 2*, 1-13. https://doi.org/10.1007/s42521-019-00014-x

[3] Sun, A., Lachanski, M., & Fabozzi, F. J. (2016) Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis, 48*, 272-281. https://doi.org/10.1016/j.irfa.2016.10.009

[4] Jing, N., Wu, Z., & Wang, H. (2021) Hybrid Model Integrating Deep Learning with Investor Sentiment Analysis for Stock Price Prediction. *Expert Systems with Applications, 178*(3), 115019. https://doi.org/10.1016/j.eswa.2021.115019

[5] Wang, G., Chen, G., & Chu, Y. (2018). A new random subspace method incorporating sentiment and textual information for financial distress prediction. *Electronic Commerce Research and Applications, 29*, 30-49. https://doi.org/10.1016/j.elerap.2018.03.004

[6] Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y., & Wu, X. (2019) Chinese text sentiment analysis based on extended sentiment dictionary. *IEEE Access, 7*, 43749-43762. https://doi.org/10.1109/ACCESS.2019.2907772

[7] Park, S., Lee, W., & Moon, I. C. (2015) Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters, 56*, 38-44. https://doi.org/10.1016/j.patrec.2015.01.004

[8] Wu, S., Xiao, Q., Gao, M., & Zou, G. (2020) A construction and self-learning method for intelligent domain sentiment lexicon. *International Journal of Information Technology and Management, 19*(4), 318-333. https://doi.org/10.1504/IJITM.2020.110235

[9] Oliveira, N., Cortez, P., & Areal, N. (2016) Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems, 85*, 62-73. https://doi.org/10.1016/j.dss.2016.02.013

[10] Turney, P. D. & Littman, M. L. (2003) Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315-346. https://doi.org/10.1145/944012.944013

[11] Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th annual meeting on association for computational linguistics (ACL)*, 417-424. https://doi.org/10.3115/1073083.1073153

[12] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems, 26*, 3111-3119.

[13] Ortony, A., Clore, G., & Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press, New York.

[14] Zhang, S., Wei, Z., Wang, Y., & Liao, T. (2018). Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future Generation Computer Systems, 81*, 395-403. https://doi.org/10.1016/j.future.2017.09.048

[15] Dai, L., Liu, B., Xia, Y., & Wu, S. (2008). Measuring semantic similarity between words using HowNet. *Proceedings of the 2008 International Conference on Computer Science and Information Technology Singapore*, 601-605. https://doi.org/10.1109/ICCSIT.2008.101

[16] Ku, L., Liang, Y., & Chen, H. (2006). Opinion Extraction, Summarization and Tracking in News and Blog Corpora. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 100-107.

[17] Li, J. & Sun, M. (2007). Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques. *Proceedings of the IEEE International Conference on Natural Language Processing & Knowledge Engineering*, 393-400. https://doi.org/10.1109/NLPKE.2007.4368061

[18] Loughran, T. & MeDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance, 66*(1), 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

[19] Johnman, M. T., Vanstone, B. J., & Gepp, A. (2018). Predicting FTSE 100 returns and volatility using sentiment analysis. *Accounting & Finance, 58*, 253-274. https://doi.org/10.1111/acfi.12373

[20] Xie, D. & 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.

[21] Bian, S., Jia, D., Li, F., & Yan, Z. (2018) A New Chinese Financial Sentiment Dictionary for Textual Analysis in Accounting and Finance. Available at SSRN 3446388. https://doi.org/10.2139/ssrn.3446388

[22] Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics, 132*(1), 126-149. https://doi.org/10.1016/j.jfineco.2018.10.001

[23] Yao, J., Feng, X., Wang, Z., Ji, R., & Zhang, W. (2021). Tone, sentiment and market impacts: The construction of Chinese sentiment dictionary in finance. *Journal of Management Sciences in China, 24*(5), 26-46.

[24] Xu T, Peng Q, & Cheng Y. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Systems, 35*, 279-289. https://doi.org/10.1016/j.knosys.2012.04.011

[25] Bravo-Marquez, F., Frank, E., & Pfahringer, B. (2016). Building a twitter opinion lexicon from automatically-annotated tweets. *Knowledge-Based Systems, 108*, 65-78. https://doi.org/10.1016/j.knosys.2016.05.018

[26] Zhao, M., Zhang, T., & Chai, J. (2015). Based on *SO-PMI* algorithm to discriminate sentimental words' polarity in TV programs' subjective evaluation. *Proceedings of the 2015 8th International Symposium on Computational Intelligence and Design. hangzhou, China*, 38-40. https://doi.org/10.1109/ISCID.2015.86

[27] Yang, A., Lin, J., Zhou, Y., & Chen, J. (2013). Research on building a chinese sentiment lexicon based on *SO-PMI*. *Applied Mechanics & Materials, 263*, 1688-1693. https://doi.org/10.4028/www.scientific.net/AMM.263-266.1688

[28] Liu, J., Yan, M., & Luo, J. (2016) Research on the construction of sentiment lexicon based on Chinese microblog. *Proceedings of the 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou, China, 2*, 56-59. https://doi.org/10.1109/IHMSC.2016.264

[29] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector. *Space*, 1-12.

[30] Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. *International conference on machine learning. Beijing, China*, 1188-1196.

[31] Rezaeinia, S., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications, 117*, 139-147. https://doi.org/10.1016/j.eswa.2018.08.044

[32] Ray, B., Garain, A., & Sarkar, R. (2021) An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing, 98*, 106935. https://doi.org/10.1016/j.asoc.2020.106935

[33] De Vine., L, Zuccon, G., Koopman, B., Sitbon, L., & Bruza, P. (2014). Medical semantic similarity with a neural language model. *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 1819-1822. https://doi.org/10.1145/2661829.2661974

[34] Mao, Y., Liu, S., & Gong, D. (2023). A Hybrid Technological Innovation Text Mining, Ensemble Learning and Risk Scorecard Approach for Enterprise Credit Risk Assessment. *Tehnički vjesnik, 30*(6), 1692-1703. https://doi.org/10.17559/TV-20230316000447

[35] Hu, K., Wu, H., Qi, K., Yu, J., Yang, S, Yu, T., Zheng, J., & Liu, B. (2018). A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. *Scientometrics, 114*, 1031-1068. https://doi.org/10.1007/s11192-017-2574-9

[36] Jia, K. (2021) Chinese sentiment classification based on Word2vec and vector arithmetic in human–robot conversation. *Computers and Electrical Engineering, 2021*(95), 107423. https://doi.org/10.1016/j.compeleceng.2021.107423

[37] Fauzi, M. (2019). Word2Vec model for sentiment analysis of product reviews in Indonesian language. *International Journal of Electrical and Computer Engineering, 9*(1), 525-530. https://doi.org/10.11591/ijece.v9i1.pp525-530

[38] Li, W., Zhu, L., Guo, K., Shi, Y., & Zheng, Y. (2018). Build a tourism-specific sentiment lexicon via word2vec. *Annals of Data Science, 5*, 1-7. https://doi.org/10.1007/s40745-017-0130-3

[39] Yuan, Z. & Duan., L. (2019). Construction method of sentiment lexicon based on word2vec. *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference*, 848-851. https://doi.org/10.1109/ITAIC.2019.8785471

[40] Li, S., Shi, W., Wang, J., & Zhou, H. (2021). A deep learning-based approach to constructing a domain sentiment lexicon: a case study in financial distress prediction. *Information Processing & Management, 58*(5), 102673. https://doi.org/10.1016/j.ipm.2021.102673

[41] Ortony, A., Clore, G., & Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press, New York.

[42] Engelberg, J., Reed, A., & Ringgenberg, M. (2012) How are shorts informed? Short sellers, news, and information

processing. *Journal of Financial Economics, 105*(2), 260-278. https://doi.org/10.1016/j.jfineco.2012.03.001

[43] Liu, H., Chen, X., & Liu, X. (2022). A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis. *IEEE Access, 10*, 32280-32289. https://doi.org/10.1109/ACCESS.2022.3160172

[44] Tao, W. & Chang, D. (2019). News text classification based on an improved convolutional neural network. *Tehnicki vjesnik, 26*(5), 1400-1409. https://doi.org/10.17559/TV-20190623122323

[45] Wang, Y., Yin, F., Liu, J., & Tosato, M. Automatic construction of domain sentiment lexicon for semantic disambiguation. (2020). *Multimedia Tools and Applications, 79*, 22355-22373. https://doi.org/10.1007/s11042-020-09030-1

[46] Cheng, M., Kusoemo, D., & Gosno, R. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction, 118*, 103265. https://doi.org/10.1016/j.autcon.2020.103265

[47] Chicco, D., Warrens, M., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access, 9*, 78368-78381. https://doi.org/10.1109/ACCESS.2021.3084050

[48] Chicco, D. & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21*(1), 1-13. https://doi.org/10.1186/s12864-019-6413-7

[49] Guo, J., Wu, H., Chen, X., & Lin, W. (2024). Adaptive SV-Borderline SMOTE-SVM algorithm for imbalanced data classification. *Applied Soft Computing, 150*, 110986. https://doi.org/10.1016/j.asoc.2023.110986

Contact information:

Wei WANG, PhD candidate
School of Economics & Management,
University of Science and Technology Beijing,
No. 30 Xueyuan Road Haidian District, Beijing 100083, China
E-mail: wangweiustb@163.com

Guiying WEI, PhD, Associate Professor
School of Economics & Management,
University of Science and Technology Beijing,
No. 30 Xueyuan Road Haidian District, Beijing 100083, China
E-mail: weigy@manage.ustb.edu.cn

Sen WU, PhD, Full Professor
(Corresponding author)
School of Economics & Management,
University of Science and Technology Beijing,
No. 30 Xueyuan Road Haidian District, Beijing 100083, China
E-mail: wusen@manage.ustb.edu.cn

Huixia HE, PhD candidate
School of Economics & Management,
University of Science and Technology Beijing,
No. 30 Xueyuan Road Haidian District, Beijing 100083, China
E-mail: 18810081025@163.com