*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
***Article Received**: 28 January 2023 **Revised**: 01 March 2023 **Accepted**: 12 March 2023*

# Enhancing Feature Extraction through G-PLSGLR by Decreasing Dimensionality of Textual Data

**Narender Chinthamu[1], Chandrasekar Venkatachalam[2], Muthuvairavan Pillai.N[3], Dr. Setti Vidya Sagar Appaji[4], M. Murali[5]**

[1]Enterprise Architect,
MIT CTO Candidate
Email: Narender.chinthamu@gmail.com
[2]Professor, Department of CSE,
Faculty of Engineering and Technology,
Jain (Deemed-to-be) University,
Bangalore, Karnataka
Email:drchandru86@gmail.com
[3]Assistant Professor,
Department of Computer Science and Business Systems,
R.M.D Engineering College,
Kavarapettai
Email: muthuvikni@gmail.com
[4]Associate Professor,
Department of Computer Science and Engineering,
Baba Institute of Technology and Sciences,
Visakhapatnam, Andhra Pradesh,
Email: sagarsetti4u@gmail.com
[5]Assistant Professor,
Department of iT,
Sona College of Technology,
muralimuthu@sonatech.ac.in

**Abstract**— The technology of big data has become highly popular in numerous industries owing to its various characteristics such as high value, large volume, rapid velocity, wide variety, and significant variability. Nevertheless, big data presents several difficulties that must be addressed, including lengthy processing times, high computational complexity, imprecise features, significant sparsity, irrelevant terms, redundancy, and noise, all of which can have an adverse effect on the performance of feature extraction. The objective of this research is to tackle these issues by utilizing the Partial Least Square Generalized Linear Regression (G-PLSGLR) approach to decrease the high dimensionality of text data. The suggested algorithm is made up of four stages: Firstly, gathering featured data in vector space model (VSM) and training it with bootstrap technique. Second, grouping trained feature samples using a Pearson correlation coefficient and graph-based technique. Third, getting rid of unimportant features by ranking significant group features using PLSGR. Lastly, choosing or extracting significant features using Bayesian information criterion (BIC). The G-PLSGLR algorithm surpasses current methods by achieving a high reduction rate and classification performance, while minimizing feature redundancy, time consumption, and complexity. Furthermore, it enhances the accuracy of features by 35%.

**Keywords**- Big Data, dimensionality reduction, vector space model, Bayesian information criterion

## I. INTRODUCTION

Data is growing at an exponential rate in various fields, such as biomedicine, IoT, and social media, reaching terabyte or petabyte levels. Despite the fact that the concept of big data and its importance have been around for a long time, recent technological advancements have enabled us to analyze vast data sets quickly and efficiently. As data continues to expand in both structured and unstructured forms, it will be collected and analyzed to discover unforeseen insights and even predict the future in the years ahead. Even small companies that collect and interpret data can benefit from this, as new and cost-effective technologies constantly emerge to make big data solutions more accessible.

The rapidly increasing size and complexity of data have surpassed the capabilities of existing computing infrastructure

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
*Article Received: 28 January 2023 Revised: 01 March 2023 Accepted: 12 March 2023*

and analysis algorithms to handle and process it effectively. Therefore, new approaches are needed to address the challenges posed by big data. Dimensionality reduction in textual big data is crucial in avoiding inconsistency, redundancy, and confusion. The essential need is to extract the significant variables that effectively capture the unique properties of varied data to prevent the negative effects of high dimensionality. The extraction of such variables not only leads to a better comprehension of the real world but also enables efficient analysis techniques to offer high-quality services to users.

In the fields of machine learning and statistics, dimensionality reduction refers to a set of methods employed to reduce the number of variables that need to be analyzed. This involves techniques such as feature extraction and feature selection. This process helps simplify and expedite the analysis of data for machine learning algorithms by eliminating extraneous variables.

Feature selection is the process of selecting a subset of the original set of variables or features to create a smaller subset for modeling purposes. This is typically accomplished using one of three methods: Filter, Wrapper, or Embedded.

The process of feature extraction is utilized to extract organized information from unstructured data in order to decrease the dimensionality of data from a high-dimensional space to a lower dimensional space. This structured information includes entities, relations, objects, and events. Supervised and unsupervised approaches can be used to extract features, with supervised approaches employing both core words and topic-related content to train a machine.

Text clustering is an important text mining method used to group vast amounts of text documents into smaller, meaningful clusters. This process has applications in sentiment analysis, text classification, text summarization, event tracking, and topic detection, among others. Webpages, microblogs, and social networks are excellent sources of information for text clustering, as they provide a wealth of valuable information for readers.

Text clustering can be broken down into two main stages: pre-processing and clustering. The initial stage involves basic steps such as tokenization, stop-word removal, and word stemming, which are used to break down sentences into words and eliminate unnecessary words or terms. The primary goal of the pre-processing stage is to convert the unstructured text documents into a structured format that can be efficiently processed by clustering approaches. The second stage is the clustering of the pre-processed text documents.

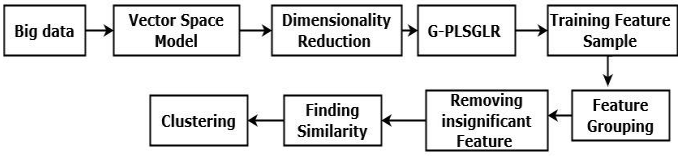



Figure 1: Proposed work's block diagram.

The complexity of dimensionality reduction poses challenges in dealing with large volumes of textual data due to factors such as high dimensionality, irrelevant words, redundancy, and sparsity, which reduce the accuracy of similarity detection in core words. As a result, there is a need to improve the quality and efficiency of feature extraction to handle unstructured data. To address this issue, this study focuses on optimizing the high dimensionality of textual data using the Partial Least Square Generalized Linear Regression (G-PLSGLR) technique. The proposed algorithm consists of four steps: (i) Collecting featured data in Vector Space Model (VSM) and training it using the bootstrap technique, (ii) Grouping trained feature samples based on Pearson correlation coefficient and graph-based technique, (iii) Eliminating unimportant features by ranking them based on PLSGR, and (iv) Selecting or extracting important features using Bayesian information criterion (BIC).

The paper makes several significant contributions:

1. The data sampling process is used to improve accuracy and reduce execution time when structuring core words.
2. Trained feature samples are grouped using the Pearson correlation coefficient and graph-based technique, which eliminates irrelevant words and reduces feature redundancy, improving overall performance.
3. G-PLSGLR, a partial least square generalized regression, is introduced to enhance accuracy, increase maximum reduction rate and classification performance, and decrease feature redundancy in feature extraction.

The study is structured with a detailed introduction on big data analytics in document clustering in section 1. Section 2 provides a literature survey, while section 3 describes the overall methodology of the proposed technique. The competency of the suggested work is demonstrated in section 4, and section 5 concludes the paper.

## II. LITERATURE SURVEY

Wang et al. [1] proposed a deep learning approach for accurate clustering through representation learning, transfer learning domain adaptation, and parameter updates. However, the paper did not provide a detailed explanation of the method for clustering large amounts of text, making it inefficient.

Brockmeier et al. [2] suggested a self-tuned descriptive clustering technique that automatically selects the number of

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
***Article Received**: 28 January 2023 **Revised**: 01 March 2023 **Accepted**: 12 March 2023*

clusters and features for each cluster. This method is appropriate for short text documents, and it selects clusters based on the topic organization of the associated feature subsets.

In their research, Zhag and Mao [3] presented a fuzzy bag-of-words model that can establish semantic correlations between words and measure the similarity between cluster words and other words using three parameters (FBoWCmean, FBoWCmini, and FBoWCmax). Although this approach is useful in reducing feature redundancy and discrimination, it does not consider the weighting term of individual documents.

According to the work of Ravindran and Thanamani [4], a K-Means document clustering technique was suggested using vector space representation (VSM) and cosine similarity. Although this method is efficient for clustering high-dimensional data, it does not perform similarity calculations between document sets during clustering.

The authors Sumathy and Chidambaram [5] proposed a clustering model that utilized idiom processing and semantic weight to determine the meaning of documents. However, their model did not show improvements in cluster accuracy or time complexity.

Xu et al. [6] proposed an algorithm for text sensitive information classification and topic tracking. The method uses a vector space model and cosine theorem for text classification and detects sensitive topics on webpages.

The researchers Liu et al. [7] suggested a combination of semantic and lexical feature extraction methods to classify questions. They utilized information gain and sequential pattern mining techniques to extract features, which were then fed into classifiers. However, this approach is restricted to question classification and cannot be used for other tasks.

Wang [8] introduced an unsupervised representative feature selection algorithm (REPFS) that filters out irrelevant features and reduces the number of meaningless features in subsequent analysis by selecting a representative feature from each redundant feature cluster.

Liwei Kuang and colleagues [20] proposed a comprehensive approach that fuses unstructured, semi-structured, and structured data into a unified model using a chunk tensor method. This method appropriately arranges all characteristics of heterogeneous data along tensor orders, and a Lanczos-based high-order singular value decomposition approach is introduced to reduce the unified model's dimensionality. The algorithm's theoretical analyses are contributed in terms of storage scheme, convergence property, and computation cost.

Cao and colleagues [9] presented a semantic feature approach based on a deep neural network (DNN) and a feature generation model for pictures and texts. They converted the retrieval of text and image data into a vector similarity calculation, improving the retrieval speed and semantic relevance of the result. However, a limitation of this work is that previously untrained or unlabelled images do not provide useful information, and the conversion of image content into text content is not always accurate.

Kiran Adnan and Rehan Akbar [10] focused on data extraction of unstructured multidimensional big data. Their work improved information extraction techniques related to data processing, data extraction, transformation, representation for a vast amount of multidimensional unstructured data, and efficiency.

## III. METHODOLOGY

*G-PLSGLR for Big Data*

The primary goal of this research is to improve the accuracy and performance of clustering in big data by utilizing the G-PLSGLR technique to simplify the identification of similar features. The suggested approach aims to reduce feature redundancy and improve the reduction rate and classification performance compared to existing methods for big data analysis.

*Text Document Preprocessing*

When dealing with a large amount of text data, the initial step is to preprocess it through parsing. This involves transforming the documents into a data model that can be effectively analyzed by a ML approach. The process of parsing involves breaking down the text into individual words, reducing them to their root form, and removing common words that do not carry significant meaning. Tokenization involves converting the document content into a sequence of terms that define it, while stemming aims to reduce the number of unique terms by converting them to their root forms. Finally, stop word removal eliminates commonly used functional words such as "the," "there," "was," and "an."

*Implementation of Vector Space Model*

After the preprocessing step, certain commonly used words will be ignored, leaving behind only the feature terms that will be used to create a document set. These feature terms are then represented in the vector space model (VSM), which is designed to have lower dimensions than existing methods. Algorithms based on the VSM approach [5] use a term-document matrix to represent text documents. By converting unstructured text into a structured VSM format, it becomes easier to cluster the data. Similarity measures and clustering algorithms can then be applied to the VSM. Assuming we have n documents ($dn$) and feature terms ($fi$) depicts in a feature document matrix, the rows of the matrix depict the feature terms or words, and the columns depict the documents. Because we are dealing with a large number of documents, the matrix is high-dimensional and needs to be reduced in dimensionality to process the data effectively.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
***Article Received**: 28 January 2023 **Revised**: 01 March 2023 **Accepted**: 12 March 2023*

*Dimensionality Reduction*

Clustering algorithms face significant performance issues due to high dimensionality and data sparsity. Extraneous features in documents can decrease the accuracy of data retrieval, making it difficult for clustering algorithms to effectively group similar documents. As a result, dimensionality reduction is crucial for text or document clustering. Dimensionality reduction [7] can be achieved through feature extraction or feature reduction. Semantic feature extraction is used for feature extraction, which involves identifying and maintaining features that are more relevant to the document. The G-PLSGLR (partial least square generalized linear regression) technique is proposed to identify relevant feature terms. This method consists of four main stages:

1. Training pertinent features
2. Grouping the features together / Feature grouping
3. Eliminating extraneous features
4. Selecting comparable features.

*Training Pertinent features*

During this step, the relevant features are gathered from the VSM model and reduced to a lower dimensionality. Then, they are trained with the bootstrap method, which analyzes the features precisely and generates a subcategory. However, high intra-class variability and limited samples can cause dimensionality issues, so the samples are labeled with both category and subcategory to incorporate human expertise. A balanced training sample is created through a bootstrap sampling strategy to counteract class imbalance. Next, for each category, similar features are extracted from the remaining samples using the current category samples.

*Feature grouping*Current feature selection algorithms mainly focus on structural features, which may lead to a loss of similarity among features, complexity for large datasets, and low accuracy. To overcome these challenges, this paper introduces a new method for feature grouping based on Pearson correlation coefficient and graph-based theory. Pearson correlation coefficient helps to measure linear dependence between two random variables and similarity between topics, as well as to identify and remove redundant features. The correlation can be represented by the following formula:

$$P_{RD} = \frac{\sum_{X=1}^{i}(R_X-\bar{r})(D_X-\bar{d})}{\sqrt{\sum_{X=1}^{i}(R_X-\bar{r})^2\sum_{X=1}^{i}(D_X-\bar{d})^2}} \quad (1)$$

The process of deriving feature groups can be illustrated as shown below, where $Rx$ and $Dx$ represent the measurements of features R and D respectively, i represents the number of measurements, r and d are their respective mean values, and r denotes the correlation between features R and D.

$$F_g = \begin{cases} 1, P_{RD} \geq th \\ 0, P_{RD} < th \end{cases} \quad (2)$$

To analyze the textual data, the correlation $P_{RD}$ between each pair of features (R and D) was computed. An edge E between two features is assigned a value of 1 if the absolute value of their correlation is greater than a predetermined threshold (th); otherwise, the edge value is set to 0. By traversing the features, a graph is constructed where the vertices V represent features and the edges indicate whether the correlation between features exceeds the threshold. Connected parts of the graph are extracted as feature groups (G).

*Eliminating Extraneous Features*

In the next section, irrelevant features are eliminated based on their topic and ranking. For this purpose, the PLSGLR regression coefficient is used as an indicator, following the PLSRGLM method [20]. Unlike similar features, the PLSGLR method selects latent components and considers the response variable in regression. The PLSGLR model with H components is expressed as follows:

$$N(Z=1) = \frac{e^{(D_0+\sum_{q=1}^{Q} D_Q I_Q)}}{1+e^{(D_0+\sum_{q=1}^{Q} D_Q I_Q)}} \quad (3)$$

The logistic regression for response variable Z includes a partial coefficient, represented as $D_Q$, for component $I_Q$, and an intercept represented by $D_0$. Component $I_Q$ comprises N features, and it can be represented as follows:

$$I_Q = \sum_{w=1}^{N} Y_{Wq} * L_W \quad (4)$$

$$N(Z=1) = \frac{e^{(\beta_0+\sum_{i=1}^{N}\beta_i L_i)}}{1+e^{(\beta_0+\sum_{i=1}^{N}\beta_i L_i)}} \quad (5)$$

The method used to select important features is based on BIC, where $\beta_i$ indicates the coefficient of feature $L_i$. The features are sorted pursuant to to their coefficients, and the intercept $\beta$ is removed for subsequent analyses. The dataset is two-class and includes N features, with a response variable that takes a value of 1 for the same category and 0 for the other category. PLSGLR technique is used to extract Q principal components ($I_Q$), and each feature has a coefficient c. The importance of each feature is calculated by multiplying its coefficient with that of the corresponding component. The resulting coefficient β represents the feature's importance, which is used to rank the features based on the absolute value of β. To estimate statistically significant variables and prevent excessive repeated features in textual data, a non-parametric bootstrap procedure is used. The significance test eliminates irrelevant features.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
***Article Received**: 28 January 2023 **Revised**: 01 March 2023 **Accepted**: 12 March 2023*

*Selecting Comparable Features*

In terms of similarity features, BIC is useful for selecting features that are representative of different categories within a specific model. This method is often applied to feature selection problems, which involve selecting the most relevant features for a given dataset as.

$$BIC = -2 \ln g + K \ln n \quad (6)$$

In the context of the PLSGLR method, the variable g denotes the probability, whereas n represents the total number of features present in the sample, and k indicates the number of features that have been added to the PLSGLR model.

Firstly, a set of ranked features is obtained for a specific category, consisting of k features, through a significance test. Then, the Bayesian information criterion (BIC)-based Partial Least Squares Generalized Linear Regression (PLSGLR) is calculated based on the selected features and n samples. Other techniques like random forests or support vector machines can also be utilized to derive BIC values for the selected features. However, this study does not cover such techniques. The features are then incorporated into the PLSGLR model in a ranked order, one by one, and the BIC is computed every time a new feature is added. The lower the BIC score, the more optimal and accurate the newly added feature(s) are in terms of their dimensionality. To reduce the high dimensionality of big data, only features that meet certain criteria are selected. Specifically, features associated with the minimum Bayesian information criterion (BIC) or those with a BIC variation of less than 3 are chosen.

*Experimental setup*

This proposed work is mainly focused on reducing the high dimensionality of big data built by the Hadoop cluster. It improves accuracy and reduction rate in collected data samples. The hardware and software environment is represented as below:

*Hardware and Software Environment Requirements*

*Hardware specifications:*

- Memory: 8G
- Hard disk: 800G
- Processor: Intel Pentium 1.5GHz

*Software specifications*

- Operating System: Windows 8.0 or later versions
- Hadoop version: V 1.0.1
- MAHOUT version: Mahout 0.7

*The Data Collection*

In practical scenarios, people can obtain a considerable amount of text data without any specific boundaries. However, there may be instances where it is necessary to collect data related to a particular topic during a specific period. For example, if reviews are required for a particular product of a brand, the available data for that category might be limited. Therefore, it is essential to test the proposed method on datasets that are of general scale. In this study, the consumer reviews dataset of Amazon products has been used for classification and dimensionality reduction. This dataset, obtained from the source https://data.world/datafiniti/consumer-reviews-of-amazon-products, comprises various attributes, such as id, name, brand, primary category, review date, added review title, review user name, and so on. There are 5000 text data in each category, with 1544776 words in the review section, 17467 words in the review title section, and 5002 words in the review user name section.

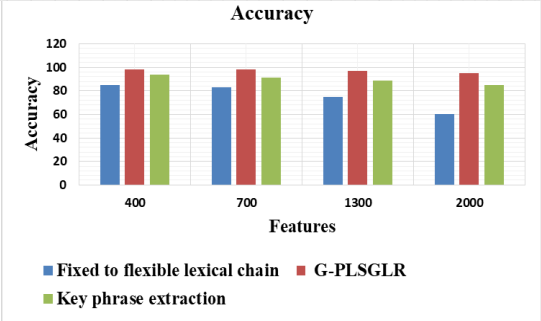
*Data Preprocessing*

Parsing tasks involving tokenization, stemming, and stop word removal are made in preprocessing to remove the basic words. In each category, nearly 5000 data are analyzed to remove the basic words, punctuations, articles, and prepositions. The Vector Space Model is utilized to enhance the data's quality by processing the remaining words.

G-PLSGLR extract and train the features so that similar features are grouped, and insignificant features are removed from the dataset. Appropriate keywords that describe the product rating are collected, and these words are taken as features. These features are trained to get an accurate result and predict the irrelevant features.

The customer review dataset's accuracy, feature redundancy, and redundancy ratio are depicted in Figures 2(a), 2(b), and 2(c). The identification of precise features in big data presents a significant challenge for the Flexible to Fixed Lexical Chain and Keyphrase Extraction techniques, as they exhibit high complexity. These methods also demonstrate inefficiency in determining feature similarity in lengthy documents. To tackle this issue, the Rank-based Ordering technique is implemented to prioritize the most relevant features. The newly proposed G-PLSGLR algorithm outperforms the existing algorithms in identifying feature similarities.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
***Article Received**: 28 January 2023 **Revised**: 01 March 2023 **Accepted**: 12 March 2023*



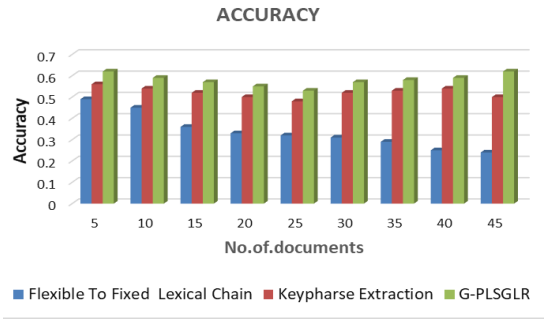




Figure 2: Accuracy

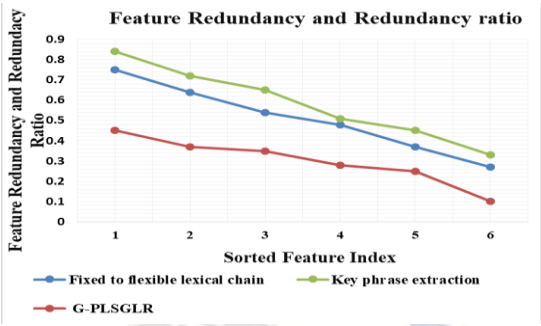



Figure 2(a): Newsdataset's accuracy standard Figure 2(b): Feature Reduction

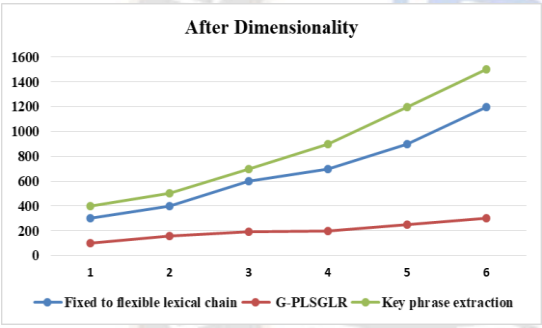



Figure 2(c): Dimensionality of news dataset reduction

## IV. Result and Discussion

### *1. Accuracy*

Table 1 compares the accuracy of the proposed method with the Flexible to Fixed Lexical Chain and Key Phrase Extraction technique, while Figure 1 depicts the accuracy levels.

| No of Documents | ACCURACY | | |
|---|---|---|---|
| | G-PLSGLR | Extraction of important phrases. | Conversion of flexible lexical chains to fixed lexical chains |
| 5 | 0.56 | 0.62 | 0.49 |
| 10 | 0.54 | 0.59 | 0.45 |
| 15 | 0.52 | 0.57 | 0.36 |
| 20 | 0.50 | 0.55 | 0.33 |
| 25 | 0.48 | 0.53 | 0.32 |
| 30 | 0.52 | 0.57 | 0.31 |
| 35 | 0.53 | 0.58 | 0.29 |
| 40 | 0.54 | 0.59 | 0.25 |
| 45 | 0.50 | 0.62 | 0.24 |

The current Flexible to Fixed Lexical Chain algorithm has some limitations. It has two steps, namely Flexible to Fixed Lexical Chain and Fixed Lexical Chain, which are used to generate a parameter for the Fixed Lexical Chain, making the algorithm complex and time-consuming when converting unstructured data to structured data.

On the other hand, the suggested G-PLSGLR approach has four steps and aims to improve the accuracy level by utilizing the Pearson correlation coefficient and graph-based theory techniques. These techniques group features and eliminate irrelevant ones, leading to reduced feature redundancy and improved accuracy compared to the existing algorithm.

### *2.Execution time*

It refers to the duration during which a system carries out a particular task or process, including the time spent running any system or run-time services related to the task. To compare the performance of G-PLSGLR with flexible to fixed lexical chain and important phrase extraction protocols, Table 2 presents their execution time. The graph in Figure 3 depicts the execution time of each algorithm.

| High dimensional data | Execution time (seconds) | | |
|---|---|---|---|
| | conversion of flexible lexical chains to fixed lexical chains | G - PLSGLR | Keyphrase extraction technique |
| 5 | 0.006 | 0.003 | 0.009 |
| 15 | 0.004 | 0.003 | 0.013 |
| 25 | 0.008 | 0.004 | 0.019 |
| 35 | 0.013 | 0.006 | 0.022 |
| 45 | 0.026 | 0.009 | 0.028 |
| 55 | 0.036 | 0.014 | 0.037 |
| 65 | 0.041 | 0.021 | 0.042 |

The term "execution time" pertains to the period that a system needs to finish a task or procedure, encompassing both runtime and system service times. Figure 3 illustrates the execution time for the dataset with high dimensionality. The suggested approach applies Pearson correlation coefficient and graph-based techniques to evaluate and organize the features based on their relevance, leading to an extracted feature set that

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
***Article Received**: 28 January 2023 **Revised**: 01 March 2023 **Accepted**: 12 March 2023*

can be accomplished in less time. This approach retrieves only the necessary features at each step and removes extraneous words. The training process is enhanced by utilizing the bootstrap technique, making it easier to predict the correct features. This approach reduces feature redundancy, resulting in less execution time than the existing methods.

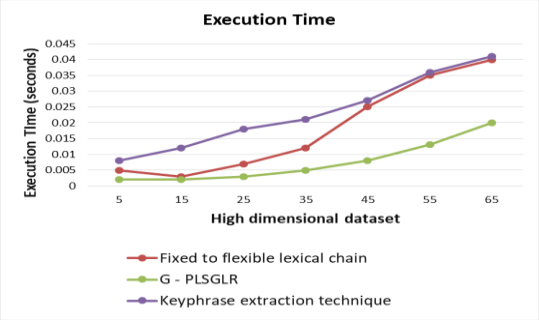



Figure 3: Execution Time

### *3. Feature extraction in big data*

In big data, feature extraction involves reducing the dimensionality of raw data to make it easier to process. Since large datasets often have many variables, they can be computationally expensive to work with. Figure 4 displays a comparison of feature extraction in G-PLSGLR, flexible to fixed lexical chain, and keyphrase extraction. The outcomes are also presented in Table 3.

| Technique | Percentage |
|---|---|
| G- PLSGLR | 35% |
| Key phrase extraction | 20% |
| Flexible lexical chain | 15% |

The objective of this research is to tackle the problems encountered by the current approach for extracting fixed lexical chains and keyphrases from flexible text. This approach faces difficulty in precisely detecting characteristics in extensive datasets and determining the similarity between features in lengthy documents.

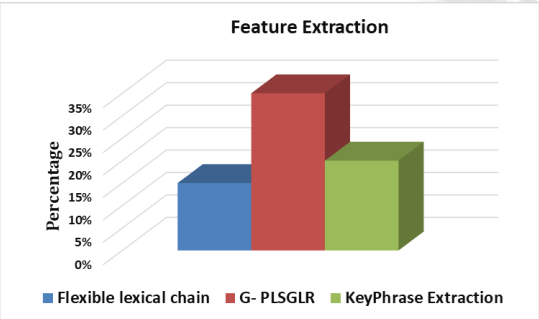



Figure 4: Feature Extraction

The suggested G-PLSGLR algorithm has a high reduction rate and utilizes feature-based PLSGLR to extract more important features. This approach results in low feature redundancy, and the Pearson correlation technique is used to group features while the rank-based order technique is employed to prioritize highly relevant features. The algorithm also improves the identification of similarities between features, ultimately achieving a 35% higher performance than the current approach.

## V. CONCLUSION

This study aims to enhance feature extraction performance by addressing the challenges caused by high-dimensionality, which include issues such as computational complexity, inaccurate features, high sparsity, extraneous words, redundancy, and noise. To address these issues, the proposed G-PLSGLR approach employs four key processes: (i) Utilizing the vector space model (VSM) to collect feature data and training it with the bootstrap technique, (ii) Grouping the trained feature samples through the Pearson correlation coefficient and graph-based method, (iii) Ranking importance group features based on PLSGR to eliminate insignificant features, and (iv) Selecting or extracting essential features using Bayesian information criterion (BIC).

The suggested approach aims to enhance the accuracy, proficiency, and classification performance of feature extraction. Compared to existing methods, it reduces the reduction rate, complexity, time-consuming process, and feature redundancy. The rank-based order technique is employed to prioritize highly pertinent features, and the approach improves the identification of similarities between features, ultimately achieving a 35% higher feature extraction performance than existing approaches.

## REFERENCES

[1] Binyu Wang, Wenfen Liu, Zijie Lin, Xuexian Hu, Jianghong Wei and Chun Liu, ""A Text Clustering Algorithm Based On Deep Representation Learning"", The Journal of engineering, 2018.

[2] Austin J. Brockmeier, Tingting Mu,Sophia Ananiadou, and John Y. Goulermas, ""Self-TunedDescriptive Document Clustering using a Predictive Network"",IEEE Transactions on Knowledge and Data Engineering, 2017.

[3] Yuanping Zhu, Kuang Zhang, ""Text segmentation using super pixel clustering"", IET journals Image Process., 2017.

[4] R. Malathi Ravindran and Dr. Antony Selvadoss Thanamani ""K-Means Document Clusteringusing Vector Space Model"" Bonfring International Journal of Data Mining, Vol. 5, No. 2, July 2015.

[5] Ms. K.L.Sumathy and Dr. M. Chidambaram ""Semantic Based Vector Space Model to Improve the Clustering Accuracy in Knowledge Repositories"", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 14, No. 9, September 2016.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 4s*
*DOI: https://doi.org/10.17762/ijritcc.v11i4s.6540*
***Article Received**: 28 January 2023 **Revised**: 01 March 2023 **Accepted**: 12 March 2023*

[6] Guixian xu, ziheng yu and qi, ""EfficientSensitive Information Classification and Topic Tracking Based on Tibetan Web Pages"", IEEE.Translations and content mining, VOLUME 6, 2018.

[7] Yaqing Liu, Xiaokai Yi, Rong Chen, Zhengguo Zhaianf and Jingxuan Gu, ""Feature extraction based on information gain and sequential pattern for English question classification"", IET journals 2018.

[8] Yintong Wang, ""Unsupervisedrepresentative feature selection algorithm based on information entropy and relevance analysis"", IEEE Translations and content mining 2017.

[9] Jiarun Cao, Chongwen Wang, Liming Gao, ""A Joint Model for Text and Image Semantic Feature Extraction"", ACAI 2018.

[10] R. Krishnana, V.A. Samaranayake & S. Jagannathan, ""A Multi-step Nonlinear Dimension-reduction Approach with Applications toBig data"",

[11] Kiran Adnan and Rehan Akbar 2019, ""An analytical study of information extraction from unstructured and multidimensional big data"", Journal of Big Data, Springer Open.

[12] Sudha Ramkumar and Dr. B. Poorna, ""Text Document Clustering Using Dimension Reduction Technique"" International Journal of Applied Engineering Research2016.

[13] Ayush Aggarwa, Chhavi Sharma, Minni Jain and Amita Jain, ""Semi Supervised Graph Based Keyword Extraction Using Lexical Chains and Centrality Measures"", ISSN 2018.

[14] Prateek Chanda and Asit Kumar Das, ""A Novel Graph Based Clustering Approach to Document Topic Modeling"" IEEE 2018.

[15] Terry Ruas and William Grosky, ""Semantic Feature Structure Extraction from Documents Based on Extended Lexical Chains"".

[16] Shabanaafreen and dr.b.Srinivasu, ""Semantic Based Document Clustering Using Lexical Chains"" International Research Journal of Engineering and Technology (IRJET) 2017.

[17] Francis Musembi Kwale ""An Overview of VSM-Based Text Clustering Approaches"",International Journal of Advanced Research in Computer Science, 2014.

[18] Halima ELAIDI, Younes ELHADDAR, Zahra BENABBOU and Hassan ABBAR, ""An idea of a clustering algorithm using support vector machines based on binary decision tree"",IEEE Translations and content mining 2018.

[19] Yaohuan Huang , Chuanpeng Zhao, Haijun Yang , Xiaoyang Song and Zhonghua Li, Jie Chen 2019, ""Feature Selection Solution with High Dimensionality and Low-Sample Size for Land CoverClassificationin Object-Based Image Analysis"", Remote Sensing.

[20] Liwei Kuang, Laurence T. Yang, Jinjun Chen, Fei Hao, Changqing LuoView, ""A Holistic Approach for Distributed Dimensionality Reduction of Big Data"", IEEE Transactions on Cloud Computing, 2018.