Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 6, June 2019, pp. 1518-1527
DOI: 10.23940/ijpe.19.06.p3.15181527

Feature Dimension Reduction Optimization Algorithm for Massive Micro-Blog Data based on Hadoop

Haodong Zhu*, Wenqi Li, and Hongchan Li

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

Abstract

For the micro-blog sentiment analysis problem in big data environments, the "dimension disaster" caused by the continuous increase in text information data brings great challenges to the emotional analysis of micro-blogs. To solve this problem, this paper proposes a fusion of the advantages of three feature dimensionality reduction algorithms, based on the traditional document frequency (DF), mutual information (MI), and chi-square test (CHI). Firstly, the document frequency factor is added to the mutual information (MI) algorithm to solve the problem of low-frequency word defects. Then, the standard score factor is added to the chi-square test (CHI) algorithm to solve the negative correlation problem. Finally, the average value is calculated and the advantages of the three algorithms are fused. An improved Proposed DF-MI-CHI fusion algorithm is proposed. The simulation results show that after using this algorithm to process the micro-blog data, the accuracy of sentiment analysis is improved and maintained at 95%. The recall rate is more than 90%, and the F value is maintained between 92% and 94%. In the % interval, it is higher than other improved algorithms and tends to be stable, which indicates that the algorithm can effectively improve the accuracy and efficiency of micro-blog emotional sentiment analysis when dealing with massive micro-blog text data.

Keywords: feature dimension reduction; micro-blogging emotion; feature selection; hadoop; HDFS

(Submitted on March 10, 2019; Revised on April 5, 2019; Accepted on June 7, 2019)

1. Introduction

With the rapid development of the Internet and mobile devices, the data volume of various Internet media platforms is increasing explosively [1]. As a part of all kinds of information in the Internet, the micro-blog takes text data as the main organizational form. Although micro-blog text data has the characteristics of being unstructured and difficult to understand and having large interference information, it has strong timeliness and a wide audience. It can provide various effective information for the government and enterprises [2]. Therefore, how to quickly obtain effective information from massive micro-blog data is still a research hotspot for many scholars. At present, the growing text data of micro-blogs has expanded the scale of data sets, and the number of features used to represent text information has also increased, resulting in an increase in data dimension and the "dimension disaster" [3].

The "dimension disaster" will lead to an increasingly sparse distribution of data and the decline of data organization effect, which brings great challenges to the computer analysis and processing of massive data information [4]. In order to solve this problem, researchers have used feature selection to reduce the dimension of the data features. The representative algorithms are document frequency (DF), information gain (IG), mutual information (MI), and chi-square test (CHI). Ye et al. [5] put forward an improved text frequency algorithm (TF-IDF), which integrates new network words into the vector space model and redistributes weights to improve the performance of classifiers. This method solves the problem of endless new words emerging in the network and improves the accuracy of feature analysis. However, it still faces the problems of filtering out rare words in text information and reducing the accuracy of classifiers. Mao [6] proposed a feature selection algorithm based on maximum conditional joint mutual information (MCJMI). When selecting features, not only is the conditional mutual information (MI) is considered, but also the integrated joint mutual information (MI) is used to enhance

* Corresponding author.
E-mail address: zhuhaodong80@163.com

the accuracy of feature selection. However, this method tends to use low-frequency words excessively and does not consider the influence of word frequency on text analysis. Qiu et al. [7] proposed a variance-based chi-square test (Var-CHI) feature selection algorithm. Three adjusting parameters were added to improve the uneven distribution of chi-square test (CHI) feature words, but there are still shortcomings of low-frequency words.

This paper mainly focuses on the low-frequency word defects in the above algorithms. The rare words are easily filtered out, and the word frequency is not considered. The proposed DF-MI-CHI fusion optimization algorithm is combined with the advantages of the three feature selections. Based on the method, the relevant calculation factors are added to make up for the shortcomings of the traditional algorithm, and the improved algorithm proposed by the predecessors is further optimized. In this paper, the text frequency factor is added to the mutual information (MI) algorithm, and the feature frequency and feature distribution are taken into account. It is mainly aimed at solving the problems that rare words are easily filtered out and the influence of word frequency on text dimension reduction is not considered. Then, the standard fraction factor is multiplied before the chi-square test (CHI) algorithm, and the concept of feature word distribution is introduced. The problem of low-frequency words is solved, and the negative correlation of the chi-square test (CHI) algorithm itself is solved to some extent. Finally, the average of the whole formula is calculated. This algorithm can not only improve the performance of the classifier as a whole, but also improve the classification accuracy and recall rate of text data information.

2. Relevant Basic Works

2.1. Relevant Feature Selection Methods

After pre-processing the text to be classified, a feature space of feature words will be obtained. A large number of meaningless words are deleted by de-noising in the pre-processing stage [8], which reduces the spatial dimension. However, for text classification, this feature space is still quite large, resulting in more and more sparse text features and reducing the efficiency and accuracy of classification [9]. Therefore, it is necessary to reduce the dimension of text feature vectors by feature selection. Feature selection involves selecting the best feature subset from the original feature. At present, the following features selection methods are widely used.

2.1.1. Document Frequency (*DF*)

Document frequency (DF) is a simple and easy to understand feature selection algorithm. It is realized by calculating the number of occurrences of a feature word in a document [10]. Its implementation process is as follows: firstly, set the maximum and minimum values of the document frequency, that is, the threshold value of each feature word. Then, the document frequency of each feature word in the corpus is calculated. Finally, retain the feature words that meet the threshold and delete those that exceed it [11]. The calculation equation of document frequency (DF) is shown in Equation (1).

$$DF(t_i, C_j) = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

In Equation (1), $n_{i,j}$ represents the number of occurrences of feature word t_i in category C_j and $\sum_k n_{k,j}$ represents the sum of the occurrence times of all characteristic words in category C_j. A larger document frequency (DF) indicates that the feature word appears too frequently in the text. This means the feature word is not discriminative, and it has little effect on the sentiment analysis. It should be deleted [12]. Although the method is simple and efficient, it is easy to filter out rare words that are of great significance to the text.

2.1.2. Mutual Information (*MI*)

Mutual information (MI) is an algorithm used to describe the correlation between two variables. The concept of information entropy is introduced in the calculation of mutual information (MI), which is used to measure the degree of data instability or burst in data sets [13]. Its calculation equation is shown in Equation (2).

$$Entropy(s) = Entropy(P_1, P_2, \cdots, P_m) = -\sum_{i=1}^{m} P_i \log_2 P_i \tag{2}$$

In Equation (2), s is used to represent the training set and P_i is used to represent the probability of target attributes appearing in all samples. The smaller the value of information entropy, the purer the distribution of target attributes. If the entropy is zero, the target attributes of all samples take the same value [14].

For mutual information (MI), it is assumed that the feature term is represented by t_i and the category is represented by C. Its equation is shown in Equation (3).

$$MI(t_i, C) = \log \frac{P(t_i, C)}{P(t_i)P(C)} \tag{3}$$

In Equation (3), $P(t_i, C)$ represents the probability of the occurrence of the feature term t_i in Category C, $P(t_i)$ represents the probability of the occurrence of the feature term t_i in the whole corpus, and $P(C)$ represents the probability of the occurrence of Category C. The greater the value of mutual information (MI), the higher the co-occurrence of feature t and category C. Therefore, when the value of mutual information (MI) reaches its maximum, the feature item is the best feature of the category [15]. The disadvantages of mutual information (MI) are that it ignores the word frequency of feature words and pays too much attention to the document frequency of feature items, which will result in the loss of some important feature items.

2.1.3. Chi-Square Test (*CHI*)

This method is also called the square fitting test X^2 statistical method. It is used to measure the degree of correlation between feature words t and category C. Its calculation equation is shown in Equation (4).

$$X^2(t_i, C) = \frac{N(AE - BD)^2}{(A+D)(B+E)(A+B)(D+E)} \tag{4}$$

In Equation (4), N denotes the total number of documents in the corpus. A denotes the frequency of the feature item t_i in the total document, and the document belongs to the category C. B denotes the frequency of feature t_i in the total document, but the document does not belong to category C. D indicates that the feature item t_i does not appear in the document, but the text category belongs to the frequency of the category C. E indicates the frequency of documents that contain neither feature t_i nor category C in the total document [16]. In the calculation, the default feature item and the category are independent of each other. After calculation, if the value of the formula is small or zero, the deviation degree is small. The default original hypothesis is that the feature items and the categories are independent. If the value is large, that is, the degree of error is large, the original hypothesis is not established, and the feature item is related to the category [17]. The final value of the chi-square test (CHI) usually uses its maximum value or average value, and the equation for calculating its maximum value is shown in Equation (5).

$$X^2_{MAX}(t_i) = \max_{j=1}^{M} \{X^2(t_i, C)\} \tag{5}$$

For multi-class partitioning, experiments show that using the maximum chi-square value as the screening condition can achieve better results. Therefore, this paper chooses the maximum of chi-square test (CHI) as the final CHI value of feature t_i [18].

2.2. Text Information Acquisition in Chinese Micro-Blog

This paper mainly focuses on the Chinese micro-blog, which is widely used as the background of data research. The main content is the text information released by users through the micro-blog account and the evaluation information of related micro-blogs. Data acquisition of micro-blog text is the first step of research. There are two commonly used data acquisition methods. One is the web crawler-based web crawling analysis method, while the other is the open API interface method provided by micro-blog platforms. The acquired data is stored in Hadoop-based HDFS for further analysis and use.

Web crawler is a kind of image naming method. It vividly describes the way to get web page data as the process of crawler searching the target along the path of the big web. This network refers to the Internet. The path is the complex web page links in the network. The specific process is divided into four steps: DNS domain name resolution, Web page crawling, web page parsing, and establishment of URL link library. The Hadoop distributed crawler tool flow chart is shown in Figure 1.

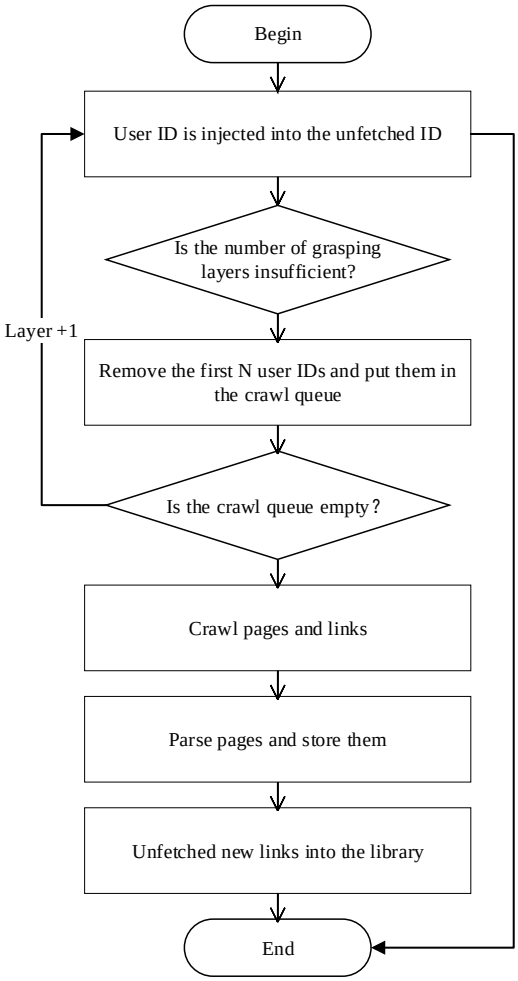

Figure 1. Hadoop distributed crawler tool flow chart

This paper uses web crawler to obtain micro-blog data for research.

2.3. Micro-Blog Text Pre-Processing

Users usually have some meaningless vocabulary when they publish or comment on micro-blogs, which brings interference to the further processing of text data information. This requires the pre-processing of stored data information. The pre-processing process includes word segmentation, part-of-speech tagging, and deactivation of text. The text pre-processing flow chart is shown in Figure 2.

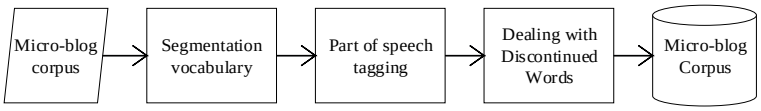

Figure 2. Text pre-processing flow chart

After obtaining text data, the text is segmented first. Chinese text data is different from English text. In English text, words can be segmented by obvious spaces or symbols, but there is no obvious division between words in Chinese text. Usually, the dictionary-based matching segmentation method and word frequency statistics-based segmentation method are

used to segment words. In this paper, we use the Python language version of the "jieba" component, which has high frequency, stable performance, and high accuracy, to segment the micro-blog text. After word segmentation, part-of-speech tagging should be made for each word after word segmentation. For example, "I" is a noun, "rope skipping" is a verb, and "beautiful" is an adjective. Rule-based part-of-speech tagging, statistics-based part-of-speech tagging, and their combination are commonly used part-of-speech tagging methods at present. This paper uses the combination of the two methods to tag parts of speech. First, rule-based part-of-speech tagging is used to tag common and simple words, and then statistical part-of-speech tagging is used to tag complex and emerging words. The efficiency also improves the comprehensiveness of the corpus. Stop word processing is also called denoising processing. It mainly deletes the words that do not have emotional meaning in the text, such as "de", "le", "ye", and so on, which are common in Chinese text information. After deleting them, the accuracy of text classification can be improved effectively, and the dimension of vector space can be reduced to a certain extent.

2.4. Hadoop

Hadoop is a software platform for analyzing and processing big data. It is the frame of open source software implemented by Appach in Java language. It realizes the distributed computing of massive data in a cluster composed of a large number of computers. The core design of its framework is HDFS and MapReduce. Among them, HDFS provides storage for massive data, while MapReduce provides computation for massive data.

HDFS has become the standard for big data disk storage and is used for online storage of large log files. After years of development, the architecture and functions of HDFS have been basically solidified. Important features such as HA, heterogeneous storage, and short circuit access of local data have been realized. As HDFS becomes more stable, the community becomes less active, and HDFS usage scenarios become more mature and fixed; file format encapsulation in the upper layer will increase. In the future, new storage formats will appear to adapt to more application scenarios, such as groups of storage to serve machine learning applications. In the future, HDFS will continue to expand support for emerging storage media and server architectures.

MapReduce is a programming model for processing large sets of semi-structured data. A programming model is a way of dealing with and structuring a particular problem. For example, in a relational database, use a collection language to execute queries, such as SQL. Tell the language what result you want and submit it to the system to figure out how to generate the calculation. You can also use more traditional languages (C++, Java) step by step to solve the problem. MapReduce and Hadoop are independent and work well together. More information about HDFS and MapReduce is introduced as follows.

2.4.1. HDFS

HDFS (Hadoop Distribution File System) is an important distributed file storage structure in the Hadoop system architecture. It is very suitable for storing massive amounts of data such as TB and PB. The HDFS distributed cluster is composed of three parts: Client, NameNode, and DataNode. A cluster usually needs a NameNode and several DataNodes. NameNode is mainly responsible for the namespace of the file system and the metadata of the file system and the client accesses the files and manages the mapping of data blocks to DataNode points. DataNode is responsible for storing and managing data blocks. Each data block is composed of two files, real data and metadata. According to the different storage locations of the data blocks, when registering with the NameNode, corresponding tags and block lengths will be generated according to the block ID of the block copy of DataNode in each block report to form an updated view of data blocks in the cluster. The main advantages of HDFS distributed file system are high fault tolerance, large storage capacity, and low cost. Therefore, this paper uses an HDFS distributed file system to store a large number of incremental data on micro-blog.

2.4.2. MapReduce

MapReduce divides the calculation process into two stages: Map stage and Reduce stage. The Map stage corresponds to the mapper processing function, which can filter and transform the original data. The Reduce stage corresponds to the reducer function, which mainly aggregates the processing results of mapper function and outputs the final processing results. The specific processing of MapReduce is as follows: firstly, the input data to be processed is divided into smaller Split sub-blocks and sent to Mapper. The Mapper calls the custom Map method to process the Split sub-block and converts it into an intermediate result. Store the results on the local disk. Then, call the Reduce function to aggregate the intermediate results. Finally, store the output in HDFS. The processing diagram of MapReduce is shown in Figure 3.

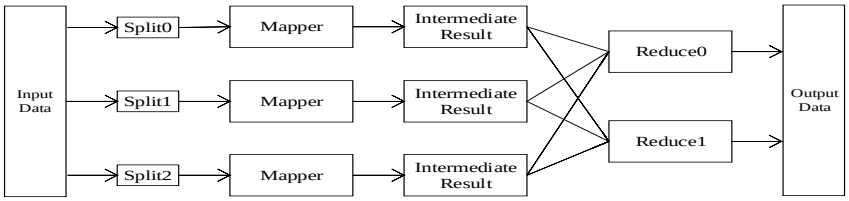

Figure 3. The processing diagram of MapReduce

3. Optimized Feature Dimension Reduction Algorithm

Most of the previous feature dimension reduction optimization algorithms are only considered from one side, focusing on solving the influence of a certain factor on the feature dimensionality reduction algorithm. The improved text frequency algorithm (TF-IDF) focuses on the influence of new network words on feature dimensionality reduction but ignores the problem that rare words are filtered out. The feature selection algorithm based on maximum conditional joint mutual information (MCJMI) focuses on solving the effect of low-frequency words on feature dimensionality reduction, but it does not consider the influence of word frequency on text analysis. The variance-based chi-square test (Var-CHI) feature selection algorithm focuses on solving the uneven distribution of feature words, but it ignores the problem of low-frequency word defects. Therefore, in this paper, based on the above problems, a feature selection algorithm based on document frequency (DF), mutual information (MI), and chi-square test (CHI) is proposed.

The algorithm first calculates the value of the document frequency and adds the document frequency factor before the mutual information (MI) algorithm. Then, it calculates the chi-square test (CHI) value and multiplies the standard score factor. Finally, it adds the values of the first two steps to calculate the average value. This algorithm can make up the shortcomings of the document frequency (DF), mutual information (MI), and chi-square test (CHI) for low-frequency word analysis. It improves the efficiency and accuracy of classification.

Based on the above analysis, the new algorithm equation presented in this paper is shown in Equation (6).

$$W_{Ct} = \frac{DF(t_i, C_i) \times MI(t_i, C_i) + \alpha \times X^2(t_i, C_i)}{2} \tag{6}$$

In the above Equation (6), $DF(t_i, C_i)$ represents the document frequency factor, $MI(t_i, C)$ represents the mutual information value, $X^2(t_i, C)$ represents the value of the chi-square test, and α is a standard fraction used to solve the negative correlation problem in the chi-square test (CHI), whose calculation method is shown in Equation (7).

$$\alpha = \frac{D_C - \bar{D}}{\sigma} \tag{7}$$

In Equation (7), D_C denotes the number of documents in Category *C* where the feature word *t* appears, $\bar{D}$ denotes the average number of documents in which the feature word *t* appears, and σ denotes the standard deviation of the frequency of documents in which the feature word *t* appears.

4. Simulation Experiments

4.1. Acquisition and Processing of Experimental Data Sources

Through the web crawler, 5,000 real-time microblog texts and comment information were obtained. Using random sampling, 70% of the micro-blog data is used as training samples and 30% is used as testing samples. All crawled data are stored in the Hadoop-based HDFS distributed file storage structure.

The crawled data is segmented, and the part of speech is marked. We use the "jieba" component Python language version to segment the acquired micro-blog text [19] and remove special symbols and meaningless words such as "@", "de", "le", and "etc". Then, tag the part of speech, and convert the processed micro-blog text information into a corresponding word frequency matrix. Stored in the Hadoop distributed file system (HDFS), it is ready for the experimentation of the

feature dimension reduction algorithm.

4.2. Evaluation Index of Experiment

In the analysis of micro-blog emotions, the precision, recall, and *F* value are usually used as criteria to identify micro-blog emotions. We usually use a table to describe the relevant variables.

According to the Table 1, the precision, recall, and *F* value equations can be defined as follows.

Precision (*P*):

$$P = \frac{TP}{TP + FP} \tag{8}$$

Recall (*R*):

$$R = \frac{TP}{TP + FN} \tag{9}$$

F Value:

$$F = \frac{2P \times R}{P + R} \tag{10}$$

Table 1. The explanation of relevant variables about Precision and Recall

	Relevant	Unrelated
Retrieved	True Positives (TP)	False Positives (FP)
Not Retrieved	False Negatives (FN)	True Negatives (TN)

In Equations (8) to (10), *TP* is used to represent the number of correctly classified micro-blog texts in the sentiment analysis of micro-blogs. *FP* represents the number of retrieved but incorrectly classified micro-blog texts. *FN* is the number of micro-blog texts not retrieved but belonging to this category. *TN* is the number of documents not retrieved and not belonging to this category. Therefore, *TP+FP* is the total number of micro-blogs in the system that should be judged as the sentiment of the category. *TP+FN* is the number of micro-blog texts that should be judged as the category [20]. Finally, the results of analysis and calculation are plotted by Matlab, which is convenient for observation and analysis.

4.3. Experimental Results and Analysis

In this paper, 1500 micro-blog text data are tested using the improved text frequency algorithm (TF-IDF) [5], which integrates new network words into the vector space model; the maximum conditional joint mutual information (MCJMI) [6]; the variance-based chi-square test algorithm (Var-CHI) [7]; and the proposed DF-MI-CHI fusion improved dimension reduction algorithm, which is improved by the previous methods. In this paper, we mainly analyze and test the positive sentiment of micro-blog text. The accuracy, recall rate, and F value of positive emotional tendency analysis are shown in Figures 4, 5, and 6 respectively.

From Figure 4, we can see that the accuracy of the four algorithms is relatively high when we analyze the emotions of small-scale micro-blogs. As the number of micro-blog texts increases, the improved TF-IDF algorithm aims to integrate new words into the network, ignoring the rare words that appear less frequently. When analyzing more complex micro-blog texts, there are rare words that will have important significance and appear less frequently. Because the improved dimension reduction algorithm synthesizes the advantages of the three algorithms and effectively solves the shortcomings of filtering out rare words, the overall accuracy of the proposed DF-MI-CHI dimension reduction algorithm is higher than the other three methods. With an increase in the number of micro-blog texts, it shows a relatively stable trend.

The recall rate is for the test results in the test set. For the emotional analysis of micro-blogs, the recall rate of positive micro-blog emotions represents the ratio between the number of correct positive emotional tendency micro-blog texts and

the number of real positive emotional tendency micro-blog texts in the test set of the micro-blog. After analysis, the recall rate of sentiment analysis of micro-blogs based on four dimensionality reduction algorithms is shown in Figure 5. The maximum conditional joint mutual information (MCJMI) algorithm has a higher recall rate when analyzing a certain number of micro-blog texts because it tends to use low-frequency words. However, because the maximum conditional joint mutual information (MCJMI) algorithm does not take into account the influence of word frequency on the sentiment analysis of micro-blogs, more texts are accompanied by impossibility. Eliminating redundancy leads to a lower recall rate. The improved proposed DF-MI-CHI comprehensive dimension reduction method adds the document frequency factor to the mutual information, which fills the existing problems of the improved maximum conditional joint mutual information (MCJMI) algorithm. The recall rate is obviously higher than the other three and is relatively stable. It effectively improves the accuracy of micro-blog sentiment analysis.

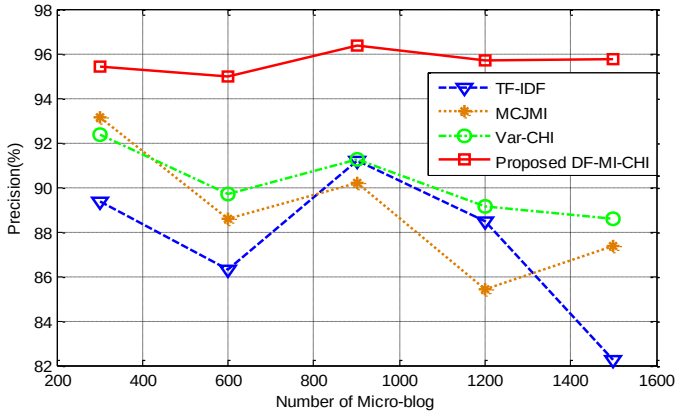

Figure 4. Comparison of precision results of four dimensional reduction methods for emotional analysis of micro-blogs

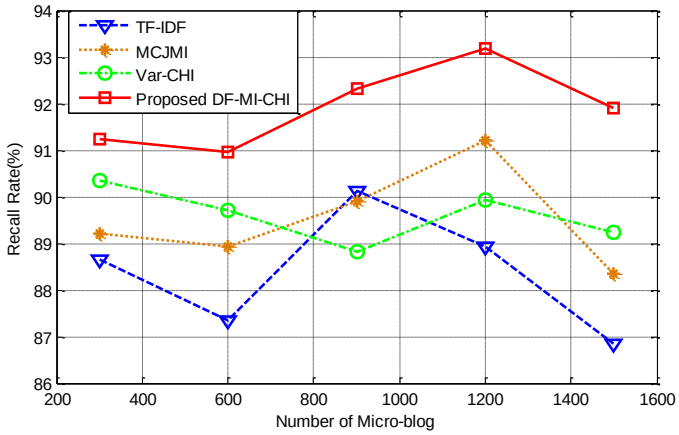

Figure 5. Comparison of recall rate of four dimensional reduction methods for sentiment analysis of micro-blogs

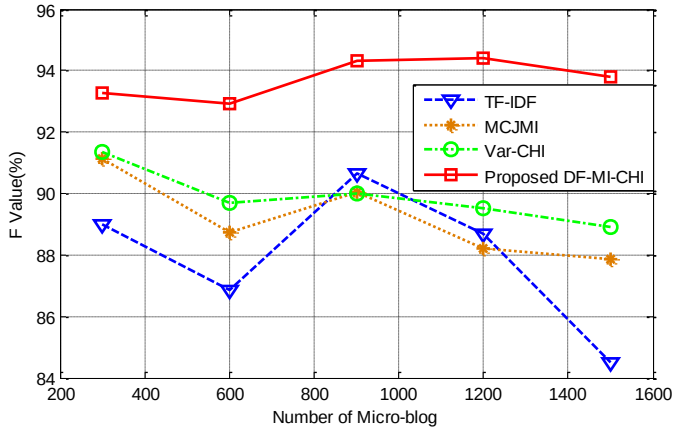

Figure 6. Comparison of F value of four dimensional reduction methods for emotional analysis on micro-blogs

The *F* value is a comprehensive index used to meet the needs of different categories of text analysis for different accuracy and recall. We can adjust the proportion of accuracy and recall by adjusting the factors in the expression. In this paper, we take the factor as 1, that is, accuracy and recall are of the same importance, as shown in Figure 6. The accuracy and the recall rate of improved text frequency (TF-IDF) algorithm fluctuate greatly. The maximum conditional joint mutual information (MCJMI) algorithm based on global joint mutual information and the improved chi-square test (Var-CHI) algorithm based on variance are relatively stable, but the accuracy and recall rate of word frequency and low frequency word processing are not high enough. Based on the improved proposed DF-MI-CHI integrated dimensionality reduction algorithm, the shortcomings of the three methods are compensated. The accuracy and recall are high and stable, and the *F* value tends to be stable.

5. Conclusions

Faced with the explosive growth of micro-blog text data, when analyzing the emotional tendency of micro-blogs, the increase in the number of feature words leads to the occurrence of "dimension disaster". This paper proposes a new improved dimension reduction algorithm for this problem. Firstly, we need to crawl the micro-blog data, use the distributed crawler tool based on Hadoop to obtain the real-time data of micro-blog, and store the acquired micro-blog data in the distributed file system HDFS based on Hadoop. Then, we pre-process the acquired data, store it into the micro-blog corpus through word segmentation, part-of-speech annotation, and deletion of stop words. Finally, we save it in the micro-blog corpus. This paper uses an improved dimension reduction algorithm of proposed DF-MI-CHI, which combines the document frequency (DF), mutual information (MI) and chi-square test (CHI). After testing and analyzing the positive emotional orientation of micro-blogs, the accuracy of the improved dimension reduction algorithm of proposed DF-MI-CHI is 95%. The recall rate is above 90%. Compared with the other three algorithms, with the increase in the number of micro-blog texts, the accuracy, recall rate, and F value tend to be stable.

Acknowledgments

The authors would like to thank the editors and the anonymous reviewers for their helpful comments and suggestions, which have improved the presentation. This work was supported in part by the Key Science Research Project of Colleges and Universities in Henan Province of China (No. 19A520009) and the National Science Foundation of China (No. 81501548).

References

1. X. N. Kong, "A Summary of Text Acquisition and Pre-Processing in Chinese Micro-Blog," *Software Guide*, Vol. 16, No. 2, pp. 186-189, 2018
2. X. B. Tang and Z. Q. Wang, "Research on Micro-Blog Topic Tracking Model based on Wikipedia Semantic Extension," *Information Science*, Vol. 35, No. 2, pp. 80-85, 2017
3. Z. G. Wang, "The Realization Method of Text Classification Processing in Micro-Blog in the Process of Network Public Opinion Monitoring," *Library and Information Guide*, Vol. 12, No. 1, pp. 129-133, 2016
4. Y. Yang, X. U. Bing, and M. Y. Yang, "An Emotional Classification Method based on Joint Deep Learning Model," *Journal of Shandong University*, Vol. 12, No. 4, pp. 652-662, 2017
5. X. M. Ye and J. C. Xia,"Improvement of TF-IDF Algorithm for Text Categorization," *Computer Engineering and Application*, Vol. 33, No. 5, pp. 65-70, 2018
6. Y. C. Mao and H. Ping, "Feature Selection based on Mutual Information of Maximum Joint Conditions," *Journal of Computer Applications*, Vol. 41, No. 20, pp. 172-178, 2018
7. Y. F. Qiu and W. Wang, "CHI Feature Selection Method based on Variance," *Application Research of Computers*, Vol. 29, No. 4, pp. 1304-1306, 2012
8. Z. G. Jin and B. H. Hu, "Multidimensional Feature Emotional Analysis of Micro-Blog based on Deep Learning," *Journal of Central South University*, Vol. 149, No. 5, pp. 1135-1140, 2018
9. J. Lv, X. Wang, and F. Huang, "TREST: A Hadoop based Distributed Mobile Trajectory Retrieval System," in *Proceedings of IEEE International Conference on Data Science in Cyberspace*, pp. 341-346, 2016
10. P. Y, Zou, J. H. Yang, and X. M. Li, "Supervised Topic Models with Weighted Words: Multi-Label Document Classification," *Frontiers of Information Technology & Electronic Engineering*, Vol. 19, No. 4, pp. 513-523, 2018
11. B. L. Li, "Using Class based Document Frequency to Select Features in Text Classification," *Abstract of Big Data Technology and Applications*, Vol. 25, No. 14, pp. 698-705, 2015
12. Z. L. Zhen, H. J. Wang, and L. X. Han, "Categorical Document Frequency based Feature Selection for Text Categorization," *Computer Engineering and Management Sciences*, Vol. 110, No. 4, pp. 526-531, 2011
13. L. F. Wang and Y. Xu, "Synergy and Redundancy in a Signaling Cascade with Different Feedback Mechanisms," *Communications in Theoretical Physics*, Vol. 70, No. 10, pp. 485-495, 2018
14. C. Zheng, Q. N. Xu, and J. P. Zhang, "Research on Recommendation System based on Mutual Information," *Microelectronics & Computer*, Vol. 35, No. 12, pp. 76-79, 2018

15. W. Liang and Y. Su, "Research on Text Classification Method based on Improved Mutual Information Function," *Bulletin of Science and Technology*, Vol. 34, No. 11, pp. 188-191, 2018
16. M. Y. Huang and X. B. Zhang, "Emotional Text Feature Selection based on CHI and Information Gain," *Journal of Xi'an Polytechnic University*, Vol. 12, No. 6, pp. 713-717, 2018
17. C. X. Song, X. H. Chen, and Q. Niu, "An Improved Feature Selection Method based on CHI in Text Categorization," *Microelectronics & Computer*, Vol. 35, No. 9, pp. 74-78, 2018
18. C. J. Fan and Y. T. Wang, "An Improved CHI Text Feature Selection Method," *Computer and Modernization*, Vol. 25, No. 11, pp. 7-11, 2016
19. K. Chen, S. J. Li, and B. Xie, "Emotional Analysis of Micro-Blog based on Semi-Supervised Learning," *Computer and Digital Engineering*, Vol. 46, No. 9, pp. 1850-1855, 2018
20. X. Kong and Q. Lin, "Summary of Research on Subjective Text Emotional Classification," *Information Technology*, Vol. 42, No. 8, pp. 126-130, 2018

Haodong Zhu received his B.S. degree from Lanzhou Jiaotong University in 2004, M.S. degree from Sichuan University of Science & Engineering in 2008, and Ph.D. from the Graduate University of Chinese Academy of Sciences in 2011. As a postdoctoral scholar, he studied image big data processing in the Postdoctoral Mobile Station of Computer Science and Technology at Tongji University from 2014 to 2016. As a visiting scholar, he studied micro-blog big data processing at Griffith University from 2017 to 2018. Since 2010, he has been an associate professor and master's tutor in the School of Computer and Communication Engineering at Zhengzhou University of Light Industry. His major research interests include cloud computation, intelligence information processing, computing intelligence, and data mining.

Wenqi Li received her B.S. degree from Zhengzhou University of Light Industry in 2017. Since 2017, she has been studying for her master's degree in the Computer and Communication Engineering College at Zhengzhou University of Light Industry. Her main research directions are intelligence information processing, micro-blog sentiment analysis, cloud computing, and data mining.

Hongchan Li received her B.S. degree from Heilongjiang Bayi Agricultural University in 2007 and her M.S. degree from Sichuan University of Science & Engineering in 2010. Since 2010, she has been a lecturer in the School of Computer and Communication Engineering at Zhengzhou University of Light Industry. Her major research interests include cloud computation, intelligence information processing, computing intelligence, and data mining.