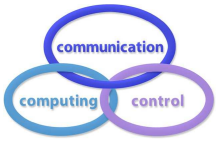
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 15, Issue: 2, Month: February, Year: 2020
Article Number: 3811, https://doi.org/10.15837/ijccc.2020.2.3811



CCC Publications

A Latent Dirichlet Allocation and Fuzzy Clustering Based Machine Learning Model for Text Thesaurus

J. Luo, D. Yu, Z. Dai

Jia Luo
NEWHUADU Business School
Minjiang, Fujian, China
luojia_minjiang@163.com

Dongwen Yu
NEWHUADU Business School
Minjiang, Fujian, China

Zong Dai*
Hunan Zhaoshan Investment & Holdings Co. Ltd. Xiangtan
Hunan, China
*Corresponding author: 14492121@qq.com

Abstract

It is not quite possible to use manual methods to process the huge amount of structured and semi-structured data. This study aims to solve the problem of processing huge data through machine learning algorithms. We collected the text data of the company's public opinion through crawlers, and use Latent Dirichlet Allocation (LDA) algorithm to extract the keywords of the text, and uses fuzzy clustering to cluster the keywords to form different topics. The topic keywords will be used as a seed dictionary for new word discovery. In order to verify the efficiency of machine learning in new word discovery, algorithms based on association rules, N-Gram, PMI, and Word2vec were used for comparative testing of new word discovery. The experimental results show that the Word2vec algorithm based on machine learning model has the highest accuracy, recall and F-value indicators.

Keywords: text, LDA, fuzzy clustering, thesaurus, Word2vec, machine learning.

1 Introduction

Text is composed of sentences consisting of continuous words and punctuation. Segmentation is usually the first step in natural language processing tasks such as text classification and clustering, sentiment analysis, and topic recognition. The effect of segmentation will directly affect the accuracy of Chinese text processing. New words can cause segmentation errors. Recognizing new words is of great significance to other natural language processing tasks. The emergence of new words reduces the coverage of the dictionary and makes analysis difficult. Existing researches on dictionaries mainly focus

on the construction of sentiment dictionaries. Blei et al. [4] proposed an implicit Dicrete's probabilistic topic model LDA in 2003. It formally introduced topics in the form of hidden variables for the first time to form a three-layer Bayesian model. Close to the PLSI model, the topics selected by LDA are not bound by the text of the training set. It is a completely unsupervised machine learning algorithm that clusters based on multiple topics. At the same time, the text can be reduced in dimension to obtain the text representation in the topic dimension, which can make the machine learning algorithm execute more efficiently. After the birth of LDA, different forms of expansion appeared. In 2004, Blei proposed a hierarchical LDA of a tree structure, where each tree node represents a topic [10]. In 2006, Blei considered the disadvantage of irrelevant LDA topics and proposed Correlated Topic Model (CTM) [6]. Wang added time attributes to the LDA model, and built a topic model where topics change with time [19]. Griffiths et al. considered the rigorous word exchange hypothesis in this model and proposed to obtain sentence structure information through HMM, use LDA to construct semantic relations, and combine the two to propose the HMM-LDA model [11]. Blei et al. proposed the Supervised Latent Dirichlet Allocation (sLDA) model for the poor performance of unsupervised clustering LDA model in text classification. This model introduced text labeling into the LDA model, and label compliance and topic probability distribution. For the normal linear distribution, the model can be imported to get the new text category [15]. Li used the LDA algorithm to perform topic mining on railway complaint texts [14].

The research on dictionary construction mainly focuses on the construction of emotional dictionaries. At present, most common emotional dictionaries are manually created by scholars, and most authoritative emotional dictionaries are mostly English emotional dictionaries, such as General Inquirer [13], Opinion Lexicon [3], Senti-WordNet and Q-WordNet [2]. There are currently three new word discovery methods. The first is a new word discovery method based on rules; the second is a new word discovery method based on statistics; the third is a new word discovery method based on rules and statistics. There are three methods for automatically building the sentiment dictionary, a knowledge base-based method, a corpus-based method, and a combination of a corpus and a knowledge base. Hu et al. extended the sentiment dictionary by considering the synonymous and antonymous relationships of words in WordNet [13]. Andreevskaia et al. extended the sentiment dictionary through the meaning of words in the dictionary WordNet and the relationship between different words [1]. Hassan et al. calculated the semantic similarity between different words through the dictionary WordNet, constructed semantic maps between different words, and finally obtained the word polarity through the graph-related algorithm [12].

Dictionary Word2vec is a set of deep learning toolkits that represent words in documents as word vectors, released by Google in 2013. Through training corpus, multi-dimensional real number vectors of words are obtained [16]. On this basis, the distance between word vectors can be calculated to measure their similarity [17]. At present, scholars often use the Word2Vec tool to study text clustering, synonym recognition, machine translation, and topic extraction [8]. It is found that public opinion classification needs to mine the topics of corporate public opinion texts and classify public opinion texts according to certain rules. The most important ones are the extraction of public opinion keywords and the establishment of public opinion dictionaries. Subsequent public opinion classification and prediction accuracy can be higher.

Common sentiment dictionaries include thesaurus such as HowNet sentiment dictionary and Dalian Institute of Technology's sentiment ontology library, which only covers thousands of Chinese words. There is no clear method or model for dictionary construction in professional fields, and text analysis in professional fields still needs improvement. In this paper, a text of corporate public opinion is taken as an example to build a model of text keyword extraction, clustering, topic mining, and new word discovery based on machine learning algorithms. First, the public opinion text data of six companies with significant public opinion dissemination are obtained through a web crawler program. Second, these text data are pre-processed to form text data that can be analyzed. Third, the LDA algorithm is used to initially extract the topic keywords, and keywords are clustered to perform topic classification based on fuzzy clustering algorithm, and it compares different new word discovery algorithms and expands the keyword thesaurus of corporate public opinion based on the seed dictionary. Finally, machine Learning Model of Text Thesaurus is constructed.

2 Machine Learning Models

The machine learning model of corporate Internet public opinion topic mining and thesaurus construction mainly includes the collection of public opinion texts, data preprocessing, text de-duplication, word segmentation, and de-duplication stop words. The pre-processed text uses the LDA algorithm to obtain the keywords and the fuzzy clustering to get the topic classification of public opinion texts. These keywords will be used as a seed dictionary for new word discovery. The Word2vec algorithm is used to form public opinion vocabulary into word vectors, and pre-train 7 million public opinion texts to form a public opinion text pre-training model. Based on the original seed dictionary, new word discovery is used to expand keywords to form a corporate public opinion lexicon, machine learning model is finally constructed (Figure 1).

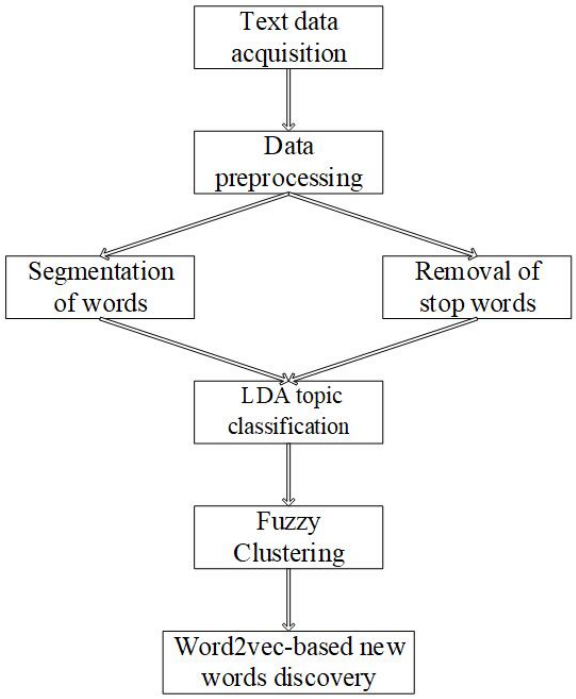


Figure 1: Machine learning model

2.1 LDA

LDA topic model belongs to the typical model of topic mining in natural language processing. It is a three-layer Bayesian probability generation model based on probability map. The main idea of the LDA topic model is to assume that each document in the document set is composed of multiple topics, and each topic is a polynomial distribution of multiple words on a fixed vocabulary. The purpose is to use efficient probability inference algorithms to process large-scale data. To extract potential topics from a text corpus and provide a method for quantifying research topics. This method has been widely used in various topic discovery, such as hotspot mining, topic evolution, and trend prediction.

The terms of LDA was defined as follows [21]: A document is a sequence of N words denoted by $w = (w_1, w_2, ..., w_n)$ where w_n is the nth word in the sequence, and a corpus is a collection of M documents denoted by $D = \{d_1, d_2, ..., d_M\}$; Since the process to generate the topic for M documents are independent of one another, we can have M conjugated structures and the generative process of probabilistic of topics in corpus is as follows:

$$p(\overrightarrow{z}|\overrightarrow{\alpha}) = \prod_{m=1}^{M} p(\overrightarrow{z}_m|\overrightarrow{\alpha}) = \prod_{\mathrm{m}=1}^{\mathrm{M}} \frac{\Delta(\overrightarrow{\mathrm{n}}_{\mathrm{m}}+\overrightarrow{\alpha})}{\Delta(\overrightarrow{\alpha})}. \tag{1}$$

The process to generate words for K topics are independent of one another, we can have K conju-

gated structures and the probabilistic of words in corpus is as follows:

$$p(\overrightarrow{w}|\overrightarrow{z},\overrightarrow{\beta}) = \prod_{k=1}^{k} p(\overrightarrow{w}_{(k)}|\overrightarrow{z}_{(k)},\overrightarrow{\beta}) = \prod_{\mathrm{K}=1}^{\mathrm{K}} \frac{\Delta(\overrightarrow{\mathrm{n}}_{\mathrm{k}}+\overrightarrow{\beta})}{\Delta(\overrightarrow{\beta})} \tag{2}$$

Thus, within a document, the probability distribution over words specified by the LDA model is given as follows:

$$p(\overrightarrow{w},\overrightarrow{z}|\overrightarrow{\alpha},\overrightarrow{\beta}) = p(\overrightarrow{w},\overrightarrow{z}|\overrightarrow{\beta})p(\overrightarrow{z}|\overrightarrow{\alpha}) = \prod_{\mathrm{k}=1}^{\mathrm{K}} \frac{\Delta(\overrightarrow{\mathrm{n}}_{\mathrm{k}}+\overrightarrow{\beta})}{\Delta(\overrightarrow{\beta})} \prod_{\mathrm{m}=1}^{\mathrm{M}} \frac{\Delta(\overrightarrow{\mathrm{n}}_{\mathrm{m}}+\overrightarrow{\alpha})}{\Delta(\overrightarrow{\alpha})} \tag{3}$$

LDA is an unsupervised machine learning method, which can effectively describe the relationship between the document, the topics contained in the document, and the vocabulary contained in the topic, reflecting the underlying semantics in the document. This method uses the co-occurrence features of the terms in the text to mine the subject features of the text without any prior knowledge about the text. In addition, the LDA topic model can map text from the "document-vocabulary" high-dimensional space to the "document-topic" and "topic-vocabulary" low-dimensional space, and has very powerful dimensionality reduction capabilities. This is very valuable in an era of huge data dimensions. At present, this model has a large number of applications in the fields of image classification, text sentiment classification, recommendation system, automatic abstraction and other fields.

2.2 Fuzzy clustering

Cluster analysis is a mathematical method for classifying things according to certain requirements. The actual classification problem is often accompanied by ambiguity, so the clustering problem is more accurately solved by fuzzy mathematics. In practical fuzzy clustering problems, cluster analysis based on fuzzy equivalence relations and cluster analysis based on fuzzy pseudo-order relations are mainly used. Among them, the former is more commonly used.

$\{X_1, X_2, ..., X_n\}$ is the entire classified object, and each object x_i is represented by a set of data $\{x_{i1}, x_{i2}, ..., x_{in}\}$. Establish a fuzzy similarity relationship R on x, R can be expressed as a fuzzy similarity matrix $R = (r_{ij})_{n\times n}$, where the similarity between x_i and $x_j r_{ij}$ can be specified by one of the following methods according to the actual situation.

Quantity product:

$$r_{ij} = \begin{cases} 1, & i = j, \\ (\sum\limits_{k=1}^{m} x_{ik}x_{jk})/M, & i \neq j, \end{cases} \tag{4}$$

where M is an appropriate positive number and satisfies

$$M \geq max_{i\neq j}(\sum_{k=1}^{m} x_{ik}x_{jk}). \tag{5}$$

Angle cosine:

$$r_{ij} = \frac{\sum\limits_{k=1}^{m} x_{ik}x_{jk}}{\sqrt{\sum\limits_{k=1}^{m} {x^2}_{ik}} \cdot \sqrt{\sum\limits_{k=1}^{m} {x^2}_{jk}}} \tag{6}$$

Correlation coefficient:

$$r_{ij} = \frac{\sum\limits_{k=1}^{m} |x_{ik} - \overline{x_i}| \cdot |x_{jk} - \overline{x_j}|}{\sqrt{\sum\limits_{k=1}^{m} (x_{ik} - \overline{x_i})^2} \cdot \sqrt{\sum\limits_{k=1}^{m} (x_{jk} - \overline{x_j})^2}}, \tag{7}$$

where $\bar{x_i} = \frac{1}{m}\sum_{k=1}^{m} x_{ik}$, $\bar{x_j} = \frac{1}{m}\sum_{k=1}^{m} x_{jk}$. Others are maximum and minimum method, arithmetic mean minimum method, geometric mean minimum method, absolute value index method and so on. Which one of the above methods is chosen depends on the characteristics of the actual problem [7].

This article collects the text data of the company's public opinion through crawlers, uses the LDA algorithm to extract the keywords of the text, and develops fuzzy clustering to cluster the keywords to form different topics. The topic keywords will be used as a seed dictionary for new word discovery.

3 Data source and preprocessing

3.1 Data acquisition

The objects collected in this study are those companies that have major Internet public opinion incidents and have experienced major untrustworthy behaviors such as operational difficulties, arrears, and breach of contract in actual operations [5, 9]. The negative information on the Internet public opinion of these companies was collected (Table 1).

Table 1: Objects collected by Internet public opinion

Company Name	Industry	Main problem
LeTV	The Internet	Debt crisis
OFO	The Internet	Debt crisis
Storm video	The Internet	Arrest of legal representative
Kangmei Pharmaceutical	Pharmacy	Arrest of legal representative
Shenzhen Hemei Group Co. Ltd.	Retail	Note 1*
Shenwu Environmental Technology Co. Ltd.	Environment	Note 2**

* Note 1: Significant decline in performance, inclusion in the list of dishonest performers, suspected violation of information and laws.

** Note 2: Huge losses, financial fraud, and information disclosure violations.

3.2 Data collection methods

The corporate public opinion data is collected using crawler software. The collected online public opinion information sources include WeChat, Weibo, forums, government websites, news websites, industry forums and other information sources.

Table 2: Collected information of corporate public opinion (PO)

Company Name	PO collection time period	Collecting PO
LeTV	2014-2019	70000
OFO	2018.6-2019.4	7200
Storm video	2018.8-2019.9	2395
Kangmei Pharmaceutical	2018.8-2019.9	1515
Shenzhen Hemei Group Co. Ltd.	2018.8-2019.9	5996
Shenwu Environmental Technology Co. Ltd.	2018.8-2019.9	3601

The collected content includes title, link, news source and time. The collected content is shown in Figure 2.

3.3 Data processing

Obtaining standard comment information is only a prerequisite for cleaning public opinion information, and it is necessary to process the standard comment again, that is, word segmentation and labeling. Unlike English, we do not use spaces or punctuation to distinguish words in Chinese. This requires Chinese word segmentation technology to accurately extract Chinese words from uninterrupted Chinese character strings. At the same time, relevant information such as part of speech should be extracted. For example, after judging a word as an adjective, you can also extract the adverbs nearby to judge the strength or emotion of the adjective. Generally speaking, the more authoritative system for analyzing Chinese morphology is the Jieba word segmentation system, which is also a widely used

大规模裁员！解散整个海外部门，小黄车还饿能撑多久？	http://baijiahao.baidu.com/s?id=16024238560834
2018年ofo爆裁员、哈罗想"上位" 共享单车将上演"新三国杀"？	http://www.xinjuren.com/news/64577.shtml
乐百家	http://www.ceweekly.cn/news/nrds/1i3sncpg.html
屡陷卖身、裁员风波ofo到底黄不黄？	https://weibo.com/ttarticle/p/show?id=230961424
屡陷卖身、裁员风波 ofo到底黄不黄？	http://p.hibor.com.cn/ecodetail_5795409.html
ofo辟谣裁员消息不实海外业务解散不符 至暗时刻坚持独立运营	http://www.investide.cn/?news=384303
传ofo裁员官方否认 共享单车上游厂家业绩已疲态尽显	https://news.cnblogs.com/m/n/598433/
ofo小黄车大裁员：戴威要学王兴，拿巨头的钱做自己的事儿	http://www.twoeggz.com/news/8941511.html
从"新四大发明之一"沦为"融资黑洞"，共享单车创业给我们带来哪些启示？	http://share.iclient.ifeng.com/vampire/sharenews
ofo要失身？是摩拜与哈罗背后捅刀，还是共享单车进入死亡模式？	http://news.tuxi.com.cn/news/119999990125163/
ofo否认大规模裁员，但共享单车新一轮的洗牌还是来了...	http://news.tuxi.com.cn/news/119999990125163/
ofo疑裁员 共享单车或将上演"新三国杀"	http://www.kaiwind.com/news/info/201806/05/t20
ofo小黄车又被传"要黄了"背后都有巨头的抗衡	http://sh.qihoo.com/pc/918889745124b7dd2?sign
自从有了共享单车根本没法走路了，严重霸占公共资源//@yuange1975:大概一年前说要把那时侯开	http://weibo.com/2041017753/GjYWonRZm
小黄车快黄了？有员工称形势严峻	http://www.cmweekly.cn/xiaohua/26397.html
摩拜单车卖身美团,哈罗单车逆势第1;水即可载舟,亦可覆舟:ofo传闻说虎嗅曝光因其融资困难而大幅	http://weibo.com/5788516931/GjZed1i0e
ofo爆裁员、哈罗想"上位" 共享单车将上演"新三国	http://www.cecc.org.cn/news/201806/526056.htm
小黄车快黄了？有员工称形势严峻 应该是实在没	http://www.cccaigang.cn/zhongguo/zhongguojuns
ofo被传大规模裁员资金链紧张 小黄车也快黄了？	http://baijiahao.baidu.com/s?id=16024189190133
别了，小黄车？是韬光养晦还是危机重重……	https://kuaibao.qq.com/s/20180605A1G40500
小黄车快黄了 员工称形势严峻应该是实在没钱了	http://www.redsh.com/ppnews/20180605/180712
小黄车真要黄？被爆裁员百分之五十，还得罪了滴滴跟阿里	https://kuaibao.qq.com/s/20180605A119QS00
ofo再起风波，资本肥料终将撑死一批创业者	https://kuaibao.qq.com/s/20180605A1N1GY00

Figure 2: Sample public opinion data collection

word segmentation technology. The system supports Chinese word segmentation, can also tag parts of speech, recognize new words, and recognize named entities.

Jieba's word segmentation results are as follows:

/ w, ofo / nx, why / ryv, repeated / d, trapped / vi, capital chain / nz, crisis / n, rumor / n,? / w, David / nr, also / d, can / v, "/ w, willful / a," / w, how long is / ryt,? / w,] / w, since / p, ofo / nx, rejection of / v, Didi / q, of / ude1, acquisition of / v, offer / v, since / f, / w, about / vn, ofo / nx, / ude1, negative / b, message / n, / d, / d from time to time, / v, appears. / w, with / p, motorcycle / b, worship / v, be / pbei, beauty / b, group / n, purchase / v, / w, harrow / nz, bicycle / n, input / v, ali / nt, embrace / n, / w, one-hearted / d, want / v, maintain / v, independent / a, status / n, / ude1, ofo / nx, / w, how / ryv, talent / n, avoid / v, repeatedly / d, fall into / v, funding chain / nz, crisis / n, / ude1, rumor / n,? / w

In general, the storage format of various texts is very different, and the text may have noisy information, the system must preprocess the text to make the text meet the input requirements of the classifier.

1. De-duplication of raw data

Due to the existence of a large amount of duplicate or similar text data in corporate public opinion, the original data is pre-processed initially and the duplicate data is de-duplicated.

2. Text segmentation

This research uses the Jieba word segmentation algorithm. Its main functions are: Chinese word segmentation, part-of-speech tagging, recognition of new words, entity name recognition, etc. The word segmentation accuracy is high, and the recall rate of new word recognition based on role tagging is higher than 90%, and the part-of-speech tagging and word segmentation processing speed reaches 543.5KB/s. The word segmentation stage is to perform Chinese word segmentation, part-of-speech tagging for each text, and output its results in a prescribed format. Because some special vocabulary algorithms cannot be directly identified, during the word segmentation process, some special words such as "Little Yellow Car" and "Davi" are manually labeled to establish a word segmentation dictionary.

3. Filter stop words

In the word collection obtained in the word segmentation stage, many words are meaningless. The impact of these words on the analysis work can be ignored, but if these words are not used as text feature words, Bringing large errors to text classification results, these words are often referred to as stop words in this article. Removed words marked as part of speech: orientation, preposition, quantifier, auxiliary, punctuation, non-morpheme. These words are like, "Yeah, this, if, that, then." If these words are added to the text analysis, it will increase the cost of our text analysis. Therefore, we introduce a stop-word dictionary and remove these stop-words in the process of word segmentation.

According to all available resources, various stop-word lists such as "HIT's Stop Words Thesaurus", "Sichuan University Machine Learning Intelligent Laboratory Stop Thesaurus", and Baidu's Stop Words will be organized to focus on extracting Chinese Words (instead of a large number of English words and Chinese punctuation marks) came out of a more comprehensive vocabulary, with a total of 1598 stop words.

4 Fuzzy clustering for public opinion topics

4.1 Topic extraction

Before feature fusion, we first need to extract features from the text. In the following, feature extraction is performed on the text in two ways. Topic features are extracted using the LDA topic model, and then word2vec is used to lexicalize the topics into word vectors. Because the original feature data of the LDA model is a bag-of-words model, a bag-of-words model is required before extracting topic features. LDA is a three-layer Bayesian model. The first and second layers represent the probability distribution of topics under the document, and the second and third layers represent the probability distribution of terms under the topic (Figure 3).

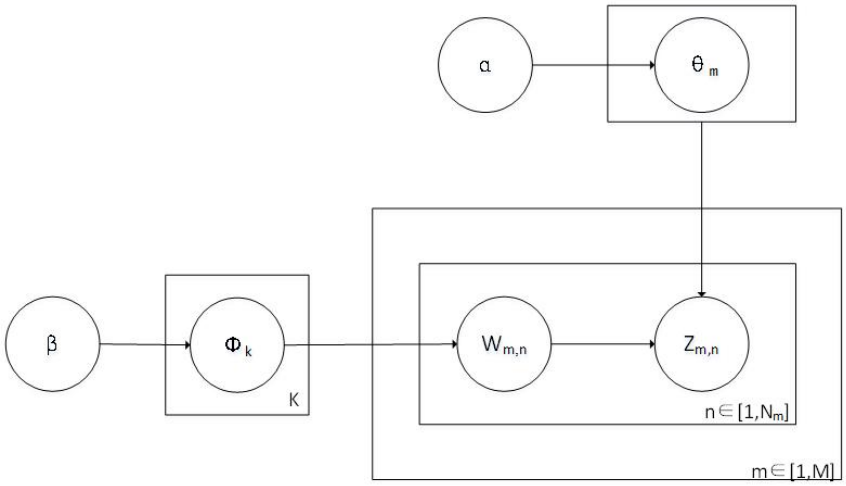


Figure 3: Bayesian model of LDA

In Figure 3, α and β are Dirichlet prior hyper parameters; the words in document M will be clustered into Z topics, for each topic $Z \in \{1, 2, ..., k\}$, and sample distribution is $\phi_k \sim Direchlet(\beta)$. Bayesian model is as follows [17]:

1. Choose $N \sim Possion(\xi)$
2. Choose a topic distribution $\theta \sim Direchlet(\alpha)$
3. For each of word $w_{m,n}$ in m th document:

(a) Choose a topic of the word $Z_{m,n} \sim Multinomial(\theta_m)$

(b) Choose a word $w_{m,n} \sim Multinimial(\phi_{z_{m,n}})$

We get the following results based on fuzzy clustering algorithm:

0, ’0.155 * "Video" + 0.080 * "Initiated" + 0.076 * "Project" + 0.041 * "Previous" + 0.028 * "Recommended" + 0.026 * "This year" + 0.021 * "Title" + 0.021 * "Youku" + 0.020 * "Nobody" + 0.017 * "Domestic" ’ 1, ’0.144 * "Problem" + 0.073 * "True" + 0.057 * "Credit" + 0.053 * "Now" + 0.045 * "Solved" + 0.022 * "Done" + 0.020 * "Maintenance" + 0.016 * "Daily "+ 0.015 *" Xiaobian "+ 0.011 *" below "’ 2, ’0.311 * "Shareholder" + 0.166 * "WeChat" + 0.092 * "Fund" + 0.047 * "Exposure" + 0.042 * "Display" + 0.017 * "Brand" + 0.017 * "Cool" + 0.015 * " Multi-site "+ 0.010 *" Company announcement "+ 0.007 *" Picture "’ 3, ’0.127 * "Published" + 0.067 * "Payment" + 0.052 * "" + 0.041 * "Encounter" + 0.040 * "Media" + 0.037 * "Data" + 0.025 * "Corruption" + 0.021 * " House Leakage "+ 0.018 *" Frender "+ 0.017 *" Recent "’, 4, ’0.086 * "Disputes" + 0.059 * "Half year" + 0.056 * "Billion dollars" + 0.042 * "This house" + 0.030 * "Late" + 0.017 * "Drag" + 0.016 * "Shuffle" + 0.015 * "One Out" + 0.015 * "Approach" + 0.012 * "Internet Biography" ’ 5, ’0.547 * "100 million yuan" + 0.081 * "goodwill" + 0.046 * "font size" + 0.026 * "look" + 0.015 * "maintenance" + 0.011 * "fall" + 0.008 * "hurry up" + 0.008 * "Finance Network" + 0.004 * "Let’s wait and see" + 0.004 * "About 100 million" ’ 6, ’0.128 * "Smart" + 0.071 * "Super" + 0.058 * "Burst" + 0.055 * "Economic Network" + 0.037 * "Someone" + 0.033 * "Cause" + 0.030 * "Tesla" + 0.015 * "attribution" + 0.015 * "channel" + 0.014 * "list" 7, ’0.214 * "Received" + 0.076 * "Reporter" + 0.073 * "Partner" + 0.053 * "Securities Daily" + 0.011 * "Burning money" + 0.011 * "Follow the trend" + 0.009 * "Business newspaper" + 0.009 * "No solution" + 0.009 * "Foundation" + 0.008 * "Suppress" 8, ’0.289 * "Executive" + 0.106 * "Roll-call" + 0.074 * "Since the month" + 0.022 * "Block" + 0.012 * "Clarification" + 0.009 * "Management" + 0.007 * "Home" + 0.006 * "Suspect" + 0.005 * "Entrepreneur" + 0.004 * "sunyuchen" ’ 9, ’0.410 * "listed company" + 0.080 * "default" + 0.070 * "one quarter" + 0.011 * "chicken feathers" + 0.011 * "web news" + 0.010 * "billions" + 0.008 * "yesterday" + 0.007 * "call it" + 0.006 * "smooth" + 0.005 * "order" ’ 10, ’0.107 * "fund" + 0.093 * "headline" + 0.087 * "ten billion" + 0.079 * "internet" + 0.053 * "multiple" + 0.029 * "only" + 0.018 * "sing empty" + 0.016 * "Focus" + 0.015 * "Include" + 0.014 * "One step" ’

4.2 Subject fuzzy clustering results

In recent years, due to the development of the Internet, corporate information disclosure is more rapid and faster than ever. The use of corporate Internet public opinion information as an indicator of corporate credit evaluation has received increasing attention from scholars in this field. There are few existing researches on the subject classification. In this paper, keywords are obtained based on LDA, and fuzzy clustering methods are used to aggregate 9 subject categories, including operations, employees, funds, justice, capital, corporate negative news, and regulatory punishment, founders and management, corporate credit, etc. The specific classification is shown in Table 3.

5 Word2vec-based thesaurus machine learning model construction

5.1 Word vectors

Computers cannot directly understand human languages. When dealing with natural language problems, scholars have used some features in the text to cleverly map them into digitized dimensional sequences. Google opened Word2vec in 2013 to train word vectors, which can express words in the form of vectors. The *Word2vec* model contains two types of word vector learning structure models: Skip-Gram and CBOW (Continuous Bag of Words Model) models. Both structures include an input layer, a mapping layer, and an output layer. When it is determined that the number of words w context words is n, the Skip-Gram model predicts the context of the current word. The CBOW model uses contextual vocabulary to predict the current word (Figure 4). Skip-Gram model predicts the context based on the current word, given the word sequence $W = \{w_1, w_2, ...w_m\}$, the model maximizes the average log probability as:

$$l(W) = \frac{1}{M} \sum_{m=1}^{M} \sum_{-L \le i \le L} \log p(w_{m+i}/w_m), \tag{8}$$

where L is the size of the context window.

The CBOW model predicts target words by specifying window words, given a word sequence $W = \{w_1, w_2, ...w_m\}$, the model maximizes the average log probability as:

$$l(W) = \frac{1}{M} \sum_{i=L}^{M-L} \log p(w_i/w_{i-L}, ..., w_{i+L}). \tag{9}$$

Table 3: Types and keywords of clustering results

Topic	Topic category	Topic keywords
Topic1	Operation	Bankruptcy, stoppages, emptying, layoffs, dilemmas, stagnation, dismissal, struggling, sinking, bad debt, defeat, defeat, winter, collapse, failure, loss of debt, debt collection, embezzlement, dissatisfaction, empty numbers, difficulties, life extension, bump, stop, shrink
Topic2	Employee	Rights protection, wages, wages, reporting, arrears of wages, compensation, complaints, banners, disputes, arrears, names of affiliated companies, payroll, comfort, claims, severance, resignation, layoffs, goodwill, debt collection, dumping, layoff
Topic3	Funds	Deposit,debt, crisis, arrears, cash flow, funds
Topic4	Judicial	Lawyer letter, case, legal representative, prosecution, enforced person, court, application, Lao Lai, judgment, claim, lawsuit, arrest, filing, law, rights protection, freeze, violation, breach of trust, breach of law, suspect, charge, overdue, fraud seized
Topic5	Capital	Mergers and acquisitions, acquisition, unicorn, equity, pledge, shrinking, financing, evaporation, stock price, liquidation, falling, limit, falling back, continuous board, flight, cash cut, liquidation, backdoor, collapse, delisting, clearance Bearish, manipulation, decline, repurchase, hollowing out, reduction, cash, shareholder name
Topic6	Corp. negative news	Smeared, survived, finished, negative news, cool, shocking, rumors, bursts, unexpected, urgent, dark, clarification, scandal, worst, failure, storm, halberd, fall, black list, shelling, stinking
Topic7	Regulatory penalties	Punishment, warning, fine, inquiry, notification of criticism, investigation, ticket, heavy penalty, prohibition of employment, prohibition of entry, supervision, interview, CSRC, China Consumers Association, MIIT, Shenzhen Stock Exchange, Shanghai Stock Exchange
Topic8	Management negative	Runaway, leave, responsibility, sale, self-help, kick out, name of manager
Topic9	Negative credit	False, liar, question, doubt, true and false, deceived, doubt, inquiry, scam, flicker, crisis of trust, recidivism, nonsense, tearing, trap, blacklist, broken faith, executed

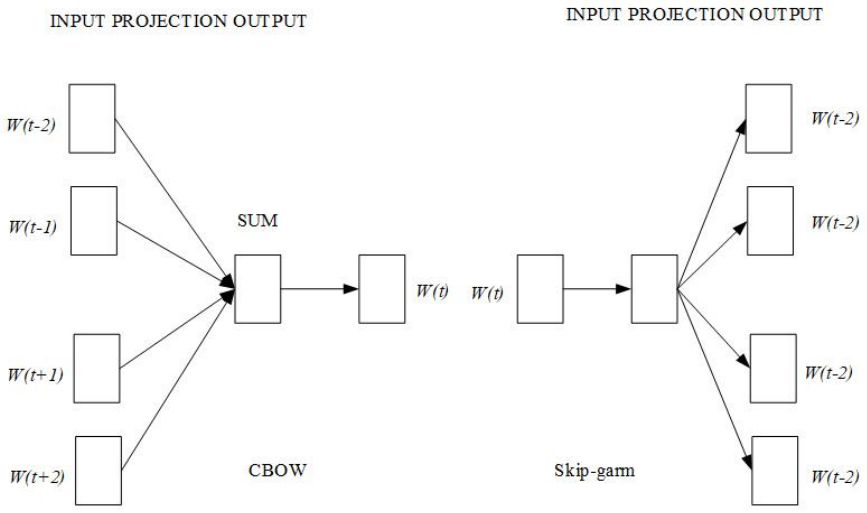


Figure 4: Principle of Word2vec algorithm

Steps for network calculation:

- Input layer: onehot of the context word. (Assuming the word vector space dim is V and the number of context words is C)
- All onehots are multiplied by the shared input weight matrix W ($V \cdot N$ matrix, N is a number set by yourself, and the initialization weight matrix W)
- The resulting vector (note the result of multiplying the onehot vector by the matrix) is added and averaged as the hidden layer vector, with size $1 \cdot N$.
- Multiplied by the output weight matrix $W'\{N \cdot V\}$
- The vector $\{1 \cdot V\}$ is obtained by processing the activation function to obtain the V-dim probability distribution PS: because it is onehot, each dimension represents a word, and the word indicated by the index with the highest probability is the predicted middle word (Target word)
- Compared with onehot of true label, the smaller the error, the better. loss function (generally a cross-entropy cost function)

Use the public opinion keywords in the seed dictionary as seed words, bring them into the trained word vector, query for words similar to the seed word, and select the first five similar words as synonyms to be included in the dictionary. These words will also be used as seeds again. Words to get more synonyms, take "false" as an example, and use the new word recognition algorithm to get the top 5 synonyms, as shown in Table 4.

Table 4: Word similarity

No	Word	Similarity
1	Imaginary	0.7475070953369141
2	Misleading	0.6770043969154358
3	Fabrication	0.6676515340805054
4	Fraud	0.659699559211731
5	False report	0.6317102313041687

Some results are shown in Table 5.

5.2 Results analysis

The topic keyword extraction structure is constructed into a star network, as shown in the Figure 5. It can be seen that under some topics, more keywords are found, and some keywords are less. This is due to the evolution of corporate network public opinion and reflects the main problems behind the public opinion of the company, according to the number of theme keywords, can predict the development trend of public opinions [22, 23].

The topic keywords can also be used to construct a co-occurrence relationship network, as shown in Figure 6. This network abstracts words into nodes in the network, and abstracts the co-occurrence relationship of words in the word set. It is called an edge in the network, that is, if 2 words appear in the same word set at the same time, it is considered that there is an edge between them. The edge is an abstraction of the co-occurrence relationship between words in a word set. Through the co-occurrence relationship network, tap the internal connection of corporate public opinion [18].

All the nodes of a subgraph are adjacent to at least k other points in the subgraph. Such a subgraph is called a K-core. K-kernel analysis can find the most closely related nodes in the network, and these nodes can summarize the structural characteristics of the network [20]. Through the K-core diagram, local problems of the public opinion network can be found, as shown in Figure 7:

The final plan identified 3092 keywords for corporate public opinion. There are three main evaluation indicators used in this article, namely accuracy rate P, recall rate R, and F value. Its formula is

$$p = \frac{s_r}{s_a}. \tag{10}$$

Table 5: Thesaurus of synonyms

Topic category	Topic keywords
Operation	'Corporate Bankruptcy', 'Closed', 'Debt', 'Liquidation', 'Creditors',' Resumption ',' Stop Construction ',' Arrears', 'Stop Production', 'Engineering Payments',' No Trace ',' Moving 'Empty', 'Empty', 'Empty', 'Empty', 'Salary Reduction', 'Closed', 'Resignation', 'Salary Reduction', 'Management', 'Dilemma', 'Crisis',' Dilemma ',' Disengagement ',' Overcoming difficulties ',' Stagnation ',' Recession ',' Imbalance ',' Intensification ',' Weakness ',' Succession ',' New appointment ',' Assignment ',' Assignment ',' Treasurer ',' Get rid of ',' Sink ',' Pain ',' Despair ',' Tear ',' Trapped ',' Unable to extricate ',' Trapped ',' Trapped ',' Distressed ',' Impairment ',' Bad debts', 'Receivables',' Accounts', 'Bad debts',' Routing ',' Battle situation ',' Great victory ',' Reversal ',' Rivals', 'Counter attack', 'Retreat', 'Retreat', 'Retreat', 'Losing Soldiers',' Early Winter ',' Cold ',' Advent ',' Cold Wind ',' Severe Cold ',' Subprime Crisis', 'Stock Market', 'Finance Crisis', 'financial market', 'slump', 'bankruptcy', 'mismanagement', 'business failure', 'layoffs',' chain of funds', 'usile loan', 'lost money', 'lossy losses', 'Run the road', 'Home broke', 'Debt repayment', 'Debt collection', 'Debt', 'Creditor', ' ',' Carry out ',' Fake false report ',' Private share ',' Illegal ',' diarrhea ',' Altitude response ',' Sickness medicine ',' Cultural conflict ',' Diarrhea ',' Dilemma ', 'Bottle Bottle', 'Crisis of Trust', 'Crisis', 'Facing', 'Blessing', 'Supernatural Power', 'Saving', 'Rebirth', 'Golden Body', 'Stop Production', 'Reproduction', 'Suppression' , 'Banned according to law', 'return to work', 'capillary', 'diastolic', 'dilatation', 'elasticity', 'weakened'
Staff	'Consumption Association', 'Disputes', 'Complaints', 'Legal Means', 'Wages In Arrears', 'Arrears of Wages', 'Migrant Workers', 'Arrears', 'Appeals', 'Monthly Wages', 'Basic Wages' , 'Wage treatment', 'Treatment', 'Performance', 'Report phone', 'Complaint', 'Reporting Centre', 'Arrears of Wages',' Salary ',' Arrears', 'Migrant Workers',' Labor Disputes ',' Compensation ',' claim ',' infringer ',' compensation for damages ',' obligation of compensation ',' report ',' consumer ',' banner ',' banner ',' slogan ',' publicity Board ',' publicity card ',' civil dispute ',' mediation ',' economic dispute ',' mediation ',' debt dispute ',' project payment ',' arrears', 'payments',' paid wages', 'Payment', 'Wages',' Arrears of Wages', 'Unable to Pay', 'Payrolls',' Processing Capital ',' Compensation ',' Mental Loss Fee ',' Mental Loss', 'Compensation', ' Lost work ',' Compensation ',' Compensation ',' Rejection of compensation ',' Plaintiff ',' Claim ',' Order ',' Removal ',' Check ',' Escort ',' Expulsion ',' Resignation ',' Resignation ',' Changing ',' Dismissal ',' Resignation ',' Salary Reduction ',' Closed ',' Resignation ',' Salary Reduction ',' Management ',' Intangible Assets ',' Trademark Rights ',' Unfair competition ',' Shareholders 'equity,' Tangible assets ',' Repayment of debts ',' Debt collection ',' Debts ',' Creditor ',' Arrears'

height

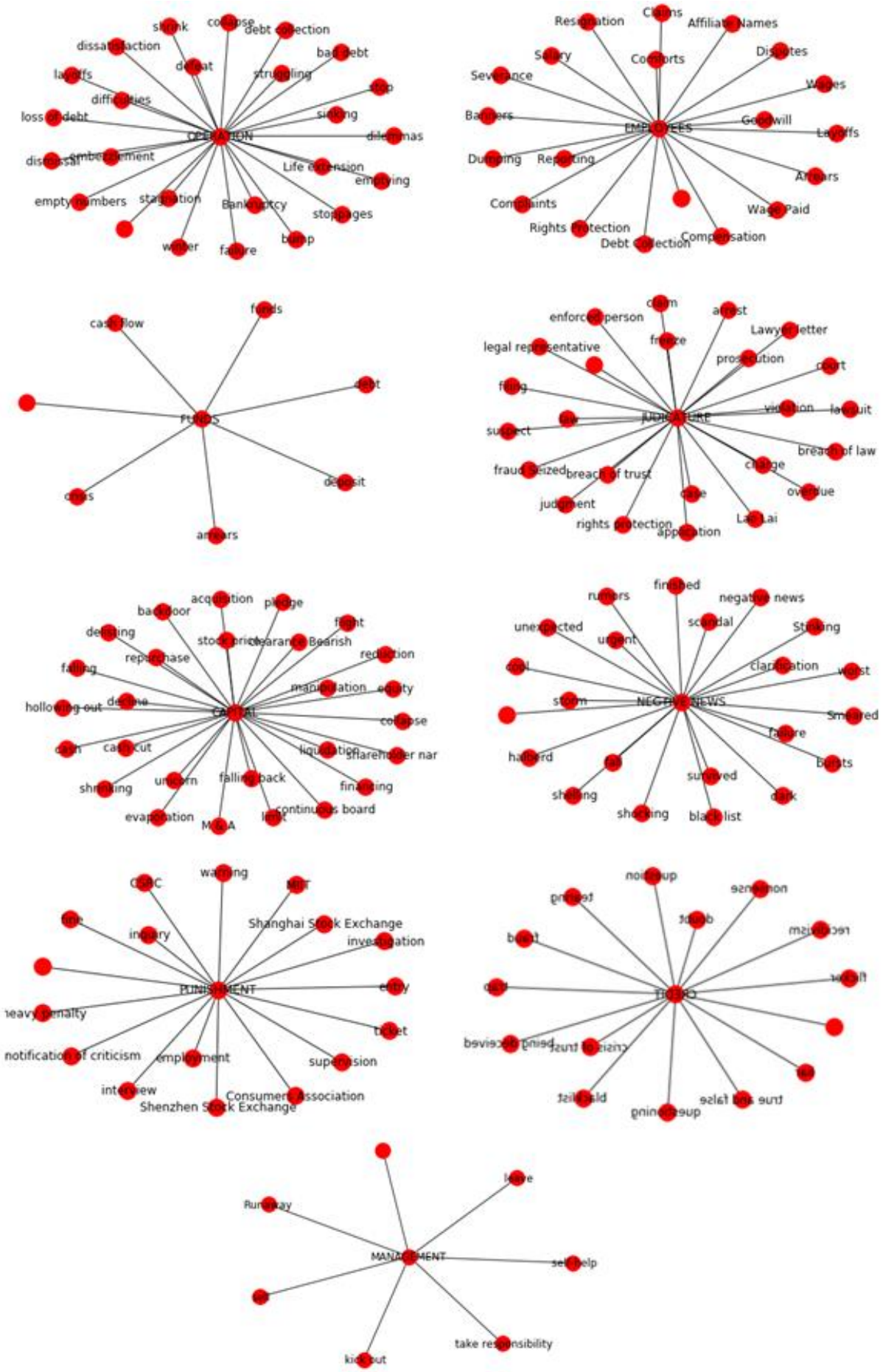


Figure 5: Star networks of subject words

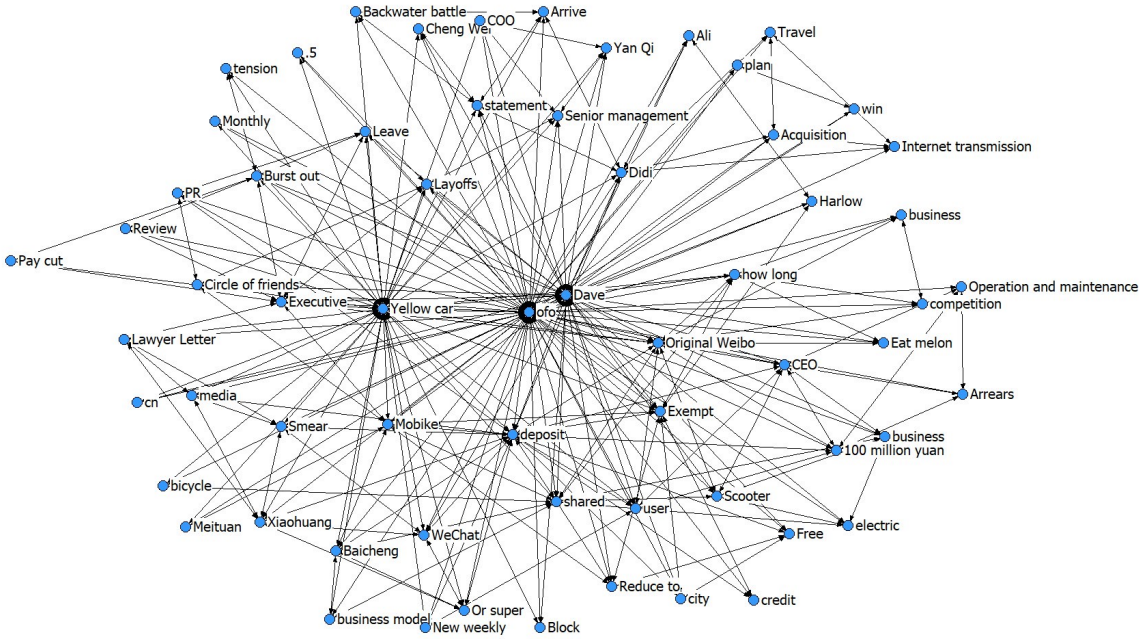


Figure 6: Keywords co-occurrence network

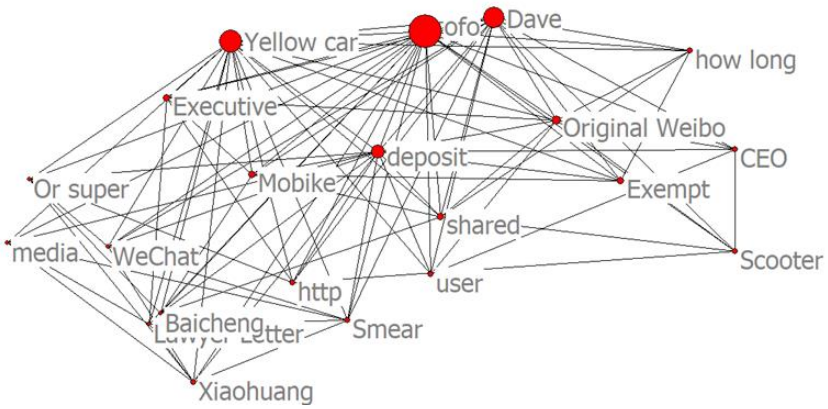


Figure 7: K-core network

$$R = \frac{s_r}{s_o}. \tag{11}$$

$$F = \frac{2 \times P \times R}{P + R}. \tag{12}$$

s_r is the number of texts classified correctly; s_a is the number of texts actually classified; s_o is the number of texts belonging to the category; F values are calculated from accuracy rate P, recall rate R.

In order to analyze the characteristics of the new word discovery algorithm in this paper, the Word2vec algorithm is compared with the method based on association rules, the method based on N-Gram, and the method based on mutual information PMI in the experiment. Analysis, the experimental results of the comparison of the last algorithms are shown in Table 6.

Table 6: Experimental results

Method	P	R	F
Association rules	0.521	0.471	0.420
N-GRAM	0.623	0.542	0.561
PMI	0.737	0.731	0.720
Word2vec	0.778	0.778	0.771

It can be seen from Table 6 that in the comparison test with the mutual information PMI, the association rule-based and the N-Gram-based algorithm, the Word2vec algorithm has achieved a good performance improvement for the topic new word discovery of corporate network public opinion. It has been improved. The association rule-based discovery algorithm has such a rapid change, poor vocabulary standardization, and poor adaptability to emerging public opinion vocabulary that has large differences in word structure, resulting in low accuracy and recall of new words. For the N-Gram-based algorithm, although the accuracy and recall rate have been improved compared with the association rule-based algorithm, in the case of multi-words, this method has a partial recognition rate for new words. low. The mutual information PMI algorithm is based on the N-Gram method, combined with the mutual information to merge multiple words, and has further improved in various indicators. The new word algorithm based on Word2vec used in this article obtained the highest index value because the algorithm uses a neural network to combine the contextual relationships between words to vectorize the words of the text and calculate the distance between words, which can quickly and accurately Discover new words.

6 Conclusion

This paper preliminarily constructs a text analysis framework for corporate Internet public opinion, and analyzes corporate Internet public opinion from text topic mining, keyword thesaurus construction. This article first combines the vocabulary mined by public opinion text topics with corporate credit evaluation models, and divides corporate online public opinion texts into operations, employees, funds, justice, capital, negative corporate news, regulatory penalties, founders and management, and companies. Credit nine categories and get the keywords under that topic as a seed dictionary. Based on the new word discovery algorithm, a keyword thesaurus for corporate public opinion keywords is constructed. The mutual information PMI, association rules, and N-Gram and Word2vec-based new word discovery algorithms are compared. The results show that Word2vec-based new word algorithms have improved accuracy, recall, and F value. This is pre-trained with Word2vec algorithm. Text, vectorized vocabulary are closely related. There are also some shortcomings in this study. The selected cases are limited and there are only six companies, so there is still room for improvement in the topic classification of public opinion texts, and further subdivided or increased classification.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Adreevskaia, A.; Bergler, S. (2006). Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In 11th conference of the European chapter of the Association for Computational Linguistics, 2006.

[2] Agerri, R.; García-Serrano, A. (2010, May). Q-WordNet: Extracting Polarity from WordNet Senses. In LREC, 2010.

[3] Baccianella, S.; Esuli, A.; Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204), 2010.

[4] Blei, D.M.; Ng, A.Y.; Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993-1022, 2003,.

[5] Chu, X.; Zhong, Q.; Li, X. (2018). Reverse channel selection decisions with a joint third-party recycler. *International Journal of Production Research*, 56 (18):5969-5981, 2018.

[6] David, M.; Blei, J.; Lafferty, D. (2005) Correlated Topic Models// Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]. MIT Press, 2005.

[7] D'Urso, P.; Leski, J.M. (2019). Fuzzy clustering of fuzzy data based on robust loss functions and ordered weighted averaging. *Fuzzy Sets and Systems*, 2019.

[8] Goldberg, Y.; Levy, O. (2014). Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722, 2014.

[9] Gong, D.; Liu, S.; Liu, J.; Ren, L. (2019). Who benefits from online financing? A sharing economy E-tailing platform perspective, *International Journal of Production Economics*, DOI: 10.1016/j.ijpe.2019.09.011, 2019.

[10] Griffiths, T.L.; Jordan, M.I.; Tenenbaum, J.B., et al. (2004) Hierarchical topic models and the nested Chinese restaurant process//Advances in neural information processing systems, 17-24, 2004.

[11] Griffiths, T.L.; Steyvers, M.; Blei, D.M., et al. (2005) Integrating topics and syntax//Advances in neural information processing systems, 537-544, 2005.

[12] Hassan, A.; Radev, D. (2010, July). Identifying text polarity using random walks. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 395-403). Association for Computational Linguistics, 2010.

[13] Hu, M.; Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM, 2004.

[14] Li, L.; Li, W. (2019) Naive Bayesian Automatic Classification of Railway Service Complaint Text Based on Eigenvalue Extraction. *Tehnički vjesnik*, 26(3): 778-785, 2019.

[15] Mcauliffe, J.D; Blei, D.M. Supervised topic models//Advances in neural information processing systems. 121-128, 2008.

[16] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[17] Mikolov, T.; Le, Q.V.; Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168, 2013.

[18] Snellman, L. (2016). Social Entrepreneurship: Making change in the world. Journal of Logistics, *Informatics and Service Science*, 3(1), 1-25, 2016.

[19] Wang, X; McCallum, A. (2006) Topics over time: a non-Markov continuous-time model of topical trends//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 424-433, 2006.

[20] Wei, K., Gou, J., Chai, R., & Dai, W. (2013, September). Creation of customer evaluation model in the catering industry supply chain ecosystem. In 2013 5th International Conference on Intelligent Networking and Collaborative Systems (pp. 751-756). IEEE, 2013.

[21] Zhang, Q.; Liu, S.; Gong, D.; Tu, Q. (2019). A Latent-Dirichlet-Allocation Based Extension for Domain Ontology of Enterprise's Technological Innovation. *International Journal of Computers Communications & Control*, Vol. 14, No.1, pp.107-123, 2019.

[22] Zhang, D. (2017). High-speed train control system big data analysis based on the fuzzy rdf model and uncertain reasoning. *International Journal of Computers Communications & Control*, 12(4), 577-591, 2017.

[23] Zhang, D.; Sui, J.; Gong, Y. (2017). Large scale software test data generation based on collective constraint and weighted combination method. *Tehnicki vjesnik*, 24(4), 1041-1050, 2017.

Cite this paper as:
Luo, J.; Yu, D.; Dai, Z. (2020). A Latent Dirichlet Allocation and Fuzzy Clustering based Machine Learning Model for Text Thesaurus, *International Journal of Computers Communications & Control*, 15(2), 3811, 2020.
https://doi.org/10.15837/ijccc.2020.2.3811