Volume 11, Issue 2
April 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Sentiment Analysis Based on Social Networks Using Support Expectation-Maximization for E-Commerce Applications

[1] M. A Jamal Mohamed Yaseen Zubeir

Assistant Professor,P.G & Research Department of Computer Science,

Jamal Mohamed College (Autonomous), Tiruchirappalli –620 020, Tamil Nadu, India.

E-mail: [1]maj@jmc.edu

ABSTRACT

In recent years, the use of social networks has increased rapidly. More people and companies are becoming dependent on Data Mining (DM) technology on social networking to promote the reform of unstructured data. It may be likely to put them in a systematic model. Social media has played an important role in promoting different products, generating reviews and highlighting customers' opinions. Reviews provided by the customer help to establish a good name for the product and also help to improve the quality of the product. Reviews also help customer select the best product among available products. For this reason, the provision of a particular product's customer reviews facilitates the false-positive product review by providing a false negative of product reviews, to make it a possible downgrade which speaks about the project organization itself. The proposed Support Expectation-Maximization (SEM) algorithm is used to analyse the reviews based on sentimental lexical words to customers select the best products to resolve this problem. It collects the dataset from E-commerce online websites or applications such as Flip-kart or Amazon. The collected datasets are trained into the pre-processing process to analyse the stemming, Tokenization and Stop-word Elimination. Then, feature selection is done to select the best features of customer reviews sentiment lexical words weightages to reduce the dimensionality. A feature selection dataset are trained into the proposed Support Expectation-Maximization (SEM) algorithm to classify the reviews about the product, and it helps select the best product.

Keywords: *Data Mining (DM), social networks, Support Expectation-Maximization (SEM), product review, sentimental lexical words, E-commerce.*

1. INTRODUCTION

Social media has become an effective way to communicate and share information like communicating global interests of the people together, a community of interests and information costs.There are many major sources for social media operation. Many of the receivers provide the previous excellent access and availability than the main source of information. The advantages of social media offer companies worldwide, e-commerce, a platform for online shopping.

The world is now reduced to a small village in the tangible impact of social media. This connects ethnic community around the world from different regions and facilitates their opinions, experiences, emotions, hobbies, photos and videos to be shared about the product opinions. The door of the benefits are open to all customers based on the analysis of data provided by the social media which improves the organisation's productivity.The structure of social media data is displayed in different formats, such as unorganized: text, audio, images, and video. In addition, social media, which performs a traditional statistical method, is not suitable for analysing the huge amount of these data to provide a large amount of continuous real-time data.

Therefore, Data Mining (DM)is the technology which can play an important role in overcoming these problems. Opinion mining is a kind of natural language processing used to track the public's awareness of a particular product. The specified text is classified as expressed polarity document, if the opinion is positive and the basic tasks of opinion mining is to determine whether the opinion is negative or positive.

Dominant work in the decision-making process, most of statistics, is how customers think about the products based on transactional data. One such form of social media data, a customer at the level of satisfaction for a particular product, has been written to generate product evaluation reports. These reviews are designed to help individuals and business organizations. This censorship prompted some people to enter their false reviews to promote certain products or downgrading some people. The main reason for this action is to make products and reliable service unfaithful to write reviews and ratings false. So, to earn more profits, these fake comments must be detected and removed. The system uses three slides of pre-processing in the development of fake review detection and classification. The proposed pre-processing method will be used to remove a review or modified review of abuse properly.

2. RELATED WORK

Dissemination and development of data mining technology bring a serious threat to the security of sensitive personal information [1].The basic idea of privacy data mining is to perform the mined data effective algorithm by changing the data in a way that does not compromise the confidential information included in the data. For each user of the type, it will adopt a method for discussing his privacy problem to protect confidential information.Technical and economic data of mining enterprise feature a multi-dimensional non-linearity [2]. It is an important economic indicator for the sale price

Volume 11, Issue 2
April 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

data-mining companies of mineral products, and geological data is an important technical data.

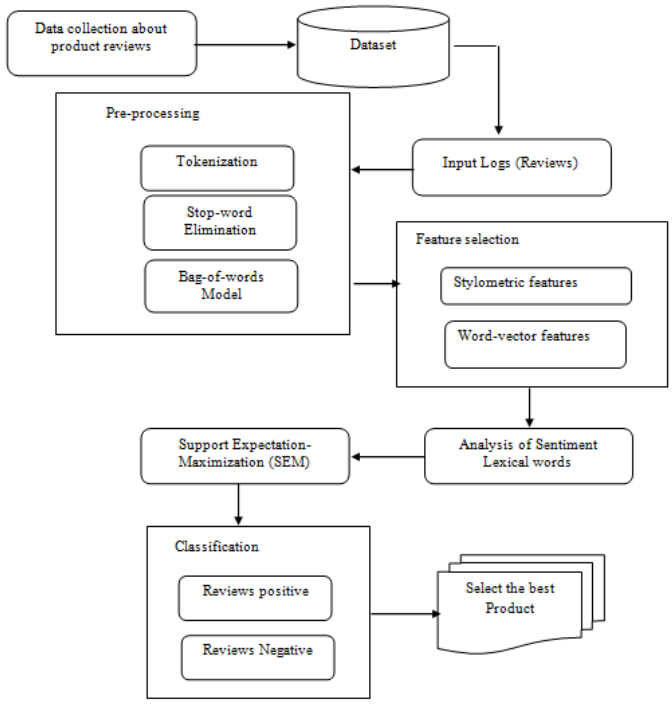

Figure 1: Proposed diagram for sentiment analysis

Due to limitations of technical conditions and equipment conditions, the course of mineral development reduces the accuracy of the estimated shape and reserves of the ore body by a large amount of geological data.

The success of learned rules in data mining depends on its compatibility: it is convenient to take the appropriate action in any of the actual business environment [3]. To improve the action rules, different researchers, initially from only a variety of data mining, which is presented in the data set of the business field framework, focused on various factors.

If knowledge is extracted from the dispersed sampled data as classification rules, it may be necessary to bond or fuse these rules in the conventional method, which is typically (e.g., in the form of a classification of the cluster)done by combining the output of the classifier by a set of complex classification rules [4].This technique is based on the multinomial distribution for classifying a multivariate normal distribution for the probability of continuous ones to generate a classifier and the input size.

The purpose of our job is to find that the only non-labeled image has given a very general set of dominant objects. This creates more challenges than the typical co-localization or weak teacher localization work [5].Effective pattern mining-based data mining methods facilitate pre-trained Convolutional Neural Network (CNN) in localized mining of objects called that uses the benefits of the function representing.

Clustering is a set of statistical learning techniques designed to find a structure that groups a homogeneous data partition, called a cluster. There are several fields of successful application clusters such as medicine, biology, finance, economy [6].The concept of clustering in a multi-factor data analysis is a problem. The Principle Component Analysis (PCA), while dealing with only the quantification especially in the medical, the problems that exist in the natural world, in many cases, is based on the qualitative and quantitative measurement of different kinds.

Position acquisition technique is used for tracking fast-forward path data generation traces of moving objects. Track is usually expressed as a sequence of the geographical position of the time stamp [7]. A wide range of applications can benefit from orbit data mining. The unprecedented opportunity as it has a large orbit data which led to enormous challenges.

Secure Multi-party Computing (SMC) makes it possible to calculate the input function parties together while maintaining confidentiality for each input. It is widely used in such privacy data mining tasks and privacy requirements, such as learning task output and when used simultaneously, input data privacy applications are also protected [8].However, the existing SMC-based solutions are temporary, because it has been proposed for a particular application and cannot be applied directly to other applications.

Big data focus on multiple autonomous sources of complex and growing data sets on a large scale. Network, data storage and data collection capabilities, with the rapid development of big data, have rapidly expanded in all areas of science and engineering, such as biomedical sciences, physics and biology [9].Data-driven model, information sources, mining, analysis, modeling of the user's interest are the need for security and privacy considerations.

Many of the data mining techniques have been existing to useful patterns in the text document. However, how to effectively discover patterns, especially in text mining, is still an open research issue [10].Many of the existing text mining techniques, the long-term methods, will come from all of the ambiguity synonymous with problems.

Traditional data mining techniques, including data distribution and function of unanalysed are not entirely derived from the association rules. As a data mining traditional techniques results, the probability of having a large redundancy, large root mean square error and the approximate time of such defects are very high [11].To address these issues, based on the density estimation distribution pattern of the customer review data stream, an Association Rule Mining (ARM) algorithm is suggested.

The author introduces a set of scalable algorithms for recognizing the day-to-day behaviour patterns of people. These patterns have been extracted from a variety of time data that has been collected from the Smartphone [12] [13].The author has developed a sensor on these devices and has often determined behaviour patterns and time granularity, which has been inspired by the individual period of time to the event.

The sparse witnesses in Internet of Things (IoT), ubiquitous computing, and the behavioural data, big data are released from various sources [14]. Frequent item set mining is the ARM, which is the first step without the excellent performance of machine learning [15].

Volume 11, Issue 2
April 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

3. MATERIALS AND METHODS

Nowadays, anyone can write or review any comments on the text, which can draw attention of individuals and organizations to give unworthy spam to promote or discredit some of the target product. Data Mining (DM) technology is the process of extracting the knowledge hidden from the data. Therefore, developing an automatic intelligent system for reviewing comments is mandatory to divide them into spam and non-spam categories. A spam analyser is to provide comments automatically which classifies the user's comments about the product as positive and negative. The proposed Support Expectation-Maximization (SEM) algorithm helps the customer to select the best quality of products.

3.1 DATA COLLECTION

In this stage, data collection from online e-commerce sites and applications, such as Amazon, Flipkart is done to create a transactional data set which has different characteristics and sizes to collect the evaluation and manual review. The product review comments dataset is a random selection of any of the records and products available on any website.

3.2 PRE-PROCESSING

This stage is an important part of Data Mining (DM) technology, a data processing technology to handle noisy and inconsistent data. This pre-processing stage contains three steps for analysing the comment reviews about the product, they are tokenization, a bag of word model and stop word elimination process.

3.2.1 TOKENIZATION

In this stage, Tokenization has split the sentence into sparse, punctuation, and symbols which are called tokens.All characters in the string are a continuous part of the token. All tokens may occur only by alphabetic characters, numeric characters or alphanumeric characters.

3.2.2 STOP-WORD ELIMINATION

This stage is going to help text mining, the most common words such as prepositions, articles, and a professional noun can be regarded as a stop word. Since each text document is being processed in this statement, it is not required for text mining applications. All of these words have been eliminated.

3.2.3 BAG-OF-WORDS (BOW)

In this stage, the Bag of Words (BOW) is one of the easiest language models used in natural language programs. This unigram model of the text can track the number of occurrences of each word. In the BOW, a specific subjective score is given to each word to account your own personal word. This will be doing the text corpus has been referred to as the vocabulary unique word list.

3.3 FEATURE SELECTION

Feature selection is done to select best features of vocabulary, lexical words based customer reviews about products. Feature selection is applied on both Stylometric features and Word-vector features using Chi-square test.

3.3.1 STYLOMETRIC FEATURES

Elimination of recursive functions has been used to select the most important functions from the stylometric feature set. This is the number of features in the weakest data eliminated from the data set which is set to be reduced to at least a specific value.

3.3.2 Word-vector features using Chi-square test

The uses of the words are removed from the vocabulary lemmatizer words and stemming port vector space or vocabulary. Both stem and Lemmatization are then used to restore their roots, but the coarse method does consider the context of the speech portion of word generation.

The proposed chi-square test is used to reduce the time complexity when dealing with large-scale vector space. In addition, Lemmatizer, stemmer, and the chi-square test, reduced to a greater extent to ensure good performance, will be combined into words. The proposed chi-square test selects the features from the dataset-based vocabulary weights.

3.4 ANALYSIS OF SENTIMENT LEXICAL WORDS

Sentiment lexical words are emotions in a common dictionary word list to improve the important lexical vocabulary of data mining techniques used in the document. Sentiment lexical words also are improving the quality of the product analysis which is negative or positive.

3.5 SUPPORT EXPECTATION-MAXIMIZATION

Sentimental analysis using Support Expectation-Maximization (SEM) is done to the most accurate classification. Weightages based reviews comments are trained into the proposed SEM algorithm and it helps to select the best product recommendation. The proposed SEM algorithm supports maximum positive review comments about the product to recommend to customer and it helps to select quality of product.

Algorithm Steps
Input:
Data collection from E-commerce website product reviews (pr) dataset
Output:
Best product recommendation Begin
Step 1:
Import the product reviews dataset
Step 2:
Read the dataset pr = pr1, pr2….prn
Step 3:
Analysis the customer's opinion
Step 4:
Classification based on feature selection weightages
Step 5:
Recommended to best product to customer
End

Let us assume that pr represents the n number of customer product reviews dataset in this algorithm steps which are provided to recommend the best product.

4. RESULTS AND DISCUSSION

Customer review data that have been collected from social media can be improved by selecting the best products available to customers with quality.

Table 1 describes the simulation parameters of the proposed implementation using python language. The following parameters like classification accuracy performance, sensitivity performance, specificity performance, time complexity are compared between the proposed and existing methods.

Table 2shows the analysis of classification accuracy performance for customer reviews based product recommendation. The proposed Support Expectation-Maximization (SEM) is compared with previous algorithms like Attribute Depth Measure (ADM), Feature Influence Measure (FIM), Functional Influence Correlation Measure (FCM), and Functional Correlation Measure (FCM).

Results of analysis of Specificity performance shown in table 4 for recommended best product based on customer reviews. Results of analysis of sensitivity performance shown in table 3 for recommended best product based on customer reviews.

Figure 1 defines the proposed diagram for sentiment analysis about particular product data collection from online e-commerce websites or applications such as amazon etc. for customer opinion about the product. The proposed Support Expectation-Maximization (SEM) algorithm is used to classifying the best product buying customer.

Figure 2 shows the analysis of classification performance for product recommendation based on customer reviews. The proposed Support Expectation-Maximization (SEM) classification's accuracy result is 97.3%. Similarly the existing methods are Attribute Depth Measure (ADM) classification's accuracy result is 95.4%, Feature Influence Measure (FIM) classification's accuracy result is 95.8%, Functional Influence Correlation Measure (FICM) classification's accuracy result is 96.2%, and Functional Correlation Measure (FCM) classification's accuracy result is 96.3%. Figure 3 defines the analysis of sensitivity performance for product recommendation based customer reviews. The proposed Support Expectation-Maximization (SEM) algorithm sensitivity performance is 92.5%, likewise the existing algorithms are similarly the existing methods are Attribute Depth Measure (ADM) sensitivity result is 89.3%, Feature Influence Measure (FIM) sensitivity result is 89.4%, Functional Influence Correlation Measure (FICM) sensitivity result is 89.5%, and Functional Correlation Measure (FCM) sensitivity result is 90.2%.

Figure 4 defines the specificity performance for product recommendation based on customer reviews. The proposed Support Expectation-Maximization (SEM) specificity performance is 93.1%, similarly the existing methods are Attribute Depth Measure (ADM) specificity result is 88.3%, Feature Influence Measure (FIM) specificity result is 89.4%, Functional Influence Correlation Measure (FICM) specificity result is 89.6%, and Functional Correlation Measure (FCM) specificity result is 90.3%.

Figure 5 defines the analysis of time complexity performance for product recommendation based on customer reviews. The proposed Support Expectation-Maximization (SEM) algorithm time complexity result is 21sec similarly the existing methods are Attribute Depth Measure (ADM) time complexity result is 35 sec, Feature Influence Measure (FIM) time complexity result is 30 sec, Functional Influence Correlation Measure (FICM) time complexity result is 26 sec, and Functional Correlation Measure (FCM) time complexity result is 23 sec.

Table 1: Details of Simulation parameters

Parameters	Values
Simulation tool	Anaconda
Language	Python
Dataset type	Content based customer reviews
Number of dataset	1000
Testing dataset	800
Training dataset	200

Table 2: Analysis of Classification accuracy performance

No of product review data	ADM in %	FIM in %	FICM in %	FCM in %	SEM in %
100	93.2	94.1	95.2	95.3	96.1
200	94.1	94.3	95.1	95.4	96.3
300	95.2	95.3	95.4	96.2	96.4
400	95.4	95.8	96.2	96.3	97.3

Table 3: Analysis of sensitivity performance

No of product review data	ADM in %	FIM in %	FICM in %	FCM in %	SEM in %
100	85.2	86.3	87.4	88.5	89.4
200	86.1	87.2	87.4	88.6	89.6
300	88.2	88.4	88.6	89.2	90.2
400	89.3	89.4	89.5	90.1	92.5

Table 4: Analysis of Specificity performance

No of product review data	ADM in %	FIM in %	FICM in %	FCM in %	SEM in %
100	84.2	85.3	86.4	87.5	88.4
200	85.1	86.4	87.3	87.6	88.6
300	87.2	88.2	88.7	89.2	90.3
400	88.3	89.4	89.6	90.3	93.1

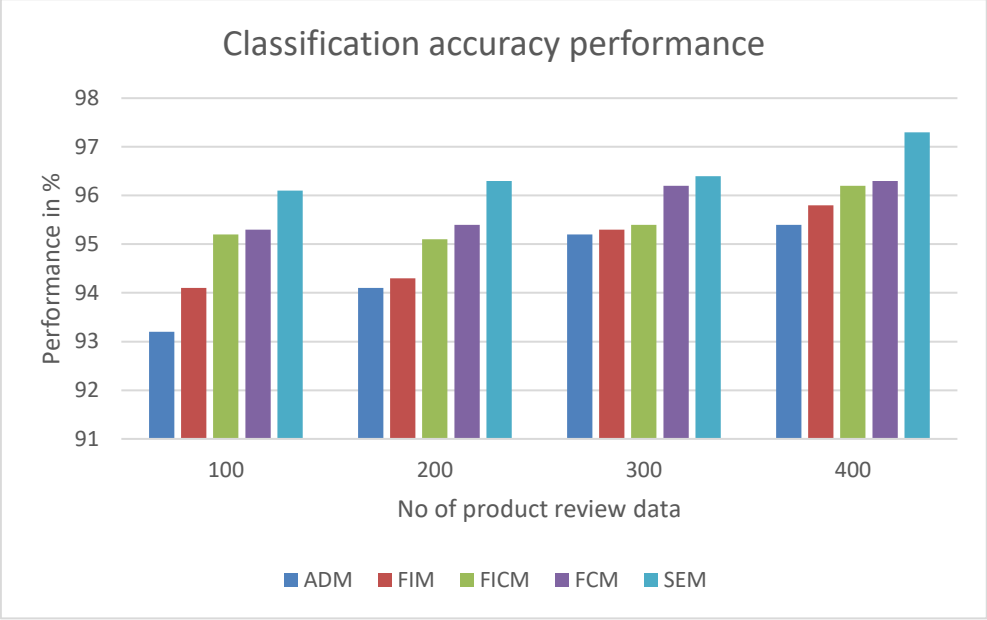

Figure 2: Analysis of classification accuracy performance

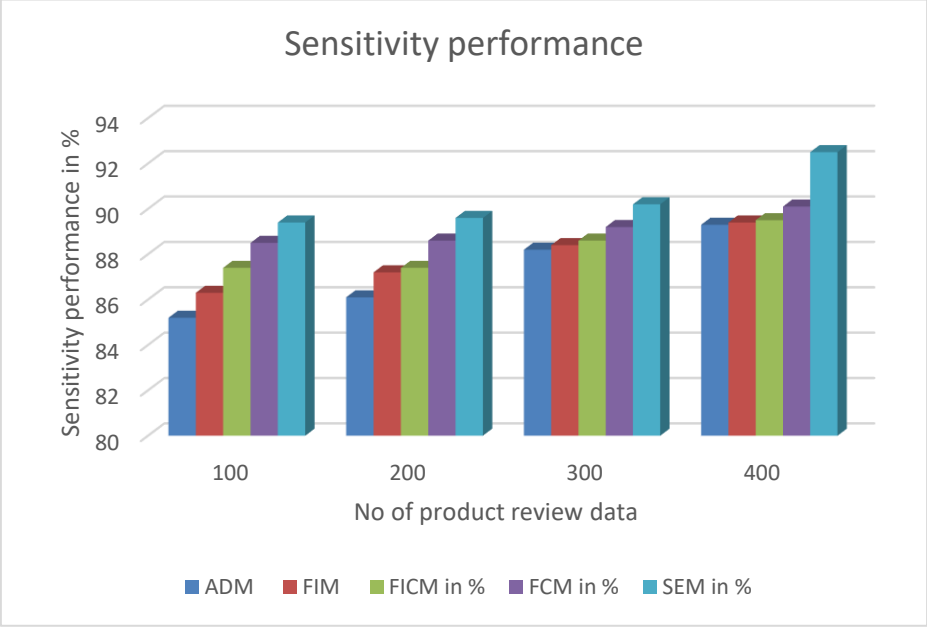

Figure 3: Analysis of sensitivity performance

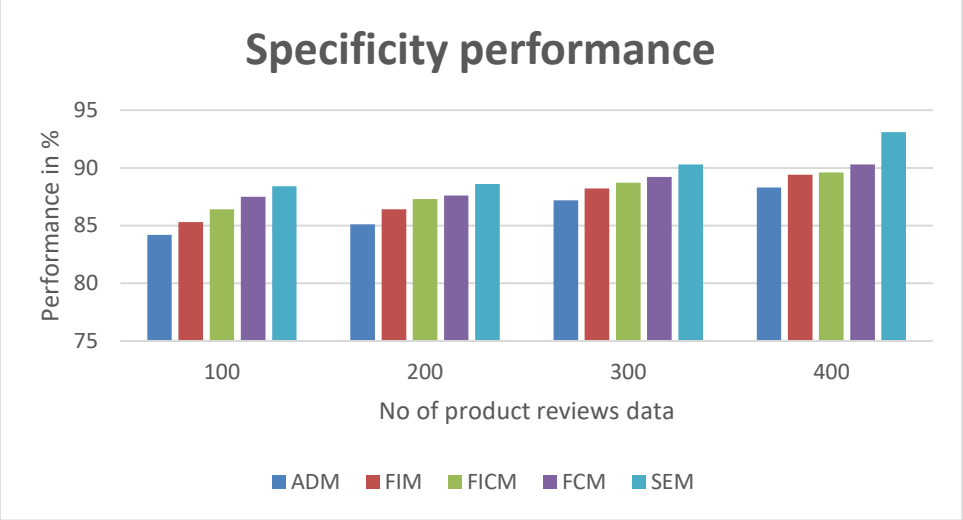

Figure 4: Analysis of specificity performance

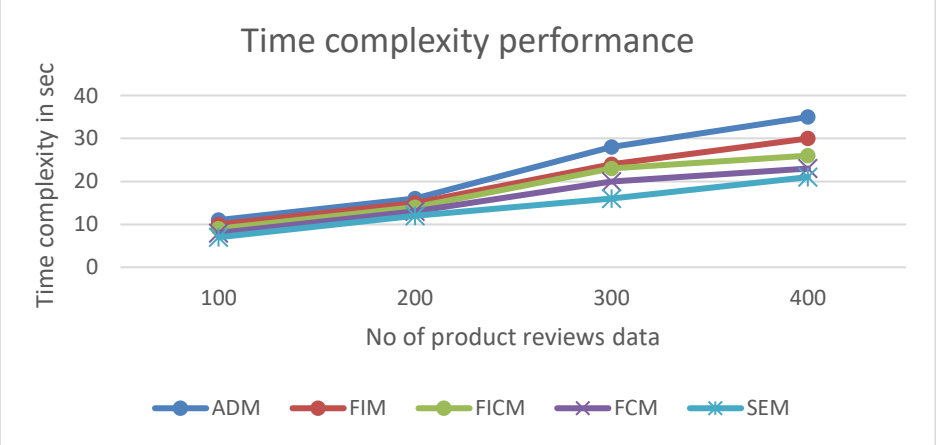

Figure 5: Analysis of Time complexity performance

5. CONCLUSION

Data Mining (DM) based product reviews are done to find the location of the thoroughly enormous amount of data and is the process of discovery which is beyond the simple analysis patterns and trends. There is a sentimental analysis to be used in product reviews for the detection of spam. Sentiment analysis plays an important role in business decisions about the product/service. The main challenge of sentiment analysis will include the weight of the essential functional features for customer opinion review classification. Decision-making about the products can lead to both business growth and de-promotion based on customer opinions in social networks. The proposed SEM algorithm has been proposed to classify the accuracy performance of products based on customer reviews or opinion. The proposed algorithm provides the classification accuracy performance of 97.3%, specificity performance of 93.5%, sensitivity performance of 92.5%, and time complexity result of 21 sec.

REFERENCES

[1] L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren, "Information Security in Big Data: Privacy and Data Mining," in IEEE Access, vol. 2, pp. 1149-1176, 2014, doi: 10.1109/ACCESS.2014.2362522.

[2] J. Ming, L. Zhang, J. Sun and Y. Zhang, "Analysis models of technical and economic data of mining enterprises based on big data analysis," 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, pp. 224-227, doi: 10.1109/ICCCBDA.2018.8386516.

[3] F. Fatima, R. Talib, M. K. Hanif and M. Awais, "A Paradigm-Shifting From Domain-Driven Data Mining Frameworks to Process-Based Domain-Driven Data Mining-Actionable Knowledge Discovery Framework," in IEEE Access, vol. 8, pp. 210763-210774, 2020, doi: 10.1109/ACCESS.2020.3039111.

[4] D. Fisch, E. Kalkowski and B. Sick, "Knowledge Fusion for Probabilistic Generative Classifiers with Data Mining Applications," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 3, pp. 652-666, March 2014, doi: 10.1109/TKDE.2013.20.

[5] R. Zhang et al., "Object Discovery From a Single Unlabeled Image by Mining Frequent Itemsets With Multi-Scale Features," in IEEE Transactions on Image Processing, vol. 29, pp. 8606-8621, 2020, doi: 10.1109/TIP.2020.3015543.

[6] F. Saâdaoui, P. R. Bertrand, G. Boudet, K. Rouffiac, F. Dutheil and A. Chamoux, "A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining," in IEEE Transactions on NanoBioscience, vol. 14, no. 7, pp. 707-715, Oct. 2015, doi: 10.1109/TNB.2015.2477407.

[7] Z. Feng and Y. Zhu, "A Survey on Trajectory Data Mining: Techniques and Applications," in IEEE Access, vol. 4, pp. 2056-2067, 2016, doi: 10.1109/ACCESS.2016.2553681.

[8] S. G. Teo, J. Cao and V. C. S. Lee, "DAG: A General Model for Privacy-Preserving Data Mining," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 1, pp. 40-53, 1 Jan. 2020, doi: 10.1109/TKDE.2018.2880743.

[9] X. Wu, X. Zhu, G. Wu and W. Ding, "Data mining with big data," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014, doi: 10.1109/TKDE.2013.109.

[10] D. Liu, W. Baskett, D. Beversdorf and C. -R. Shyu, "Exploratory Data Mining for Subgroup Cohort Discoveries and Prioritization," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 5, pp. 1456-1468, May 2020, doi: 10.1109/JBHI.2019.2939149.

[11] N. Zhong, Y. Li and S. Wu, "Effective Pattern Discovery for Text Mining," in IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 30-44, Jan. 2012, doi: 10.1109/TKDE.2010.211.

[12] X. Li, Y. Wang and D. Li, "Medical Data Stream Distribution Pattern Association Rule Mining Algorithm Based on Density Estimation," in IEEE Access, vol. 7, pp. 141319-141329, 2019, doi: 10.1109/ACCESS.2019.2943817.

[13] R. Rawassizadeh, E. Momeni, C. Dobbins, J. Gharibshah and M. Pazzani, "Scalable Daily Human Behavioral Pattern Mining from Multivariate Temporal Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 11, pp. 3098-3112, 1 Nov. 2016, doi: 10.1109/TKDE.2016.2592527.

[14] M. Yasir et al., "TRICE: Mining Frequent Itemsets by Iterative TRimmed Transaction LattICE in Sparse Big Data," in IEEE Access, vol. 7, pp. 181688-181705, 2019, doi: 10.1109/ACCESS.2019.2959878.

[15] S. Hajian and J. Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining," in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 7, pp. 1445-1459, July 2013, doi: 10.1109/TKDE.2012.72.

AUTHOR PROFILES

Dr.Jamal Mohamed Yaseen Zubeir working as Assistant professor in Department of Computer Science, Jamal Mohamed College, Trichy, Tamilnadu, India. He Completed his PG and PhD from Jamal Mohamed College Affiliated by Bharathidasan University.