Information Processing and Management 59 (2022) 102822

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Application of Informetrics on Financial Network Text Mining Based on Affective Computing

Anzhong Huang [a,c], Yuling Zhang [b], Jianping Peng [c,d,*], Hong Chen [c]

[a] School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang, 212100, China
[b] School of Economics and Management, Shangqiu Normal University, Shangqiu, 476000, China
[c] School of Management, Guangzhou Xinhua University, Dongguan, 523133, China
[d] School of Business, Sun Yat-Sen University, Guangzhou, 310003, China

ARTICLE INFO

Keywords:
informetrics
SparkR
Distributed computing
affective computing

ABSTRACT

With the rapid development of internet, text data is becoming richer, but most part of them is unstructured. So compared to statistics data, the text data is more difficult to be utilized. How to apply the informetrics on financial network text mining is a supplement to the traditional research methods of finance. The paper tries to forecast exchange rate volatility through informetrics on financial network text mining by means of affective computing. We find that if the amount of informetrics on network is used during predicting, only the peak and valley values of its volatility and are synchronous with the volatility of exchange rate. While the volatility of emotional intensity of words of informetrics on network in text data can accurately predict not only the drastic volatility of exchange rate, but also the moderate volatility.

Introduction

Recently, many currencies have seen a significant devaluation against the U.S. dollar, which has caused many investors and policymakers to worry. Therefore, how to accurately predict the future trend of the exchange rate has become the main task of current researchers. There are many traditional methods to predict the future trend of financial variables, such as probability method, signal method, Markov switching approach, network method, etc. But these methods are helpless facing online text data. So this paper tries to explore whether the exchange rate can be accurately predicting through informetrics on financial network text mining.

With the rapid development of informatization, websites are becoming more and more prosperous, and the financial industry has produced a large amount of text data, especially the unstructured data through microblogs, Twitter and so on. Some Researches show that the comments, emotional factors and other information contained in these unstructured text data have a great impact on the financial industry. Therefore, informetrics on financial network text mining has been paid more and more attention. The application of informetrics of text data in financial field mainly include:

(1) Forecasting the future trend of the stock market. For example, (Zhang, Fueshres and Gloor, 2011), (Nguyen, Shirai and Velcin, 2015) and (Krishna *et al.*,2020) found that informietrics on Twitter through affective computing could predict stock movements. (Ishijima, Kazumi and Maeda, 2015) and (Shahi *et al.*,2020) accurately predicted the movement of Nikkei index through informietrics by means of affective computing. (Runpeng, Wengming Lingyan, 2015), (Kaishen, Hao and Weining, 2014) and

[*] Corresponding author.
E-mail address: mnspjp@mail.sysu.edu.cn (J. Peng).

https://doi.org/10.1016/j.ipm.2021.102822
Received 11 August 2021; Received in revised form 2 November 2021; Accepted 9 November 2021
Available online 3 December 2021

(Cuiqing, Kun and Yong, 2015) studied the relationship between informetrics on network through affective computing of text data and Shanghai index from different angles.

(2) Forecasting the financial risks. informetrics on internet by means of affective computing is often used to predict financial risks too. For example, (Anzhong, 2018) and (Anzhong, Lening and Zheng,2020) predicted fraud risk in e-commerce through affective computing of text data. (Haworan, Xun and Yuefeng, 2013), (Lei, Yanpen and Qunfeng, 2014) had studied the application of affective computing of text data in financial risk prediction. Similar studies also include (Eiman,Lulu and Alan,2015), (Jin, xing, Ruiqiao,2015) and(Anzhong and Fei,2020) act..

(3) Application in other field. Informetrics on financial network text mining has been used in many field, even including human behavior researches, such as (Lie and Yairi, 2015) and (Anas *et al.*, 2017) and (Anzhong, Jie and Huimei,2020). However, there are few Scholars who paid attention to its application in forecasting the future movement of exchange rate. Although (Peramunetilleke and Wong, 2002) predicted the exchange rate volatility through informetrics on network text mining, the text data they used only came from money market. The quality of text data coming from money market is relatively high, but the quantity is relatively insufficient. So it is more difficult for us to use online text data than to use text data only coming from money market.

Therefore, the research of the paper is an expansion of the application field of the informetrics on network text mining. Simultaneously, the research methods of the paper are different from the existing literature, because the paper tries to use the low quality online text data from online rather than the high quality text data from money market.

Through introducing the distributed computing technology, this paper constructs a SparkR platform, which not only can make use of the convenience and ease of operation of R language, but also make use of the SparkR platform for distributed computing to solve the big data challenges in text financial data analysis. Based on the SparkR platform, this paper tries to explore the application of affective computing of the online financial text data in predicting the future movement of exchange rate.

2. Informetrics on Internet by Means of Affective Computing

This paper tries to examine whether the informetrics on network text mining through affective computing can be used to forecast exchange rate volatility. In order to accomplish this task, the following two problems need to be solved: (1) how to select key text data through feature selection so as to distinguish the emotional attitude through emotion classification algorithm; (2) after classifying text data into emotional attitudes, it is needed to quantify the emotional attitudes and examine their impact on exchange rate volatility.

In order to solve these two problems, this paper designs three models to discuss them from different angles. All the models take the exchange rate of RMB as the research objects and the online text data are get on each website.

In order to solve the problem proposed in the paper by using text information, the paper deals with text information by affective computing. The frame of affective computing is shown in Fig. 1.

Model 1: The Influences of the amount of text data on exchange rate volatility

Definition 1: The amount of text data refers to the total amount of relevant news generated by an enterprise over a period of time. When calculating the amount of text data of the currency, the relevant web pages searched by the subject name of the currency shall be attributed to the news with weight α given to the news. While the news searched by the subject name of other currency but containing the subject name of the currency are weighted β. Sum of the news for each currency shall be obtained to get the amount of text data for each currency. The specific formula is as follows:

$$I_i = \alpha P_i + \beta N_i \tag{1}$$

Where0 $< \beta < \alpha < 1$, and $P_i$is the amount of the news of thei^{th}foreign exchange, and$N_i$is the amount of the news which contain the

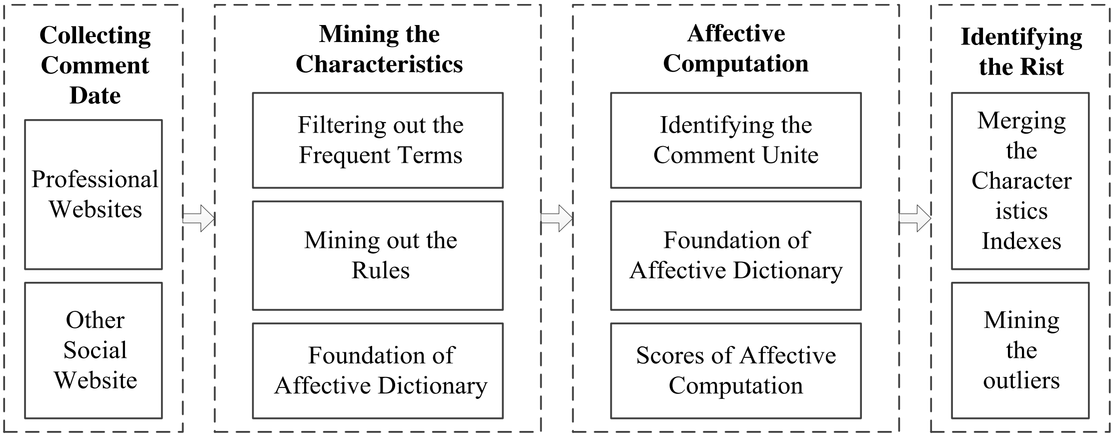
Collecting Comment Date
Professional Websites
Other Social Website
Mining the Characteristics
Filtering out the Frequent Terms
Mining out the Rules
Foundation of Affective Dictionary
Affective Computation
Identifying the Comment Unite
Foundation of Affective Dictionary
Scores of Affective Computation
Identifying the Rist
Merging the Character istics Indexes
Mining the outliers

Fig. 1. Framework of the Affective System

name of the currency but does not belong to the currency.

Model 2: The Influences of the amount of news emotional information

Definition 2: The amount of news emotional information refers to the difference between the number of positive news and the number of negative news over a period of time.

The influence of news on exchange rate volatility is positive and negative. So it is necessary to classify the news to distinguish the different influences of news exchange rate volatility. Therefore, the model 2 introduces news emotion to differentiate news, and separately counts the number of positive news and negative news in the same period of time, and then examines the influence of news emotion information on exchange rate volatility.

In model 2, the paper still adopts the method of attribution and empowerment used in model 1. In order to screen out the news that affects the exchange rate volatility and classify this news emotionally, this paper uses K-means clustering to obtain the required feature words and training sets. For improving the clustering accuracy and fully extract classified polar words, the paper selects enterprise news texts with weekly stock price increases and decreases of more than 100 basis points as the clustering training set, and after removing some commonly used financial words and stop words contained in the training set, the news is grouped into positive and negative categories. The paper selects respectively positive and negative categories containing a large number of texts as a classification training set, and deletes polar words appearing frequently in both positive and negative news as a feature word candidate set after calculating the polar word frequency. In order to further extract, the final feature words, this paper uses the following methods to realize, the realization process is divided into the following three steps:

Step 1: assuming that the initial feature set of the training sample is ϕ, and selecting a feature word from the candidate word set to add to the feature set according to word frequency from big to small, and training the training sample to obtain a preliminary model.

Step 2: comparing the classified news with the previous clustering results, manually labeling news samples with different classifications and clusters, and then adjusting the feature words accordingly. The adjustment process is to extract key words according to the feature set and remove the key words with the highest word frequency from the feature set if the manual annotation result is consistent with the clustering result. If the result of manual annotation is consistent with the result of classifier, the article will be deleted from the training set, and the corresponding word frequency will be adjusted to complete the whole adjustment process.

Step 3: repeating step 1 and step 2 until the accuracy is stable and then stops training to obtain the final model. Finally, the trained classifier is used to classify the news, determine the emotion of the news, and remove some news that cannot be classified due to mixed information or insufficient emotional information as news that has no obvious influence on the stock price. Through the above steps, we can obtain the emotional information of news for each enterprise. The specific formula is:

$$I_i = \alpha(L_i - J_i) + \beta(M_i - K_i) \tag{2}$$

Where L_i is the amount of positive emotional news belongs to the currency, and J_i is the amount of negative emotional news, and M_i is the amount of positive emotional news in news that includes the name of the currency but does not belong to the currency e, and K_i is the amount of negative emotional news in news that includes the name of the currency but does not belong to the currency. The flow chart of news emotional information is shown in Fig. 2.

Model 2 uses clustering algorithm to select features and emotion classification is carried out through a classification algorithm. In this paper, we use biased training samples to improve the accuracy of clustering, and the results of clustering include two kinds. One kind is the news that needs special information. This kind of news has a great influence on the foreign exchange market. We can obtain the main characteristics that we need through word segmentation, that is, to solve the problem of feature selection. The other is news that contains insufficient information and has little or no clear impact on the s foreign exchange market. This kind of news will be deleted. Through sample biased selection and clustering algorithm, we can make full use of a large number of news to accurately and quickly extract key information, so as to adapt to the actual needs of data explosion and rapid response.

Model 3: The Influence of the intensity of emotional word in text data

Definition3: The intensity of emotional words in news refers to the difference between the number of positive emotional words and negative emotional words in news generated by an enterprise for a period of time. Model 2 obtains the emotion of news through classification algorithm, and can also use the number of emotional words to express the emotion of news. In order meet the sufficiency

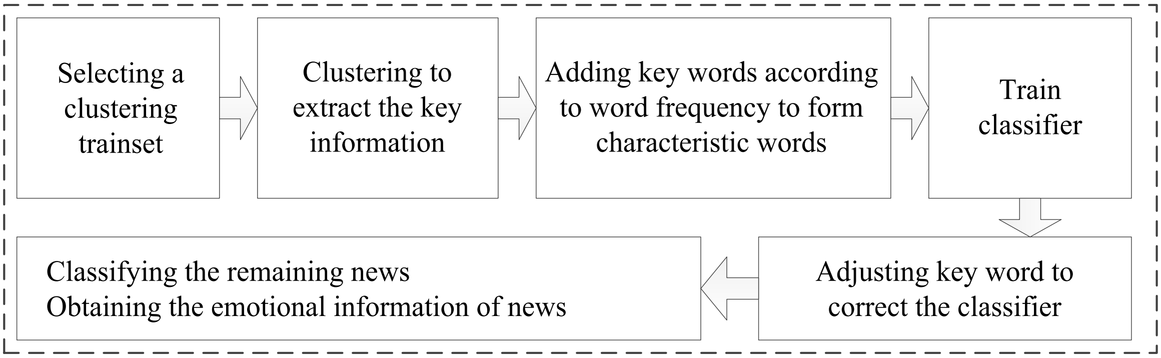
Selecting a clustering trainset
Clustering to extract the key information
Adding key words according to word frequency to form characteristic words
Train classifier
Adjusting key word to correct the classifier
Classifying the remaining news
Obtaining the emotional information of news

Fig. 2. Flow chart of news emotional information

of this research, model 3 introduces the intensity of emotional words in news. This article still adopts the news attribution and empowerment method of model one, introduces Reuters financial dictionary and manually modifies it to select the emotional words as the emotional words dictionary. Separately calculate the number of positive and negative emotional words contained in each piece of news, and calculate the intensity of emotional words in each enterprise. The formula is as follows:

$$I_i = \alpha(W_i - X_i) + \beta(Y_i - Z_i) \tag{3}$$

Among them, $W_i$is the amount of positive emotional words in the news, and X_i is the amount of negative emotional words in the news, and Y_i is the amount of positive emotional words in the news that includes the name of the foreign exchange but does not belong to the foreign exchange, and $Z_i$is the amount of negative emotional words in the news that includes the name of the foreign exchange but does not belong to the foreign exchange. Model 3 selects the characteristics by means of an emotional dictionary, and distinguishes emotional attitudes by calculating word frequency. This distinction is relatively simple, and emotional results can also be obtained. Because there are many news texts and the frequency matrix of the feature word formed after filtering text is a high-dimensional sparse matrix, so the corresponding clustering and classification algorithms need to be implemented on a distributed platform in order to adapt to the timeliness and accuracy requirements of big data financial analysis. The following discusses the implementation of these algorithms on SparkR platform.

3. SparkR Platform and the Algorithm of Network Mining

SparkR is an R development package exploited by AMPLab, which combines Spark with R language and not only has data structure and application architecture of Spark, but also makes use of the convenience of R language to complete data analysis. So it is very suitable for completing big data analysis tasks. The R package is the R lightweight front end of ApacheSpark, which provides the RRDD of API and the users can run job interactively through the R shell on the cluster. SparkR provides PipelineRDD optimization, which can submit Execute to RVM to calculate and then return the results to Execute in a unified way without frequent gradual communication between the two, thus greatly saving the time of transmission and deserialization. Finally, SparkR also supports common closure functions, and variables referenced in user-defined functions are automatically sent to other machines in the cluster. SparkR combines R language with SparkContext through JN interface to start a javaSparkContext, and then connect with Executor of Worker through JavaSparkContext. The operation principle of SparkR is shown in Fig. 3.

Because it is convenient to use this package to realize the corresponding algorithms such as crawler, clustering and classification, which are needed for big data analysis. According to the description in the previous section, K-means algorithm is selected in clustering algorithm, and random forest algorithm is selected in classification algorithm. The specific implementations of algorithms are described below.

3.1. Preprocessing Network Information

After obtaining the news corpus, the news needs to be preprocessed, whose core of text preprocessing is to segment (filter) Chinese documents and form a frequency matrix of characteristic words. The word segmentation tool adopts Rwordseg package provided by R. The text preprocessing task can be completed in three stages. In the first stage, the text captured by the crawler is stored on each node in a distributed way, and all the text is segmented in a distributed way by using a segmentation tool to form a frequency matrix of characteristic words. In the second stage, the custom ordered characteristic words of financial dictionary are initialized to RDD and broadcast so that through character matching, the serial numbers (*key*) of the matched characteristic words are recorded and the value of *Value* is 1.

In the third stage, the operator reduceBekey is used to process each document, adding the values of *Value* with the same *Key* values, and adding the news label to get the frequency matrix (*i key_Value*) of the corresponding document, wherein *i* indicates that this is the i^{th} document, *Key* is the serial number of the key word in the financial dictionary, and *Value* is the number of times the characteristic word appears in the document. The relevant steps are shown in Table 1.

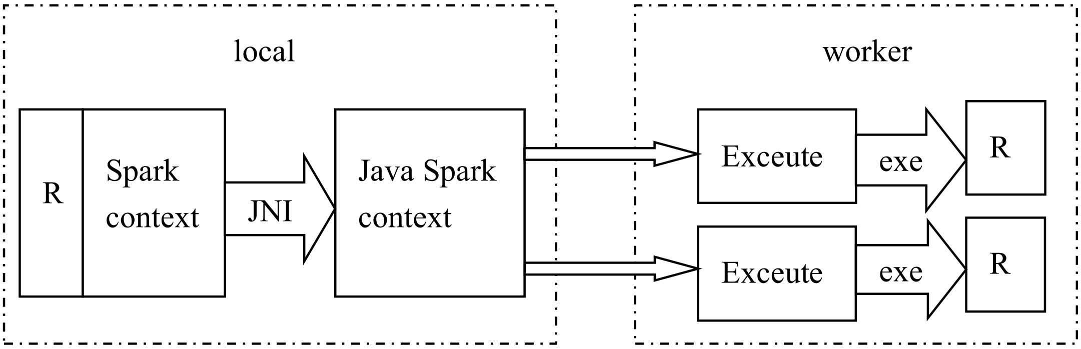

Fig. 3. Operating principle of SparkR

3.2. Realization of Distribution of CART and Random Forest Algorithms on SparkR

CART and random forest algorithms are mainly used for classification. Among them, CART algorithm is a widely used learning method of decision tree. CART algorithm assumes that the decision tree is a binary tree in which the characteristic value of the internal node is only "yes" or "no". The algorithm is a learning method which will output conditional probability distribution of predicted values under the given input sample conditions and it is consisted of two parts: spanning tree and pruning, among which spanning tree is the process of recursively constructing binary decision tree. CART method uses Gini index to select splitting characteristics.

Definition 4: In the classification problem, it is assumed that there are$K$Classes and the probability of sample point belonging to theK^{th}class isP_K, so the probability distribution of Gini index is defined as:

$$\mathrm{Gini}(p) = \sum_{k=1}^{K} P_k(1-P_k) = 1 - \sum_{k=1}^{K} P_K^2 \tag{4}$$

As for the two-class classification problem, if the probability of sample point belonging to the first class is$p$, then the Gini index of probability distribution is defined as follow:

$$\mathrm{Gini}(p) = 2p(1-p) \tag{5}$$

As for the given sample set$D$, the Gini index is:

$$Gini = 1 - \sum_{k=1}^{K} \left(\frac{C_k}{|D|}\right)^2 \tag{6}$$

In which, $C_k$is thek^{th}class sample subset in the sample set,and$K$is the number of class. If the sample set D could be divided into two subsetD_1andD_2 according to whether the characteristic$A$of Dcould take the possible value$a$:

$$D_1 = \{(x,y) \in D | A(x) = a\} D_1 + D_2 = D \tag{7}$$

Then under the characteristic value (A) of D, the Gini index of sample set D is defined as:

$$\mathrm{Gin}(D,A) = \frac{|D_1|}{|D|}\mathrm{Gin}(D_1) + \frac{|D_2|}{|D|}\mathrm{Gin}(D_2) \tag{8}$$

Random forest algorithm is an integrated-classification algorithm, which was first proposed by Leoreiman and Adele Cutler. Random forest is a classifier that integrates multiple pruning-free CART, in which the type of output depends on the mode of the type of output by the individual tree. The training set used by each tree is extracted from the total training set through extracting and returning again. This means that some samples in the overall training set may appear in the training set several times, or may never appear in the training set of a tree. When training the nodes of each tree, the used features are randomly extracted from all the features according to a certain proportion without being put back. According to Leo Breiman's suggestion, if the total number of features is$M$, then this ratio can be defined as $\mathrm{sqrt(M)}, 1/2\mathrm{sqrt(M)}$.

3.2.1. Generating algorithm of SparkR distribution of CART

If we assuming that the values of all the characteristic variables are discrete, then format of the sample data which are read into by row will be DataFrame. The characteristic variables and class labels are respectively extracted and converted into feartureRDD and labelRDD through operator toRDD, in which each element in featureRDD and labelRDD is of type$\mathrm{c(key, value)}$, where key is the line number of the elements in the original DataFrame and value is the corresponding value. Then CART can be generated in three steps. In the first step, a combination, which is consisted of the characteristic value of the each column in which the row lays and the name of the characteristic variable, will be generated to be regarded as *Key* and 1 can be used as the key value. Then the number of sample of each variable will be calculated through operator reduceByKey. In the second step, in order to count the number of samples of every variable belonging to specific class through reduceByKey, the new *key* will be combined the *key* generated in step 1with the labelRDD, and 1

Table 1
Algorithm 1: Algorithm of Preprocessing Text Data

Step	
Input:	News text to be filtered
Output	Matrix of word frequency of text
Step1	Initializing news text to RDD and applying to filter function to filter the words in news text.
Step2	Defining function filter $=$ function $(x)\{$if $(x = =$ dictRDD $[i])$ return $- c(i,$ $1)\}$,
lapply (fileRDD, fun $=$ filter) And the result is recorded as featureRDD whose type belongs to$\mathrm{c(key, value)}$ where keyis iand valueis 1.	
Step3	Calling the function reduceByKey(featureRDD, $+$,2L, and the result is the frequency of the every non-zero key word in every text.

Among the three steps, step2 completes the task of filtering word and step3 get the matrix of Characteristic word through calculating by reduceByKey.

will be used as the key value. In the third step, the corresponding Gini value will be calculated and the sample will be segmented by split function and record the line number of the original sample to ensure sample consistency. The specific algorithm steps are shown in table 2.2.

The probability (p) of Gini index calculated in step 10 is the quotient of numRDD divided by denRDD, and the ratio of values of separation of variables in sample is value/min denRDD. The eleventh step selects the variable whose Gini value id min and split the sample set into two subset according to value of variable, which it integrity is assured by the value of key.

The termination condition of the whole algorithm is that the number of samples in the node is less than the predetermined threshold, or the Gini index of the sample set is less than the predetermined threshold, or there are no more features. After the CART being generated, we can utilize distribution of the generated CART to classify under the condition of initializing the data samples to RDD and then classify them in parallel according to the selected characteristics.

3.2.2. Realizing the Distribution of Random Forest Algorithm

This article first completes the task of initializing the sample to RDD, and at the same time, all the feature variable names are initialized to a new RDD, which is called fearturenameRDD. Then the article will determine the number (t) of used CART and generate tCART in different node distributions. Sampling is done through operator sampleRDD, and the sampling method, namely return sampling and non-return sampling, is assigned by *with Re placement* parameter in operator sampleRDD to complete the return sampling of sample and non-return sampling of features. After sampling, the split function is used to change the corresponding column in feartureRDD in feartureRDD into a new RDD for training to complete the random forest algorithm.

3.3. Realization of Clustering Algorithm on SparkR Distribution

Means is an unsupervised learning algorithm, which is a typical object function clustering method based on prototype. It takes a certain distance from the data point to the prototype as the optimization objective function, and uses the method of finding the extremum of the function to obtain the adjustment rule of the iteration operation. The algorithm needs to first determine the parameter k, i.e. how many classes the entire sample data needs to be divided into. Then the algorithm divides the ndata sets into these *k* classes, which makes the similarity difference between classes larger and the similarity within classes lower. In this paper, the similarity between samples is measured by Euclidean distance. The next section will discuss how to implement K-means algorithm on SparkR on distribution. means clustering algorithm is solved in three steps. The first step is to define the distance function as the Euclidean distance between the sample points and find out which cluster center the sample is closest to. In the second step, the samples were divided into *k* samples according to the nearest gathering point. in the third step, the clustering center was recalculated. In the second stage of calculation, the number of samples contained in each class needs to be calculated at the same time. The specific implementation steps are shown in Table 3.

In step 1, the operator takeSample is used to complete the sampling task without putting back. Step 3 and step 4 respectively calculate the distance from the sample point to the aggregation point and the nearest aggregation point. Step 8 calculates a new center point. The algorithm ends until tempDISTis less than the given threshold convergeDist.

Table 3
Realizing the distribution of K-means algorithm

Step	Main tasks and algorithms
Input:	biased sample set, the clustering number(k),threshold distance (convergeDist)
Output	kclass centers
Step1	Reading in the data set, initializing to RDD and recording as pointRDD.
Step2	Calling function takeSample(PointRDD, with Replacement = F, and recording the result as kpoint and beginning to circulate and iterate.
Step3	Calling function lapply(pointRDD, fun = function(p){dis.fun(point s, kpoint s)}),
Defining, dis.fun = function(P, C){apply(C, 1, function(x){colsuns($(t(p) - x)^2$)))} and recording as matRDD.	
Step 4	Calling functionlapply(matRDD, fun = function(x){max.col(− x)}), and recording the result as cp.
Step 5	Calling function lapply(pointRDD, fun = function(x){cbind(1, x)}), and filling 1 into every sample and still recording as pointRDD.
Step 6	calling function
Step 7	Calling function cpreduceByKey(splitRDD, + , 2L).
Step 8	Defining function to calculate the new center point lapply(splitRDD, function(x){sum = x[[2]][, − 1], count = x[[2]][− 1]sum /count}), and recording the result as newpoint.
Step 9	calling function
Step 10	when tempDist > convergeDist, iterating from step 4 to step 10 untiltempDist ≤ convergeDist

Experiments and Analysis of Results

4.1. The Impact of Informetrics on Internet on Volatility of Exchange Rate

According to the model described in section 1, this paper carried out three groups of experiments. The middle prices of intermediate exchange rate between RMB and USD from June 1, 2017 to June 29, 2018 are selected as the research sample. We Use the news of exchange rate between RMB and USD headlines from June 1, 2017 to June 29, 2018 from Sina finance and economics, Hutchison finance and Tencent finance as the corpus, and saving us dollars as currency subject names as feature words dictionary 1, the weights $\alpha$and β are set as 0.75 and 0.25 respectively, and the exchange rate volatility is defined as$\frac{(e_t - e_{t-1})}{e_{t-1}}$.

4.2. Experiment 1

The experiment will test the relationship between the amount of the text data and exchange rate volatility, and the experimental steps are as follows:

Step 1: filtering the news text set and storing the distribution on SparkR platform.

Step 2: find out the feature words, which are included in dictionary 1, in each news text, through algorithm 1, and obtaining the subject of RMB and dollars contained in each news text.

Step 3: calculating the amount of text data about exchange rate and getting the final amount of text data through weighting.

In experiment 1, the total amount of financial text data in the same period is calculated by the same weighting method because no news emotion was introduced. Considering that the impact of the same amount of text data on the exchange rate should be similar, that is, when the exchange rate continues to rise or fall by a similar margin, the corresponding amount of text data should be roughly the same instead of increasing or decreasing incrementally, this paper uses the amount of news from 31 May 2017 as the base point to draw the corresponding volatility trend of the amount of text data for a year. The results are shown in Fig. 4.

It can be seen from Fig. 4 that in most cases the volatility of the amount of text data and of the exchange rate are not always consistent and sometime opposite. While the peak values of the volatility of the amount of text data and these of the exchange rate volatility are synchronous, so are the valley values.

There are two obvious characteristics in the relationship between the amount of text data and the exchange rate volatility. First, when the exchange rate fluctuates obviously, the amount text data also fluctuates obviously. The reason behind this characteristic may be that model 1 does not filter the news text. When the exchange rate fluctuates gently, some news text that has little influence on the exchange rate still be generated and result in the amount of the news fluctuates greatly than that of exchange rate. Second, during some time periods, the volatility of text data is much higher than that of exchange rate. The reason behind this characteristic may be that although the amount of text data fluctuates greatly, the news contained has both positive and negative directions, so its impacts on the volatility of exchange rate will be insignificant.

4.3. Experiment 2

Through clustering, the second experiments will search the news which is described by key information and this news could materially affect the volatility exchange rate. Then, the experiment will distinguish between the positive and negative emotions of all news by classification. The experimental steps are as follows:

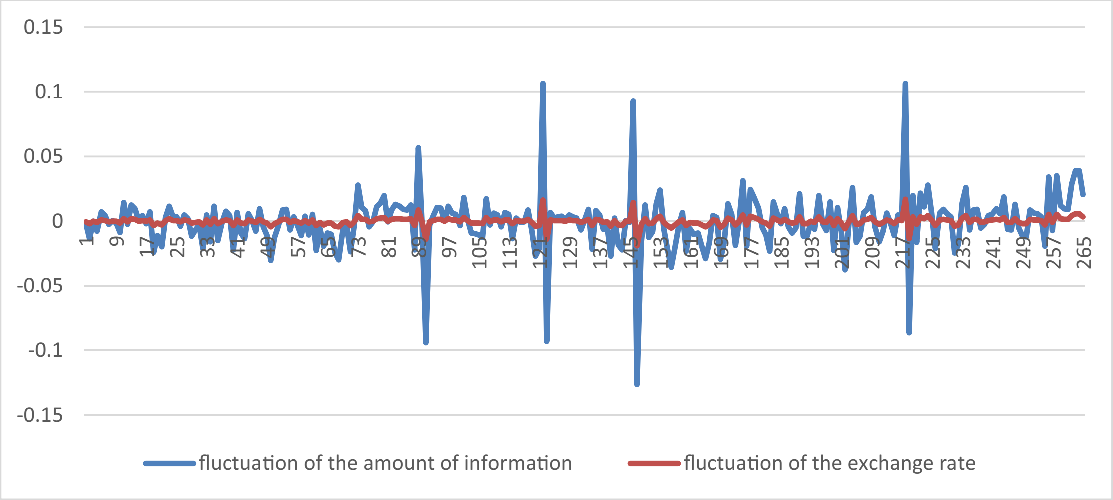

Fig. 4. The volatility of the amount of text data and the exchange rate

Step 1: segmenting the news text set, storing the distribution on SparkR platform and filtering the clustering training set.

Step 2: extracting a classification training set and a feature word candidate set through clustering.

Step 3: sequentially adding polar words in the feature word candidate set to the feature words according to the word frequency, and training the classification model until the classification accuracy does not change obviously.

Step 4: adjusting the classification model through manual labeling.

Step 5: classifying the remaining news and calculating the amount of emotional information of the.

According to the definition of volatility of exchange rate in the experiment 1, the time series charts of the amount of emotional information and volatility in the past year are drawn respectively. The results are shown in figure5.

Compared with Fig. 3, Fig. 5 shows that the volatility of the amount of emotional information of the text data is closer to the volatility of exchange rate, but there is also some daily volatility of exchange rate that are not so close to the amount of emotional information of the text data.

4.4. *Experiment 3*

The third experiments will explore the influence of the intensity of emotional words in text data on the volatility of exchange rate. The paper introduces *the Reuters Financial Dictionary* and its emotional words are selected through manual revision, which are stored as feature word dictionary 2. The specific experimental steps are as follows:

Step 1: filtering the news text set and storing the distribution on SparkR platform.

Step 2: calculating the number of feature words which are contained in the feature word dictionary 1 in each news text, and calculating the difference between the number of positive and negative emotional words in each news text through algorithm 1.

Step 3: finding out all the feature words which are contained in the feature word dictionary 1and the main subject of RMB or dollar included in each news text through algorithm 1.

Step 4: calculating the intensity of the emotional words about the exchange rate, and weighting them to get the final intensity of the emotional words. Operating as experiment 2, the results are shown in Fig. 6.

The volatility of emotional intensity of words in text data can accurately predict not only the drastic volatility of exchange rate, but also the moderate volatility. While are still some volatility will be falsely predicted. After examining the samples, we found that investors' expectations about the RMB/USD exchange rate fluctuations sometimes appear surprisingly consistent, such as after major decisions by two monetary authorities, after business frictions between two countries, etc. Therefore, the emotional calculation value based on text information is highly consistent with the real exchange rate fluctuations.

Conclusion

Through informetrics on financial network mining by means of affective computing, which is realized through the SparkR platform, in which the distributed computing technology is introduced, the paper tests whether text data can be used to forecast the volatility of exchange rate between RMB and U.S. dollar. It finds that:

(1) In most cases the volatility of the amount of informetrics on network are not always consistent and sometime opposite. While the peak values of the volatility of the amount of information of the exchange rate volatility is synchronous, so are the valley values.

(2) The volatility of the amount of emotional informetrics on network is more closer to the volatility of exchange rate, but there are also some daily volatility of exchange rate that are not so close to the amount of emotional information of the text data.

These may be resulted from rebounding after over-falling or selling after profit, etc., which will result in the volatility of amount of emotional information of text data and exchange rate will not completely match. Therefore, this article believes that apart from these

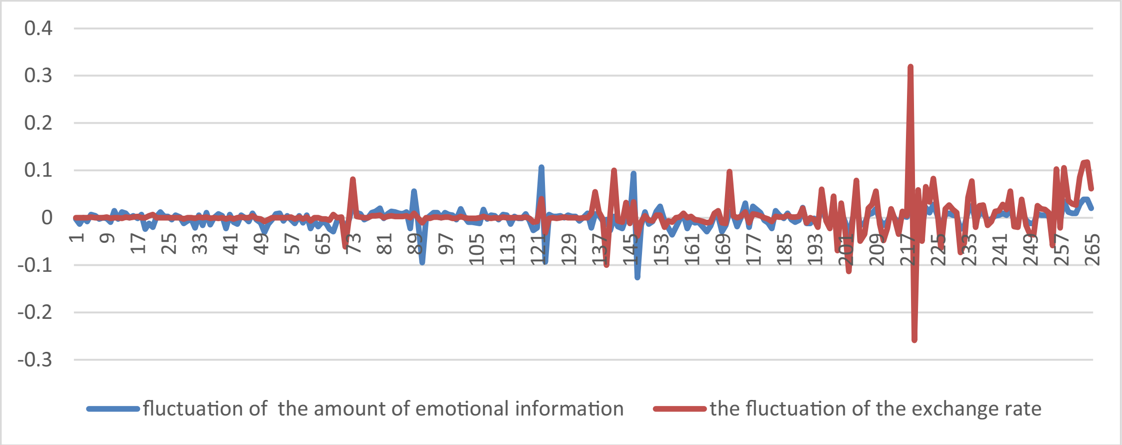

Fig. 5. The volatility of the amount of emotional information and the exchange rate

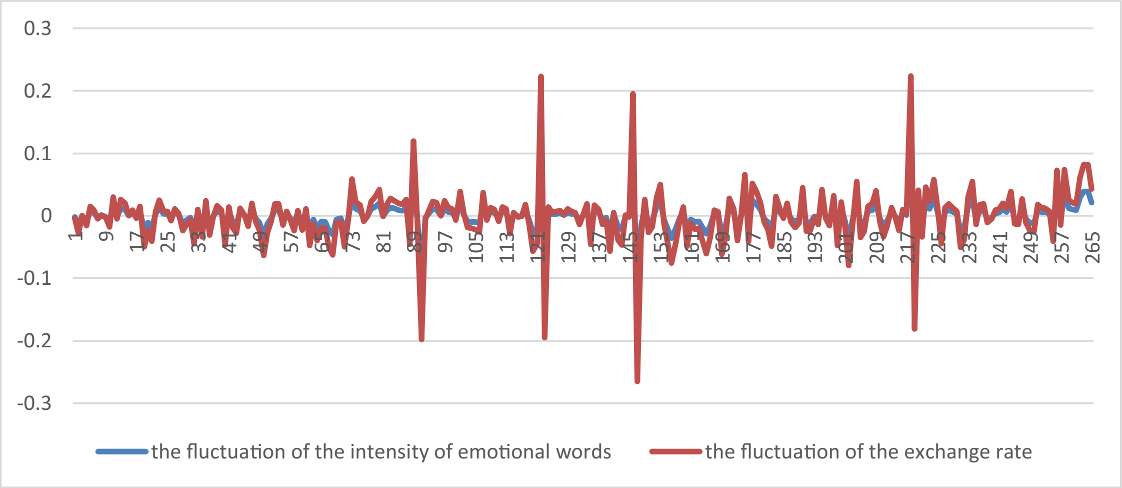

Fig. 6. The volatility of the intensity of emotional words and the exchange rate

daily volatility, the volatility of the amount of emotional information of the text data is basically related to the volatility of exchange rate.

(3) The volatility of emotional intensity of words of informetrics on network in text data can accurately predict not only the drastic volatility of exchange rate, but also the moderate volatility. While are still some volatility will be falsely predicted.

It may be limited by the emotion dictionary and the large amount of text data, which make it difficult for the emotion dictionary to accurately express news emotion. This leads to less text data that contains specific emotional words when exchange rate volatility is small and cannot accurately measure the volatility range of exchange rate.

So we can conclude that the amount of informetrics, the amount of emotional informetrics on network by means of affective computing can be used in predicting the volatility of the exchange rate, but the predicted results of the emotional intensity of informetrics on network are the best.This shows that the text information mining method and algorithm proposed in this paper are feasible. However, we can't predict the application fields and application prospects of this method and algorithm, because their success still has some particularity. Therefore, the following problems still exist in this study.

Although the emotional intensity of information is the best in the experiment, we can't conclude that the other two indicators are dispensable. In this paper, the exchange rate is taken as the experimental object, and the amount of information is extremely huge. The excellent performance of emotional intensity of information is probably based on the huge amount of information. Therefore, we can't judge whether the emotional intensity of information can perform well if the amount of information decreases.

It is not clear whether the method and algorithm proposed in this paper can predict the low frequency data. The experimental object of this paper is exchange rate, which is a high-frequency data. The difference in application value between high-frequency data and low-frequency data makes us unable to think that this method and algorithm can be applied to low-frequency data prediction, such as interest rate.

We can't judge whether the method and algorithm are optimal.

Data Availability

The datasets used to support the results of this study are available from the corresponding author upon request.

Authors' Contributions

Anzhong Huang and Jianping Peng contributed equally to this work.

Author introductions

Author: Anzhong Huang (1971.11…), male, a professor of College of Economics and Management, Jiangsu University of Science and Technology (Address: 666#, Changhui Road, Dantu District, Zhenjiang City, Jiangsu Province, China, Postcode 212100) and School of Management, Guangzhou Xinhua University (Address: Guangzhou Xinhua University, Dongguan City, Postcode 523133). Mainly studying fields include financial theories and engineering as well as Fintech. Email: az311@126.com.

Yuling Zhang(1983.07…),male, a Lecturer, School of Economics and Management, Shangqiu Normal University, mainly studying social capital and enterprise organization. Address: 55# Pingyuan Road, Shangqiu City, Henan Province, Postcode 476000. Email: ZHANGXYZYL@163.com.

Corresponding Author: Jianping Peng, male, a professor and doctoral supervisor of school of Management, Sun Yat-sen University, and the president of School of Management, Xinhua University, Guangzhou. Address: School of management, Guangzhou Xinhua University, Dongguan City, Postcode: 523133. Email: mnspjp@mail.sysu.edu.cn.

Hong Chen (1979.09…), male, a lecturer of School of management, Guangzhou Xinhua University, mainly studying financial technology and entrepreneurship. Address: School of management, Guangzhou Xinhua University, Dongguan City, Postcode: 523133. Email: 395778065@qq.com.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgement

This research is sponsored by the National Nature Science Foundation of China(Grand Number:71572196) and social sciences project of the Ministry of Education of PRC., named Researches on the Mechanism of Microcredit for Poverty Alleviation and Innovation of Jiangsu Practice Model(20YJA790028) as well as the Anhui Social Science Fund Project(AHSKY2020D53)

References

Anas, S., Leo, G., Raymond, B., & Hui, W. (2017). Affective State Detection via Facial Expression Analysis within a Human-computer Interaction Context. *Journal of Ambient Intelligence and Humanized Computing, 4*, 1–17.

Cuiqing, J., Kui, L., & Yong, D. (2015). Stock Behavior Prediction Based on Social Media. *Chinese Journal of Management Science, 23*(1), 17–24.

Eiman, K., & Luluah, A. (2015). Alan C... Emotions in Context: Examining Pervasive Affective Sensing Systems, Applications, and Analyses. *Pers Ubiquit Comput, 19*, 1197–1212.

Ishijima, H., & Kazumi, T. (2015). Maeda A... Sentiment Analysis for the Japanese Stock Market. *Global Business and Economic Review, 17*(3), 237–255.

Krishna Kumar, Md (2020). Tanwir Uddin Haider. Enhanced Prediction of Intra-day Stock Market Using Metaheuristic Optimization on RNN–LSTM Network. *New Generation Computing, 9*(11), 1–42.

Shahi Tej, Bahadur (2020). Shrestha Ashish;Neupane Arjun; Guo William. Stock Price Forecasting with Deep Learning: A Comparative Study. *Mathematics, 8*(9), 1441–1449.

Anzhong, H., & Fei, W. (2020). Two-stage adaptive integration of multi-source heterogeneous data based on an improved random subspace and prediction of default risk of microcredit. *Neural Computing and Applications* (on line:https://doi.org/10.1007/s00521-020-05489-z).

Anzhong, H., Jie, C., & Huimei, Z. (2020). Construction of patient service system based on QFD in internet of things. *The Journal of Supercomputing* (published on line: https://doi.org/10.1007/s11227-020-03359-y.

Jin, H., Xing, R., & Ruiqiao, J. (2015). Analysis of Public Opinion in Financial Field Based on SVM and Dependency Syntax. *Computer Engineering and Application, 51*, 230–235.

Kaishen, L., & Hao, C. (2014). Weining Q... Microblog Sentiment and China's Stock Market. *Journal of Systems Science and Mathematical Sciences, 34*(5), 565–575.

Lei, H., Yanpeng, W., & Qunfeng, Z. (2014). Research and Improvement of Automatic Keyword Extraction Method. *Computer Science, 41*(6), 204–207.

Liew, C. F., & Yairi, T. (2015). Facial Expression Recognition and Analysis: a Comparison Study of Feature Descriptors. *IPSJ Trans Comput Vis Appl, 7*, 104–120.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment Analysis on Social Media for Stock Movement Prediction. *Expert Systems with Applications, 42*(24), 9603–9611.

Peramunetilleke, D., & Wong, R. K. (2002). Currency Exchange Rate Forecasting from News Headlines [J]. *Australian Computer Science Communications, 24*(2), 131–139.

RunPeng, H., Wenming, Z., & Lingyan, B. (2015). Stock Market Forecasting Based on Microblog Emotional Information. *Journal of Industrial Engineering and Engineering Management, 29*(1), 47–52.