Quantitative Finance and Economics

http://www.aimspress.com/journal/QFE

QFE, 2(2): 279–293.
DOI:10.3934/QFE.2018.2.279
Received: 29 November 2017
Accepted: 17 January 2018
Published: 19 April 2018

Research article

Topic generation for Chinese stocks: a cognitively motivated topic modeling method using social media data

Wenhao Chen[1,*], Kinkeung Lai[2] and Yi Cai[3]

[1] Department of Management Science, City University of Hong Kong, Hong Kong
[2] Department of Industrial and Manufacturing Systems Engineering, Hong Kong University, Hong Kong
[3] School of Software Engineering, South China University of Technology, Guangzhou, China

* **Correspondence:** Email: wenhachen2-c@my.cityu.edu.hk; Tel: +66452907.

Abstract: With the explosive growth of user-generated data in social media websites such as Twitter and Weibo, a lot of research has been conducted on exploring the prediction power of social media data in financial market and discussing the correlation between the public mood in social media and the stock market price movement. Our previous research has demonstrated that the topic-based public mood from Weibo can be used to predict the stock price movement in China. However, one of the most challenging problems in topic-based sentiment analysis is how to get the relevant topics about a stock. The relevant topics are also considered as concepts about a stock which can be used to build the ontology of stock market for semantic computing and behavioral finance research. In this paper, motivated by the basic level concept in cognitive psychology, we present a novel method using Latent Dirichlet Allocation (LDA) to generate topics about a stock based on the social media data. The experimental results show that the proposed method is effective and better than other topic modeling methods. The topics generated by our method are more interpretable and could be used for topic-based sentiment analysis.

Keywords: topic modeling; cognitive psychology; semantic computing; text mining; financial engineering
JEL classification numbers: C55

1. Introduction

In recent years, social media websites, such as Facebook, Twitter and Weibo become more and more popular. Large volume of data is published by online users in these websites. Even some organizations and companies use social media websites as their official channels to publish their opinions and announce new products to the public. In China, Weibo is the largest microblog social meida website which provides Twitter-like service. As of mid 2017, there are around 361 million monthly active users, more than 159 million daily active users based on the company financial report of Sina Weibo. Weibo is consisted of large volume of posts. A post in Weibo is a short text including words, sentences and emojis. A lot of users publish their opinions about some special topics such as daily life, new products, government policy and financial market in the website through posts (Gao et al., 2012). Most of the posts are written in Chinese. In addition, Weibo has an approval mechanism for registration which can help to confirm the user in the system is a real person instead of a robot program. As a result, the data collected from these social media websites, to some extent, represents users' opinion which provides a data set for semantic computing and human behavior analysis.

As the stock market in China has already been one of the largest stock market around the world, it is an important research topic that to find out how the stock price movement is affected by human behavioral factors in China. The social media website, Weibo, provides a useful data source for understanding human behaviors as users post their opinions about Chinese stock market and comment on each other's post in the website. Previous research (Chen et al., 2016) indicated that the large scale Weibo posts represent the real public opinions about certain topics in the stock market and the topic-based mood from Weibo can be used to predict the stock price movement in China. However how to generate topics is not discussed in the research. As a result, in this paper, we want to solve the problem of topic generation for Chinese stocks. The topics about a stock could be some financial terms, relevant news, economic or other industry factors in addition to features about the stock. These topics can help investors to understand the stock better and collect more information which could impact the stock price based on the topics. In addition, these topics can be used as concepts to build an ontology for stock markets which will help to explore the semantics from other user-generated multimedia resources.

As shown in Figure 1, there are many topics related to 'Shanghai Stock Exchange Composite Index"(SSECI). "sh000001" is the instrument code for SSECI. "Interest rate cut" and "Circuit breaker" is the policy related to SSECI and will impact the price of the index. Furthermore "Volume" and "Industry Sector" are the features of SSECI. To understand the complete public opinion about a stock, we need to collect users' opinion about all topics related to SSECI. To solve the topic generation problem, motivated by previous topic modeling and cognitive psychology research, we propose a novel term weighting LDA method in this paper. The topics shown in Figure 1 is part of the result when using our proposed method to generate topics about SSECI from Weibo data. The reason of using Weibo as the source is because of its large volume posts about stocks in China. If we don't know the topics about SSECI, sentiment analysis can only be conducted on the text including the keyword "Shanghai Stock Exchange Composite Index" in social media websites. As topics about SSECI can be generated by our proposed method, the keyword list is extended to the topic expressions and words. Using these topic words such as "interest rate cut" to search in social media websites, we can get more useful data about SSECI for sentiment analysis and opinion mining.

Topic modeling methods such as LDA are normally used for generating topic or aspect expressions (Jo and Oh, 2011) (Mimno et al., 2011) from textual documents. However, Mimno et al. (2011) found that the topics generated by LDA sometimes are not interpretable and have no meaning. The reason is that some common words are included in different topics. As a result, LDA should involve term weights in the model. In this paper, we extend the work presented in (Yang et al., 2016) on

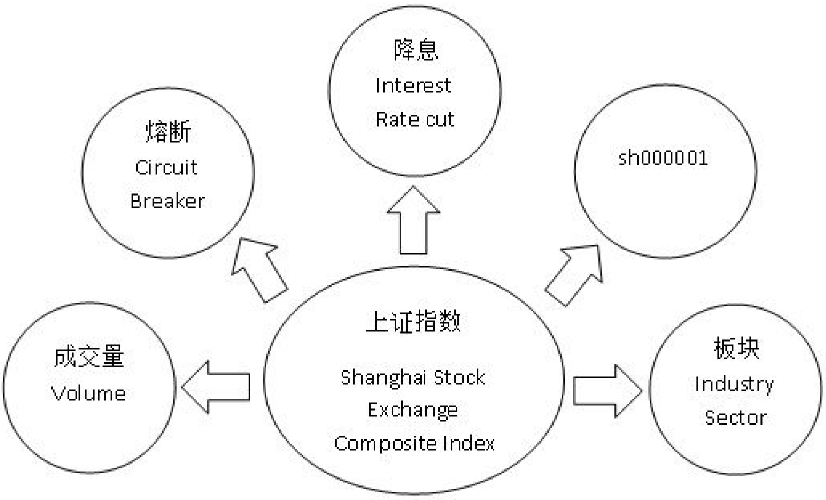

Figure 1. Words related to “Shanghai Stock Exchange Composite Index”.

term weighting LDA and propose an unsupervised method to extract the latent stock topics. Different term weighting schemes are discussed in our research which are used to measure the power of words for topic discriminating. To enhance the topic modeling ability of LDA, we include the theory of basic level concepts and introduce a new category utility like term weighting scheme. In cognitive psychology, there is a family of concepts named basic level concepts which are most differentiated from one another and most human knowledge is organized by basic level concepts (Rosch et al., 1976). Basic level concepts are demonstrated to have the highest category utility (Gluck, 1985). The contribution of our work is as follows:

(a) To analyze Chinese social media text data, we provide an method to enhance the Chinese segmentation ability.

(b) How to use term weighting schemes in LDA is discussed and we introduce a new term weighting scheme based on the theory of basic level concepts in cognitive psychology to determine topic discriminating power of words. Based on the new term weighting scheme, we propose a novel term weighting LDA method in this paper.

(c) To demonstrate the effectiveness of the proposed method, we conduct experiments comparing the performance of our method with standard LDA and term weighting LDA using different term weight schemes. The results show that the proposed method outperforms other methods. The topics from our model are proved to be more differentiated from one another and self-explained through manual tagging and classification approach.

2. Related work

2.1. Social media analysis

Over the past several years, as the development of Internet, social media websites such as Twitter and Weibo have received much attention due to their enormous users and user-generated content. Abbasi and Chen (2008) have demonstrated that the information retrieval and automated analysis technology are useful for understanding the online content such as forum posts and interactions in social media. Through analyzing the content in social media, Liang et al. (2009) indicated that a company can get the first hand knowledge or feedback from its clients. And it can also help to understand how the online customer networks appear and evolve (Chau and Xu, 2007). In business intelligence, this kind of data generated from social media websites can help organizations to make business decisions (OLeary, 2011). Gruhl et al. (2005) used the sentiment from twitter to forecast spikes in book sales. Mishne and Glance (2006) also used the twitter sentiment to predict the revenues of box-office for movies in North America.

For financial market, some research has been conducted to use the sentiment analysis result from social media to forecast the stock price. Schumaker and Chen (2009) applied machine learning methods to financial news articles and found that the sentiment in news articles has an immediately impact on the market price. Based on the sentiment score retrieved from the posts on the Yahoo Finance Forum, Liu et al. (2006) has indicated the correlation between the sentiment score and the stock price. Gibert and Karahalio (2010), using the LiveJournal as a source, have extracted the anxiety, worry and fear mood from the posts in that website. They found that the increase on expression of anxiety has indicated that the S&P 500 Index will move downward soon. Bollen et al. (2011)has investigated the correlation between the collective mood states from large-scale twitter feeds and the value of the Dow Jones Industrial Average (DJIA) over time. Using twitter posts as well, Zhang et al. (2011) found that the twitter sentiment can be used to predict NASDAQ and S&P500 index as well. Li et al. (2014) discussed the problem in using news to predict stock price in Hong Kong.

As Chinese text is complicated in terms of segmentation, the number of research on Chinese social media mining is limited. In previous research, Gao et al. (2012) indicated the difference between Weibo and Twitter. Yang et al. (2012) have proposed a classifier to automatically detect the rumors from the posts in Weibo. For sentiment analysis, Fan et al. (2014) found out that the correlation of anger among users is significantly higher than that of joy in Weibo. In addition, Chen et al. (2016) demonstrated that public mood states extracted from the large scale Weibo posts represent the real public opinions about some special topics of the stock market. These topic-based public mood states from Weibo are used to predict the stock price movement.

2.2. *Topic modeling*

In general, topic modeling is a text mining method which could be used to identify conceptual topics and generate topic expressions. The two widely-used topic modeling methods are PLSA (Probabilistic Latent Semantic Analysis) (Hofmann, 2001) and LDA (Latent Dirichlet Allocation). PLSA is an unsupervised learning method which is based on a statistical latent-class model. To solve the overfitting problem in PLSA, Blei et al. (2003) proposed LDA which is a generative method introducing Dirichlet prior distributions over document-topic and topic-word distributions. Guo et al. (2009) used multilevel LDA to categorize product aspects. Although LDA is widely used, it has the limitation that some topics generated will mix general words which makes the topics difficult to be interpreted. To enhance the ability of standard LDA in generating unmixed topics, Andrezejewski et al. (2009) designed a new model called DF-LDA which takes domain knowledge given by users into consideration. Wilson and Chew (2010) claimed that LDA should involve weights of words in document in the model as the words which scatter across more documents are less important and should have lower weights. Yang et al. (2016) proposed a Term Weighting LDA algorithm based on the discussion of the topic discriminating power of words.

2.3. Basic level concept

In cognitive psychology, in a hierarchical category structure such as a taxonomy of plants, there is one special level of concepts named the basic level, at which the categories are cognitively basic. The basic level categories, defined by Rosch et al. (1976), carry the most information and are the most differentiated from one another in human minds. They are the categories easier than others to learn and recall by humans as concepts. In psychology, generally a concept holds the common features of a category of instances and is the abstraction of that category. Basic level concepts are the abstraction of basic level categories. Objects are identified as belonging to basic level categories easier than other categories, and recognized as the basic level concepts faster than other concepts. For example, in classifying life forms, basic level categories tend to be at the level of the genus (maple, dog etc.). When we see a maple, we could call it a "plant", a "maple" and a "sugar maple", but most people will identify it as "maple". The concept "maple" is a basic level concept.

Through a lot of experiments, psychologists (Gluck, 1985) gave the metric, category utility, to characterize basic level concepts. They demonstrated that the character of basic level categories was that they had the highest category utility. Category utility was intended to supersede more limited measures of category goodness such as cue validity and collocation index. It provides a normative information-theoretic measure of the predictive advantage gained by the person who possesses knowledge of the given category structure over the person who does not possess this knowledge. Some research has been conducted on using category utility for semantic computing. Chen et al. (2010) and Cai et al. (2016) used category utility to generate basic level concepts and build ontologies from collaborative tags.

3. The proposed model

3.1. Chinese segmentation

Unlike English, there is no space between Chinese words in a Chinese document. The first step for Chinese text mining is Chinese words segmentation (Peng et al., 2004). In our research, we first use two famous tools for Chinese segmentation: Jieba and Standford CoreNLP. Jieba uses the dynamic programming to find out the most probable combination and hidden markov model to detect new words. Similar as Jieba, Stanford CoreNLP is designed by the Stanford Natural Language Processing Group for text analysis. As the finance domain is different with other Chinese domains, some specific words cannot be identified through these 2 methods. For example, the Chinese word with the meaning "Hong Kong and Shenzhen Stock Connect" is divided into meaningless characters by Jieba and Stanford CoreNLP. As a result, we use association rule mining to improve the segmentation result (Hu and Liu, 2004) and find out meaningful Chinese phrases. Let $C = \{c_1, \ldots, c_n\}$ be a set of Chinese items, and D be a set of transactions. Each transaction consists of a subset of items in C. The problem of mining association rules is to generate all association rules in D that have support and confidence greater than the minimum requirement. For example, although the Chinese characters for "Hong Kong and Shenzhen Stock Connect" is divided by CoreNLP, our method find out that these characters are associated with each other and always presented together in the same sentence with high confidence. As a result, the Chinese phrase with the meaning "Hong Kong and Shenzhen Stock Connect" is found by our method. The result is given in Figure 2.

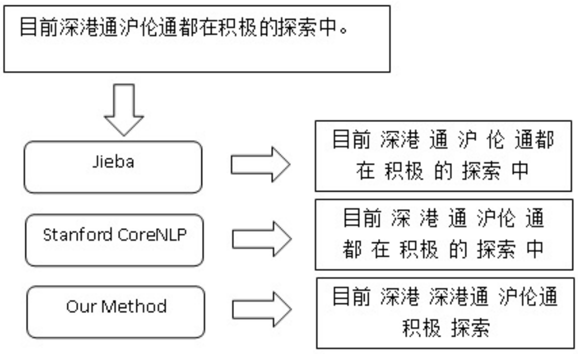
目前深港通沪伦通都在积极的探索中。
Jieba
目前 深港 通 沪 伦 通都 在 积极 的 探索 中
Stanford CoreNLP
目前 深 港 通 沪伦 通 都 在 积极 的 探索 中
Our Method
目前 深港 深港通 沪伦通 积极 探索

Figure 2. Chinese segmentation results comparison.

Although our method has better performance in terms of new word detection. It is possible that some redundant words are generated as well. To prune the redundancy, p-support is used. P-support of a word w is the number of the sentences including the word and these sentences don't contain any superset word phrase of w. If a word has a p-support lower than the minimum requirement which we set to 3 and it is a subset of other word phrases, it will be pruned.

3.2. *Category utility*

Basic level concepts constructed from basic level categories are considered as easier for human to understand. Basic level categories are the categories with the highest category utility which representing the predictive advantage gained by a person who know the given category structure over a person who does not possess this knowledge. Category utility is also a tradeoff between the intra-category similarity and inter-category dissimilarity of instances. The instances are clustered into different categories and normally described by a set of features or properties. Intra-category similarity is reflected by conditional probability of the form $p(f_i|c_k)$ where f_i is a feature and c_k is a category. The probability is high means a large number of category members sharing the same features. For information retrieval, the features can be represented by tags or keywords. For topic modeling, these features are the terms including in each documents.

Given a set C categories and a set F of features, the category utility is defined as follows:

$$cu(C,F) = \frac{1}{m}\sum_{k=1}^{m} p(c_k)[\sum_{i=1}^{n} p(f_i|c_k)^2 - \sum_{i=1}^{n} p(f_i)^2] \tag{1}$$

where $p(f_i|c_k)$ is the probability that a member of category c_k has the feature f_i, $p(c_k)$ is the probability that an instance belongs to category c_k, $p(f_i)$ is the probability that an instance has feature f_i, n is the total number of features, m is the total number of categories. In topic modeling method, a topic can be considered as a category which includes all the documents related to that topic. As a result, c_k represents topic k. f_i represents a term in the document. m is the total number of topics. n is the total number of terms or words. $p(f_i|c_k)$ is the probability that a document related to topic c_k has the term f_i. $p(f_i)$ is the probability that a document has the term f_i. $p(c_k)$ is the probability that a document is related to topic c_k.

3.3. Term weighting LDA

As discussed in the related work section, a lot of research has been conducted to enhance the topic modeling ability of LDA by including term weights into the model. Different term weighting schemes will generate different results. Term weighting schemes are widely used to measure the importance of words in documents. For information retrieval tasks, there are many unsupervised schemes such as tf and $tf * idf$. tf is term frequency which is the count of a term in a document. idf is inverse document frequency which is the log value of dividing the total number of documents by the number of documents that contain a specific term. $tf * idf$ is the combination of tf and idf. In addition, there are also entropy based term weights such as *bdc*. *bdc* is based on the entropy of terms in categories. Wang et al. (2015) declared that *bdc* outperforms the state-of-the-art schemes, e.g. $tf * idf$ in text categorization tasks. However *bdc* method needs to know the categories labels of each document. As there are a set of documents related to a topic generated by LDA method, in our method these documents are considered as in the same category labeled with the topic. However as the topics and the topic related documents are unknown in the beginning, we need to first get the initial topics and categories based on standard LDA and then the weights of words could be calculated. Topic-indiscriminating words will be considered as unimportant and get relatively low weights.

The framework of our method is shown in Figure 3. First of all, the Weibo posts are subject to the pre-processing mechanism. After the process of segmentation, stemming and words pruning, each post is transformed to a list of Chinese words. Then we could use our proposed term weighting LDA model to generate topics about a certain stock. The proposed model has 5 steps. In the first step, as shown in Figure 3, LDA is executed in different subsets and we can get the initial information about the topics. After that, based on the result of step 1, a category utility like term weighting scheme is applied to calculate the weights of words. Different term weight schemes can be used in this step. The *bdc* term weights are calculated as follows in our method:

$$bdc(t) = 1 + \frac{\sum_{i=1}^{k} \frac{p(t|c_i)}{\sum_{i=1}^{k} p(t|c_i)} \log \frac{p(t|c_i)}{\sum_{i=1}^{k} p(t|c_i)}}{\log(k)} \tag{2}$$

Where k is the number of topics defined. $bdc(t)$ is the weight of term t. $p(t|c_i)$ is the proportion of the term t in the topic c_i.

According to the definition of basic level concepts and category utility, the topic modeling result with higher category utility means the result is closer to the basic level concepts and easier be learnt and interpreted by humans. As a result, a category utility like formula is given as the metric to calculate the weights and topic discriminating ability of different words. Category utility like term weighting scheme is calculated as follows:

$$\sigma_i = \max_{k=1}^{m} [p(t_i|c_k)^2 - p(t_i)^2]^2 \tag{3}$$

where $p(t_i|c_k)$ is the probability that a member of topic c_k has the word t_i and $p(t_i)$ is the probability that a document has the word t_i. m is the number of topics. Similar as category utility, this formula also represents the predictive advantage of a person who know the given category structure ($p(t_i|c_k)^2$) over a person who does not possess this knowledge ($p(t_i)^2$). In addition, the intra-category similarity is represented by $p(t_i|c_k)^2$. $p(t_i|c_k)^2$ is high means most of the documents in the topic is similar and has the same word. $p(t_i|c_k)^2 - p(t_i)^2$ represents the inter-category dissimilarity and the difference is larger when the word is more topic specific and has high topic discriminating power.

In step 3, the number of words is diminished proportionally according to weights of words. Hence, the total discounted number of words in document *m* under topic *k* is calculated as follows:

$$n_m^k = \sum_{t=1}^{t=V} \sigma_t n_{mkt} \tag{4}$$

where σ_t denotes the weight of word *t*, which is ranging from 0 to 1. The weight of a word is calculated based on its topic discriminating capability. n_{mkt} is the number of word *t* belonging to topic *k* in document *m*. Similarly, the total discounted number of word *t* under topic *k* is calculated as follows:

$$n_k^t = \sum_{m=1}^{m=M} \sigma_t n_{mkt} \tag{5}$$

Step 4 is to execute the LDA again, using the discounted values calculated in step 3. Conditional probability of word *i* in document *m* under topic *k* is calculated as Equation 6 which is same as standard LDA. The difference is that the counting variables are replaced with the discounted values. The word *t* will have less probability to be assigned in topic *k* if the weight of word *t* is lower. Finally, in step 5, topics about a stock could be generated from the LDA result. Each topic will have a list of related words. Based on the ranking of words, the top 20 words of a topic are defined as expressions that describe the topic of the stock. These final topics and the related topic words are described as aspects about a stock in the model.

$$p(z_i = k|\vec{z}_{k,\neg i}, \vec{\omega}, \vec{\alpha}, \vec{\beta}) = \frac{n_{m,\neg i}^k + \alpha_k}{\sum_{k=1}^{k=K}(n_{m,\neg i}^k + \alpha_t)} \frac{n_{k,\neg i}^t + \beta_k}{\sum_{t=1}^{t=V}(n_{k,\neg i}^t + \beta_t)} \tag{6}$$

4. Experiments

In this section, we describe experiments conducted to evaluate the performances of our proposed term weighting LDA method. In the quantitative experiment section, we apply different term weighting schemes in LDA such as bdc, tf-bdc, tf-idf and standard LDA without term weighting schemes is discussed as well. In the qualitative experiment section, we test the performance of our method and standard LDA, 10 judges are asked to label the topics generated from different methods.

4.1. Data set and experiment setup

The datasets used in our experiments are the collections of Weibo posts in different periods. Weibo is one of the most famous social media websites in China which has similar functions as Twitter. As we want to learn the related topics about a certain stock, we collect the posts related to the stock "Shanghai Stock Exchange Composite Index" (SSECI) from Jan 1st 2015 to Mar 31st 2017. The posts are divided into 3 datasets based on the posted time which are 2015 dataset (6112 posts), 2016 dataset (7579 posts), 2017 dataset (2186 posts).

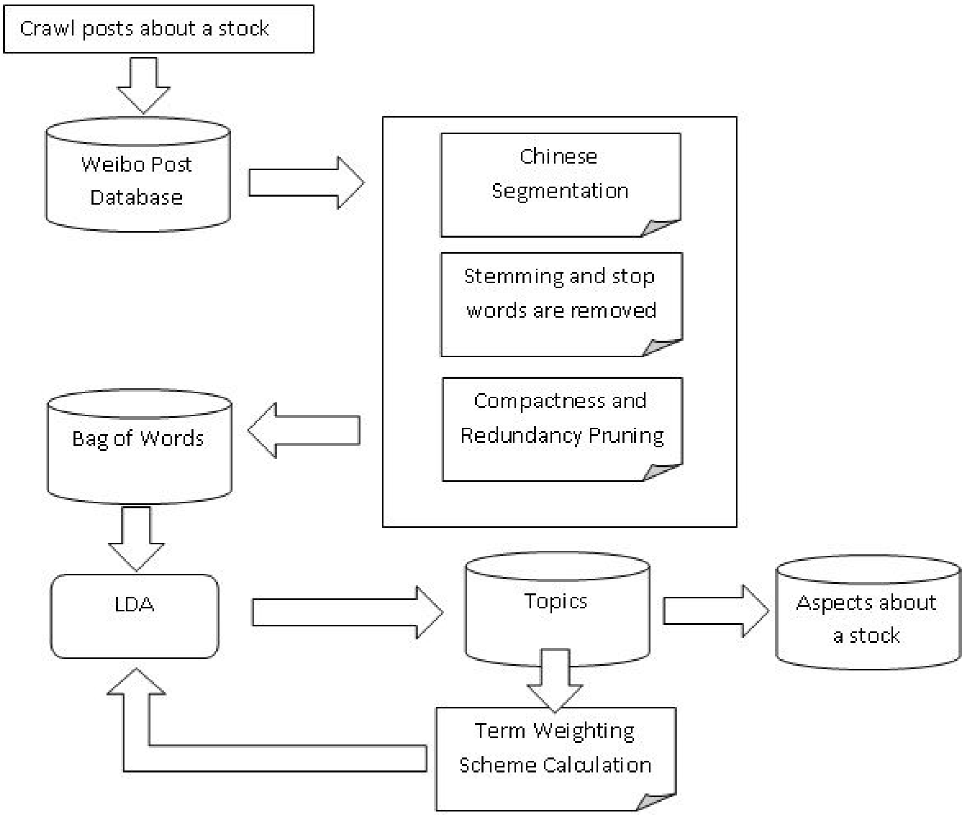
Crawl posts about a stock
Weibo Post Database
Chinese Segmentation
Stemming and stop words are removed
Compactness and Redundancy Pruning
Bag of Words
LDA
Topics
Aspects about a stock
Term Weighting Scheme Calculation

Figure 3. Framework of our term weighting LDA model.

All posts are subjected to the same pre-processing mechanism: (1) Words are converted into lower case; (2) Posts are segmented into sentences; (3) Each Chinese sentence are segmented and tagged for part-of-speech; (4) Punctuation is removed; (5) Words are stemmed and stop words and strings containing numbers and URLs are removed; (6) Compactness and Redundancy pruning is conducted to remove unnecessary words.

For all models, we take a single sample as posterior inference after 2500 iterations of Gibbs sampling. The aspect number K was set to 20. As small changes of α and β will not have big impact, we set $\alpha = 1$ and $\beta = 0.1$. For the term weighting LDA, the iterations of preceding LDA model are set to 1000 which is the same as the setting in (Yang et al., 2016).

4.2. Quantitative evaluation

As discussed in previous research, topic coherence is used to evaluate the performance of different LDA models. Higher topic coherence means the performance is better and the words included in the topic are correlated with each other. Through the experiment result, as shown in Figure 4, we can find out that the entropy based term weighting LDA such as bdc-LDA has better performance than standard LDA and tf-idf LDA in 2015 and 2016 datasets. In addition, our proposed method using category utility like term weighting scheme performs best in both years and improve the result of bdc-LDA.

To further demonstrate that the topics are more interpretable and distinguishable with each other, we use survey to ask 10 judges whether the topics could be interpretable. The 10 judges are students selected from different departments of a business school. We use Precision@n (or p@n) to check

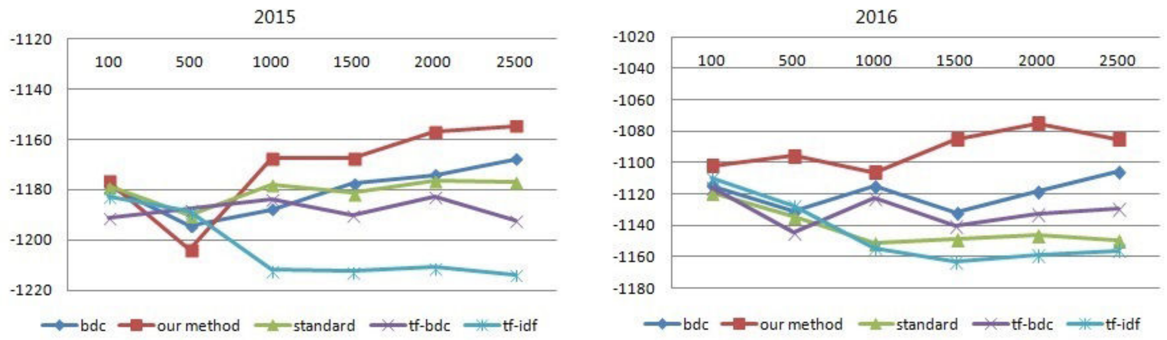

Figure 4. Topic coherence comparison of TWLDA with different term weighting schemes.

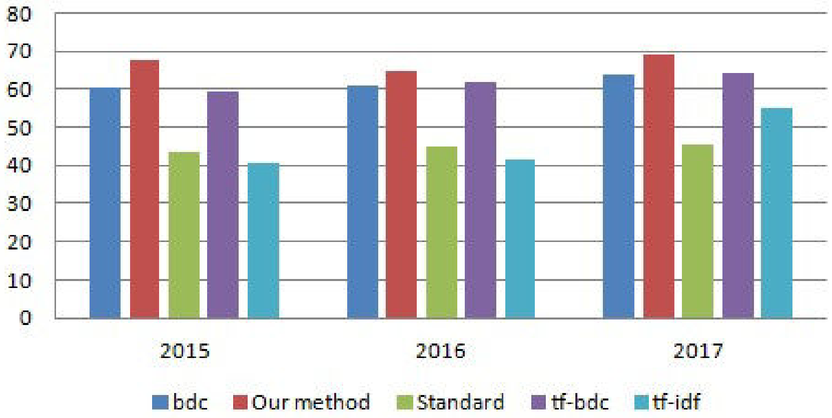

Figure 5. Precision@10 of topics generated from 3 datasets.

whether the words related to different topics are interpretable and could be used to describe the topic. It is commonly used as a metric for information retrieval (Mukherjee and Liu, 2012). Top words are selected for comparison, here the n is set to 10. First of all, Judges will be asked to find out correct topics in which half of the top 20 words are related to each other. Then they are required to label each word based on their understanding of the correct topics. If the word is consistent with the topic, it will be labeled as correct, otherwise incorrect. The experiment result is shown in Figure 5, the y-axis is the precision percentage. According to the precision, our method has the best performance in all three datasets (67.5% in 2015, 65% in 2016, 69% in 2017). LDA using bdc and tf-bdc term weight scheme also have higher precision value comparing to the standard LDA. However using the term weight scheme tf-idf , the performance is even worse than standard LDA for 2015 and 2016 datasets. The reason why tf-idf has worse result than standard LDA is that tf-idf ignores the category label of each document. In addition, the words have high term frequency could be some general words. If the term weighting LDA model gives more weights to these kind of words, more general words or topic indiscriminating words will be involved in each topic. As a result, the topics are mixed up and cannot be interpreted which makes the precision lower. For other term weighting LDA model, the categories or topics are generated in the first LDA calculation and the term weights are calculated based on the categorization result. The topic discriminating words are given high weights and their numbers are increased in next LDA calculation. As a result, more topic specific words are involved in each topic. These words make the topic easier be recognized by human and higher precision. In addition, our proposed method using category utility like term weight method has more correct words than other methods. The Precision@10 on average is more than 60%. The reason is that category utility like term weight calculation method considers both intra-category similarity and inter-category dissimilarity, more topic specific words are identified and given high weights.

4.3. Qualitative evaluation

Table 1 and Table 2 are the results of standard LDA and our term weighting LDA method using category utility like term weighting scheme for 2015 dataset. Top 5 words are selected for each topic and we translate the Chinese phrases to English words. As mentioned in last section, we ask 10 judges to label the topics as bad or good through survey. If more than half of the judges mark the topic as bad, we will consider that the topics are un-interpretable by human. These bad topic words are shown in red in the two tables. Our method has 4 un-interpretable topics while standard LDA has 11 un-interpretable topics. In addition, for standard LDA method there are many general words in different topics, for example "Shanghai Composite Index" is involved in more than 15 topics.

Table 1. Topics generated by standard LDA.

Topic	Top Words
1	Today, Shanghai Composite Index, Price shocks, market, nowadays
2	Market, Shanghai Composite Index, Funds, Stocks, Stock Market
3	Index, Shanghai Composite Index, Facebook, trade, Friday
4	Stock Market, China, Shanghai Composite Index, Time, Fall Below
5	Shanghai Composite Index, today, sh000001, fall, forecast
6	Shanghai Composite Index, tomorrow, possible, if, The Main Market
7	Stock, sh000001, Shanghai Composite Index, today, Tape Reading
8	Link, Shanghai Composite Index, bullish, bearish, microblogging
9	Shanghai Composite Index, drop in percentage, brokers, limit down, time
10	Shanghai Composite Index, bull, index, trend, start
11	Shanghai Composite Index, China stock market, 2015, today, sh000001
12	A-share, price-earnings ratio, stock index, stock market, Shanghai Stock Exchange
13	Blogs, Shanghai Composite Index, market close, 100 million yuan, volume
14	Index, Growth Enterprise Market, Shanghai Composite Index, funds, markets
15	Rebound, The main market, price shocks, adjustments, continue
16	increase in percentage, Growth Enterprise Market, Shanghai Composite Index, up, down
17	Bank, Market, Shanghai Composite Index, Growth Enterprise Market, Industry Sector
18	Shanghai Composite Index, the stock market, a-share, market, market price
19	Forecast, trend, today, popular, point of view
20	Moving average, minutes, rebound, trends, daily k-line

Table 2. Topics generated by our method.

Topic	Top Words
1	Today, Everybody, Time, Now, The year of the goat
2	Time, 3000, United States, Reasonable, Factor
3	Stock, Everyone, Chance, Reduce your holdings in the fund, Oil
4	1-minute K-line, Daily K-line, Down, MACD, Chase the down trend
5	Sh000001, stock, company, stock market, finance and economics
6	Support position, Shanghai Composite Index , rebound, positions, Horizontal consolidation
7	Volume, 10 Billion RMB Trading Volume, Shenzhen Component Index, Total volume after market close, before market close
8	Moving average, rebound, trend, upward trend, weekly K-line
9	Post, blog, drop in percentage, increase in percentage, decline
10	Breakthrough, upward, pressure, long position, downward
11	Industry sector, ChiNext, Banks Sector, Small and Medium Enterprise Board, The subject shares
12	Interest rate cut, reserve rate cut, revenue, Finance, China
13	Government, China Securities Regulatory Commission, Beijing, Circuit Breaker, Circuit Breaker Mechanism
14	Institutional investors, retail investors, A Shares, Investors, The national investors
15	Index, Shanghai and Shenzhen Composite Index, China Securities Index, long positions, Shenzhen Composite Index
16	Historic high, bull market, trend, leverage, high
17	Market, signals, trading, Japanese candlestick charting techniques, sub-new shares
18	Trillion, IPO, Financing, Market capitalization, Bubbles
19	Growth Enterprise Market, Industry sector, Main board, Stock, Index
20	Market Close, Close Price, Above, Price standing above, Quantitative

The result demonstrates that our method can reduce the impact of topic-indiscriminating words. The topics from our method are interpretable and inline with human understanding, for example, top words about a topic from 2015 datasets are Chinese phrases with the meaning of “Interest rate cut” “Reserve rate cut” “Revenue” “China” “Finance”. This topic describes a policy change of China in 2015 which has a great impact on “Shanghai Stock Exchange Composite Index”. In our previous research, Chen et al. (2016), we have demonstrated that the public opinion about the topic “interest rate and reserve rate cut” is correlated with the price movement of “Shanghai Stock Exchange Composite Index”. As a result, our term weighting LDA method could be used to find the topics about a certain stock and these topics are related to the price of the stock. In addition, comparing with other LDA methods, only our method could extract the topic “Circuit Breaker” which is a relevant topic of SSECI. Most of the posts discussing this topic are published at the end of 2015. Other methods cannot find this topic as the weights of relevant words are not high.

5. Conclusion and future work

To address the challenge of topic generation for sentiment analysis, in this paper, we propose a novel topic modeling method to generate topics about a stock based on the data in a Chinese social media website (Weibo). Motivated by basic level concept in cognitive psychology, we use a category utility like metric to calculate the term weights in LDA model. Our results demonstrate that our method is more effective than standard LDA and other term weighting LDA methods. The topic expressions produced through our method are seldom mixed together, more interpretable, and have less general words comparing to other methods. These topics could be used in topic-based sentiment analysis model for predicting stock price movement.

There are several potential extensions to our current research. First, the proposed method is based on Chinese. Thus how to derive the model to other languages and extract topics for other stock markets e.g Tokyo Stock Exchange is a problem. Second, the topic generation method could be incorporated with the research of domain specific ontology building, the topics could be used to build an ontology for financial market and explore the semantics from other user-generated multimedia resources in the big data era.

Conflict of Interest

All authors declare no conflicts of interest in this paper.

References

Abbasi A, Chen H (2008) CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication. *MIS Quart* 32: 811-837.

Andrzejewski D, Zhu X, Craven M (2009) Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *Proc Int Conf Mach Learn* 382: 25–32.

Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. *J Mach Learn Res* 3: 993–1022.

Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *Comput Sci* 2: 1–8.

Cai Y, Chen W, Leung H, et al. (2016) Context-aware ontologies generation with basic level concepts from collaborative tags. *Neurocomputing* 208: 25–38.

Chau M, Xu J (2007) Mining Communities and Their Relationships in Blogs: A Study of Hate Groups. *In J Human-Computer Studies* 65: 57–70.

Chen W, Cai Y, Lai K, et al. (2016) A topic-based sentiment analysis model to predict stock market price movement using Weibo mood. *Web Intelligence* 14: 287-300.

Chen W, Cai Y, Leung H, et al. (2010) Generating ontologies with basic level concepts from folksonomies. *Procedia Computer Sc* 1: 573–581.

Fan R, Zhao J, Chen Y, et al. (2014) Anger is more influential than joy: Sentiment correlation in Weibo. *PloS one* 9.

Gao Q, Abel F, Houben GJ, et al. (2012) A Comparative Study of Users? Microblogging Behavior on Sina Weibo and Twitter, In: Masthoff J., Mobasher B., Desmarais M.C., Nkambou R. (eds), *User Modeling, Adaptation, and Personalization*, Springer, Berlin, Heidelberg, 88–101.

Gilbert E, Karahalio E (2010) Widespread worry and the stock market. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, Washington, DC, 59–65.

Gluck M (1985) Information, uncertainty and the utility of categories. *Proceedings of the Seventh Annual Conference on Cognitive Science Society*, 283–287.

Guo H, Zhu H, Guo Z, et al. (2009) Product feature categorization with multilevel latent semantic association. *Proceedings of the 18th ACM conference on Information and knowledge management*, 1087–1096.

Gruhl D, Guha R, Kumar R, et al. (2005) The predictive power of online chatter. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* 41: 78–87.

Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Mach learn* 42: 177–196.

Hu M, Liu B (2004) Mining opinion features in customer reviews. *National Conference on Artifical Intelligence,* AAAI Press, 755–760.

Jo Y, Oh A (2011) Aspect and sentiment unification model for online review analysis. *Proceedings of the fourth ACM international conference on Web search and data mining*, 815-824.

Li X, Xie H, Chen L, et al. (2014) News impact on stock price return via sentiment analysis. *Know-Based Syst* 69: 14–23.

Liang H, Tsai F, Kwee A (2009) Detecting Novel Business Blogs. *Proceedings of the 7th International Conference on Information.* IEEE Press, 1–5.

Liu A, Gu B, Konana P, et al. (2006) Predicting stock price from financial message boards with a mixture of experts framework. *Intelligent data exploration & analysis laboratory*, 1–14.

Mimno D, Wallach H, Talley E, et al. (2011) Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Mishne G, Glance N (2006) Predicting Movie Sales from Blogger Sentiment. *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, 155–158.

Mukherjee A, Liu B (2012) Aspect extraction through semi-supervised modeling. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 339–348.

OLeary D (2011) Blog Mining-Review and Extensions: From Each According to His Opinion. *Decision Support Syst* 51: 821–830.

Peng F, Feng F, McCallum A (2004) Chinese segmentation and new word detection using conditional random fields. *Proceedings of the 20th international conference on Computational Linguistics*, 562.

Rosch E, Mervis C, Gray W, et al. (1976) Basic objects in natural categories. *Cogn Psychol* 8: 382–439.

Schumaker R, Chen H (2009) Textual analysis of stock market prediction using breaking financial news: the AZFin text system. *ACM T Informa Syst* 27: 1–19.

Wang T, Cai Y, Leung H, et al. (2015) Entropy-based term weighting schemes for text categorization in VSM. *Tools with Artificial Intelligence*, 325–332.

Wilson A, Chew P (2010) Term weighting schemes for latent dirichlet allocation. *The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, 465–473.

Yang F, Liu Y, Yu X, et al. (2012) Automatic detection of rumor on Sina Weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 1–7.

Yang K, Cai Y, Chen Z, et al. (2016) Exploring Topic Discriminating Power of Words in Latent Dirichlet Allocation. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2238–2247.

Zhang X, Fuehres H, Gloor P (2011) Predicting stock market indicator through twitter 'I hope it is not as bad as I fear'. *Procedia-Soc Behav Sci* 26: 55–62.

AIMS Press