Stock Investors' Preferences on Stock Forum Topics Based on FNS-LDA2vec

Ganglong Duan[1,a], *Jinkai Zhang[2,b] , Mingyue Jiang[3,c]

[a]58406775@qq.com, [b*]zjk13934720743@163.com, [c]1103926122@qq.com

Xi'an University of technology School of Economics and Management Xi'an, Shaanxi, China

Abstract—This article constructs an investor topic preference mining model through stock bar text data and the FNS-LDA2vec method. Firstly, dynamic topic mining of the LDA model is achieved by dividing user documents in the stock community by time. Then, by combining the improved LDA topic model and the Word2vec word vector model, a dynamic mining model of stock bar user topic preference based on FNS-LDA2vec is constructed, and topic representation is learned through the joint learning of document vectors and word vectors. Finally, empirical results show that the topic extraction model constructed in this article is superior to the comparison model. The model has broad application value in personalized recommendation for investors and stock prediction.

Keywords-Share bar discussion; Topic model; LDA model; Data mining;

1 Introduction

The stock market is a platform full of information exchange and investor interaction, as the main information exchange platform in the stock market, its topics can directly affect investors' decision-making and judgment. Therefore, by analyzing the preferences of investors in the stock market for stock topics, we can more accurately understand the needs and psychology of investors, so as to better provide services and guidance for them. It is important to understand the demographics and topic preferences of investors. By mining investors' statements and comments on social media platforms such as stock bars, it is possible to gain insight into their views and attitudes towards different stock and market events. However, due to the massive amount of text data and diverse language expressions, how to efficiently extract valuable information from this data is a very challenging issue. Based on this, this study proposes a stock investor stock bar topic preference mining method based on FNS-LDA2vec. Among them, the LDA2vec model is used for semantic representation of topics. By combining LDA and Word2vec, investors' topic preferences can be mined more accurately, and then provide a reference for more intelligent stock investment decisions. In this study, the method is also applied to the stock bar discussion data of stock investors, and empirical research is carried out to prove the effectiveness and feasibility of the method. This study has certain reference value for in-depth understanding of investors' group characteristics and topic preferences, and improving the accuracy and efficiency of stock investment decisions.

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China

DOI 10.4108/eai.27-10-2023.2341922

2 Related research

FNS-LDA2vec is a follow-up research progress based on LDA2vec model, which has wide application in the field of theme model, especially in the topic preference mining of stock investors. The model combines FastText, Negative Sampling, and LDA2vec to convert text data into low-dimensional vectors, and can find similar topics and identify keywords and topics in the text, so as to help investors understand hot topics and market trends, and further optimize investment strategies. FNS-LDA2vec can build vector representations of words and documents at the same time, and combines the advantages of word embeddings and topic models. It can be used to analyze topics and semantics in natural language text[7].

Stock investors' topic preference mining refers to the use of theme models and other methods to analyze the remarks made by stock investors in communities such as stock bars, explore their interests and attitudes in different stocks or industries, and provide reference for investment decisions. The FNS-LDA2vec model is used to model and vectorize the text data in the stock bar and other communities to obtain the topic distribution and document vector of investors, and then analyze their topic preferences and similarity. Stock Bar is a securities community under Oriental Wealth Network, mainly providing investors with real-time market analysis, stock information, stock information, etc[1].

For stock investor topic preference mining, researchers often need to process massive amounts of text data to obtain valuable information. In practice, researchers can use techniques such as Web Data Mining, sentiment analysis, topic modeling, and other techniques to analyze stock topics. These technologies can help investors identify useful stock investment information and extract hot topics and market trends from it[16].

In addition, machine learning algorithms, such as support vector machines, naïve Bayes classification, logistic regression, etc., can be combined to classify and predict text data, so as to better identify market trends and investment directions, and improve investment returns. On this basis, the use of FNS-LDA2vec model to mine the topic preferences of stock investors in the stock bar is a research direction with high application value and prospects[15].

3 Designed Based on the Fns-Lda2vec Model

3.1 LDA topic model

LDA is a hierarchical Bayesian model. In this model, this paper assumes that there are a total of T topics in the document collection, and each topic $Z$is represented as a unilanguage model on a dictionary Vthetaθ_z, that is, a polynomial distribution on the dictionary. It is further assumed that each document d corresponds to this Ttopic and has a document-specific polynomial distribution d .The shaded circles in the figure represent observed variables, the unshaded circles represent latent variables, the arrows represent conditional dependencies between the two variables, and the boxes represent repeated sampling, where the number in the lower right corner of the box indicates the number of repeats. Figure1 shows the LDA generation process[5].

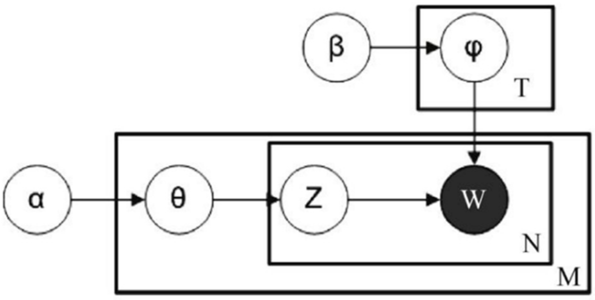


Figure1 Diagram of the LDA generation process

LDA models involve Gamma functions, binomial distributions, multinomial distributions, beta distributions, Dirichlet distributions, conjugate prior and Bayesian frameworks, Gibbs sampling, etc., and these mathematical concepts and techniques play an important role in the derivation and implementation of LDA.

(1) Gamma function: essentially generalization of factorial functions on real numbers, Gamma functions play an important role in probability density functions for normalizing distributions. The factorial formula when it is an integer and the generalized formula when it is real are respectively:

$$\begin{gathered}\tau(n) = (n-1)!\,，\ integer \\ \tau(t) = \int_0^{\infty} x^{t-1} e^{-x} dx\,, real\ number\end{gathered} \tag{1}$$

(2) Binomial distribution: used to model the probability distribution of discrete data, such as modeling the occurrence of words. Generally, there are only two values, which refer to the frequency of the desired result occurrence of event A in k experiments with two outcomes. Each experimental result is independent of the last time, if the probability of B occurring is P(B), and P(A) + P(B) = 1.

$$P(A) = \binom{n}{k} P^k (1-P)^{(n-k)},\ \ (k = 0，\ 1，\ 2，\ ..n) \tag{2}$$

(3) Multinomial distribution: used to model the probability distribution of multi-category discrete data, which is used in LDA to represent the distribution of words in the document. Just as the six points of the shaker have a probability of shaking, the multinomial distribution will also have multiple results. Suppose that an experiment has K possible scenarios of 1, 2...k, then when n occurrences, 1 occurs n1 times.... k occurs nk times, then the probability of P(1, 2, ..k) is shown in Equation 3:

$$\begin{aligned}&P\left(x_1，\ x_2, \dots, x_k; n, p_1, p_2, \dots, p_k\right) \\ &= {p_1}^{x1} {p_2}^{x2} \dots {p_k}^{xk} \frac{n!}{x_1! \dots x_k!}\end{aligned} \tag{3}$$

thereinto$\sum_{i=1}^{k} p_i = 1，\ p_i > 0$

(4) Conjugate priori: Understanding requires understanding several keywords such as prior distribution, posterior distribution, and likelihood estimation. Prior distributions are used to express the degree of uncertainty of an uncertain quantity, i.e. subjective estimates of probability

distributions for unknown parameters before data or evidence is available. A posterior distribution is a parameter probability distribution obtained by combining a prior distribution and a likelihood function after obtaining data or evidence. It represents the probability of updating the parameters after the data are observed, and the prior and posterior distributions have the same functional form and become conjugate priors[8].

where the prior distribution probability is:

$$\mu^{a}\ （1-\mu）^{b} \tag{4}$$

The posterior distribution probabilities are:

$$\mu^{a}\ （1-\mu）^{b} \tag{5}$$

(5) Beta distribution: A continuous probability distribution with a value interval of [0,1] of the Beta distribution. The beta distribution is also a conjugate distribution of the binomial distribution. The values of its arguments alpha and beta are greater than 0, and its probability density formula is expressed as Equation 6.

$$\begin{aligned} f(x,\alpha,\beta) &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}\,(1-u)^{\beta-1}du} \\ &= \frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \end{aligned} \tag{6}$$

where τ is a function of $\tau(x)$.

(6) Dirichlet distribution: also known as multivariate beta distribution, is its distribution in high dimension, is a k-dimensional conceptual function, according to the introduction on Wikipedia, dimension K geq2(x1, x2... xK-1 dimension, of K total) on the parameter a1,...,ak> 0, based on the Lebesgue measure in the Euclidean space RK-1, has a probability density function defined by the formula expressed as Equation 7and8:

$$f(x_1,\ldots,x_{k-1};\alpha_1,\ldots,\alpha_k) = \frac{1}{B(\alpha)}\prod_{i=1}^{k} x_i^{\alpha_i-1} \tag{7}$$

where $B(\alpha)$ is equivalent to a multinomial beta function.

$$B(\alpha) = \frac{\prod_{i=1}^{k}\tau(\alpha i)}{\tau(\sum_{i=1}^{k}\alpha i)} \tag{8}$$

and $\alpha = (\alpha_1,\ldots,\alpha_k), x_1 + x_2+\ldots+x_{k-1} + x_k = 1, x_1，\ x_2,\ldots,x_{k-1} > 0$， and on simplex of dimension (k-1), the probability density of the other regions is 0.

(7) Gibbs sampling: Regarding the estimation of the parameters of the document-subject and the parameters of the subject-word, this paper uses the Gibbs sampling method to determine, this paper needs to solve the conditional probability distribution of each feature dimension, then this article must obtain the joint distribution of the topic Zm, n and the vocabulary Wm, n, and then obtain the probability distribution P(W) of a word under the corresponding theme, until

Gibbs sampling converges, at this time this paper finds the theme of all words in turn and obtains the distribution of each topic word, Then, the number of topics is calculated to obtain the topic distribution of each document.

This paper aims to study the dynamic changes in the topic preferences of stock investors by dividing the time period of the documents published by users in the securities community[9]. The sliding method of time window is used to divide the document into different time periods according to its characteristics. There are two ways to split documents: one is equal time slicing, that is, a fixed time span. This method may lead to an uneven distribution of the number of documents, if the number of documents in a certain time slice is small, which will affect the user portrait modeling effect; The second is to divide the number of equal documents, that is, the number of fixed documents. This approach may overlook more nuanced changes in the user portrait because the time span of the document set may be too large[4].

In order to make better use of the time dimension to segment user documents, this paper adopts two combined methods. First, this article slices the document into equal time periods and sets a threshold. If the number of documents in one time period does not reach the threshold, it is automatically merged into the next time period. However, given the continuity of user preferences, this article can only exchange documents within the same time slice, which may lead to poor continuity of the excavated topics[12]. So this article ensures the continuity of the theme by adding a smoothing factor lambda. To expand it, the prior probability of the subject-term in the t −time period comes from the posterior probability of the subject-term output in the $t-1$time period, and is achieved using the smoothing process in Equation 9. The topic preference mining graph model of stock investors based on dynamic LDA is shown in Figure 2.

$$\varphi_t = \lambda\varphi_{t-1},\ \lambda = \frac{a\mathrm{N}_t + (1-\mathrm{a})\mathrm{N}_{t-1}}{\mathrm{N}_{t-1}} \tag{9}$$

where φ_t is the probability distribution of terms in the t time period, N_t is the total number of terms in the $t$time period, and a is the custom parameter to adjust the proportion of the number of terms in the t time period and the $t-1$ time period.

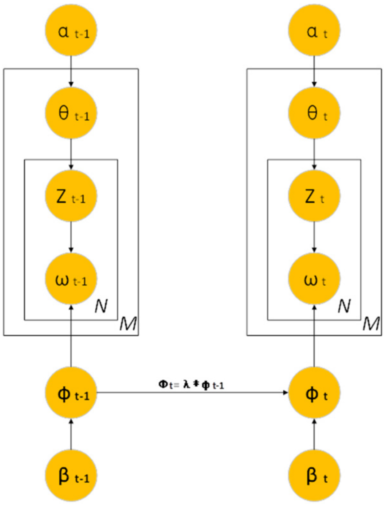


Figure 2 Dynamic LDA Graph Model for Mining User Topic Preferences in Securities Com-munity

Among them, θrepresents the probability distribution of the topic, Z represents the topic, ω represents the word item, φ represents the probability distribution of the word item, M represents the total number of documents, and N represents the total number of word items in the m-th document.

Dynamic topic modeling generation process for stock investors' topic preferences:

Step1: According to the set time dimension segmentation method, the document is divided into document sets in t time slices.

Step2: Randomly extract a time slice t.

Step3: Determine whether t is the first time slice:

1) t is the first time slice, that is $t-1$, then:

a) For selected document d, extract its topic probability distribution $\theta_1 \colon P(\theta_1|\alpha_1)$

b)For each term in document d, select a subject: $Z_1 \colon P(Z_1|\theta_1)$, generate each term $\omega_1 \colon P(\omega_1|Z_1, \beta_1)$;

2) t is not the first time slice, i.e. t ≠ 1, then:

a) Calculated according to Equation $\varphi_t = \lambda\varphi_{t-1}$;

b) For selected document d, extract its topic probability distribution;

c)For each term in document d, select a subject$Z_t \colon P(Z_t|\theta_t)$, generate each term$\omega_t \colon P(\omega_t|Z_t, \beta_t)$.

There are many methods for parameter estimation of LDA topic models, but Gibbs sampling has attracted widespread attention because of its advantages of conciseness and easy implementation. Gibbs sampling method uses the conditional distribution sampling method to simulate the joint distribution, and then the simulated joint distribution is used to obtain the conditional distribution. This process loops continuously until it reaches a convergent state. Gibbs sampling is very effective in dealing with some multivariate probability distributions that are difficult to sample directly.

Gibbs Sampling learning process based on LDA model:

Step 1: Initialize the subject $Z_0$for each word item w in the document;

Step 2: Calculate the number of terms w under each topic, and the number of terms in topic Z in each document d_i;

Step 3:compute $P(Z_{i=}j|Z_{-t}, w_i$,That is, the subject of the current word is excluded, and a new topic Z_1 is extracted for that term;

Step 4: Repeat Step 3 until the two probability distributions of document-subject and subject-term converge;

Step 5: Enter the parameters to estimate the values of φ and θ.

Because this paper time-slices the document and adds a smoothing factor to the modeling process, the probability distribution of the $t-1$time period will directly affect the probability distribution of the t time period, therefore, in the Gibbs sampling process of this paper, the posterior probability of the time period t $P(Z_{i=}j|Z_{-i}, w_i, d_i \cdot)$is calculated as shown in Equation 10.

$$P(Z_{i=}j|Z_{-i}, w_i, d_i \cdot) = \frac{\dfrac{(n_{-i,j}^{(w_i)})_t + v(n_{-i,j}^{(w_i)})_{t-1} + \beta(n_{-i,j}^{(d_i)})_t + \alpha}{(n_{-i,j}^{(\cdot)})_t + v(n_{-i,j}^{(\cdot)})_{t-1} + V\beta(n_{-i,\cdot}^{(d_i)})_t + T\alpha}}{\sum_{j=1}^{T} \dfrac{(n_{-i,j}^{(w_i)})_t + v(n_{-i,j}^{(w_i)})_{t-1} + \beta(n_{-i,j}^{(d_i)})_t + \alpha}{(n_{-i,j}^{(\cdot)})_t + v(n_{-i,j}^{(\cdot)})_{t-1} + V\beta(n_{-i,\cdot}^{(d_i)})_t + T\alpha}} \tag{10}$$

where $Z_{i=}j$ means to assign j to the vocabulary $w_i$as the subject, and Z_{-i} means the subject of all words except the current vocabulary w_i; bullet represents all other known or visible information, including all other words W_{-i}, documentation d_{-i}, and hyperparameters α and β; $v(n_{-i,j}^{(w_i)})_{t-1}$ represents the sum of the number of words assigned to subject$j$and vocabulary$w_i$in the $t - 1$ time period; $v(n_{-i,j}^{(\cdot)})_{t-1}$represents the sum of all vocabulary assigned to topic j during the $t - 1$ time period; $n_{-i,j}^{(w_i)}$ represents the sum of the number of words assigned to topic j with the same vocabulary w_i, $n_{-i,j}^{(\cdot)}$ means the sum of all vocabulary quantities assigned to topic j, $n_{-i,j}^{(d_i)}$represents the sum of the number of words assigned to topic j in the document d_i, $n_{-i,\cdot}^{(d_i)}$ means that the $Z_{i=}j$ is removed from the document d_i the sum of the number of words assigned to the topic for all the other . Based on this, the values of φ and θ can be reestimated for the document set within each time slice, as shown in Equations 11 and 12.

$$(\varphi_w^{(Z=j)})_t = \frac{\left(n_j^{(w)}\right)_t + v(n_j^{(w)})_{t-1} + \beta}{\left(n_j^{(\cdot)}\right)_t + v(n_j^{(\cdot)})_{t-1} + V\beta} \tag{11}$$

$$(\varphi_{Z=j}^{(d)})_t = \frac{\left(n_j^{(d)}\right)_t + \alpha}{\left(n_\cdot^{(d)}\right)_t + T\alpha} \tag{12}$$

where the vocabulary w represents the unique vocabulary, $n_j^{(\cdot)}$represents the sum of all the words assigned to subject j in the document set within the time slice, $n_j^{(w)}$ represents the sum of the number of times the word w in the document set in the time slice is assigned to subject j, and $n_j^{(d)}$ represents the sum of all the words assigned to subject j in a document d in the document set within the time slice. The $n_\cdot^{(d)}$represents the sum of all the words assigned to topics in document d.

3.2 Word2vec word vector model

The LDA model does not consider the relationship between words and context logic, while the word vector model emphasizes describing the relationship between words and pays more attention to the context context, making the word vector more relevant to semantics. Different from the LDA theme model, the Word2vec model uses local features to predict words, thereby preserving the semantic relationship between words and showing excellent generalization ability[10].

Word2vec is a model for generating word vectors, the algorithm can map words into a continuous vector space, such that in this vector space, the distance between semantically similar words is close. Word vectors are one of the important technologies in natural language processing, which can capture the semantic and grammatical relationships between words, providing strong support for tasks such as text analysis, sentiment analysis, and text classification[11]. The core idea of the Word2vec model is to learn the vector representation of words through the contextual information of the vocabulary. Specifically, Word2vec trains a neural network model to predict the context of a word itself when given the context of the word (CBOW model), or the context of a word given it (skip-gram model).

Both the CBOW model and the Skip-gram model training process relies on large-scale text corpus, which continuously adjust the word vector to achieve optimal performance on the prediction task[8]. After the training is completed, the Word2vec model can generate a high-dimensional vector representation for each word, so that the geometric distance of semantically similar words in the vector space is closer. The architecture diagram of the two types of Word2vec training model is shown in Figure 3.

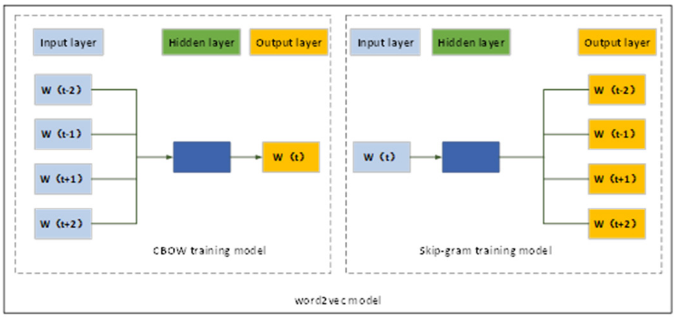


Figure 3 Architecture diagram of Word2vec training model

CBOW model: The CBOW model is a contextual lexicon-based word vector generation method that uses contextual lexicon to make predictions on target vocabularies. Specifically, the CBOW model takes the background window of a specific word (such as the vocabulary in front of the vocabulary and the vocabulary behind the vocabulary) as input, and uses a neural network to train it, so as to realize the prediction of the vocabulary in the center of the background window[6].

Skip-gram model: The Skip-gram model is the opposite of the CBOW model, which uses target words to predict contextual words coming from the corpus. Specifically, given a word, the goal of the Skip-gram model is to make predictions by making predictions about the vocabulary of the vocabulary in the context window.

Overall, the LDA topic model is used to learn the vector representation of a document and thus predict the vocabulary in the document, but it does not consider the correlation between words in a local scope. In contrast, the Word2vec word vector model predicts its neighboring words given the vocabulary and generates word vectors by repeating the process over and over again. The main differences between the LDA topic model and the Word2vec word vector model are shown in Table 1. In this paper, it is intended to combine the LDA topic model and the Word2vec word vector model to mine unstructured text information of stock investors. Using the LDA topic model, the relationship between topics and words will be described from a global perspective and the subject information of documents will be captured. Through the Word2vec word vector model, words will be predicted in the form of local features, and the semantic relationship between words will be preserved. Such comprehensive application will give full play to the advantages of LDA and Word2vec models to analyze and understand text information from two dimensions: global and local.

Table 1 Main differences between the LDA topic model and the Word2vec word vector model

model	LDA topic model	Word2vec word vector model
output	Document-Subject and Subject-Term two probability distributions	The word vector corresponding to the word
training	Using the co-occurrence relationship of words in documents, the vocabulary is clustered by theme, and the document-vocabulary matrix is decomposed into two matrices: document-subject and subject-vocabulary	A three-layer neural network is used to learn a "top-down-vocabulary" matrix, where the context consists of several nearby words, from which a representation of word vectors is obtained
function	Generate implicit topics that describe the relationship between words and topics, topics, and documents globally	Local features are used to predict words, and the semantic relationship between words is preserved

3.3 LDA2vec model

3.3.1. LDA2vec model construction

LDA2vec is an extension of word2vec and LDA, which is a model for learning vocabulary, documents, and topic vectors together. LDA2vec is built on the skip-gram model of word2vec, which is mainly used to generate vocabulary vectors. Skip-gram and word2vec are essentially neural networks that learn word embeddings by using input words to predict the surrounding context. In the LDA2vec model, it processes document-sized pieces of text and decomposes the document vector into two different components, which is a step forward on the paragraph vector approach. Following the same principle as the LDA model, LDA2vec decomposes document vectors into docu-ment weight vectors in a topic matrix. Document weight vectors represent the alloca-tion of proportions of different topics, while topic matrices are composed of different topic vectors. Therefore, LDA2vec constructs the context vector by combining the dif-ferent subject vectors that appear in the document[2].

Each topic has a distributed representation vector that is in the same vector space as the word vector. This means that topic vectors and vocabulary vectors coexist in the same vector space, so that topic vectors can be directly compared and related to specific vocabulary vectors. However, not every topic vector has to consist of specific words that are expected, and some topic vectors may be semantically similar to other words[17]. For example, one topic vector might be associated with words like "pitching," "catcher," "braves," and so on, while another topic vector might be associated with words like "Jesus," "God," "faith," and so on. Each document vector is a weighted sum of topic vectors.

The purpose of calculating the negative sampling loss is to distinguish between observed context-objective pairs (German+document, airline) and negative-sampling pairs (German + document, bear).

Word2vec's skip-gram negative sampling (SGNS) target is modified to utilize document-wide feature vectors while learning sequential document weights loaded into topic vectors. The total loss clause L is the sum of the negative sampling loss (SGNS) of the jump graph L_{ij}^{neg} adds the Dirichlet likelihood term to the document weight, L_d the loss is made by the context vector, $\vec{c}_j$, the pivot word vector vec $\overrightarrow{w_j}$, the target word vector $\vec{w}_i$, and the negative sampled word vector $\vec{w}_l$:

$$L = L^d + \Sigma_{ij} L_{ij}^{neg} \tag{13}$$

$$L_{ij}^{neg} = \log_{\sigma}\left(C_j \vec{w}_i\right) + \sum_{l=0}^{n} \sigma\left(-\vec{c}_j \vec{w}_l\right) \tag{14}$$

By using LDA2vec, contextual vocabulary is predicted in a different way than traditional methods. In LDA2vec, context vectors are used for vocabulary prediction instead of using lexical vectors directly. A context vector is made up of two different vectors: a vocabulary vector and a document vector. Vocabulary vectors are generated by the skip-gram word2vec model to capture semantic relationships between words[3]. A document vector is a weighted combination of two components. First, the document weight vector represents the weight distribution for each topic in the document. Second, the topic matrix represents each topic and its corresponding vector embedding. In this way, document vectors are able to consider both the topic distribution of the document and the semantic representation of the topic[8]. Document vectors and vocabulary vectors work together to generate context vectors for each vocabulary in the document. These context vectors are used in the input-hidden layer process of the LDA2vec model, enabling prediction of the context vocabulary. The advantage of LDA2vec is that it is able to learn not only word embeddings of vocabulary (and embeddings of context vectors), but also the representation of topics and documents. Through this integrated learning approach, LDA2vec can more comprehensively capture semantic and thematic information in text data, providing richer and more accurate vector representations for natural language processing tasks.

The LDA topic model describes the relationship between topics and terms, generating implicit topics, but using the bag-of-word model assumption, words are independent of each other, resulting in ignoring the relationship between terms, thus affecting modeling accuracy. The word2vec word vector model predicts word terms from local features, which has good generalization performance while maintaining the semantic connection between words. Therefore, in the case of unstructured text, this paper combines the global implicit information in the LDA theme model with the local implicit information in the Word2vec word vector model, and uses the whole corpus information and the contextual implicit semantic space of the word to make the text modeling more accurate and effective[14].

This paper combines the dynamic LDA topic model with the Word2vec word vector model to build a dynamic mining model of stock investor topic preference based on FNS-LDA2vec, the framework of which is shown in Figure 4.

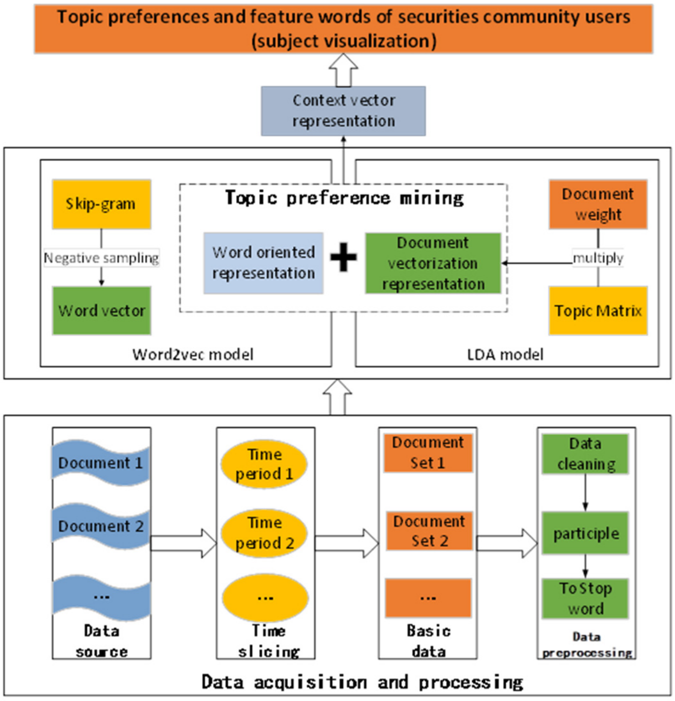


Figure 4 Dynamic Mining Model for quity Investor Topic Preference Based on FNS-LDA2vec

As can be seen from Figure 4, the dynamic mining model of stock investors' topic preference based on FNS-LDA2vec is mainly divided into three parts, namely text data acquisition and processing, stock investor topic preference (theme) mining, and topic visualization on this basis.

3.3.2. Stock bar investor topic preference (theme) mining

The FNS-LDA2vec model proposed in this paper is mainly composed of two core parts. Firstly, this paper improves the LDA topic model and obtains the subject-term distribution matrix and document weight. Secondly, this paper uses the skip-gram model to derive word vectors based on semantic relationships. Secondly, this paper adds the document vector and the word vector to obtain a context vector containing the semantic relationship between words. This vector is able to more accurately express the subject of the document[20].

Word vectorized representation

The word vector is mainly trained by the Skip-gram model in the Word2vec word vector model, which contains two parts, namely modeling and acquiring word vectors[19]. The modeling process of the skip-gram model is similar to that of auto-encoder, first using the training data to build a neural network, encoding and compressing the input in the hidden layer, decoding the data to the initial state in the output layer, and when the model training is completed, remove the output layer, retain the hidden layer, and obtain the parameters learned by the model during the training process (such as the weight matrix of the hidden layer), which is the word vector that this article needs to learn[13].

Step1: Use each word in the sentence as an input word in turn to build a training word pair;

Step2: The Skip_window parameter and num_skips parameter are defined as follows: the skip_window parameter is used to specify the number of words selected from the left or right of the current input word; The num_skips parameter specifies the number of output words selected from the entire window.

Step3: Based on the training data provided, the neural network will generate a probability distribution that represents the likelihood that each word in the dictionary will appear at the same time as the output word.

Predicting output words by input words can be seen as a multi-classification problem, and in natural language processing, predicting output words usually uses the softmax function to calculate the probability of output words. However, due to the large number of words in the vocabulary, the amount of computation using the softmax function will be large, resulting in serious impact on computational efficiency. In order to reduce this effect, the Negative sampling (NEG) method can be used to estimate the error, thereby reducing the amount of calculation. Negative sampling is a technique that can be used to improve the speed of model training and the quality of word vectors. Compared to the original method of updating all weights per training sample, negative sampling makes each training sample update only a small part of the weights, which greatly reduces the amount of computation. In negative sampling, the sampling probability is determined according to the word frequency, and for small-scale datasets, 5-20 negative samples are more effective. For large-scale datasets, only 2-5 negative samples can be selected. The probability of sampling each vocabulary is calculated as shown in Equations 15.

$$len(w) = \frac{count(w)^{3/4}}{\sum_{u\epsilon vocab} count(u)^{3/4}} \tag{15}$$

where $count(w)$is the word frequency of the word w, $u$is the word frequency in the vocabulary $vocab$, and sfrac$^{3}/_{4}$in the formula is empirical, and empirically shows that this formula is far more effective than other formulas.

Document vectorized representation

The document vector is mainly obtained by the product of the subject-term matrix and document weight trained by the improved LDA topic model, and the specific construction and training method of the improved LDA topic model is described above. Among them, the document weight a_{jk} refers to the probability that document J belongs to the subject k, and the value range of a_{jk} is generally between 0~1 to ensure the credibility of the research results. After improving the LDA theme model and obtaining the document vector representation of the corresponding document, and then adding it to the corresponding word vector obtained by the skip-gram model, and the vector after adding the two is used as the initial value of the context vector, the initial value calculation method of the corresponding context vector of the word j is shown in Equation 16.

$$\overrightarrow{c_j}=\overrightarrow{w_j} + \overrightarrow{d_j},\overrightarrow{d_j} = a_{j0} \cdot \overrightarrow{t_0} + a_{j1} \cdot \overrightarrow{t_1} + \cdots a_{jk} \cdot \overrightarrow{t_k} \tag{16}$$

where $\overrightarrow{w_j}$ represents the word vector of the word j, and the aforementioned $\overrightarrow{d_j}$ represents the vector representation of all word-context pairs for the word j, which makes the mixing vector easier to understand; $\overrightarrow{t_k}$ represents the vector representation of the corresponding topic k, which is consistent with the length of the word vector obtained by matrix decomposition after obtaining the subject-term matrix according to the improved LDA topic model. The topic vector representation in this article is common to all documents, but the specific topic distribution in different documents is determined by the document weight a_{jk}. On this basis, after obtaining $\overrightarrow{t_k}$, the subject word vector of the subject topic is extracted according to the similarity between the word vector and the subject's topic.

The objective function of this model

Similar to the Word2vec word vector model based on negative sampling, this method first randomly extracts word pairs from the (context j, vocabulary i) word pairs in the corpus an3 Topic visualizations and their measurements d from the context "negative" phrase. The objective function is shown in Equations 17:

$$L_{SGNS} = \sigma\left(\overrightarrow{c_j} \cdot \overrightarrow{\omega_i}\right) + \sigma\left(-\overrightarrow{c_j} \cdot \vec{w}_{negative}\right) \tag{17}$$

Secondly, by defining the Dirichlet possibility of document weight, the document weight is closer to the sparse Dirichlet distribution, which improves the convenience of weight measurement and the optimality of weight. Its calculation is shown in Equation 18:

$$q(d_k|\alpha) = \sum_k (a_k - 1) \log a_k \tag{18}$$

Finally, the objective function of this model, shown in Equations 19, represents the distinguished observed word-context pairs from negative sampling and adds regularization to the document weights:

$$L = \sum\nolimits_{wordpairs} \left[\sigma(\overrightarrow{c_j} \cdot \overrightarrow{\omega_i}) + \sigma(-\overrightarrow{c_j} \cdot \vec{w}_{negative}) + \sum\nolimits_{documents} \log q(d_k|\alpha) \right] \tag{19}$$

3.3.3Topic visualizations and their measurements

According to the topic preferences of stock investors, this paper can conduct topic mining and visually display the identified themes and their characteristic words. This visual display method can visually show the probability of different themes and the relationship between them, which is helpful for this paper to better analyze the themes preferred by stock investors. The traditional way to visualize topics is to use a tag cloud, where the size of the label represents its importance, and the importance is determined by the frequency or probability distribution of the label. In addition, this paper can use Python's built-in drawing function to make topic distribution and topic strength map, and can also use pyLDAvis toolkit and other tools to visualize the topic recognition results.

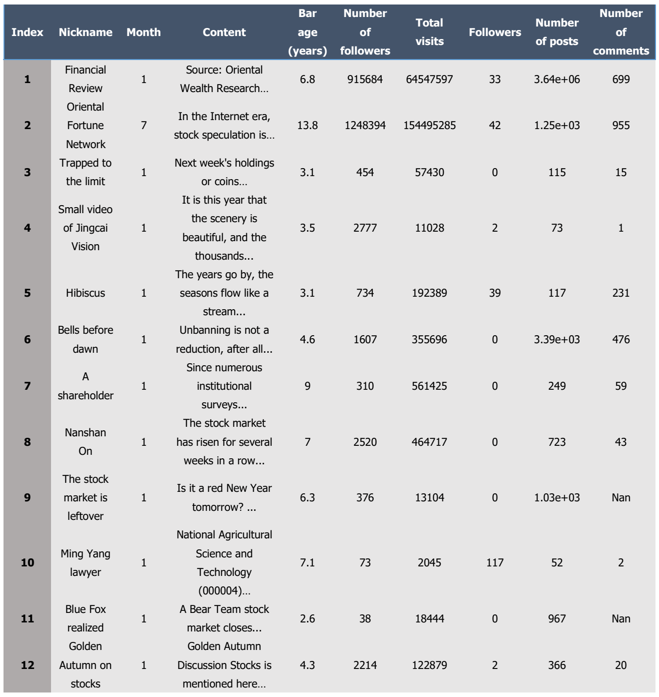
4 EXPERIMENT AND RESULT ANALYSIS

4.1 Selection and Processing of Data

The trading data of stock investors is usually stored in the trading systems of brokers, and due to the high confidentiality of this data, there is a lack of sufficient open source data. Therefore, this article only provides an effective analysis of social media-related concepts. Although the customer's trading behavior and the customer itself have relevant information about the broker, the bank customer customization model proposed in this article is not completely credible. Based on this, this paper connects the one-dimensional information of the three elements, and uses the methods of this paper to obtain a more complete user portrait of stock investors[[18]].

4.1.1Data source selection and data acquisition

Online media have the advantage of better communication skills and easier dissemination of information than traditional paper media. With the increasing popularity of the Internet and the reduction in the price of mobile phones, people began to be more inclined to obtain and share information through the Internet. In China, Oriental Fortune-Huaxia Stock is one of the most commonly used stock trading platforms, which provides professional services as a professional website. Therefore, in the second part of the Oriental Wealth Project website, this article selects "actual combat quotes" as the main source for users to obtain security information. The original data table (partial) is shown in Figure 4.

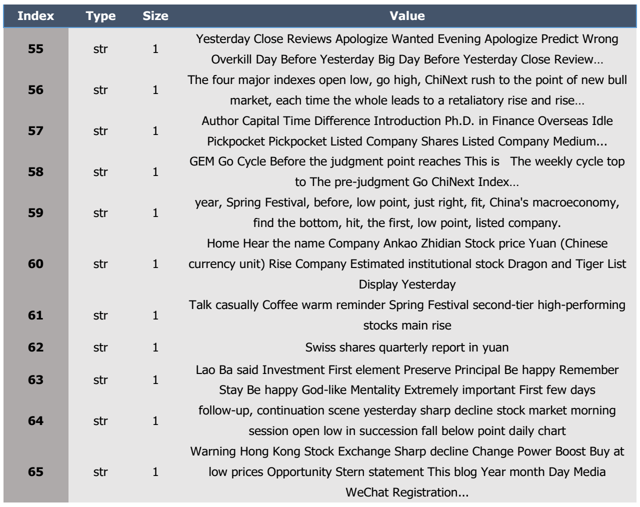


Index	Nickname	Month	Content	Bar age (years)	Number of followers	Total visits	Followers	Number of posts	Number of comments
1	Financial Review	1	Source: Oriental Wealth Research...	6.8	915684	64547597	33	3.64e+06	699
2	Oriental Fortune Network	7	In the Internet era, stock speculation is...	13.8	1248394	154495285	42	1.25e+03	955
3	Trapped to the limit	1	Next week's holdings or coins...	3.1	454	57430	0	115	15
4	Small video of Jingcai Vision	1	It is this year that the scenery is beautiful, and the thousands...	3.5	2777	11028	2	73	1
5	Hibiscus	1	The years go by, the seasons flow like a stream...	3.1	734	192389	39	117	231
6	Bells before dawn	1	Unbanning is not a reduction, after all...	4.6	1607	355696	0	3.39e+03	476
7	A shareholder	1	Since numerous institutional surveys...	9	310	561425	0	249	59
8	Nanshan On	1	The stock market has risen for several weeks in a row...	7	2520	464717	0	723	43
9	The stock market is leftover	1	Is it a red New Year tomorrow? ...	6.3	376	13104	0	1.03e+03	Nan
10	Ming Yang lawyer	1	National Agricultural Science and Technology (000004)...	7.1	73	2045	117	52	2
11	Blue Fox realized	1	A Bear Team stock market closes...	2.6	38	18444	0	967	Nan
12	Golden Autumn on stocks	1	Golden Autumn Discussion Stocks is mentioned here...	4.3	2214	122879	2	366	20

Figure 4 Original data of Oriental Fortune stock bar - stock market practice bar(part)

4.1.2Data Preprocessing

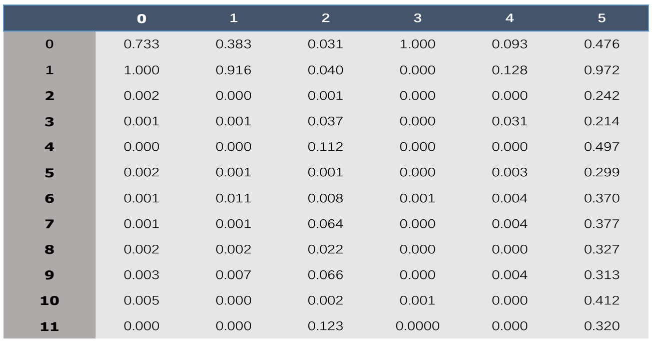
Since users contain a lot of noise information in the post, such as HTML, English, special symbols, etc., it is necessary to first Chinese the word segmentation and remove the end on the main body, so as to build a database at the back of the website for the investor's favorite website. Mine testing. In this experiment, the vocabulary is separated by selecting the appropriate vocabulary verification tool, and the closing word is extracted from the end vocabulary database. The four popular Chinese are the end. This dictionary also includes Chinese suffix table, HIT suffix table, Baidu suffix table, etc. The content (partial) of the user's post after the above operation is shown in Figure 5.



Index	Type	Size	Value
55	str	1	Yesterday Close Reviews Apologize Wanted Evening Apologize Predict Wrong Overkill Day Before Yesterday Big Day Before Yesterday Close Review…
56	str	1	The four major indexes open low, go high, ChiNext rush to the point of new bull market, each time the whole leads to a retaliatory rise and rise…
57	str	1	Author Capital Time Difference Introduction Ph.D. in Finance Overseas Idle Pickpocket Pickpocket Listed Company Shares Listed Company Medium...
58	str	1	GEM Go Cycle Before the judgment point reaches This is The weekly cycle top to The pre-judgment Go ChiNext Index…
59	str	1	year, Spring Festival, before, low point, just right, fit, China's macroeconomy, find the bottom, hit, the first, low point, listed company.
60	str	1	Home Hear the name Company Ankao Zhidian Stock price Yuan (Chinese currency unit) Rise Company Estimated institutional stock Dragon and Tiger List Display Yesterday
61	str	1	Talk casually Coffee warm reminder Spring Festival second-tier high-performing stocks main rise
62	str	1	Swiss shares quarterly report in yuan
63	str	1	Lao Ba said Investment First element Preserve Principal Be happy Remember Stay Be happy God-like Mentality Extremely important First few days
64	str	1	follow-up, continuation scene yesterday sharp decline stock market morning session open low in succession fall below point daily chart
65	str	1	Warning Hong Kong Stock Exchange Sharp decline Change Power Boost Buy at low prices Opportunity Stern statement This blog Year month Day Media WeChat Registration...

Figure 5 Post content of Guba users after text preprocessing (part)

Since the number of different data types is very different, this article first combines the number units of the website into "numbers", and then prepares for future cluster analysis through the comparison of six data types. Figure 6 shows the user account data in the above operation process (partial):



	0	1	2	3	4	5
0	0.733	0.383	0.031	1.000	0.093	0.476
1	1.000	0.916	0.040	0.000	0.128	0.972
2	0.002	0.000	0.001	0.000	0.000	0.242
3	0.001	0.001	0.037	0.000	0.031	0.214
4	0.000	0.000	0.112	0.000	0.000	0.497
5	0.002	0.001	0.001	0.000	0.003	0.299
6	0.001	0.011	0.008	0.001	0.004	0.370
7	0.001	0.001	0.064	0.000	0.004	0.377
8	0.002	0.002	0.022	0.000	0.000	0.327
9	0.003	0.007	0.066	0.000	0.004	0.313
10	0.005	0.000	0.002	0.001	0.000	0.412
11	0.000	0.000	0.123	0.0000	0.000	0.320

Figure 6 Normalized numerical data of Guba users (part)

The FNS-LDA2vec model constructed in this paper mainly includes two core parts, one is the topic-term distribution matrix and document weight obtained based on the improved LDA topic model, and the other is the word vector based on semantic relationship based on the skip-gram model. On this basis, the obtained document vector and word vector are added to finally obtain a context vector containing the semantic relationship between words and can more accurately represent the document subject.

The research content of this paper is divided into item extraction of FNS-LDA2vec and comparison with traditional LDA project extraction. This question is the most common evaluation method that can illustrate the subtraction effect of engineering models from multiple aspects. This article uses problem variables to determine the optimal sample size for subjectively extracting objects, and the corresponding calculation formula is shown in Equation 20:

$$Perplexity = exp[-\frac{\sum_{d=1}^{M} log(w)}{\sum_{d=1}^{M} N_d}] \tag{20}$$

The steps of FNS-LDA2vec based on FNS-LDA2vec stock bar user topic preference mining experiment are as follows:

Step 1: Representation of word vectors. Through the Skip-gram model in the Word2vec word vector model, the processing of text content is converted into K-dimensional vector calculation by using the context information of words. This step generally uses the Word2vec word vector template to generate vector words of words in the test group as subsequent entries. The test was implemented using the Gensim toolkit, and its relevant important parameter settings are shown in Table 2:

Table 2 Word2vec word vector model related parameter settings

parameter	value	meaning
sg	1	The trained model isskip-gram
size	100	Word vector dimension
window	5	Training window size
min_count	5	The dictionary truncates the lowest frequency
hs	0	The HS method is not used
negtive	5	Number of noise words

Step 2: Document vector representation. It adopts the dynamic LDA project mode (the duration of the time block is set to 3 months) and supports the verb matrix function of function parameters in the sklearn toolkit, using the LDA.inference class to calculate document weights and multiply them. Gets the document vector for subsequent insertions. The selection of the number of topics in this test is based on the following two criteria: the content needs of this article and the topic questions, and the topic confusion of the model in this paper is shown in Figure 7:

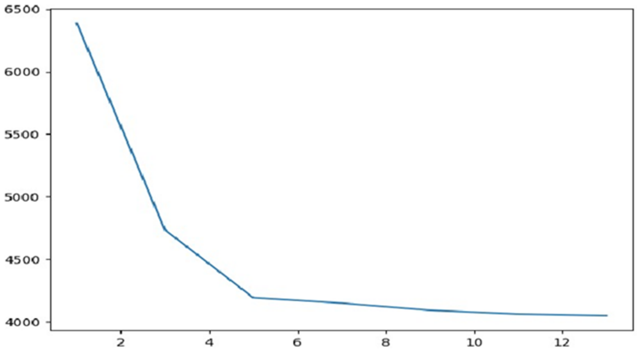


Figure 7 confusion degree (fine granularity) of this model

Considering that the distribution of topics should become one of the characteristics of user gathering, the number should not be too large. From Figure 7, it can be seen that after the number of questions k=5 in this chapter of the experiment simulation, the puzzle curve gradually stabilizes or flattens and maintains consistency, so the number k of questions in both the LDA model and the improvement model used in this chapter is set to 5.

Step 3: Iterate on the core algorithm. Combined with the above above, the most basic method of FNS-LDA2vec problem extraction model, the problem features are learned through noun vector and document vector, and iteration is achieved through the objective function.

4.1.3 Analysis of results

By using the wordcloud library, people can form a word cloud in the stock bar and visualize the preprocessed text information. Portfolio, technology, market, etc. are the three keywords used by stock bar users this year, and entrustment, company, sharing, creation, purchase, etc. are also more commonly used, which also means that stock bar users are most concerned about the problems in the actual operation of stock investment.

Through the word cloud graph, we found that the two theme words "market" and "company" were among the top two in each time period, with slightly different proportions in different time periods. However, the theme word 'shares' appears more frequently during a certain period of time, but its frequency is much lower during another period of time than the former; The above results fully indicate that users' topic preferences have continuity, but they may have different focuses during a specific period. When conducting thematic preference analysis, it is necessary to pay attention to the dynamic evolution of their preferences in order to more accurately understand and analyze users.

5 Conclusion

This article focuses on the unstructured data in the user indicator system, namely the content of user posts in securities related communities. Firstly, by adding time slices, dynamic topic mining of the LDA topic model is achieved; Then, the improved LDA model is combined with the Word2vec word vector model to jointly learn topic representation using document vectors and word vectors. The advantages of LDA preserving document level information and Word2vec preserving semantic relationships between words are fully combined to improve topic extraction efficiency. Based on the above experimental results, this article points out that the role of knowledge leaders in securities communities in marketing promotion should be fully utilized. Different textual data such as stock introductions, stock reviews, and stock names exert differentiated recommendation effectiveness. Using the extracted more accurate topics as the topic preferences of stock investors has certain reference value for investors and stock market analysts.

References

[1] Sun T. Stock bar public opinion and stock price synchronicity[J]. Financial Engineering and Risk Management,2023,6(4).

[2] Zhang T,Cui W,Liu X, et al. Research on Topic Evolution Path Recognition Based on LDA2vec Symmetry Model[J]. Symmetry,2023,15(4).

[3] Jianfeng X,Yunhe Z,Zhiqiang L, et al. A Recognition Method of Truck Drivers' Braking Patterns Based on FCM-LDA2vec[J]. International Journal of Environmental Research and Public Health,2022,19(23).

[4] Jingxian G,Yong Q. Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example[J]. Entropy,2021,23(10).

[5] Bo J. Digital Media VR Design Recommendation Analysis Relying on LDA Model[J]. Journal of Physics: Conference Series,2021,1992(2).

[6] Kang H,Yang J. Performance Comparison of Word2vec and fastText Embedding Models[J]. Journal of Digital Contents Society,2020,21(7).

[7] Xin Z, Hum Chang, Lee J U S. Collaborative Filtering Recommendation Algorithm Based on LDA2Vec Topic Model[J]. Proceedings of the Korean Computer Information Society, 2020, 28(2).

[8] Chuan C,Agres K,Herremans D. From context to concept: exploring semantic relationships in music with word2vec[J]. Neural Computing and Applications,2020,32(4).

[9] Zhang H,Nie J,Ruan Z. The Users Emotional Study of Netease Cloud Music Based on LDA Model[C]// Changchun Normal University,IEEE, Jilin University, Northeast Normal University.Proceedings of 2019 4th International Conference on Cloud Computing and Internet of Things(2019 CCIOT).2019:32-35.DOI:10.26914/c.cnkihy.2019.049461.

[10] Shao T,Chen H,Chen W. Query Auto-Completion Based on Word2vec Semantic Similarity[C]//Asia Pacific Institute of Science and Engineering.Proceedings of 2nd International Conference on Machine Vision and Information Technology (CMVIT 2018).IOP Publishing,2018:146-151.

[11] Gao J,He Y,Zhang X, et al. Duplicate Short Text Detection Based on Word2vec[C]//The Institute of Electrical and Electronics Engineers,IEEE Beijing Section.Proceedings of 2017 IEEE 8th International Conference on Software Engineering and Service Science.Institute of Electrical and Electronics Engineers,2017:557.

[12] Liu Y,Xu S. A local context-aware LDA model for topic modeling in a document network[J]. Journal of the Association for Information Science and Technology,2017,68(6).

[13] Moody E C. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec.[J]. CoRR,2016,abs/1605.02019.

[14] Xu Weijun, Huang Jinglong. Research on Black-Litterman portfolio model based on financial text sentiment mining Evidence from the posting text of eastmoney stock forum and the A share market[J]. Operations Research Transaction, 1007-6093(2022)26:4<1.

[15] Xu Weijun, Peng Zijin. The Design of Mean Reversion Strategy Considering Investor Sentiment Basedon Textual Information---Evidencefrom the Posting Text of Eastmoney Stock Forum and the A-Share Market[J]. Operations Research and Management Science[J], 1007-3221(2022)31:3<19

[16] Cen Yonghua, Tan Zhihao .Impacts of Financial Media Information on Stock Market:An Empirical Study of Sentiment Analysis. Data Analysis and Knowledge Discovery[J], 2096-3467(2019)3:9<98.

[17] Zhang Weiwei, Hu Yaqi. Academic Abstract Clustering Method Based on LDA Model and Doc2vec.Computer Engineering and Application[J], 1002-8331(2020)56:6<180.

[18] Wei-Jia Q , Yan-Qing W . Empirical research on customer segmentation of securities based on clustering[J]. Journal of Computer Applications, 2010.

[19] Mikolov T , Sutskever I , Chen K , et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 201

[20] Zhang P , Yang T , Zhou C , et al. Modeling Multi-factor Sequential User Behavior Data over Social Networks[J]. Chinese Journal of Electronics, 2016, 25(2):364-371.