Computing and Informatics, Vol. 41, 2022, 689–713, doi: 10.31577/cai_2022_3_689

A NOVEL DATA ANALYTIC MODEL FOR MINING USER INSURANCE DEMANDS FROM MICROBLOGS

Chun Yan

College of Economics and Management
Shandong University of Science and Technology
Qingdao 266590, China
&
College of Mathematics and System Science
Shandong University of Science and Technology
Qingdao 266590, China
e-mail: yanchunchun9896@sina.com

Lu Liu

College of Economics and Management
Shandong University of Science and Technology
Qingdao 266590, China
e-mail: 617128404@qq.com

Wei Liu*

College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao 266590, China
e-mail: liuwei_doctor@yeah.net

Man Qi

School of Engineering, Technology and Design
Canterbury Christ Church University
CT1 1QU, UK
e-mail: man.qi@canterbury.ac.uk

Abstract. This paper proposes a method based on LDA model and Word2Vec for analyzing Microblog users' insurance demands. First of all, we use LDA model to analyze the text data of Microblog user to get their candidate topic. Secondly, we use CBOW model to implement topic word vectorization and use word similarity calculation to expand it. Then we use K-means model to cluster the expanded words and redefine the topic category. Then we use the LDA model to extract the keywords of various insurance information on the "Pingan Insurance" website and analyze the possibility of users with different demands to purchase various types of insurance with the help of word vector similarity. Finally, the validity of the method in this paper is verified against Microblog user information. The experimental results show that the accuracy, recall rate and F1 value of the LDA-CBOW extending method have been proposed compared with that of the traditional LDA model, respectively, which proves the feasibility of this method. The results of this paper will help insurance companies to accurately grasp the preferences of Microblog users, understand the potential insurance needs of users timely, and lay a foundation for personalized recommendation of insurance products.

Keywords: LDA model, Word2Vec, insurance demand, preferences

1 INTRODUCTION

With the rapid development of the Internet financial industry, a new insurance management method, Internet insurance, came into being. In the era of big data, comprehensive market expansion is taking place in Chinese Internet insurance. At the same time, the focus of business competition is gradually shifting from "products" to "customers". In order to enhance its core competitiveness, an insurance company must keep abreast of customer dynamics and deeply tap the potential insurance needs of customers. As a social media platform, Microblog has the advantages of concise information, open interaction, and fast communication speed, and is favored and supported by the majority of netizens. Microblog provides netizens with a platform for understanding entertainment news and interacting with celebrities. At the same time, the Microblog users' behavior of publishing, reposting, commenting can also reflect their personalities, preferences and demands indirectly. In the context of smart insurance, insurance companies should vigorously develop advanced science and technology such as big data, cloud computing, and artificial intelligence to make them ingeniously integrated with insurance marketing, and strive to expand service areas, broaden sales channels, and improve service levels so that the competitiveness of insurance companies can be improved.

In the current research, many scholars have studied the data in the Microblog platform, but their research directions are mostly focused on public opinion analysis [1, 2, 3, 4, 5], sentiment analysis [6, 7, 8, 9, 10], social network analysis [11, 12, 13]

[*] Corresponding author

and Microblog content recommendation [14, 15, 16]. As for the research in the insurance field, most of the literature is based on insurance product pricing [17, 18, 19, 20], fraud identification [21, 22, 23], and risk management [24, 25]. However, there are few research to extract user insurance demands from Microblog content by natural language processing methods. In this paper, we innovatively combine the topic model and the word vector model to analyze the potential insurance demands of Microblog users. First of all, the Latent Dirichlet Allocation (LDA) topic model is used to obtain the candidate topics in the Microblog content, and the word vector model training corpus is used to obtain the trained Continuous Bag-of-Words (CBOW) model; secondly, we use the CBOW model to obtain the candidate topic word vector and use the word vector correlation to expand the word vector data set; then we use the K-means algorithm [26] to cluster the expanded word vector data set and redefine the topic category, and compare the results with the original Microblog text theme. Finally, we extract keywords of various insurance protection content from the "Pingan Insurance" website as a classification basis, and the keyword vectors are tested for similarity between the keywords and the subject words in the redefinition subject to obtain the possibility of users to purchase various insurance types, and finally achieve the purpose of mining the insurance needs of Microblog users. The specific contributions are highlighted below:

- The method we proposed can simultaneously use the global semantic extraction capabilities of the LDA model and the local semantic analysis capabilities of the word vector model. While solving the problem of semantic sparsity of input text, it can improve the accuracy of text topic extraction.
- While extracting keywords in the text content of Microblog users, we also extracted keywords based on the feature descriptions of different types of insurance on the Internet insurance website. By matching the two types of keywords, we analyzed the possibility of users buying different types of insurance, which is an insurance product. This provides support for the personalized recommendation of insurance products.

This paper is organized as follows. Section 2 introduces the related literature of theoretical knowledge. Section 3 presents the system framework, details the relevant theories of the two models used, and combines the LDA model with the CBOW model to achieve text expansion. Section 4 is an empirical analysis of data by the proposed method. Section 4.5 takes user A as an example to verify the effectiveness of the proposed method. Section 5 concludes the paper.

2 RELATED LITERATURE

The literature review in this section is divided into three parts: analysis and research on customer demands, analysis on text data by topic model and analysis on text data by word vector model. Next, we will introduce them in turn.

2.1 Analysis and Research on Customer Demands

In the era of big data, the insurance industry is transitioning from “product-centric” to “customer-centric”. The customer’s demand is the key to insurance product sales. The demands of each customer will be different in types and levels due to their own subjective conditions and objective conditions such as the environment. Excavating the demands of customers in depth helps to achieve the company’s goals of precise marketing and personalized services to consumers. Chen et al. used a language processing tool to preprocess the extracted demand information to help customers to express their demands more accurately, and proposed a new method of demand expression [27]. Jiang et al. used user online comments as input data to analyze user demands, so that the accuracy of the obtained demands was improved [28]. Dou and Zong added hesitation calculation on the basis of interactive genetic algorithm, the result showed the uncertainty of user demands can be solved [29]. Huo adjusted the marketing model to be “customer demand oriented” in order to optimize the business model of insurance companies [30]. Yuan et al. established a multi-objective genetic algorithm, which can quickly configure the optimal product according to customer demands [31]. Qiu et al. proposed the DFSI model to describe customer demands. Experimental results show that this method is more accurate than other traditional model prediction results [32]. Yu et al. combined the information in the platform to extract task features and analyze the user’s interest perception features to realize the developer’s personalized recommendation [33].

2.2 Analysis on Text Data by Topic Model

As an important model for semantic analysis, the topic model has the functions of collecting, classifying and dimensionality-reducing text according to the topic. Among them, the LDA model is the most common topic model. The LDA model is a three-layer Bayesian probability model with a structure of “document-topic-word”. This model was proposed by Blei et al. and was developed based on the probabilistic hidden semantic analysis (PLSA) model [34]. In recent years, more and more scholars have applied the LDA model and its improved methods to the interest mining of Microblog users. Gerrish and Blei extended the LDA model to obtain the DIM model, which was able to identify the most influential documents in the collection [35]. Weng et al. used the LDA model to process the integrated Microblog user history information to obtain the interest preferences exhibited by users [36]. Ramage et al. expanded the LDA model to four dimensions and obtained the Labeled LDA model, which improved the recommendation effect [37]. Xu et al. used the LDA model to cluster hashtags on Microblog users to implement the function of recommending friends to Microblog users based on similar tags [38]. Zhang et al. constructed an improved LDA model based on user tags based on user microblogging behavior [39]. Wang combined the social curatorial network with the LDA model to extract the user’s potential interests, and obtained the user community, and made effective user recommendations [40]. Wang and Li used the LDA model to analyze online hotel reviews, making online user reviews more expressive [41]. Zhang and Eick used the LDA model to analyze the evolution of Twitter topics [42]. Bao et al. proposed a text classification method based on topic model and transfer learning to provide a new perspective for research in the field of text mining [43]. Li et al. model the conversations or comments in online social networks, divide them into different topics, and combine the time dimension to achieve social network group division [44]. Li et al. improved the LDA model. They used the textual information in financial reports as the research object to identify important risk points faced by the insurance industry and analyze their evolutionary laws. It could help regulators and insurance companies judge current and future risks and make effective risk supervision, prevention and control [45]. Li and Zhao took the LDA model as an example to sort out the application and extension research of the topic model. They introduced the advantages and disadvantages of the LDA model for mining text topics. At the end of the paper, they analyzed the development trend of topic models [46]. This article analyzes the shortcomings of the LDA model in processing short texts and introduces the Word2Vec model to improve the accuracy of Microblog text topic mining.

2.3 Analysis on Text Data by Word Vector Model

The above methods all belong to the model of research at the level of "text" granularity. In recent years, the method of research at the level of "word" granularity has received widespread attention. The basic idea of this method is: express words with vectors firstly, then stack word vectors to form short text vectors. This method can reduce the dispersion of short text topics and improve the focus of short text. Nikfarjam et al. used social network word vectors and text clustering to extract adverse drug reactions to test the level of public health surveillance [47]. Lilleberg et al. used word vector and support vector machine methods to classify the text, and obtained good classification results [48]. Xia used the word vector model to transform Wikipedia Chinese data and clustered it, and applied the results to TextRank keyword extraction to improve the extraction effect [49]. Zhou and Zhang converted online user comments into word vector tables and clustered candidate attribute word sets to obtain fine-grained product attribute sets [50]. Vargas-Calderón and Camargo used citizen tweets as input data and used Word2Vec to obtain potential topics [51]. Zhang and Liu combined word vectors with the BTM topic model to analyze the evolution of Microblog topics, improving the prediction accuracy of topic development [52]. Gu et al. used the comment data from the automotive industry as an example. They used the Word2Vec model to construct a set of product feature words, and realized the recognition of user comment topic features. This approach can help companies improve their competitiveness [53].

It can be seen that the LDA model has great potential for mining information of users. Using the LDA method has strong operability and practicality to analyze the user's Microblog globally to obtain potential needs, while the word vector model is better at obtaining local semantics and can solve the problem of poor semantic sparseness of short text on Microblog. There is a strong complementary relationship between the two methods. This paper takes Microblog users' behavior data on the platform as input, and combines the traditional implicit Dirichlet distribution (LDA) model with the Continuous Bag-of-Words (CBOW) model as an expansion method to analyze customer insurance preferences and a potential demand. We also take Microblog user A as an example to verify the LDA-CBOW model's ability to mine customers' potential insurance needs, and lay the foundation for personalized insurance recommendation.

3 INTRODUCTION OF LDA MODEL AND WORD2VEC EXTENDING METHOD

This paper aims to use the topic analysis capability of the LDA model and the similarity calculation function between words in the CBOW word vector model to achieve the expansion of potential keywords of Microblog users, and by clustering the expanded keywords, the accuracy of mining potential demands of Microblog users can be improved.

3.1 Theoretical Framework

The research framework of this paper is: firstly, train the corpus to obtain the corresponding word vectors; then, use the topic model to mine the candidate topics of Microblog text as a control group; then construct the candidate topic word vectors, and use the word vector model to calculate the similarity between the words. It is expanded to obtain an expanded word vector data set; then, the expanded word vector data set is clustered to obtain a redefined topic, and the obtained theme is compared with the control group; finally, the word vector is used to calculate the similarity between the keywords of various insurance protection content on the "Pingan Insurance" website and the words in the redefined theme to get the possibility of users with different demands to purchase various types of insurance, and analyze the insurance needs of users.

The specific process is shown in Figure 1.

3.2 Principle of LDA Theme Model

As a typical bag-of-words model, there is no succession between words in the LDA model. The document does not reflect the order of words, but only represents a simple combination of words. In general, each document corresponds to a different number of topics, and individual words in the document are generated by a certain topic. As an unsupervised learning algorithm, the LDA model does not require computer language training or artificial labeling in advance. It only needs to provide documents and specify the number of topics during data text training.

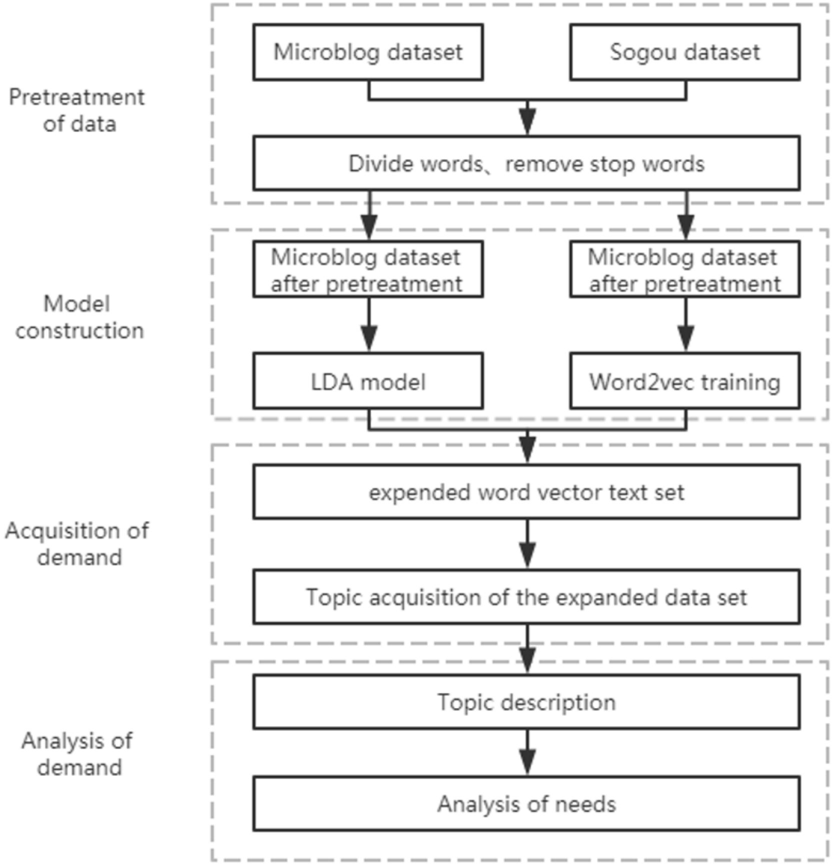

Pretreatment of data
Microblog dataset
Sogou dataset
Divide words、remove stop words
Model construction
Microblog dataset after pretreatment
Microblog dataset after pretreatment
LDA model
Word2vec training
Acquisition of demand
expended word vector text set
Topic acquisition of the expanded data set
Analysis of demand
Topic description
Analysis of needs

Figure 1. The theoretical framework based on LDA model and word vector expansion method

1. Analysis process of LDA topic model

The method of generating a document in the LDA model is shown in Figure 2.

In Figure 2, D represents the entire document collection, K represents the total number of topics in the document, N represents the total number of words in each document, the hidden variable Z represents a certain topic, ω is the word of the text, and the parameter α is the prior distribution hyperparameter of the document-topic probability θ, the larger the α, the more topics included in the document, and the less conversely; the parameter β is the prior distribution hyperparameter of the topic-word probability distribution φ, the larger the β, the more the words included in the topic. Where ω is the only observable variable.

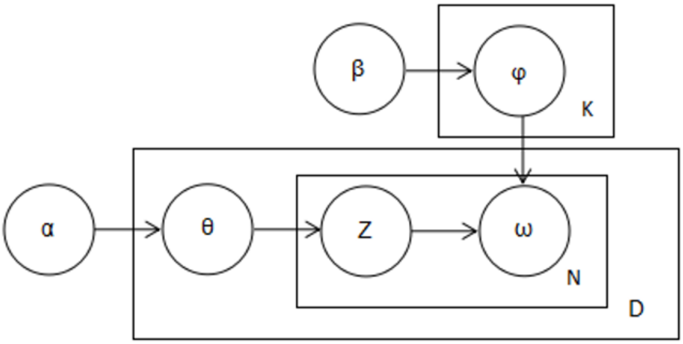


Figure 2. Bayesian network diagram of LDA topic model

The specific process is as follows:

- Firstly, select a document d_i according to probability;
- Secondly, the topic distribution θ_i of the document d_i is generated from the Dirichlet distribution with the hyperparameters;
- Thirdly, the topic $Z_{i,j}$ of the j word of the document d_i is sampled from the theme's polynomial distribution θ_i ;
- Fourthly, the distribution of words corresponding to the topic $Z_{i,j}$ is generated from the Dirichlet distribution with the parameter;
- Finally, sampling from word distribution to generate word $\omega_{i,j}$.

The joint probability distribution of Microblog topics can be expressed by Equation (1).

$$p(w_i, z_i, \theta_i, \varphi|\alpha, \beta) = \prod_{j=1}^{N} p(\theta_i|\alpha)p(z_{i,j}|\theta_i)p(\varphi|\beta)p(w_{i,j}|\theta_{z_{i,j}}). \tag{1}$$

In order to obtain the maximum likelihood estimate of the distribution of d_i words in the document, we integrate θ_i and φ in Equation (1) and sum the subject $Z_{i,j}$. After obtaining the maximum likelihood estimation results, it is necessary to perform parameter training on the LDA topic model.

2. Calculation of perplexity of LDA topic model

The division effect of the LDA model is directly determined by the number of topics, so we need to enter the appropriate number of topics. When the number of input topics is too small, the result of topic division will be too general and cannot fully reflect the topic characteristics of the data. At the same time, when the number of input topics is too large, the results of topic division will be repeated. This paper uses the perplexity index to determine the reasonable number of subject divisions.

Perplexity can measure the pros and cons of the language probability model. The smaller the perplexity, the better the clustering effect of the model. The definition formula of perplexity is shown in Equation (2).

$$perplexity(D) = \exp\left(-\frac{\sum \log P(\omega)}{\sum_{d=1}^{M} N_d}\right). \tag{2}$$

Among them, the denominator represents the number of all words in the test set; $P(\omega)$ represents the probability of each word in the test set, and the calculation formula is shown in Equation (3).

$$P(\omega) = P(z|d) * P(\omega|z). \tag{3}$$

Calculate the perplexity of the models trained on different topics, and select the number of topics with the smallest perplexity as the optimal number of topics.

3.3 Word Vector Construction Based on CBOW Model

Word vectors can express the semantic features of words in a digital way, and the semantic similarity between words can be reflected by the distance or similarity between the vectors. The traditional word vector representation method One-Hot coding cannot describe the similarity between words well, and this method requires a longer vector when representing a word, which is prone to dimensional disaster. The Word2Vec model overcomes the shortcomings of One-Hot model well. Since it was proposed by the Google team in 2013, it has been widely used in natural language processing.

Word2Vec is mainly divided into two categories: the CBOW model for guessing the target word from the original sentence and the Skip-gram model for guessing the original sentence from the target word. Combined with the actual situation of the article, this article selects the CBOW model for research.

1. Principle of CBOW model

The structure of the CBOW model is shown in Figure 3. From left to right, the model is a three-layer neural network model of the input layer, the projection layer, and the output layer.

The principle of the CBOW model is: set $\nu_i \in R$ as the vector of the word i, when i is a background word, set $\mu_i \in R$ as the vector of the word i, when i is a center word, c is the index in the dictionary of the center word ω_c, O_1, O_2, ..., O_{2m} is the index of background word ω_{O_1}, ω_{O_2}, ..., $\omega_{O_{2m}}$ in the dictionary, then the conditional probability of generating the central word given the background word is shown in Equation (4).

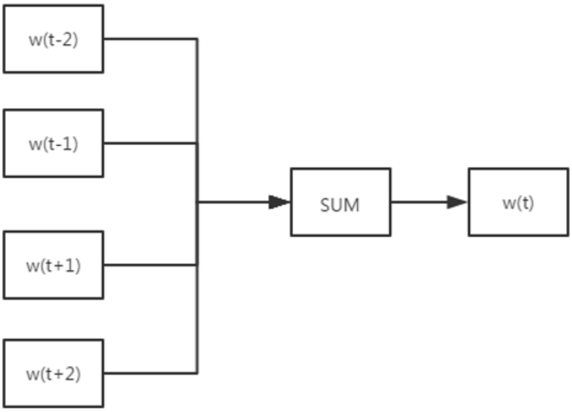


Figure 3. Structure diagram of CBOW model

$$P(\omega_c|\omega_{O_1},\ldots,\omega_{O_{2m}}) = \frac{\exp\left[\frac{u_c'(\nu_{O_i}+\cdots+\nu_{O_{2m}})}{2m}\right]}{\sum_{i\in\nu}\exp\left[\frac{u_i'(\nu_{O_i}+\cdots+\nu_{O_{2m}})}{2m}\right]}. \tag{4}$$

Set $W_O = \omega_{O_1},\ldots,\omega_{O_{2m}}$, $\bar{\nu_O} = (\nu_{O_1}+\cdots+\nu_{O_{2m}})/2m$. Then Equation (4) can be simplified as shown in Equation (5).

$$P(\omega_c|W_O) = \frac{\exp\left(\frac{u_c' * \bar{\nu_O}}{2m}\right)}{\sum_{i\in\nu}\exp\left(\frac{u_i' * \bar{\nu_O}}{2m}\right)}. \tag{5}$$

Given a text sequence of length T, let $\omega^{(t)}$ be the word of time step t, and the size of the background window be m. The likelihood function of the CBOW model is shown in Equation (6), that is, the probability that the background word generates an arbitrary central word.

$$P(\omega_c) = \prod_{t=1}^{T} P\left(\omega^{(t)}|\omega^{(t-m)},\ldots,\omega^{(t-1)},\omega^{(t+1)},\ldots,\omega^{(t+m)}\right). \tag{6}$$

2. Training of CBOW model

Since the maximum likelihood estimation of the model is equivalent to minimizing the loss function, the training of the CBOW model is shown in Equation (7).

$$L_{\log} = -\sum_{t=1}^{T} \log P\left(\omega^{(t)}|\omega^{(t-m)},\ldots,\omega^{(t-1)},\omega^{(t+1)},\ldots,\omega^{(t+m)}\right). \tag{7}$$

Due to

$$\log P(\omega_c|W_O) = \mu_c'\bar{\nu_O} - \log\left(\sum_{i\in\nu}\exp(\mu_c'\bar{\nu_O})\right). \tag{8}$$

Differentiate Equation (8), we can get the gradient of log of conditional probability with any background word vector $\nu_{O_i}(i = 1, 2, \ldots, 2m)$ in Equation (7), as shown in Equation (9).

$$\begin{aligned}\frac{\partial \log P(\omega_c|W_O)}{\partial \nu_{O_i}} &= \frac{1}{2m}\left(u_c - \sum_{j\in\nu}\frac{\exp(\mu_j' \cdot \bar{\nu}_O)\mu_j}{\sum_{i\in\nu}\exp(\mu_j' \cdot \bar{\nu}_O)}\right)\\ &= \frac{1}{2m}\left(u_c - \sum_{j\in\nu}P(\omega_j|W_O)\cdot \mu_j\right).\end{aligned} \tag{9}$$

3.4 Extension of Topic Word Vector and Definition of the Topic

The method of combining the topic model and the word vector model is as in Algorithm 1.

Algorithm 1 LDA-CBOW Algorithm

```
Input: Microblog text data
Output: Redefined topic categories
Input the Microblog text data into the LDA model to obtain candidate topics, and take the first $n$ keywords with the highest probability under the candidate topics to form a vocabulary $F_1 = \{\omega_{t1}, \omega_{t2}, \ldots, \omega_{tn}\}$.
Input the keywords in $F_1$ into the CBOW model to get the vector representation of each word to form a word vector data set $T_1 = \{\mu_{t1}, \mu_{t2}, \ldots, \mu_{tn}\}$.
Obtain the first $m$ most similar words of each word in the vocabulary $F_1$ through the CBOW model, and fill them into the vocabulary $F_1$ to form the vocabulary $F_2$.
redInput the keywords in $F_2$ into the CBOW model to get the vector representation of each word to form a word vector data set $T_2 = \{\mu_{t1}, (\alpha_{11}, \ldots, \alpha_{1m}), \ldots, \mu_{tn}, (\alpha_{n1}, \ldots, \alpha_{nm})\}$.
Cluster $T_2$ by K-means algorithm.
Define the topic terms of each category according to the clustering results, and get the redefined topic category.
```

4 EMPIRICAL ANALYSIS

In this section, after preprocessing Sogou Chinese corpus data and Microblog data, we train the word vector model and LDA model separately, and the two are combined to achieve the expansion of subject words and clustering by the K-means model to obtain the fusion topic and Keywords, and then extract all kinds of insurance keywords in the "Pingan Insurance" website to match, calculate the probability of Microblog users to buy different types of insurance, and finally take "User A" as an example to verify the accuracy of the LDA-CBOW model to predict the insurance demand of Microblog users.

4.1 Source of Data

The data in this paper is divided into two categories: word vector training data and analysis data. The word vector training data comes from the Sogou Chinese corpus, which contains 396 500 news in different fields and is used to train the word vectors required in this paper. The main analysis data in this article is Microblog text data, using the "Octopus" crawler software, on the Sina Microblog platform to crawl the data, first collect the first 20 pages of the blog post that published the keyword "insurance" on January 10, 2020, select the author of the blog post as the analysis sample, a total of 248, crawl all the blog posts within the 248 authors, a total of 12 568.

4.2 Pretreatment of Data

For the analysis text, first remove the link in the blog post, @other Microblog users' information, after preliminary cleaning and deduplication, get 10 754 data. The "jieba" tool in Python can segment sentences more accurately in full mode, and is capable of preprocessing tasks such as word segmentation and stop words (including conjunctions, adjectives and other high-frequency words that have little relationship with the topic) in text processing. The data obtained after preprocessing such as cleaning, word segmentation, and stop word removal can be used for the training and experiment of the model in this paper.

4.3 Experimental Results of the LDA-CBOW Model

1. **Analysis of traditional LDA model**

   Firstly, train the model parameters by Gibbs sampling, and the initial parameter settings are shown in Table 1.

   In Table 1, K represents the number of topics generated by the document. In determining the optimal number of topics, this paper selects the degree of perplexity as the criterion. The smaller the perplexity level, the higher the accuracy of the topic prediction of the potential topic model for the term. Keep other parameters unchanged, observe the change of the model's perplexity degree under different K values, change the K value, observe the change of the model's perplexity degree, and take the K value when the perplexity degree is the smallest as the optimal value. After calculation, this paper takes $K = 5$, that is, when generating 5 potential topics, the perplexity is the smallest.

Parameter	Meaning of Parameters	Value
α	Hyperparameters of "document-topic" multinomial distribution θ	0.1
β	Hyperparameters of the "topic-word" polynomial distribution φ	0.1
Niters	Gibbs sampling iterations	100
K	Number of topic	–
T-words	Number of output words of each topic	20

Table 1. Initial parameter settings of LDA theme model

The relationship between perplexity and the number of topics is shown in Figure 4.

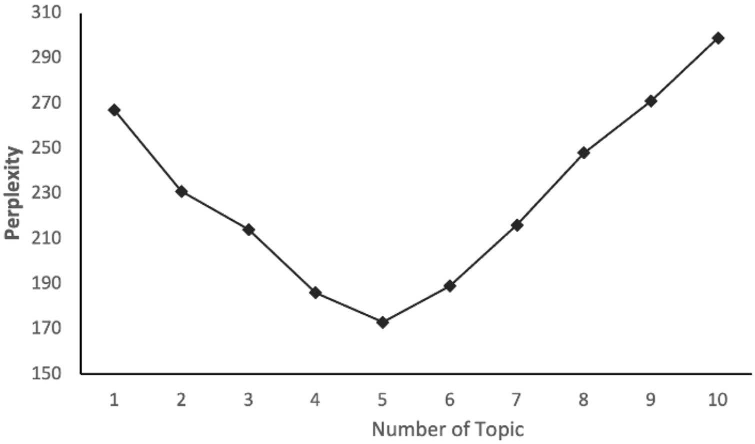


Figure 4. Relationship between perplexity and the number of topics

The results of topic division obtained by LDA model are shown in Table 2.

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
hardwork	0.007952	model	0.010998	insurance	0.013528	market	0.007359	no	0.008955
school	0.007758	car	0.008628	microblog	0.010593	stock	0.005599	hear	0.006055
baby	0.00679	engine	0.006797	hospital	0.006679	city	0.004976	microblog	0.005088
education	0.005918	formal	0.004966	do	0.006413	see	0.004879	not	0.004242
study	0.005918	refit	0.00475	accident	0.006413	individual	0.004461	airplane	0.004121
children	0.004659	piggyback	0.004535	video	0.006057	gem	0.004705	landscape	0.003879
think	0.004465	month	0.004212	children	0.004723	shock	0.003974	will	0.003396
not	0.004368	system	0.004104	look	0.004456	funds	0.003943	might	0.003396
like	0.003787	year	0.003996	minute	0.004278	obvious	0.004124	love	0.003396
go	0.003787	bumper	0.003996	guarantee	0.004278	rise	0.004345	refit	0.003396
...	...	...	...	...	...	...	...	...	...

Table 2. Topic division results of LDA model

As can be seen from Table 2, in addition to Topic 5, the remaining four topics can basically understand the topics expressed by the sample data, in order:

education, auto, insurance, stocks, but the proportion of unrelated words within the topic is relatively high. Next, the sample data is expanded by word vectors and then the topic analysis of the LDA model is carried out.

2. Topic word vector expansion and fusion analysis

Firstly, enter the top 20 words with the highest probability under each topic in Table 2 into the trained word vector model to obtain the top 20 words with the highest similarity to each word, to achieve the expansion of the topic word vector, and the expanded vocabulary, as shown in Table 3.

Type of Topic	Topic Words	Words for Expansion
Topic 1	hardwork	life, happy, have, others, teacher, thing, very, blessed, ...
	school	care, child, rigorous, hurt, angry, threaten, bear, familiar, ...
	...	...
Topic 2	model	add, increase, temporary, produce, quality, proxy, practice, ...
	car	industry, duty, danger, manage, dispute, improve, cultivate, ...
	...	...
Topic 3	insurance	business, year, number, director, medical, accident, risk, ...
	insurance	news, agency, media, website, audiovisual, video, rating, ...
	...	...
Topic 4	market	predecessor, fund, repeated, filed, cities, dollar, money, ...
	stock	add, data, company, push, choose, climate, reputation, ...
	...	...
Topic 5	no	emotion, attitude, only, with, tell, always, sacrifice, need, ...
	hear	meet, loss, memory, protect, change, around, key, hard, ...
	...	...

Table 3. Vocabulary expanded by the word vector

According to the word vector model, enter Tables 2 and 3 respectively, that is, we need to get a word vector table of all words in the thesaurus before and after the expansion, and use the K-means algorithm to cluster the data in the two word vector tables. And calculate the accuracy rate, recall rate and harmonic average F1 respectively for the results. The comparison results are shown in Table 4.

The experiment in this article randomly divides the data set into training set (80 %), test set (10 %) and validation set (10 %) according to the ratio of 8:1:1. We use the training set and test set to train the model. We evaluate the clustering effect of the algorithm with the experimental results of each method on the validation set. And we use the precision rate, recall rate and F1 value to evaluate the quality of the recommended results. The expression is as follows

Equations (10), (11) and (12):

$$\text{Precision} = \frac{\text{NUM(TP)}}{\text{NUM(TP + FP)}}, \tag{10}$$

$$\text{Recall} = \frac{\text{NUM(TP)}}{\text{NUM(TP + FN)}}, \tag{11}$$

$$F_1 = \frac{\text{NUM(2TP)}}{\text{NUM(2TP + FP + FN)}}. \tag{12}$$

Among them, TP means that the clustering result shows that the word belongs to a certain type of topic (P), which actually belongs to this kind of topic, and the prediction is correct (T); FP means that the clustering result shows the word belongs to a certain type of topic (P), which does not actually belong to this kind of topic, and the prediction is wrong (F); TN means that the clustering result shows that the word does not belong to a certain type of topic (N), and it does not actually belong to this type of topic, and the prediction is correct (T); FN means that the clustering result shows that the word does not belong to a certain type of topic (N), and it actually belongs to this topic, prediction is wrong (F).

Types of Word Tables	Precision (%)	Recall (%)	F_1 (%)
Before expansion	53.3	66.8	59.3
After expansion	60.4	72.6	65.9

Table 4. Comparison of vocabulary clustering results before and after expansion

Topic 1	study, school, baby, number, worry, child, like, ...
Topic 2	shape, car, design, engine, cost, systerm, auto, ...
Topic 3	insurance, compensation, recommend, guarantee, ...
Topic 4	market, stock, shock, individual, funds, rise, ...
Topic 5	refit, landscape, airplane, safe, happy, hotel, tourist, ...

Table 5. Topic keyword description after word vector expansion

Table 5 is the description of topic keywords obtained by clustering after word vector expansion. It can be seen that relative to Table 2, the proportion of related words in each topic of Table 5 is higher, and it can be seen that Topic 5 represents the theme of “travel”.

4.4 Analysis Results of Insurance Demands of Microblog Users

There are six major types of insurance on the “Pingan Insurance” website: health insurance, accident insurance, corporate insurance, travel insurance, house insurance and auto insurance. Compared with other insurance types, less people would have the demand to buy house insurance and enterprise insurance. House insurance is mainly for the protection of family property, bank cards, mobile phone screens and pets, and corporate insurance is mainly for enterprise managers or employers. The protection provided, because this article studies the insurance needs of Microblog users, belongs to the category of personal insurance, so corporate insurance is not included in the insurance recommendation, and only matches the other five types of insurance with the subject terms. Table 6 shows the keywords of five major types of insurance in "Pingan Insurance" extracted through the LDA model.

Type of Insurance	Keywords
Health insurance	insurance, accident, death, disability, disease, medical, vaccine, ...
Accident insurance	accident, death, disease, hospital, ambulance, infection, delay, ...
Travel insurance	travel, accident, death, plane, train, boat, acute, disease, hurt, ...
House insurance	cost, house, treatment, indoor, duty, accident, hurt, steal, robery, ...
Car insurance	car, burn, scald, lost, hospital, accident, allowance, responsibility, ...

Table 6. Keywords of different type of insurance in "Pingan insurance" website

Use the "similarity" function of the word vector to calculate the similarity between the keywords of each theme after expansion and the keywords of "Pingan Insurance", calculate the average and calculate the percentage to obtain the possibility of different themes for different types of insurance, as shown in Table 7.

Type of Insurance	Health	Accident	Travel	House	Car
Topic 1	23.48 %	20.63 %	20.34 %	15.63 %	19.92 %
Topic 2	24.71 %	22.13 %	18.79 %	14.31 %	20.06 %
Topic 3	24.63 %	20.67 %	9.52 %	14.64 %	20.54 %
Topic 4	20.77 %	21.34 %	18.68 %	19.39 %	19.82 %
Topic 5	21.82 %	19.75 %	22.23 %	16.52 %	19.68 %

Table 7. Possibility of purchasing various types of insurance under different topics

According to Table 7, the probability of health insurance purchase is higher in the five themes, indicating that regardless of the potential needs of users, the success rate of recommended health insurance will be higher. The probability of home insurance purchases in the five themes are all the lowest, so the success rate when recommending such insurance will be lower. Next, we will analyze the probability of each type of insurance purchased by combining Tables 5 and 7.

From the topic keywords in Topic 1 in Table 5, it is mainly related to children's educational issues. According to the data in Table 7, users have the highest probability of purchasing health insurance, which is 23.48 %. We might guess that users would pay more attention to themselves and to the next generation when they have the next generation. The health of the family is high, so the demand for health insurance is high; while the probability of users buying accident insurance, travel insurance and car insurance is similar, respectively 20.63 %, 20.34 % and 19.2 %, guessing that users may be doing some things such as travel, tourism, etc. During parent-child activities, there is a demand for these three types of insurance; users have the lowest probability of buying house property insurance at 15.63 %.

Topic 2 is mainly related to cars. It is guessed that users may pay more attention to car issues. According to experience, such users may have demand for accident insurance and car insurance. From the results in Table 7, the probabililty of users buying accident insurance and car insurance is indeed high, respectively 22.13 % and 20.06 %, this result is in line with reality; and the insurance type with the highest probability of user purchase is still health insurance, which is 24.71 %, which shows that people's willingness to insure health under the premise of economic permit is the strongest; the probability of users purchasing travel insurance is 18.79 %. It is guessed that some users will tend to drive by themselves when traveling, so there is a demand for travel insurance.

Topic 3 is obviously related to the purchase of insurance products, but it can only be seen that the user needs to purchase insurance, and the actual type of insurance demand of the user cannot be understood. In this case, according to the data provided in Table 7, according to the purchase probability of each insurance type recommend users in order of high to low: health insurance is recommended first, followed by accident insurance, and then car insurance, travel insurance and house insurance. Of course, a closer analysis of the Microblog content of these users can be made to speculate on the user's real needs and make insurance recommendations more accurately.

Topic 4 is related to stocks, indicating that such users may have more idle funds. From the data in Table 7, the probability of such users buying various types of insurance is basically the same. This situation may occur because there is no that type of insurance that matches this type of user in "Pingan Insurance" official website. From the user's actual situation, based on experience, wealth management insurance with lower stock risk may be more suitable, so the possibility of purchasing wealth management products will be higher.

Topic 5 is mainly related to travel, indicating that users may like to travel or plan to travel in the near future. In Table 3, users have the highest probability of purchasing travel insurance, 22.23 %, followed by health insurance with a probability of 21.82 %, This is followed by 19.75 % of accident insurance and 19.68 % of car insurance. The higher probability of these two types of insurance may be for users who tend to drive by themselves, and there is a demand for accident insurance and car insurance.

4.5 Analysis of Insurance Demand with "User A" as an Example

This section takes Microblog user A as an example to verify the accuracy of the model used to analyze the insurance needs of Microblog users.

First crawl all blog posts that "User A" has posted since registering the Microblog account, excluding irrelevant information such as likes and forwarding Microblog, a total of 428 articles. The preprocessed data such as cleaning, word segmentation, and stop word removal are substituted into the LDA model for candidate topic analysis. The calculation shows that when the number of topics is 3, the degree of confusion is the smallest, so the number of candidate topics is 3, and the results of the topic model division are shown in Table 8.

Topic 1		Topic 2		Topic 3	
like	0.01298	time	0.013319	sleep	0.016717
form	0.010435	feel	0.010708	bedroom	0.011236
many	0.00789	meniscus	0.008096	today	0.011236
beauty	0.00789	life	0.008096	life	0.008495
once	0.00789	surgery	0.008096	no	0.008495
cold	0.00789	lifetime	0.005484	know	0.008495
young	0.005345	anxious	0.005484	nature	0.005755
face	0.005345	travel	0.005484	recommend	0.005755
happy	0.005345	relax	0.005484	forever	0.005755
story	0.005345	stagnate	0.005484	freedom	0.005755

Table 8. Topic division results of LDA model

According to the results in Table 8, it can be roughly seen that, except that Topic 2 may be related to "medical" or "travel", the other two thematic features are not obvious. Next, use word vectors to expand the vocabulary, as shown in Table 9.

Type of Topic	Topic Words	Words for Expansion
Topic 1	like	friend, power, have, scene, know, human, father, mother, ...
	form	data, number, degree, fund, math, federation, union, exam...
	...	...
Topic 2	form	data, number, degree, fund, math, federation, union, exam, ...
	feel	communicate, confidence, character, congnition, intuition, ...
	...	...
Topic 3	sleep	clear, rumor, match, literature, describe, recoup, indeffierent ...
	bedroom	rest, manage, enable, link, equipment, tourist, station, road, ...
	...	...

Table 9. Vocabulary expanded by the word vector

The words in Table 9 are input into the word vector model to obtain the corresponding word vector table, and clustering is performed using the K-means algorithm, and the subject keyword description after clustering is obtained, as shown in Table 10.

Type of Topic	Keywords
Topic 1	surgery, hospital, meniscus, success, worry, anxious, forever, ...
Topic 2	travel, freedom, recommend, tourist, station, equipment, nature, ...
Topic 3	like, time, feel, story, rest, communicate, cold, form, ...

Table 10. User A's keyword vector description after clustering

It can be seen from Table 10 that Topic 1 means "medical" related topics, presuming that "User A" or his relatives are sick or hospitalized, and there may be a need for related medical insurance; Topic 2 means "travel" related topics, presuming that "User A" may be interested in travel, you can recommend travel insurance or accident insurance, etc.; Topic 3's keyword description is relatively vague, so you cannot see the meaning of the theme. This situation occurs because "User A" is more active on the Microblog platform, and may like to have no details. Share your daily life, this habit provides us with the data mining "User A" needs, but there is also a lot of useless data. Calculate the similarity between the words in each topic and the keywords in Table 6, and obtain the possibility of "User A" buying various types of insurance, as shown in Table 11.

Type of Insurance	Health	Accident	Travel	House	Car
Topic 1	26.75 %	23.13 %	17.29 %	16.12 %	16.71 %
Topic 2	22.83 %	20.47 %	23.79 %	13.79 %	19.12 %
Topic 3	23.83 %	21.25 %	19.24 %	14.66 %	21.02 %
Average	24.47 %	21.62 %	20.11 %	14.86 %	18.95 %

Table 11. Possibility of purchasing various types of insurance under different topics

Since this is an analysis of the insurance needs of the specified users, the average value of each theme to purchase the specified insurance type is finally calculated as the basis for finally recommending insurance to the user. According to Table 11, "User A" has the highest probability of purchasing health insurance at 24.47 %; followed by accident insurance with a probability of 21.62 %; then travel insurance with a probability of 20.11 %; followed by car insurance with a probability of 18.95 %; the lowest probability is 14.86 % of home insurance. Therefore, when recommending insurance to "User A", they can be recommended in this order.

5 CONCLUSION

In the increasingly competitive insurance market, keeping abreast of customer demands and preferences in real-time is the key for insurance companies to enhance their competitiveness. As one of the most popular social networking platforms in China, Microblog has many users. The blog posts published by Microblog users may express their potential interests and demands. How to use massive microblog content to analyze the potential insurance needs of users is the focus of this paper. We propose the LDA-CBOW model, which makes full use of the topic analysis ability of the LDA model and the word vectorization ability of the CBOW model to analyze the microblog user text, and obtains more accurate potential topics in the microblog user blog post through the method of subject word vector expansion. Furthermore, the potential insurance demands of users are analyzed through their potential themes. Finally, we use the existing insurance sales website content to match the extracted user theme and use Microblog "User A" as an example to verify the accuracy of this model. It proves the feasibility of obtaining customer insurance demands through Microblog content, and lays the foundation for constructing insurance customer portraits from social data and making personalized insurance recommendations.

Acknowledgement

This research was supported by the National Key Project of Statistical Scientific Research of China (No. 2019LZ10). Besides, this research was supported by Natural Science Foundation of Shandong Province (No. ZR2020MF033). The corresponding author is Wei Liu.

REFERENCES

[1] Gao, C. S.—Rong, X.—Chen, Y.: Research on Public Opinion Monitoring Index-System in Micro-Blogging. Journal of Intelligence, Vol. 30, 2011, No. 9, pp. 66–70, doi: 10.3969/j.issn.1002-1965.2011.09.013 (in Chinese).

[2] Lan, X. Y.: Research on the Diffusion Law Model of Weibo Public Opinions in Emergencies. Information Science, Vol. 31, 2013, No. 3, pp. 31–34, doi: 10.13833/j.cnki.is.2013.03.019 (in Chinese).

[3] Li, M. D.—Meng, S. J.—Zhang, H. B.: Communication Mode of Microblog: A Study Based on the Process Analysis. Journal of Intelligence, Vol. 33, 2014, No. 2, pp. 120–127, doi: 10.3969/j.issn.1002-1965.2014.02.023 (in Chinese).

[4] Wang, Y. F.—Hang, W. L.—Ding, J.: Identification and Application of Microblog Public Opinion Social Network Critical Node. Information and Documentation Services, Vol. 37, 2016, No. 3, pp. 5–11, doi: 10.3969/j.issn.1002-0314.2016.03.001 (in Chinese).

[5] Liu, J. R.: Study on Evolutionary Pathways of Microblog Public Opinion from the Perspective of Crisis Communication. Journal of Intelligence, Vol. 31, 2012, No. 7, pp. 21–24, doi: 10.3969/j.issn.1002-1965.2012.07.005 (in Chinese).

[6] Xie, L. X.—Zhou, M.—Sun, M. S.: Hierarchical Structure Based Hybrid Approach to Sentiment Analysis of Chinese Micro Blog and Its Feature Extraction. Journal of Chinese Information Processing, Vol. 26, 2012, No. 1, pp. 73–83, doi: 10.3969/j.issn.1003-0077.2012.01.011 (in Chinese).

[7] Liang, J.—Chai, Y. M.—Yuan, H. B.—Zan, H. Y.—Liu, M.: Deep Learning for Chinese Micro-Blog Sentiment Analysis. Journal of Chinese Information Processing, Vol. 28, 2014, No. 5, pp. 155–161, doi: 10.3969/j.issn.1003-0077.2014.05.019 (in Chinese).

[8] He, Y. X.—Sun, S. T.—Niu, F. F.—Li, F.: A Deep Learning Model Enhanced with Emotion Semantics for Microblog Sentiment Analysis. Chinese Journal of Computers, Vol. 40, 2017, No. 4, pp. 773–790, doi: 10.11897/SP.J.1016.2017.00773 (in Chinese).

[9] Chen, K.—Liang, B.—Ke, W. D.—Xu, B.—Zeng, G. C.: Chinese Micro-Blog Sentiment Analysis Based on Multi-Channels Convolutional Neural Networks. Journal of Computer Research and Development, Vol. 55, 2018, No. 5, pp. 945–957, doi: 10.7544/issn1000-1239.2018.20170049 (in Chinese).

[10] Li, Y. H.—Xie, M.—Yi, Y.: Sentiment Analysis of Micro-Blogging Based on DAE and Its Improved Model. Application Research of Computers, Vol. 34, 2017, No. 2, pp. 373–377, doi: 10.3969/j.issn.1001-3695.2017.02.012 (in Chinese).

[11] Zhang, M.: Social Network Analysis of Library Microblog in China. Journal of Information Resources Management, Vol. 4, 2014, No. 3, pp. 80–87, doi: 10.13365/j.jirm.2014.03.080 (in Chinese).

[12] Liu, X. P.—Tian, X. Y.: Analysis of the Social Network Structure and Influence of Media Microblog. Information Science, Vol. 36, 2018, No. 1, pp. 96–101, doi: 10.13833/j.issn.1007-7634.2018.01.017 (in Chinese).

[13] Qiu, M. W.—Jiang, Y. H.: Social Network Analysis of Weibo Users – Taking the Official Microblog of the National Library of China on Sina Microblog as an Example. Sci-Tech Information Development and Economy, Vol. 28, 2015, No. 19, pp. 137–139 (in Chinese).

[14] Di, L.—Du, Y. P.: Application of LDA Model in Microblog User Recommendation. Computer Engineering, Vol. 40, 2014, No. 5, pp. 1–6, doi: 10.3969/j.issn.1000-3428.2014.05.001 (in Chinese).

[15] Tang, X. B.—Fang, X. K.: Research on Weibo Recommendation Model Based on Implicit Dirichlet Assignment. Information Science, Vol. 33, 2015, No. 2, pp. 3–8, doi: 10.13833/j.cnki.is.2015.02.001 (in Chinese).

[16] Ma, H. F.—Jia, M. H. Z.—Li, X. H.—Lu, X. Y.: A Microblog Recommendation Method Based on Label Correlation Relationship. Computer Engineering, Vol. 42, 2016, No. 4, pp. 197–201, doi: 10.3969/j.issn.1000-3428.2016.04.035 (in Chinese).

[17] Mao, H.—Luo, S. C.—Tang, G. C.: The Development of Foreign Insurance Pricing in the Western Countries and Its References to China. Operations Research and Management Science, Vol. 12, 2003, No. 2, pp. 77–82, doi: 10.3969/j.issn.1007-3221.2003.02.018 (in Chinese).

[18] Meng, S. W.: An Application of Generalized Linear Model to Automotor Insurance Pricing. Journal of Applied Statistics and Management, Vol. 26, 2007, No. 1, pp. 24–29, doi: 10.3969/j.issn.1002-1566.2007.01.006 (in Chinese).

[19] Xu, J. L.: The Price Elasticity of Insurance Demand and the Policy Choice of China's Insurance Regulation. Journal of Guandong University of Finance, Vol. 22, 2007, No. 5, pp. 94–96, doi: 10.3969/j.issn.1674-1625.2007.05.016 (in Chinese).

[20] Zhang, Y. J.—Shi, B. S.—Wei, Y.—Jin, D. X.: Loans Defaulting Prematurely and the Pricing of Loan Insurance. System Engineering – Theory and Practice, Vol. 39, 2019, No. 10, pp. 2502–2511, doi: 10.12011/1000-6788-2018-0381-10 (in Chinese).

[21] Ye, M. H.: Research on Insurance Fraud Recognition Based on Bp Neural Network – Taking China Motor Vehicle Insurance Claim as an Example. Insurance Studies, 2011, No. 3, pp. 79–86, doi: 10.13497/j.cnki.is.2011.03.012 (in Chinese).

[22] Li, X. F.—Huang, Z. G.—Chen, X. W.: Application Research of Bagging Integration Method in Insurance Fraud Recognition. Insurance Studies, Vol. 4, 2019, pp. 66–84, doi: 10.13497/j.cnki.is.2019.04.006 (in Chinese).

[23] Yan, C.—Li, Y. Q.—Sun, H. T.: Research on Auto Insurance Fraud Recognition Based on Ant Colony Algorithm Optimized Random Forest Model. Insurance Studies, 2017, No. 6, pp. 114–127, doi: 10.13497/j.cnki.is.2017.06.011 (in Chinese).

[24] Tian, L.—Wang, Z. W.—Xu, Y. F.: Research on the Allocation of Investment Risk Limits of China's Insurance Companies Based on Economic Capital. Insurance Studies, 2011, No. 11, pp. 31–38, doi: 10.13497/j.cnki.is.2011.11.005 (in Chinese).

[25] Dong, K. X.—Xie, Z. X.—Zhen, J.—Lin, R. H.: An Analysis of the Optimal Decision of Insurance Companies Investing in Information Security Software under Dependent Risks. Insurance Studies, 2019, No. 6, pp. 66–80, doi: 10.13497/j.cnki.is.2019.06.006 (in Chinese).

[26] MacQueen, J. et al.: Some Methods for Classification and Analysis of Multivariate Observations. In: Le Cam, L. M., Neyman, J. (Eds.): Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. 1967, pp. 281–297.

[27] Chen, X.—Chen, C. H.—Leong, K. F.—Jiang, X.: An Ontology Learning System for Customer Needs Representation in Product Development. The International Journal of Advanced Manufacturing Technology, Vol. 67, 2013, No. 1, pp. 441–453.

[28] Jiang, W.—Zhang, L.—Dai, Y.—Jiang, J.—Wang, G.: Analyzing Helpfulness of Online Reviews for User Requirements Elicitation. Chinese Journal of Computers, Vol. 36, 2013, No. 1, pp. 119–131, doi: 10.3724/SP.J.1016.2013.00119 (in Chinese).

[29] Dou, R.—Zong, C.: Application of Interactive Genetic Algorithm Based on Hesitancy Degree in Product Configuration for Customer Requirement. International Journal of Computational Intelligence Systems, Vol. 7, 2014, No. Supplement 2, pp. 74–84, doi: 10.1080/18756891.2014.947118.

[30] Huo, L. H.: A Study of the "Customer's Demand-Oriented" Transformation and Innovation of Insurance Companies. Finance Forum, Vol. 21, 2016, No. 1, pp. 32–39, doi: 10.16529/j.cnki.11-4613/f.2016.01.003 (in Chinese).

[31] Yuan, J. J.—Huang, M. M.—Cao, L.: Modeling and Optimization of Product Configuration Rebuilt Driven by Customer Requirement Dynamic Change. Computer Integrated Manufacturing System, Vol. 37, 2015, No. 13, pp. 134–140, doi: 10.13196/j.cims.2018.10.020 (in Chinese).

[32] Qiu, P. P.—Huang, X. Y.—Zeng, Q. S.: Predicting Customer Demand with the Side Information Incorporated Hierarchical Bayesian Model. Computer Integrated Manufacturing Systems, Vol. 26, 2020, No. 1, pp. 191–201, doi:

10.13196/j.cims.2020.01.020 (in Chinese).

[33] Yu, X.—He, Y. D.—Liang, H. T.—Jiang, F.—Du, J. W.: A Top-K Crowdsourcing Developer Recommendation Method Considering Interest Preference. Journal of Shandong University of Science and Technology (Natural Science), Vol. 40, 2021, No. 3, pp. 58–70, doi: 10.16452/j.cnki.sdkjzk.2021.03.008 (in Chinese).

[34] Blei, D. M.—Ng, A. Y.—Jordan, M. I.: Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, 2003, No. 1, pp. 993–1022.

[35] Gerrish, S.—Blei, D. M.: A Language-Based Approach to Measuring Scholarly Impact. Proceedings of the 26th International Conference on Machine Learning (ICML 2010), 2010, pp. 375–382.

[36] Weng, J.—Lim, E. P.—Jiang, J.—He, Q.: TwitterRank: Finding Topic-Sensitive Influential Twitterers. Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10), 2010, pp. 261–270, doi: 10.1145/1718487.1718520.

[37] Ramage, D.—Dumais, S.—Liebling, D.: Characterizing Microblogs with Topic Models. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, pp. 130–137.

[38] Xu, B.—Yang, D.—Zhang, Y.—Li, F.—Gao, K. N.: Relationship Bind Topic Model Toward Tag Recommendation for Micro-Blog Users. Journal of Frontiers of Computer Science and Technology, Vol. 8, 2014, No. 3, pp. 288–295, doi: 10.3778/j.issn.1673-9418.1306051 (in Chinese).

[39] Zhang, R.—Jin, Z. G.—Wang, Y.: Recommendation Model of Microblog User Tags Based on Hybrid Grain. Computer Science, Vol. 43, 2016, No. 4, pp. 192–196 (in Chinese).

[40] Wang, D.: Research on User Profiling Based on Topic Model. Ph.D. Thesis. Beijing University of Technology, 2016 (in Chinese).

[41] Wang, T.—Li, M.: Study on an Improved Keyword Extraction Algorithm. Journal of Chongqing Normal University (Natural Science), Vol. 36, 2019, No. 3, pp. 98–104, doi: 10.11721/cqnuj20190312 (in Chinese).

[42] Zhang, Y. L.—Eick, C. F.: Tracking Events in Twitter by Combining an LDA-Based Approach and a Density–Contour Clustering Approach. International Journal of Semantic Computing, Vol. 13, 2019, No. 01, pp. 87–110, doi: 10.1142/S1793351X19400051.

[43] Bao, X.—Wang, M. R.—Liu, G. F.: A Novel Text Classification Method Based on Topic Model and Transfer Learning. Journal of Shandong University of Science and Technology (Natural Science), Vol. 40, 2021, No. 3, pp. 80–88, doi: 10.16452/j.cnki.sdkjzk.2021.03.010 (in Chinese).

[44] Li, J. P.—Cao, N.—Zhang, Q.—Zhang, W. P.—Ji, S. J.: Online Social Network Groups Discovery Algorithm Considering Themes and Time. Journal of Shandong University of Science and Technology (Natural Science), Vol. 40, 2021, No. 4, pp. 94–102, doi: 10.16452/j.cnki.sdkjzk.2021.04.011 (in Chinese).

[45] Li, B.—Wang, Y. H.—Zhu, X. Q.—Li, J. P.: Identification and Evolution Analysis of Important Risks in Insurance Industry Based on the Textual Risk Disclosures in

Financial Reports. System Engineering – Theory and Practice, Vol. 42, 2022, No. 2, pp. 333–344, doi: 10.12011/SETP2020-2520 (in Chinese).

[46] Li, L. P.—Zhao, X.: A Research Summary of Topic Discovery Methods Based on Topic Model. Journal of Minzu University of China (Natural Sciences Edition), Vol. 30, 2021, No. 2, pp. 59–66, doi: 10.3969/j.issn.1005-8036.2021.02.010.

[47] Nikfarjam, A.—Sarker, A.—O'connor, K.—Ginn, R.—Gonzalez, G.: Pharmacovigilance from Social Media: Mining Adverse Drug Reaction Mentions Using Sequence Labeling with Word Embedding Cluster Features. Journal of the American Medical Informatics Association, Vol. 22, 2015, No. 3, pp. 671–681, doi: 10.1093/jamia/ocu041.

[48] Lilleberg, J.—Zhu, Y.—Zhang, Y.: Support Vector Machines and Word2Vec for Text Classification with Semantic Features. 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC), IEEE, 2015, pp. 136–140, doi: 10.1109/ICCI-CC.2015.7259377.

[49] Xia, T.: Extracting Keywords with Modified TextRank Model. Data Analysis and Knowledge Discovery, Vol. 1, 2017, No. 2, pp. 28–34, doi: 10.11925/infotech.2096-3467.2017.02.04 (in Chinese).

[50] Zhou, Q. Q.—Zhang, C. Z.: Fine-Grained Aspect Extraction from Online Customer Reviews. Journal of the China Society for Scientific and Technical Information, Vol. 36, 2017, No. 5, pp. 484–493, doi: 10.3772/j.issn.1000-0135.2017.05.006 (in Chinese).

[51] Vargas-Calderón, V.—Camargo, J. E.: Characterization of Citizens Using Word2Vec and Latent Topic Analysis in a Large Set of Tweets. Cities, Vol. 92, 2019, pp. 187–196, doi: 10.1016/j.cities.2019.03.019.

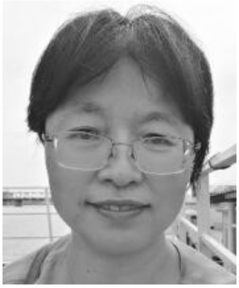
[52] Zhang, P. Y.—Liu, D. S.: Topic Evolutionary Analysis of Short Text Based on Word Vector and BTM. Data Analysis and Knowledge Discovery, Vol. 3, 2019, No. 3, pp. 95–101, doi: 10.11925/infotech.2096-3467.2018.0625 (in Chinese).

[53] Gu, Y.—Li, H.—Li, Y. Y.—Liu, J. Y.: Research on Demand Mining of Enterprise Competitive Intelligence Based on Online Reviews. Modern Information, Vol. 41, 2021, No. 1, pp. 24–31, doi: 10.3969/j.issn.1008-0821.2021.01.003 (in Chinese).

Chun Yan received her B.Sc., M.Sc. and Ph.D. degrees from the Shandong University of Science and Technology, Qingdao, China, in 2000, 2003 and 2011, respectively. She is presently Associate Professor of the Shandong University of Science and Technology, Qingdao, China. Her research interests include applied mathematics, statistics and economic management. She has about 20 technical papers published in journals and conference proceedings in her research areas.

Lu Liu received her B.Sc. degree from the Shandong University of Science and Technology, China, in 2016. She is now pursuing her Ph.D. degree in the Shandong University of Science and Technology, Qingdao, China. Her research interests include economic management and data analysis.

Wei Liu is engaged in research work in workflow, service computing, Petri net theory and application, software formal analysis and verification, and big data intelligent analysis and processing.



Man Qi is Senior Lecturer in computing at Canterbury Christ Church University, UK. Her main research interests are in the areas of intelligent systems and applications, data analytics, cyber security and HCI. Her research has been founded by EPSRC, EU and QR etc. She is Fellow of the British Computer Society (FBCS) and Fellow of the Higher Education Academy (FHEA). She has published over 70 research papers and is on the editorial boards of five international journals. She has been an external Ph.D. examiner for a number of universities in the UK and Australia. She has served as Chair and Program Committee member for over 50 international conferences and has been a long-term reviewer for many international journals.