Dr. K. Sundravadivelu[1]

Improved Semantic Information and Extraction based Effective Pattern Discovery Mining in Bigdata Using Latent Semantic Indexing Model

Journal of Electrical Systems

Abstract

In the era of Big Data, extracting meaningful patterns and insights is pivotal for decision-making. Latent Semantic Indexing (LSI) emerges as a promising approach for uncovering semantic relationships in vast datasets. This document explores how LSI can be effectively applied for pattern discovery, emphasizing its capability to extract and represent semantic information from unstructured and structured data. Key challenges, methodologies, and potential applications are discussed to highlight the role of LSI in Big Data mining. Numerous data mining methods have been proposed for mining to find useful patterns in text documents. Though, how toeffectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since mostexisting text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. This paper proposed semantic information and extraction for effective pattern discovery mining in big data using LSI algorithm. Semantic Information and extraction stages' using Latent Semantic Indexing (LSI) algorithm and patterns are organized in specific format then evaluates the term weights and discovered specific patterns in the set of documents.

***Keywords*:** VSM, LSI, SVD, LDA, Pattern Discovery, Semantic Information and Extraction, etc.,

1. INTRODUCTION

Big Data has revolutionized data analytics, enabling organizations to derive insights from massive volumes of information. However, the complexity and heterogeneity of Big Data necessitate advanced techniques for pattern discovery. Semantic information extraction focuses on identifying meaningful relationships and patterns beyond mere syntactic data analysis. Latent Semantic Indexing (LSI), a dimensionality reduction technique based on singular value decomposition (SVD), is particularly suited for capturing latent structures in textual and multidimensional data. Huge amounts of data are nowadays collected and storedby databases, with the hope of being useful in the future. This positions the challenge of managing such loads of data and extracting from it appropriate knowledge for mining. Big Data is currently globally spread and widely accepted, representingalso a synonym of vanguard in terms of information management,although this does not come without controversy argued, practitioners need to step forward "from Big Data to Big Impact" foreffectively benefiting from the advantages provided by Big Data[15].

Big Data is everywhere these days, whether in the form ofstructured data, such as organizations traditional databases, for example customer relationship management or else unstructured data, drivenby new communication technologies and user editing platforms for example text, images and videos. Socialnetworks such as Facebook and Twitter are having a huge impact oninfluencing customers' decisions, leading organizations and brandsto incorporate information originated in such platforms in their mining techniques [16, 17].

2. LITERATURE REVIEW

Big data platforms, such as Hadoop MapReduce and ApacheSpark have been adopted as one stop solution for most of big dataproblems. However, jobs running under these frameworks requirea careful parameter tunings to improve processing performances. Moreover, bad parameter tunings may lead to poor performancesor even job failures. In [1] authors introduced an adaptive framework, named mr. Moulder, for automatic tuning parameters of new jobs. Mr. Moulder exploits a dynamically extended configuration repository to recommend

[1]Assistant Professor, Department of Computer Science, School of Information Technology, Madurai Kamaraj University, Madurai-625 021, Tamil Nadu, India.
Email: svadivelu2021@gmail.com1

near-optimalconfiguration for big data jobs in a short time.The processing of a huge amount of data in a reasonable timewith the use of huge computing resources will increase the energyconsumption and consequently this will lead to the increase ofgreenhouse gas emissions and environmental impacts. Authors indeal with this problem and analyse the relevanceamong green measures and big data. In the same context, authors in [2] presented a state-of-the-art on solutionsaiming to find correlations among sustainable developmentgoals and information and communications technologies.

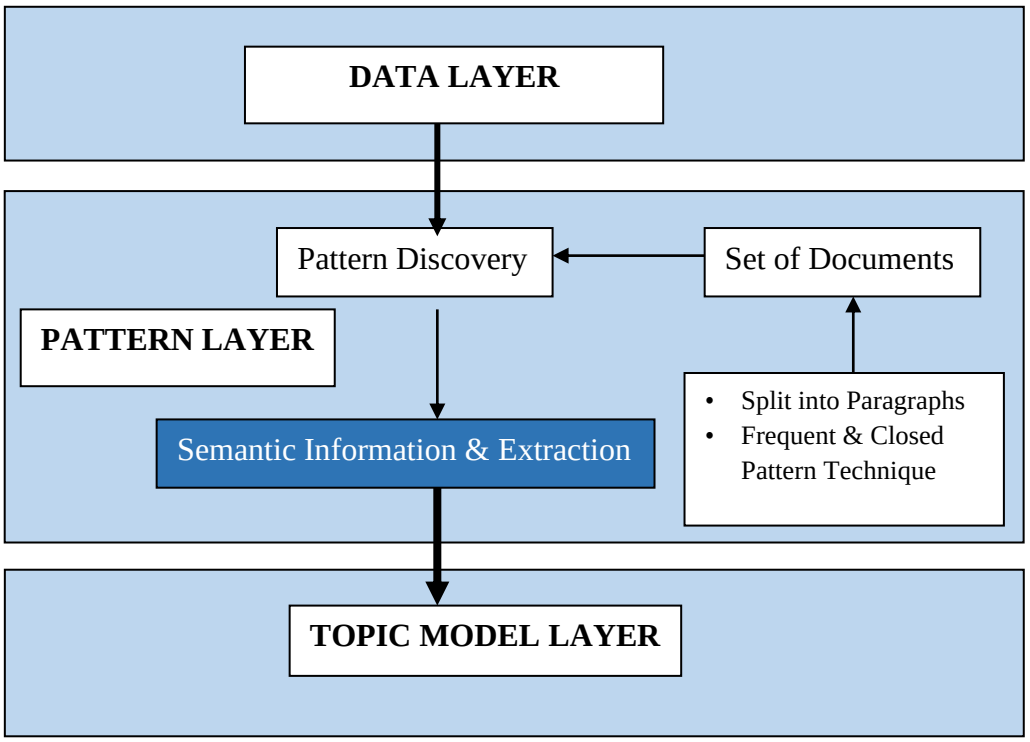
Mining techniques have been used for textanalysis by extracting co-occurring terms as descriptivephrases from document collections. However, the effectivenessof the text mining systems using phrases as textrepresentation showed no significant improvement. Thelikely reason was that a phrase-based method had "lower consistency of assignment and lower document frequency for terms" as mentioned in.Term-based ontology mining methods also providedsome thoughts for text representations. For example, hierarchicalclustering was used to determine synonymyand hyponymy relations between keywords. Also, thepattern evolution technique was introduced in in order toimprove the performance of term-based ontology mining.Pattern mining has been extensively studied in datamining communities for many years. A variety of efficient algorithms such as Apriori-like algorithms PrefixSpan FP-tree, SPADE, SLPMiner, and GST have been proposed.

These research workshave mainly focused on developing efficient mining algorithmsfor discovering patterns from a large data collection.However, searching for useful and interesting patterns andrules was still an open problem. In the field oftext mining, pattern mining techniques can be used to findvarious text patterns, such as sequential patterns, frequent item sets, co-occurring terms and multiple grams, for buildingup a representation with these new types of features.Nevertheless, the challenging issue is how to effectively dealwith the large amount of discovered patterns [4]. Term-Document Matrix: LSI begins with constructing a term-document matrix where rows represent terms and columns represent documents. Singular Value Decomposition (SVD): The matrix is decomposed into three components: Here, represents the term space, contains singular values, and represents document vectors. Dimensionality Reduction: By retaining only the top singular values, LSI captures the most significant patterns, reducing noise and redundancy.

3. PROPOSED WORK

The semantic information and extraction, text chunks become data bits, data bits become semantic metadata and semantic metadata become knowledge bytes – data pieces, ready to leverage for insights, decisions and actions. The semantic data will be utilized in the example scientific categorization to work on the presentation of analysis utilizing closed patterns in text mining. Semantic annotation is the process of tagging documents with relevant concepts [3]. The documents are enriched with metadata: references that link the content to concepts, described in a knowledge graph. This makes unstructured content easier to find, interpret and reuse. The Semantic information and extraction is the process of extracting information from unstructured textual sources to enable finding entities as well as classifying and storing them in a database.

The semantic information are adding metadata to the extracted concepts, these technology need to solves more challenges in enterprise content management and knowledge discovery. Clean and normalize data (e.g., tokenization, stop-word removal, stemming). To create a term-document matrix from structured / unstructured datasets. Apply SVD to identify significant dimensions representing semantic relationships. Extract clusters, correlations, or trends from the reduced dimensionality representation. Utilize similarity measures (e.g., cosine similarity) for identifying semantic patterns. Use tools like heat maps and clustering diagrams for presenting insights [19].



- **Latent Semantic Indexing (LSI) model**

LSI model is a type of algebraic model. The main idea of the LSI model is to map each document and query vector into a lower dimensional space associated with documents, which is used for semantic information retrieval extraction. LSI is to map each document and query vector into a lower dimensional space which is associated with documents. The documents are achieved by mapping the index terms vectors into this lower dimensional space. The LSI proposes to decompose a term-document association matrix in three components using singular value decomposition.

The first one is the matrix of eigenvectors derived from the term-to term correlation matrix; the second one is the matrix of Eigen vectors derived from the transpose of the document-to document matrix; the third one is a $r \times r$ diagonal matrix of singular values where r is the minimum between the row and the column of the original matrix, and the rank of the term-document association matrix. Consider only the largest singular values of the third matrix are kept, along with their corresponding columns in the first and the third matrix while the rest singular values are deleted. LSI transforms high-dimensional data into a lower-dimensional semantic space, improving the identification of latent patterns. LSI uses Singular Value Decomposition (SVD) to decompose a term-document matrix AAA into three matrices: $A=U\Sigma V^T$. U represents the term space, Σ contains singular values indicating importance, and V^T captures document similarities.

The resultant matrix is the matrix of rank *s*, which is closest to the original matrix in the least square sense. The relationship between two documents in the reduced space of dimensionalities can be obtained from the multiplication of the resultant matrix and its transpose. To rank documents with regards to a query, the query is modeled as a pseudo document in the original term-document matrix. Assume the query is modelled as the document with number 0. Then the first row in the multiplication of the resultant matrix and its transpose provides the ranks of all documents with respect to this query. Based on the index term list, each concept c is formed as an array in which each element is obtained by tf-idf, and all the concepts in the ontology are formed as a term-concept matrix A.

The term-concept matrix is then decomposed by the SVD (Singular Value Decomposition) approach, which can be mathematically denoted as Equation (1)

$$A = U\sum V^T \quad (1)$$

where U is the matrix derived from the term-to-term matrix given by A^T, V^T is the matrix derived from the transpose of the concept to concept matrix given by A^TA, and $\sum$ is a r × r diagonal matrix of singular values where r = min(t, N) is the rank of A. Considering that now only k largest singular values of $\sum$ are kept along

with their corresponding columns in U and V^T, the resultant A_k matrix is the matrix of rank k which is closest to the original matrix A in the least square sense[18]. This matrix is given by Equation (2)

$$A_k = U_k \sum_k V_k^T \quad (2)$$

Where k (k < r) is the dimensionality of a reduced concept space.

Analogous to the concept, a query q can be formed as an index term-based array in which each element is the tf-idf weight between the query and a term from the index term list. The array can then be translated into the concept space by Eq. (3), and then compared with A_k by the cosine algorithm to calculate the similarity values of each concept, which can be denoted by Eq. (4)

$$q' = \sum_k^{-1} U_k^T q \quad (3)$$

$$sim(c, q') = \frac{|A_k \cap q'|}{|A_k| \times |q'|} \quad (4)$$

LSI forms an efficient indexing scheme for the documents in the collection, and it supports for elimination of noise and removal of redundancy [6-15]. Captures semantic relationships by analyzing co-occurrence patterns. Handles synonymy (different terms with similar meanings) and polysemy (same term with multiple meanings).Improves computational efficiency in high-dimensional data.

4. RESULTS AND DISCUSSION

In this section, we report our experimental results. First, present the quality measures for the proposed work then, we discuss the scalability performance of the algorithms.We used performance indicators for evaluating results namely precision, recall, f-measure and mean average precision.These parameters are adopted in the experiment, a proper threshold values need to be decided to filter the inappropriate concepts for metadata.

The framework is executed using Intel Core (TM) i7 with R - Tools and its relating libraries in Windows 10 operating system having a RAM limit of 32 GB. The observational examination means to advance the boundaries of the model utilizing our proposed procedure. It likewise assesses the appropriateness of LSI Model. The data set is collected from an internet based research article data set. The collection of reports taken for the survey is 1, 00,000 articles from Google Research Scholar, PubMed, NCBI, Elsevier, and IEEE, Scopus Database, Web of Science.

Comparative Model:

LSI outperforms traditional pattern discovery techniques by focusing on latent semantics rather than surface-level statistics. However, it faces competition from advanced machine learning methods such as Latent Dirichlet Allocation (LDA) and deep learning, which can sometimes offer greater flexibility and accuracy. Traditional data mining approaches often rely on syntactic patterns, which fail to capture deeper semantic meanings. In contrast, LSI effectively uncovers latent relationships by analyzing the structure of data in reduced dimensions. Latent Dirichlet Allocation (LDA**)**: While LDA is a probabilistic model that excels in topic modeling, LSI is computationally simpler and often faster for smaller datasets. Deep Learning: Neural networks can outperform LSI in capturing complex patterns but require extensive computational resources and large datasets for training. This study of LSI technique includes three processes, semantic data, and extraction by utilizing pattern improvement, Topic layer demonstrating, and group task, to reduce the over fitting of found designs in text utilizing data information.

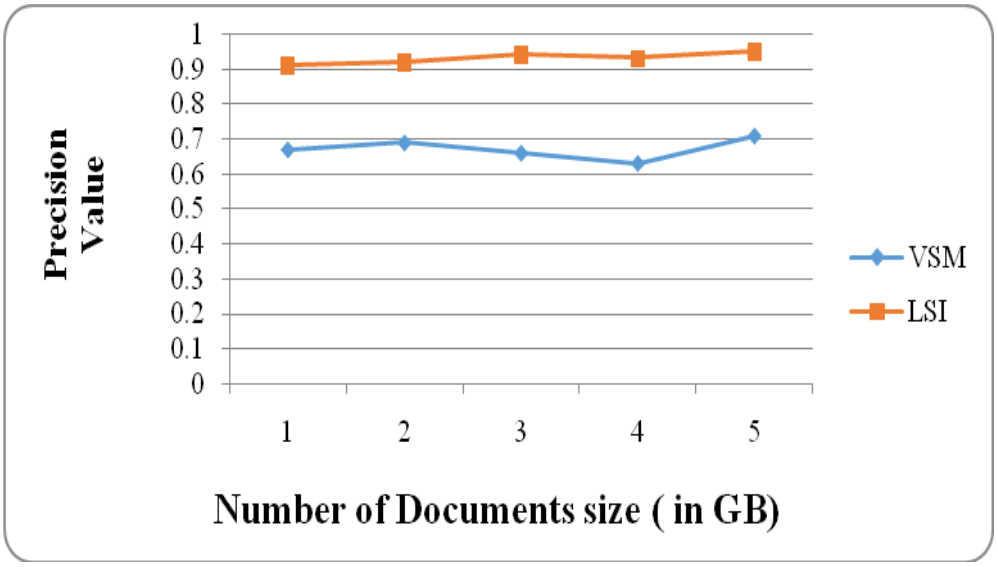
Accuracy:

Accuracy measure ascertains the extent of accurately anticipated things to the absolute number of expectations made for the whole corpus. It surveys the topic assignments for documents and their terms:

Precision:

Precision refers to amount the precision of a examine method. This research precision p is denoted as the amount of regained related data among the rescued data.

$$precision = \frac{No.of\ retrieved\ relevant\ information}{no.of\ retrieved\ information} \quad (5)$$



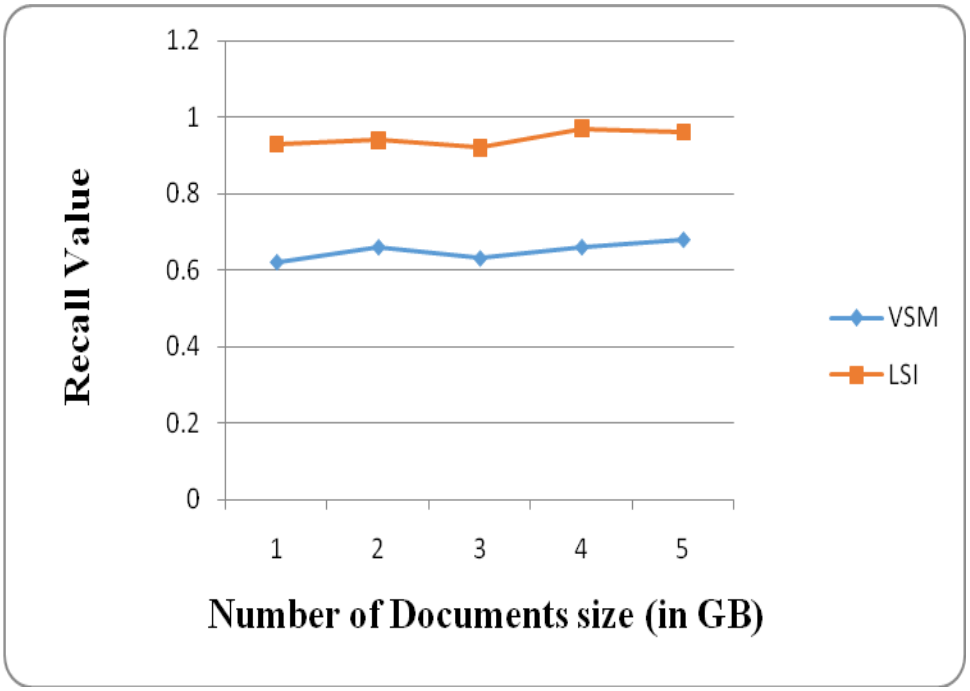
In Fig2. Precision Value

It estimates generally sure expectations that have a place with the positive class among all expectations. It gives the valid positive proportion of the model. The LSI model gotten an accuracy score of 93% though the VSM model got an accuracy score of 67% as found in Fig. 2. This demonstrates that the proposed model outflanked the benchmark model. To find precisions, the information with various sizes beginning up to 10 TB was taken. In Fig. 8 X-axis indicates Archive size, Y-axis indicates accuracy esteem. The chart was attracted correlation with VSM Model. The result of the LSI in the large information model shows preferred execution over the accomplishment of the Result % is 93%.

Recall:

Recall refers to degree the efficiency of a search system. Recall is denoted as the number of retrieved relevant information to total number of relevant information.

$$recall = \frac{no.of\ retrieved\ relevant\ information}{no.of\ relevant\ information} \qquad (6)$$



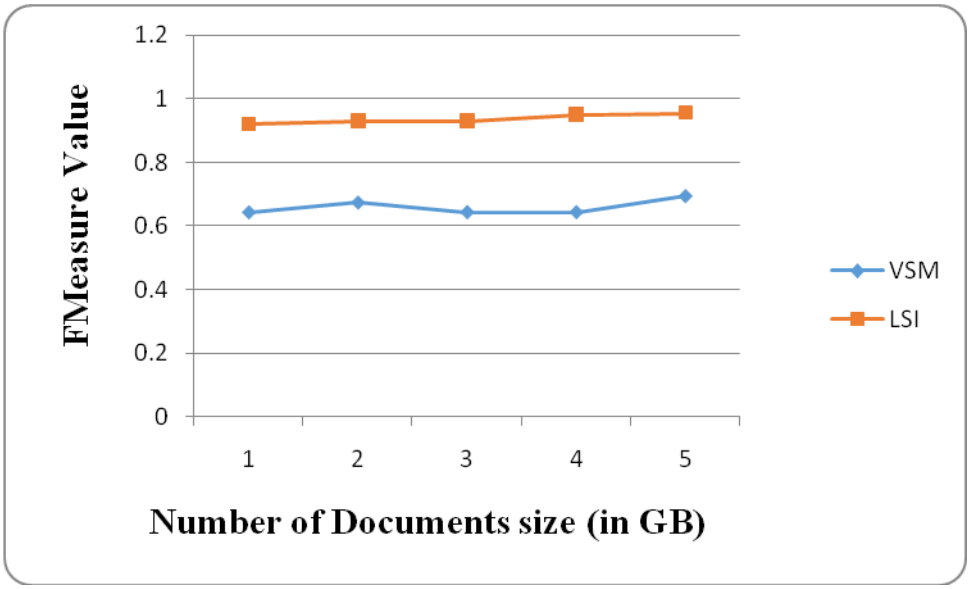
In Fig. 3 Recall Value

The recall is the modest number of recuperated records among each appropriate document. TP (True Positive) addresses positive records, FP (False Positive) addresses false records from the current framework and FN (False Negative) addresses disappointment reports from the proposed framework. In Fig. 3 X-axis means Record size, Y-axis means recall esteem. The VSM model has gotten a Review score is 65%. The proposed LSI model has gotten a Recall score which is 94% the improvement is fruitful.

F-Measure:

F-measure syndicate's precision and recall of a combined performance measure for searchers and users can specify the desired on recall or precision by forming altered hefts. When the F-measure value extents the uppermost, it means the integrated value between precision and recall influences to the maximum at the equal time.

$$Fmeasure = \frac{2 * percision * recall}{(precision + recall)} \qquad (7)$$



In Fig. 4 F-Measure Value

The existing VSM Model has acquired an F-Measure score is 66%. The proposed LSI model has gotten an F-Measure score which is 94%. The improvement is successful. To find the F-Measure esteem, the information with various sizes beginning up to 10 TB was taken. In Fig. 4 X-axis means Report size, Y - axis indicates F-Measure. The accompanying diagram results were contrasted and design disclosure in huge information with VSM Model. The result of the LSI model shows better performance than VSM Model.

TABLE 1: F-MEASURE COMPARISON RESULTS OF PROPOSED WORK

Threshold Value	Proposed Algorithm (LSI)	VSM
>0	70.98	69.38
>0.1	70.97	70.66
>0.2	71.22	73.94
>0.3	71.83	76.77
>0.4	78.18	81.84
>0.5	78.4	87

>0.6	90.65	87.95
>0.7	91.15	79.24
>0.8	90.66	87.83
>0.9	90.69	87.83

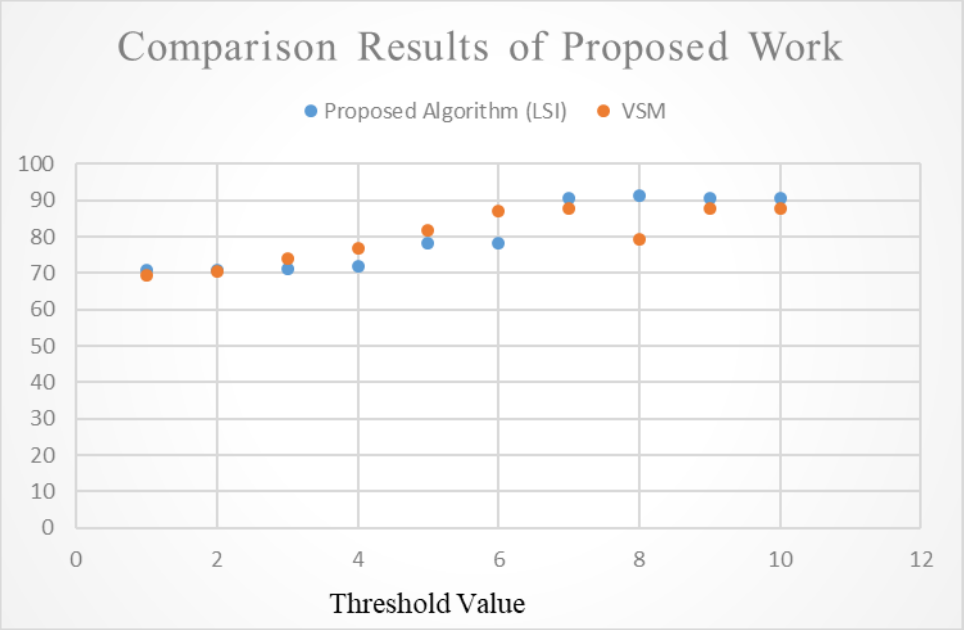


Fig. 5 Testing Results

In this section, we compare the performance of proposed work with VSM model based on the three performance parameters precision, recall, and f-measure. Table 1 and Fig.5 shows the comparison of two models on f-measure. It enhances the performance of LSI algorithm [6-15].

5. CONCLUSION & FUTURE DIRECTIONS

Various data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent itemset mining, sequential pattern mining, maximumpattern mining, and closed pattern mining. However, using these discovered knowledge or else patterns in the fieldof text mining is difficult and ineffective.This paper proposed semantic information and extraction for effective pattern discovery mining in big data using enhanced LSI algorithm. Latent Semantic Indexing is a powerful tool for semantic information extraction and pattern discovery in Big Data.

By leveraging its ability to reduce dimensionality and capture latent relationships [16], LSI aids in making sense of complex datasets. The experimental results show that the proposed model gives highest performance. Future advancements in computational power and hybrid techniques can further enhance its applicability in Big Data mining. Optimizing LSI for distributed Big Data frameworks. Combining LSI with machine learning for enhanced accuracy. Implementing LSI in streaming data environments.

REFERENCES

[1] Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.

[2] Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics, 24(1), 1-7.

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, *3, 993-1022.*

[4] Cai, L., Qi, Y., Wei, W., Wu, J., Li, J., 2019. mrMoulder: A recommendation-based adaptive parameter tuning approach for big data processing platform 93, 570–582.

[5] Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.

[6] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[7] Ding, C., & He, X. (2005). K-means clustering via principal component analysis. Proceedings of the 21st International Conference on Machine Learning, *225-232.*

[8] Hicham, R., &Anis, B. M. (2021). Processes meet Big Data: Scaling process discovery algorithms in Big Data environment. Journal of King Saud University-Computer and Information Sciences.

[9] Kamatchi Sundravadivelu, Mr. S. Muthukumar,Ms. A. Sirin Vifakga, Aditi Chaudhary,Dr. Shabnam Gulati, A Systematic Literature Review On Speech Emotion Recognition Approaches Using Different Methodologies, Library Progress International Vol.44 No. 3, Jul-Dec 2024: P. 11487-11496.

[10] Sundravadivelu, K, Senthilvel, P.G., Duraimutharasan, N., Esther T, H.R., Kumar. K, R.”_Extensive Analysis of IoT Assisted Fake Currency Detection using Novel Learning Scheme”, ICAISS 2023, pp. 1469–1477, 10.1109/ICAISS58487.2023.10250560.

[11] Sundravadivelu, K, Senthilvel, P.G., Thirupurasundari, D.R., Rajesh Kumar, K., Palani, H.K., “Automated Drone-Based Imaging Systems For Plant Health Monitoring Using Deep Learning Techniques”,2023, Intelligent Computing and Control for Engineering and Business Systems, ICCEBS 2023, IEEE , 10.1109/ICCEBS58601.2023.10449190.

[12] Sundravadivelu, K., Thangaraj, M., Gnanambal, S. (2022). An extensive work on comparing sentiment patterns in twitter archives between two persons. International Journal of Health Sciences, 6(S7), 5170-5180. https://doi.org/10.53730/ijhs.v6nS7.13104.

[13] Sundravadivelu. K, M. Thangaraj, “Analyzing Educational Tweets using LDA Model”, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), Volume 10, Issue 4, ISSN:2147-6799, PP. 100- 104, Dec. 2022.

[14] Sundravadivelu. K, M. Thangaraj, ―A Novel Approach for Discovering the Patterns by using PDBD Model in Big Data‖, Journal of Computer Science, (Science Publications), Volume 18 Issues 5, DOI: 10.3844 / jcssp.2022.382.395, ISSN: 1549-3636, pp.382-395, May 2022.

[15] Sundravadivelu. K, Suraj Rajesh Karpe, Harish V Mekali, Shital Nalgirkar, K. Abdul Rasak, Dr. V S Narayana Tinnalur, Information Theory and Coding: Techniques for Error Control and Data Compression, Journal of Electrical Systems 20-10s (2024): 5665-5674.

[16] Sundravadivelu. K, Thangaraj. M,“Mining effective patterns from text data - a survey”, International Journal of Scientific and Technology Research, Volume 9, Issue 1, Pages 1930 – 1934, January 2020.

[17] Thangaraj. M., & Sundaravadivelu, K.,‖ Mining effective patterns from text data-a survey‖ International Journal of Scientific & Technology Research. ISSN-10: 2277-86161930 IJSTR, 2020.

[18] Wu, J., Guo, S., Huang, H., Liu, W., Xiang, Y., 2018. Information and Communications technologies for sustainable development goals: state-of-the-art, needs and perspectives. IEEE Communications Surveys Tutorials 20, 2389–2406.

[19] Zhong, N., Li, Y., & Wu, S. T. (2010). Effective pattern discovery for text mining. IEEE transactions on knowledge and data engineering, 24(1), 30-44.