Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SCIBERT- DRIVEN GNN MODELS FOR DETECTING COMMUNITIES FROM SCOPUS KEYWORD CO-OCURRENCE NETWORK

KIRUTHIKA. R.*, KRISHNAVENI SAKKARAPANI

*Ph.D. Research Scholar, Department of Computer Science,

PSGR Krishnammal College for Women, Coimbatore, India

Assistant Professor, Department of Data Analytics (PG),

PSGR Krishnammal College for Women, Coimbatore, India

E-mail: *kirthikamole@gmail.com, krishnavenis@psgrkcw.ac.in

ABSTRACT

Identifying the research communities from the large-scale bibliometric data is essential for understanding and interpreting the trends among recent research topics. This article presents a novel framework for identifying communities from the Scopus bibliometric data by constructing a keyword co-occurrence network. The deep learning based scientific articles extracted from the Scopus bibliographic database were selected for this work. These collected data are segmented into five different time frames named as Scopus Bibliographic Dataset (SBD), SBD_1 (2006-2013), SBD_2 (2014-2016), SBD_3 (2017), SBD_4 (2018) and SBD_5 (2019). This work is developed as a hybrid framework by integrating the traditional Louvain algorithm with various Graph Neural Network (GNN) models, which helps to improve the performance through the extraction of the best textual information features from the keywords by applying the SciBERT model as node features. GNN techniques like GCN, GraphSAGE, FeaStConv, APPNP, WLConvContinuous and AGNN were implemented to compare among the models to find the best model for developing this novel framework. These integrated frameworks are named as SciBLoGCN, SciBLoGS, SciBLoFSC, SciBLoWLC, SciBLoAPPNP, and SciBLoAGNN. The experimental findings determine the presence of meaningful and effective research communities based on the structure and interconnection between the nodes within the network. This work supports scholars to understand the interconnection among their domains based on recent research topics, which enables strategic prioritization of their research areas.

Keywords: *Community Detection, Scopus Bibliometric Data, Keyword co-occurrence Networks, SciBERT, Complex Network.*

1. INTRODUCTION

Communities are referred to as groups of nodes that share common properties between the nodes [1–5]. Community detection is one of the significant tasks in network analysis that detects the meaningful groups of interconnected nodes that are linked to each other within the network structure [6–12]. Community detection has extensive and broad applications in many areas such as social network analysis, information theory, biological networks, recommendation systems and fraud detection [13–18]. There are different traditional algorithms, including Louvain, modularity maximization, infomap, label propagation, spinglass, leiden, and spectral clustering, that are applied for identifying the communities from the network [19–23]. To effectively capture the complex relationships and structural patterns within the network, Graph Neural Networks methods are utilized [24–27].

Graph Neural Networks are neural networks that can be applied to a graph data structure. GNNs are applied in the prediction of nodes, edges, and graph-based tasks [28,29]. One of the most important and prominent types of Graph Neural Network tasks is community detection, an approach that models and reflects the inherent structural patterns and relationship structures of complex networks [30,31]. Graph Neural Network methods are integrated to enhance the community detection for complex networks [32]. This method also offers useful information on the structure, functionality and

dynamics of interconnected systems by revealing the underlying community structure of complex networks. This allows the researchers to gain a more insightful understanding of network properties and apply them to diverse networks for analytical and predictive purposes. Some of the most popular GNN methods are Graph Convolutional Networks (GCN), Message Passing Neural Networks (MPNN), Graph Attention Networks (GAT), Graph Auto-Encoder Networks and GraphSage [33–35].

Community detection in a Scopus-driven keyword co-occurrence network is one of the more challenging processes for understanding bibliometric network data analysis [36,37]. These are used to map the intellectual structure of a research field by identifying its main thematic clusters [38]. Significant gaps remained in the field of community detection using keyword co-occurrence networks in bibliometric research [39]. There are research gaps in the application of the Louvain and GNN model to real-time large networks of co-occurring keywords [40–42]. SciBERT as a node feature extractor in bibliometric networks is understudied [43,44]. There are fewer studies on how to make these algorithms precise and scalable to large datasets [45–47]. The idea of integrating GNNs with transformer models in bibliometric tasks for detecting communities in Python has not been explored in detail [48–50]. There are limited works addressing the co-integration of new GNN models with traditional community detection algorithms on bibliometric data with evaluation metrics [51–53]. These gaps indicate that a new framework is necessary to identify communities based on the Scopus bibliometric data, with Louvain and GNN models having nodes in the form of SciBERT-based node features [54].

The main goal of this work is to identify the communities based on Scopus bibliometric data via keyword co-occurrence networks using the Louvain algorithm and Graph Neural Network (GNN) models with SciBERT node features as a novel framework. The keyword co-occurrence network is constructed by using the Index keywords field obtained from Scopus, which are represented as nodes in a network and if two keywords appear together in the articles form an edge. and the degree of this relationship is the edge weight [55]. An increasing edge weight indicates a stronger conceptual relationship between the keywords. The Louvain algorithm is one of the efficient algorithms for operating large networks that can be applied within this network to determine the community structures [56,57]. This divides the network into communities based on the strong interconnection between keywords, which means that keywords in one group are much more likely to co-occur than the keywords in other groups. The resulting clusters represent each a different research theme or sub-field, which is informative of the structure of the research domain. In order to better detect the community, GNN models are applied, with node features obtained with the help of SciBERT, a model designed with the specific goal of detecting science-related text. Combining the textual information with the network structure allows developing more profound insight into the semantic connections between keywords, which will create more accurate and meaning-based detection of communities. This framework provides a new direction for comprehending research trends and patterns of collaboration in the field of computer science. Such an approach integrates the reliability of the traditional network analysis methods with the sophisticated natural language processing and deep learning algorithms. The method identifies the connections among the research topics and research areas through the construction of a keyword co-occurrence network based on Scopus data.

The key contribution of this proposed framework is the integrated graph models to detect communities of research based on keywords and promotes the knowledge of scientific areas progressing. Scholars get a better understanding of the interrelations among disciplines. Bringing semantic models such as SciBERT to GNNs enhances the process of studying the relationships across disciplines, allowing interdisciplinary collaboration and setting strategic priorities regarding the areas of research. Scalability with large datasets will be guaranteed because optimized algorithms can monitor global research trends in good time and allow knowledge transfer to be effective, supporting the education, research and innovation network.

1.1 Organization of Manuscript

The manuscript is organized to provide a structured overview of this work, which gives a systematic description of the research study, with an introduction that gives the research background and research purpose. The work section reviews the existing studies and identifies research gaps. The methodology part describes the approach used in the community detection based on keywords and the use of graph neural networks. This section outlines the process of collecting the data, preprocessing, building the network of keyword co-occurrence, the SciBERT model, and several graph neural network methods such as GCN, GraphSAGE, FeaStConv, APPN, WLConvContinuous, and AGNN, and finally the hybrid framework of community

detection. The experimental results section will show the results of the suggested method and evaluate its effectiveness. The section contains the performance of the communities detected, the discussion of hyperparameters, performance metrics (precision, recall, F1-score, and accuracy) and analytical information. The conclusion summarizes the major findings and contributions of the research.

2. RELATED WORK

The study of current literature brings out the advancement and evolution of tools used in community detection. The concept of community detection represents a significant issue in network science and Scientometrics. The node features, improved graph learning models, and higher-order network structures have become more popular among researchers. Numerous studies are relevant to this area, and multiple gaps exist, which encourage the present study. Closely related work has investigated both the deep learning and traditional approaches to community detection. These studies offer a starting point, but issues of accuracy and scalability remain.

Liu et al. (2024) [1] used Louvain community detection and Graph Neural Networks (GNNs) together to enhance the prediction of links in scientific literature networks. Their results revealed that the integration of Louvain communities with GNNs steadily enhanced the predictive performance of AUC in the GAT model to 0.823. But this has its limits since the algorithm is also likely to combine smaller communities into bigger ones, thus possibly missing finer, but equally important, communities.

Chandrasekharan et al. (2022) [58] suggested an algorithm (Markov Clustering + Leiden clustering) of citation graphs to identify clusters of articles and corresponding communities of authors and showed that convergent clustering might identify communities of practice in immunology and beyond. Although effective, their approach was very dependent on citation links and expert appraisal of small clusters, which limits semantic depth and the scale of large bibliometric data.

Ahmed and Alzubaidi (2022) [59] suggested a Community Detection Graph Autoencoder (CDGAE), which integrates the node features and the network topology, which does not require the use of classical clustering. Their findings showed that Louvain was more efficient than K-means and Gaussian Mixture Models, especially in featureless data with the use of artificial features. This weakness is the strength of artificially created node features and dealing with dynamic networks. The present work exploits it by utilizing the power of SciBERT embeddings, which offer strong semantic node characteristics without producing artificial features, which makes the identified communities of keywords more reliable.

Garrido-Cardenas et al. (2020) [60] examined Scopus (85,370 articles on malaria) and found primary scientific communities: one dedicated to the topic of vector transmission and the other to drug resistance, which revealed such important themes as mosquito control, the development of vaccines, and studies of parasites. The collaboration patterns identified by the use of ResNetBot, OpenRefine, and genetic algorithms were confined by the reliance on the Scopus data and the specific limitations of the method they used to identify the community.

Zhang, Levina, and Zhu (2015) [61] proposed a combined community detection criterion by combining network edges and node features. Their approach learnt the power of various features within each community with the needed flexibility and worked on simulated and real networks. This had issues with high dimensionality or heterogeneity of features due to the generalizability.

Wang et al. (2023) [62] introduced a Self-supervised community detection algorithm based on node feature convolution, shortly named as SGCN-NFC. This self-supervised community detection algorithm achieves better node feature learning by models through a feature convolution layer, which minimizes the use of earlier labels. Their results indicated that they performed better than current GCN-based approaches.

Yuan, Zeng, and Wang (2022) [63] have suggested a community detection algorithm on the basis of node relationship classification. Their findings showed better performance in comparison with representative algorithms on synthetic and real-world networks. Its weakness is that it cannot be generalized to rich-semantic-featured attributed networks. The current work provides this through the creation of networks of keyword co-occurrence complemented with SciBERT embeddings, which guarantees semantic representation of academic knowledge and the elimination of the drawbacks of purely relational modelling.

A study by Fiallos et al. (2017) [51] assessed 4,552 scientific articles in Scopus (29 research areas in Ecuador) through a Social Network Analysis (SNA)and Natural Language Processing to identify collaboration communities and the most popular topics to guide government policies on resource allocation and research prioritization. They were effective, but could be limited in scalability and more precise semantic representation due to their regional data and the traditional method of SNA.

Guangliang et al. (2022) [64] developed a triangle-based approach that adds closed feature triangles and two-level constraints to identify balanced communities and minimize the free-rider phenomenon. They demonstrated that both overlapping and non-overlapping community detection were improved. But the given framework addresses the issue that very large datasets were not completely scalable.

Noor et al. (2025) [46] came up with an approach of academic community detection that involves content using BERTopic and fine-tuning a SciBERT model to generate topic-based research networks and proposed a cloning strategy that can be used to overcome the publication imbalance, where each cluster of a researcher's work can serve as a node and consequently enables exposing more collaboration prospects. Although this technique successfully addresses the dominance of prolific authors, it is limited by its use of author-level contents of publication data, making it less applicable to extensive bibliometric networks of keywords.

By observing all the important related works on the whole, a combination of these studies proves that community detection can be improved with the help of node features and effective learning models. But these current literatures are limited in their scalability, semantic representation and applicability to complex bibliometric data. There were studies investigating GNN-based bibliometric analysis and studies using BERT-based models to analyse scientific text. Extremely limited literature merged modularity-based pseudo labels with domain-specific language models to detect keyword communities. There is limited existing literature that compares across various GNN architectures and time-segmented Scopus datasets. This paper fills this gap by providing an excellently integrated model that brings together community structure learning and semantic feature representation. From the observation, this study provides a novel approach for community detection. The past research was primarily involved in the identification of research communities based on conventional algorithms like Louvain, Leiden and Infomap on citation networks. A number of works made use of the simple text measures, such as TF IDF, Word2Vec and topic models to aid the clustering. Recent works used Graph Neural Networks, although they were based primarily on network structure or generic embeddings and shallow semantic information. Numerous studies involving GNN were only tested on benchmark datasets like Cora, CiteSeer and PubMed with minimal utilization of real-world bibliometric data. This shows that there is a lack of application of integrated deep learning frameworks to large-scale real-time data. The proposed work fills this gap by demonstrating a complete community detection framework on Scopus bibliographic data based on the integration of SciBERT, Louvain and numerous GNN models.

The primary focus behind this work is to go beyond structure-based community detection and extract the semantic meaning of scientific keywords. Most of the current methods primarily cluster keywords by co-occurrence frequency strength without knowing the semantic context of these research keywords. This model is inspired to integrate semantic representation with the network structure in order to derive meaningful communities of research. The results indicate that when the semantic keyword representations are combined with the graph learning enhance the clarity and consistency of the detected communities during the various time periods. The present study contributes to the research area by using the Louvain algorithm with numerous GNN models enhanced by SciBERT embeddings, which gives a scalable, semantically rich and highly accurate framework to identify keyword communities using Scopus bibliometric information. This makes the current work an important advancement over the earlier research.

The research questions are clear and direct the study. The first research question explores the hypothesis of whether the semantic representations of scientific keywords enhance community detection in keyword co-occurrence networks based on Scopus bibliographic data. The second research question examines the effect of various Graph Neural Network (GNN) architectures on the stability and quality of discovered keyword communities with modularity-based initial partitions. This research hypothesis will be: The incorporation of domain-specific semantic keyword representations in combination with graph structural information will result in more structurally consistent and coherent research communities as opposed to the topology-driven community detection methods.

Community identification in scholarly networks assists in exposing patterns in themes and emergent tendencies in the presence of keywords in bibliometric data. Most of the currently available approaches are limited to addressing large data sets, modelling complex relationships and semantic information modelling. Scientific key on networks needs methods that combine meaningful features into the nodes with powerful community detectors. This work aims at determining the coherent keyword communities based on Scopus data with the help of SciBERT embeddings as the semantic representation

and integrating them with the GNN models and the Louvain algorithm. The study looks at whether such integration positively affects the accuracy and clarity of communities detected. The main issues are related to the high-dimensional features that should be managed efficiently and the quality of community structures that should be ensured. The methodology builds a network of keyword co-occurrences, uses SciBERT to create node features, and uses GNN models with Louvain to do community detection. Findings show the identification of accurate and interpretable keyboard communities, which offer a scalable and successful model that enhances the knowledge structure discourse in the scientific literature.

The proposed analysis is based on the up-to-date state of the art literature integration with bibliometric data for community detection using graph neural networks and scientific language models. The recent research has shown how the modularity-based Louvain approaches are effective as a strong baseline to initialize community in large-scale networks. Recent developments in Graph Neural Networks like GCN, GraphSAGE, FeaStConv, APPNP, AGNN and WLConvContinuous have demonstrated better representation learning of complex graph structures. Parallel domain-specific transformer models like SciBERT have shown to be effective at extracting semantic meaning in scientific text. The work is based on these known results, combining modularity-based community detection semantic keyword representations with a variety of GNN architectures into a single framework that is run on large-scale Scopus bibliographic data.

3. METHODOLOGY

This work includes several significant steps involved in developing this framework. This methodology section contains only the outline process of the framework, and detailed explanations are expanded in the subsections. The first step is to obtain bibliometric information of deep learning based scientific academic publications from the Scopus database. This information is pre-processed and then converted to a keyword co-occurrence network where each node corresponds to a single keyword and an edge corresponds to their co-occurrence in publications. This network is applied through the Louvain algorithm to find and distinguish separate communities with related keywords. Community structures of the keyword co-occurrence network are then identified through the Louvain algorithm, which maximizes modularity to identify dense subgraphs where nodes tend to be more interconnected. At the same time, a pre-trained language model, namely SciBERT, which was specifically trained with scientific text, is used to produce vector representations for the node features extraction of each keyword from the KCN. The main purpose of using SciBERT is to extract the semantic context of the keywords as features in the scientific data. The knowledge of the community detection is then incorporated into the Graph Neural Network (GNN) model to improve the community assessment. GNN models use both network structure and the node features generated by SciBERT to analyze the community detection process further and refine it. This study employs various GNN approaches, including GCN, GraphSAGE, FeaStConv, APPNP, WLConvContinuous, and AGNN. The GNN model uses the features generated by SciBERT as the input to improve the representation of each node with semantic richness. These hybrid frameworks are named SciBLoGCN, SciBLoGS, SciBLoFSC, SciBLoWLC, SciBLoAPPNP, and SciBLoAGNN. The GNN model is learned using the community-structured data, where the model learns meaningful graph representations and enhances the detection. The framework utilizes the advantages of both traditional community detection algorithms for pseudo-labelling and SciBERT for these features to enhance the working of the GNN model to detect the communities from the Scopus keyword co-occurrence network effectively. The performance of the GNN-based model is also measured in terms of accuracy, precision, recall and accuracy scores. The dataset is separated into a training 80% and a test 20 % sets at the node level, where the model is trained on the first set and tested on the second set. Training and testing with Louvain-derived pseudo-labels and SciBERT features allowed the GCN to be tested on unknown data, such that its performance represented true generalisation and enhanced community detection and not reproduction of the original split. The performance of community detection is evaluated by NMI and ARI scores.

The detailed GNN community detection framework based on the Scopus keyword co-occurrence network and driven by SciBERT is presented in Figure 1. In summary, the figure below shows that this research is based on a systematic research design to be used in detecting large-scale bibliometric communities. Scopus bibliographic databases are gathered and pre-treated to extract high-quality indexed keywords. The keyword co-occurrence network is built, in which nodes signify keywords and the edge signifies their simultaneous occurrence in the scientific papers. A modularity-based Louvain technique is used to identify the first

level of community structures to identify the dense groups of keywords. A domain-specific scientific language model is used to create semantic representations of keywords. These representations are combined with several Graph Neural Networks to improve and analyze community structures. Ensuring consistency is provided by the systematic way, with fixed data segmentation, training and testing strategies, and standard evaluation metrics.

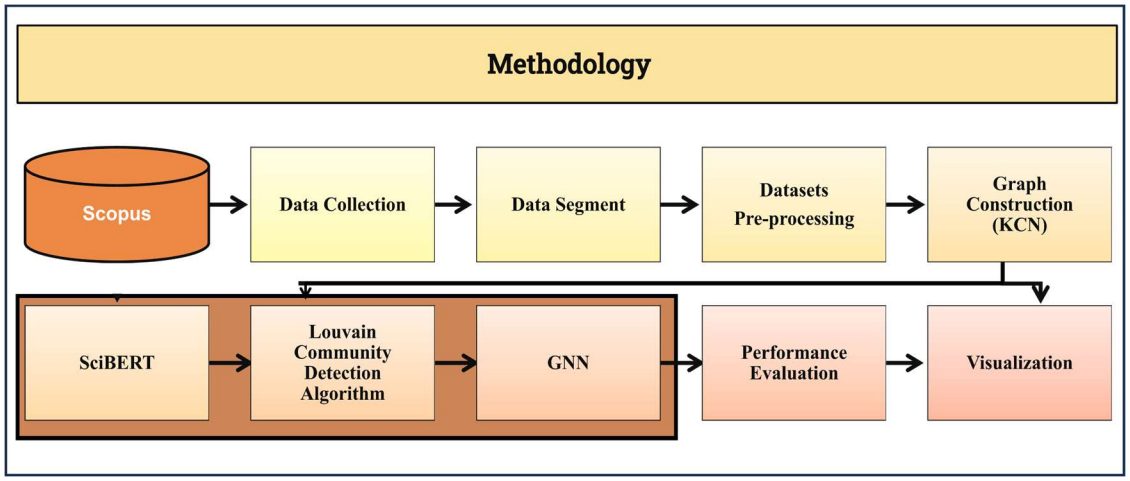


Figure 1: SciBERT-driven GNN Community Detection Framework

3.1. Datasets Description and Data Collection

Scopus bibliometric data is a large-scale scientific bibliographic database that is popular in complex community detection, network analysis, and keyword co-occurrence network (KCN) construction. Though a range of academic bibliographic databases exist (including Web of Science, PubMed and Google Scholar), Scopus is selected due to its overall coverage of subjects, wide metadata and reliable indexing of the keywords that facilitate the formation of better complex co-occurrence networks. Experiments on models are commonly evaluated on benchmark network datasets, such as Cora, CiteSeer, and DBLP, which are restricted in their scope and pre-structured based on citation data. But Scopus offers real-time and domain-specific information with deep associations of keywords, which makes it more appropriate for creating practical and scalable KCNs that are more likely to reflect the current trends in research and help to identify communities and discover knowledge more accurately.

The Deep Learning (DL) based scientific articles retrieved in the Scopus bibliographic database were chosen for this work. The main reason for choosing DL data from 2006 onwards for this research is to identify the importance, role of deep learning, its futuristic trend direction and how it has evolved within the computer science field. Scopus stores very rich scientific publication metadata, such as authors, titles, abstracts, countries, journals, institutions, citations, author keywords, etc., but in this work, the indexed keywords field is used to build keyword co-occurrence networks. The field of indexed keywords in Scopus is significant in that it captures the main ideas of each publication, which permits determining the topic accurately, identifying trends, and detecting communities within research networks. The first step is to extract data in Scopus through a specific search query (e.g., TITLE-ABS-KEY (Deep Learning)). The retrieved documents contain all types of open-access English scientific articles from the computer science field, with all attributes of metadata only. These gathered data are divided into five various periods called Scopus Bibliographic Dataset (SBD), SBD_1 (2006-2013), SBD_2 (2014-2016), SBD_3 (2017), SBD_4 (2018) and SBD_5 (2019). The data from 2020 to 2023 were reserved for future work. These five datasets are segmented in such a way that they should cover the varying sizes, including small, medium and large, of data that have been used to facilitate thorough analysis and scalability of models. The dataset wise DL based scientific articles count from Scopus are statistically visualized in Figure 2. These data, based on these time frames, were extracted in CSV file format and the sample extracted in CSV file is shown in the figure 3.

Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

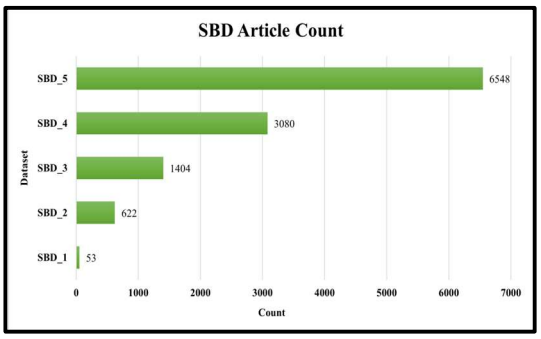


Figure 2: Article Count across Datasets

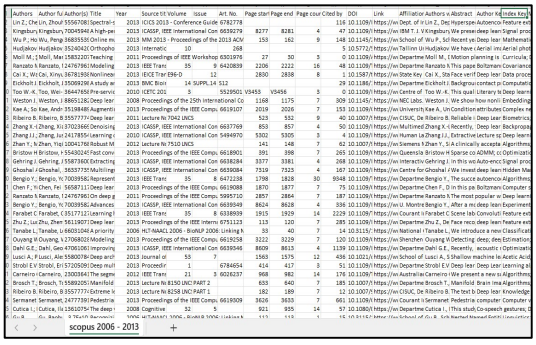


Figure 3: Extracted Dataset Sample of SBD_1

3.2 Data Pre-processing

Data preprocessing and cleaning are the most important steps after the extraction. Individual data pre-processing in Scopus bibliometric data within the Index Keyword field consists of a number of very important steps. The raw data on keywords is then retrieved from the extracted CSV file of the Scopus database. The collected raw dataset proceeds with an extensive preprocessing technique to guarantee both consistency and quality of the keywords. These raw keyword data tend to have inconsistencies and thus may skew the outcome. The detailed step-by-step process of data pre-processing methods is detailly illustrated in Figure 4.

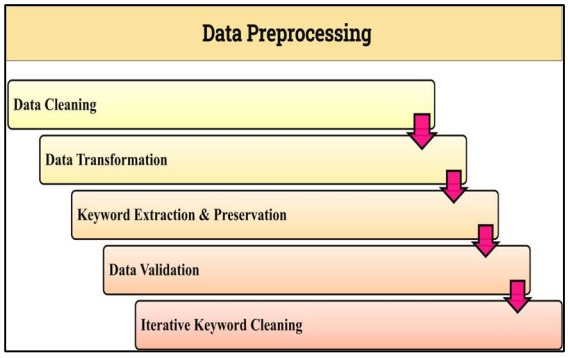


Figure 4: Data Preprocessing Methods

Data Cleaning is the first step that eliminates redundant records to ensure unique records, increasing model efficiency and reliability. Data Transformation lemmatizes keywords, de-duplicates keywords and homogeneous entries, and standardizes text representation. Lemmatization is used to abbreviate words to a root form to increase uniformity. This includes standardization, in which single and plural forms or spelling variants (e.g., the forms of optimization and optimization) are consolidated and normalization, which is the use of acronyms with their expanded forms (e.g., the acronym AI means Artificial Intelligence, and the acronym ANN means Artificial Neural Network). By consolidating semantically unrelated terms, like "Approximate inference", "Approximate inferences", and "Auto encoders", "Auto-encoders", "Autoencoders" are handled as a single record, guaranteeing the optimized network for demand domain knowledge.

Keyword Extraction & Preservation includes co-occurrence filtering to keep an important pair of keywords, semi-colons dividing the keywords and the multi-word phrases retained, like "Deep Learning" and "Artificial Intelligence" from DL, AI, Etc. The keywords are then standardized, that is, reduced to lower case and any special characters or punctuation removed, to ensure the data validation process. Same keywords are then detected and combined to cut off redundancy. More precise techniques such as iterative cleaning of the keys, fuzzy matching by Levenshtein Distance, semantic grouping by Context Boosts and standardization with external ontologies further refine the topics and analysis the results. Frequency analysis can also be conducted in order to determine the most frequently used keywords and possible outliers. Noise words are eliminated to enhance the quality of the data. Lastly, the processed and cleaned keywords are then grouped in a structured form (e.g. list or matrix) to be analyzed or visualized subsequently in bibliometric research.

3.3. Keyword Co-Occurrence Network Construction

The pre-processed keywords are now used to construct the keyword co-occurrence network. Constructing the keywords Co-occurrence Networks is very effective for numerous fields, including bibliometric analysis, content analysis, knowledge mapping and especially for keyword analysis to understand how concepts are organized, and how research topics develop over time. These networks may be extended with the help of different visualization tools and statistical methods in order to reveal recent patterns and relationships between

complex data. These networks visually show the relationships of keywords that are commonly used in academic works. The keyword co-occurrence networks could be studied with the help of network analysis methods, including centrality metrics and community detection algorithms, to identify powerful ideas, groups of similar research topics and gaps in the literature, which prove the future directions of research. The detailed diagrammatic representation of how the co-occurrence matrix is derived from the data is illustrated as the sample Co-occurrence Matrix Construction Process in Figure 5.

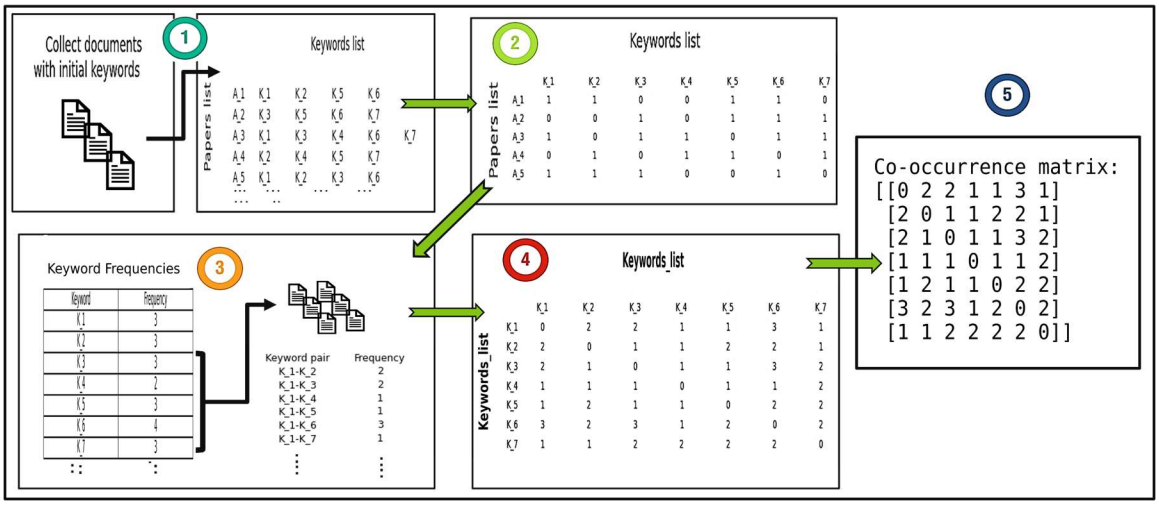


Figure 5: Sample Co-occurrence Matrix Construction Process

For building the KCN, the process is initiated by the construction of a co-occurrence matrix. The First step is to collect the list of keywords and arrange them in an article-wise keyword form. Then convert them into a binary form to represent whether the if keywords are present in that article or not. A co-occurrence matrix was constructed by counting pairs of keywords that appeared between the articles. Each column and each row in this symmetrical matrix correspond to one of the unique, pre-processed keywords that have been obtained from the whole dataset. The value a_{ij} signifies the frequency of co-occurrence of both keywords i and j in a document. When performing network analysis, the diagonal elements where the number of documents containing a single keyword is counted are typically ignored. The emphasis is placed on the off-diagonal elements, which capture the associations among various keyword pairs. This is a quantitative framework for effective keyword retrieval for constructing the KCN network. The matrix is then converted to a network graph structure. Every distinct keyword is represented in a node of the network. An undirected edge is then made between two nodes when the co-occurrence number in the matrix corresponding to each is more than zero. The intensity of the relationship between two edges is represented by an edge weight, which is literally allocated the value in the overall co-occurrence frequency of the keywords. The key components for building an effective keyword co-occurrence network are highlighted in Figure 6.

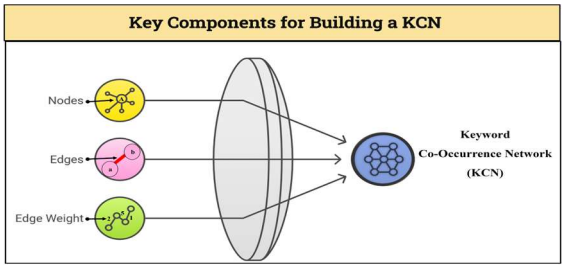


Figure 6: Key Components for Building a Keyword Co-Occurrence Network

In summary, the experiment builds a keyword co-occurrence network with keywords as nodes, co-occurrence of the keywords between the documents as the edges and the frequency of co-occurrence as edge weight. Finally, a last procedure is to apply thresholds to make the network meaningful and computationally feasible. There is a minimum frequency threshold on the nodes to eliminate node-specific or infrequent terms and a minimum co-occurrence threshold on the edges to remove weak and possibly non-significant edges and showcase the strongest and most central thematic relationships in the research area. The overall transformation from a sample co-occurrence matrix to effective sample

KCN construction with a minimum frequency threshold for sample data is visualized in Figure 7.

The construction of the co-occurrence network of the key words was extended to each of the segmented data sets, i.e. SBD_1, SBD_2, SBD_3, SBD_4 4 and SBD_5. This included forming a sample co-occurrence matrix, determining the most important elements to form a network, and converting the matrix into the final keyword co-occurrence network. The methodology guaranteed a regular and systematic representation of the keyword interaction in all periods of time and allowed for a clear analysis of the trends of research and to compose discrete communities of each dataset. After these steps, a KCN graph was created for each dataset, which captures the relationship between the keywords. The comparison of actual and optimized nodes a) & edges b) is recognized statistically in Figure 8. After the optimization of nodes and edges, the visualization of each optimized KCN for SBD_1 to SBD_5 is pictured by highlighting their nodes and edges count within the graph, as figure 9 - a) to e).

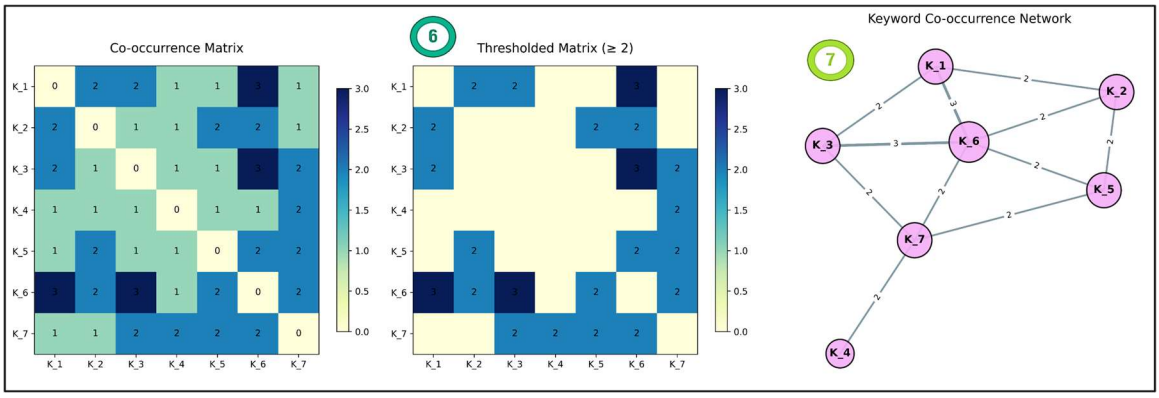
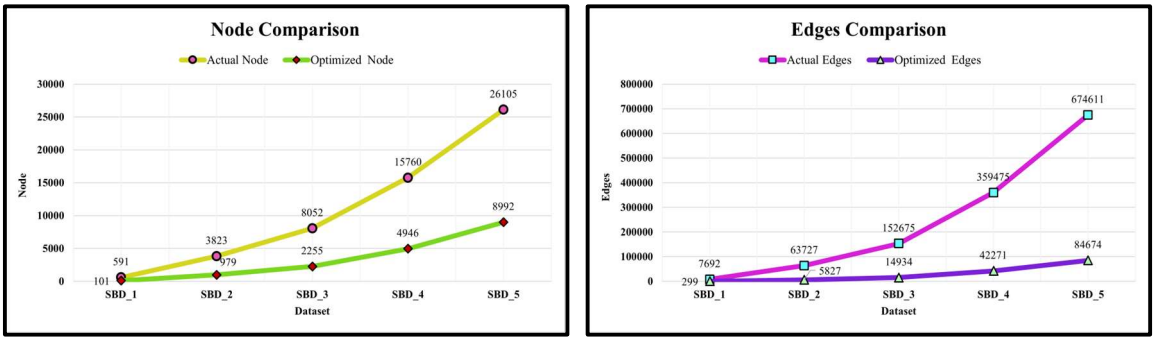
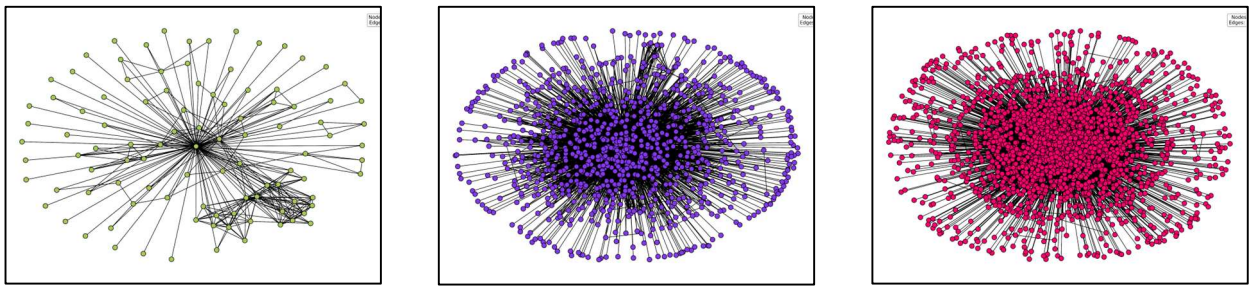


Figure 7: Transformation from Sample Co-occurrence Matrix to KCN Construction

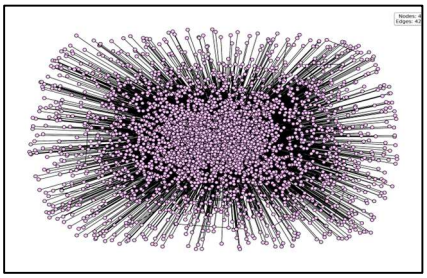


a) *b)*

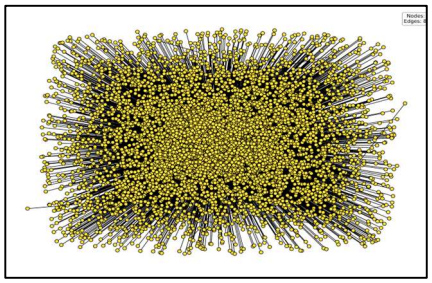
Figure 8: Node and Edge Comparison



a) SBD_1 *b) SBD_2* *c) SBD_3*



d) SBD_4



e) SBD_5

Figure 9: Graph Construction using SBD KCN

3.4. SciBERT

SciBERT is a pre-trained language model, which is specialized in scientific text and it can be exploited as an influential tool for generating node features in the community detection tasks from the scientific networks. Through encoding of scientific documents, keywords or abstracts using SciBERT, researchers can get high-dimensional representations of the vectors that represent the semantic content and context of individual nodes. These features derived by SciBERT can be fed into many different community detection algorithms, including spectral clustering or modularity optimization. The domain-specific embeddings afforded by the SciBERT allow the identification of communities due to the semantic closeness of scientific text that may be able to uncover clusters of related research problems, methodologies or cross-disciplinary interactions. This can be especially successful in large-scale scientific networks, where the relationships between nodes might be subtle enough not to be well represented by traditional structural features. Using node features added by SciBERT, community detection algorithms are able to identify more discernible and understandable scientific communities that aid in exploring research landscapes and possible collaborations.

The SciBERT language model converts scientific keywords in each document into numerical representations by capturing their meaning and context. In this work, SciBERT is used as a semantic encoder, converting every keyword in a textual form into a refined, context-sensitive numerical signal that becomes its starting node feature. These vectors are used as features to describe each node in a scientific network, and community detection algorithms group similar nodes together. This approach helps identify communities based on similarity of scientific content, rather than just structural connections. With the aggregated list of all co-occurring index keywords of a given term, SciBERT learns a dense vector representation of that word that encodes the specific meaning of that word and its research context in the dataset. This enables the Graph Neural Network to not only consider the topology of the co-occurrence network but also the rich semantic connection between concepts, so that the communities identified by the model represent clusters of keywords that are not only frequently connected but also cognitively coherent. This helps in revealing groups of related research topics, methods, or collaborations across disciplines, especially useful for large scientific networks. Using SciBERT features leads to more meaningful and understandable scientific communities, helping researchers better understand the research landscape and potential collaborations.

3.5. Community Detection

Community detection is the process of determining the densely connected groups of nodes within the complex network. This process can be achieved by applying numerous traditional algorithms on the network to understand the structures of the network and the interconnection between the nodes. The Louvain algorithm is one of the best-known traditional algorithms among other algorithms, which is chosen from the existing works for obtaining better results when compared with other methods. The Louvain algorithm is a quick and scalable algorithm to identify communities in large networks, with low computation complexity and high efficiency. The Louvain algorithm is based on first assigning nodes to their own communities and then repeatedly merging nodes in a modularity-optimal way. Such groups are then combined into super nodes, which create a smaller graph. This is repeated until no more modularity can be achieved, and a clear and effective community structure is achieved. In this work, the Louvain algorithm is used to identify communities on the Scopus keyword network of co-occurrence and then the detected communities are visualized using Python.

The difference between community detection and traditional supervised learning tasks is that community detection is not based on the availability

of node labels. This makes common evaluation measures like accuracy, precision, recall and F-measure inapplicable. As an alternative, unsupervised metrics of structural coherence and separation on the network are used to measure the quality of the identified communities. The quality of communities identified is measured by such metrics as the ARI and NMI Scores. These measures are the most appropriate since they measure the structural and relational quality of communities and do not concentrate only on the performance of predictions. Graph Neural Networks are used to strengthen this evaluation because they produce enriched representations that enhance the process of detection and evaluation of community structures.

More advanced models in Graph Neural Networks can be combined with community detection to improve community detection performance and to benefit community detection. The communities identified by the Louvain algorithm generate the pseudo labels, which can be applied to the GNN models for training. Since the Ground truth labels are not available for this network, it is real-time data. The analysis of keywords in each community is used to find the latent research themes and simulate the relationship between communities to acquire insights into the overall framework of the research field. Incorporating GNN with the Louvain algorithm gives meaningful community structures as further features to learn. The main purpose of this integration is to overcome the inapplicable traditional evaluation measures. Such community knowledge boosts the level of representation to achieve further tasks. This combination enhances accuracy to a greater extent and the interpretability of complex networks.

3.6. Graph Neural Networks (GNN)

Graph Neural Networks are deep learning frameworks specifically designed to model and analyze graph-structured data. They expand neural networks by adding graph topology to the learning process. GNNs are useful in non-Euclidean data in which relationships are more important than absolute positions. They allow structural and semantic information of nodes and edges to be retained in representations of learning. GNNs can deal with graphs of arbitrary scale and structure, which allows them to be trained to work with new graph data without requiring a complete retraining. They can be easily interpreted since they can capture the nature of patterns of relationships that are normally ignored in a conventional machine learning model.

GNNs have also been designed in different architectures to solve different graph learning tasks. Graph learning tasks are categorized as node, edge and graph level according to the focus of the task. Node-level tasks forecast the characteristics of nodes. Relationships between nodes are predicted by edge-level tasks. Graph-level tasks, generating information about the graph as a whole and community detection are one such example of a graph-level task. The GNN model used is selected for the reason that it is precise, scalable and appropriate to community detection in complex data. This provides robustness while maintaining the model computationally feasible on large-scale graph data. To achieve a good level of local and global structural information and provide high levels of computational efficiency, these GNN models were chosen. When graph topology is paired with node characteristics, its capability provides appropriate embeddings that illustrate community separations. In addition, GNNs have been extensively shown in the literature on citation and co-authorship network tasks. But this work is based on a keyword co-occurrence network, which makes it a comprehensive starting point to assess community structures. This is work. GNN Methods like GCN, GraphSAGE, FeaStConv, APPN, WLConvContinuous and AGNN were used and as listed in Figure 10.

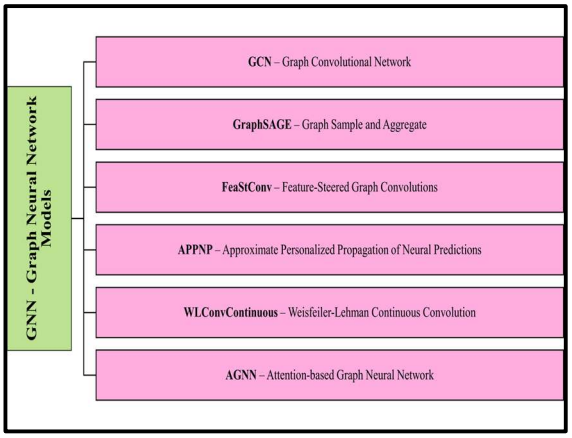


Figure 10: Implemented GNN Methods

3.6.1. Graph Convolutional Networks (GCN)

GCN represent a general type of graph neural networks that process graph-structured data through incorporating the information of the neighbours of a node. GCNs are worked out based on the layer-wise propagation rule, in which each node is expected to update its feature representation with a combination of the features of the node itself and a weighted sum of features of its neighbours. The aggregation

enables the model to learn local graph structures and similarities of nodes. To perform community detection, GCNs are trained to embed nodes so as to capture patterns of connectivity, where nodes within the same community will share similar representations, and can be clustered or labeled to recognize communities.

3.6.2. Graph Sample and Aggregation (GraphSAGE)

GraphSAGE is an inductive GNN that forms node embeddings by fixed-size neighbourhood and aggregating neighbour features using mean, LSTM or pooling functions. GraphSAGE is not like a traditional GCN since it can process unseen nodes, which makes it useful for dynamically evolving graphs. GraphSAGE embeddings are used in community detection to encode structural and attribute-based similarities, and nodes with similar roles within the graph or connectivity behaviour are grouped and hence communities in large-scale graphs are revealed.

3.6.3. Feature-Steered Graph Convolutions (FeaStConv)

FeaStConv is a GNN that learns to dynamically compute the convolution weights of the node pairs depending on their feature and facilitates adaptive message passing. This uses the learned attention to modify the contribution made by neighbouring nodes to the target node representation instead of using fixed aggregation weights. This method gives FeaStConv the ability to learn the intricate graph structures better. In order to detect communities, it discovers subtle relationships among nodes, generating embeddings that display dense and well-connected clusters.

3.6.4. Approximate Personalized Propagation of Neural Predictions (APPN)

APPNP is one of the GNN techniques which used to build on neural network predictions based on personalized PageRank-based propagation, which enables the flow of information over long distances but prevents over-smoothing. This initially calculates initial node features with a neural network and subsequently makes a propagation across the graph with a teleport probability. In community detection, APPNP embeddings represent local and global structural information. They determine communities in which nodes are well connected within the same community and have clear boundaries between various communities.

3.6.5. Weisfeiler-Lehman Continuous Convolution (WLConvContinuous)

WLConvContinuous is one of the GNN methods that is based on the Weisfeiler-Lehman graph isomorphism test. This recursively trains node features by fusing neighbor features in a way that reflects the topology of the graph only in a unique way, and thus provides the model more capacity to discern nodes with diverse structural functions. In the case of community detection, WLConvContinuous will produce similar embeddings that capture fine-grained structural differences to ensure that similar nodes belong to the same community, whereas distinct communities can be separated.

3.6.6. Attention-based Graph Neural Network (AGNN)

AGNN adds an attention mechanism to graph neural networks, enabling nodes to attach varying weights to their neighbours when aggregating features. AGNN can learn attention scores using them to concentrate on important neighbours and disregard irrelevant ones to enhance the learning of representation in heterogeneous or noisy graphs. To detect communities, AGNN embeddings emphasize high intra-community ties and attenuate the effect of weak inter-community edges and thus, it is simpler to distinguish cohesive groups of nodes into observable communities.

3.7. Hybrid Framework

A hybrid framework is a combination of several approaches aimed at utilizing their strong points to perform better. This publication combines the classic Louvain algorithm with several Graph Neural Network models and applies SciBERT to get meaningful features of keywords. This method is more accurate, robust and flexible in detecting the community and in doing this, the limitations of the other methods are overcome and the best textual information is efficiently captured. The Hybrid Framework combines different approaches to analysis to increase the accuracy and insights in analyzing complex data. The hybrid framework uses both structural and feature-based information to detect a community through the combination of community detection algorithms and the use of graph neural networks. It improves the detection of the community based on the keywords using modularity optimization with node feature learning. The framework helps to detect research clusters and identify more precise insights into complex bibliometrics networks.

The hybrid model is used to identify communities of keywords with the help of traditional community detection, semantic features extraction from SciBERT and graph neural networks. The Louvain algorithm is used first on the co-occurrence network of the keyword to draw the initial communities maximizing modularity. This process

identifies communities that are closely related to keywords densely for creating pseudo-labels of possible community membership of each node. Simultaneously, the scientific text-specific pre-trained language model SciBERT obtains the vector representation of each keyword. These embeddings represent the semantic meaning and contextual relationships of the keywords and they offer the rich node features that represent the content of the scientific publications. The SciBERT-generated node features and the Louvain-generated pseudo-labeled communities are then given as input to different GNN models, including GCN, GraphSAGE, FeaStConv, APPNP, WLConvContinuous and AGNN. These integrated hybrid framework models are named as SciBLoGCN, SciBLoGS, SciBLoFSC, SciBLoWLC, SciBLoAPPNP and SciBLoAGNN. The GNN models combine both the structure of the network and the semantic features to optimize the community detection. They establish more exact representations for each node, enabling them to recognize communities that are both structurally sound and semantically meaningful. The combination of these three elements brings the framework the accuracy, generalization and interpretability of keyword community detection in large-scale bibliometric networks.

4. EXPERIMENTAL RESULTS

Experimental Results highlight the findings of the proposed methods that are presented systematically by evaluation and analysis. The effectiveness and reliability of the developed models are supported by experimental Results, which may be evaluated by experimental evidence. The network contains nodes expressed in the form of a keyword and Edges represent relationships of co-occurrence of keywords between documents. The preprocessing was used to eliminate duplicates, self-loops and noise to ensure a quality of the network of keyword co-occurrence was obtained. Scientific term representations were used to create node features in the form of SciBERT embeddings, which are rich context representations of scientific terms.

The community detection process in this hybrid model is based on the concept of having an initial community partition that is generated using the Louvain algorithm community detection algorithm based on SciBERT Features, which produces the pseudo-labels. The framework performs on different GNN models, namely GCN, GraphSAGE, FeaStConv, APPNP, WLConvContinuous and AGNN. These proposed frameworks are named as SciBLoGCN, SciBLoGS, SciBLoFSC, SciBLoWLC, SciBLoAPPNP and SciBLoAGNN. These numerous GNN models are used to compare them and evaluate their performance through performance evaluation metrics. The adjacency matrix and the node features based on SciBERT were used to train the GNN on the training set. The Louvain algorithm produces pseudo-labels, which are only used on the training nodes, and the GNN is left to learn the patterns of the communities without having access to the test set. These pseudo-labels were applied as supervisory labels in training the GNN. The dataset was separated into training and testing datasets using a standard 80:20 ratio at the node level, where the labels on communities of test nodes are not disclosed in the training phase to test the performance of the generalization. The training set also includes nodes that have pseudo-labels produced by the Louvain algorithm that direct the GNN models to learn the community structure. The features are embedded using SciBERT-based node embeddings with the keywords in each node, which enables the models to learn semantic and structural information. The test node labels will represent an objective test of the capacity of the model to extrapolate and reveal the true structures of communities outside the original pseudo-labeling. The effectiveness and strength of the framework in reducing circular validation are further proved by comparing and contrasting it with various GNN models. The GNN models during training optimize the representations of the nodes in order to identify the communities without altering the patterns between the nodes in the graph. The test nodes were then identified to evaluate the performance of the models, which was compared to the pseudo-labels by detecting the community. These are evaluated by measuring them using evaluation metrics (Precision, Recall, F1-score, Accuracy) and partition similarity measures, Normalized Mutual Information Score (NMI), and Adjusted Rand Index Score (ARI). This experimental design shows that GNNs, augmented with the features of SciBERT, are able to identify meaningful communities of keywords in co-occurrence networks, and confirm the methodology with both supervised and structural evaluation measures.

4.1. Evaluating the Performance of Detected Communities

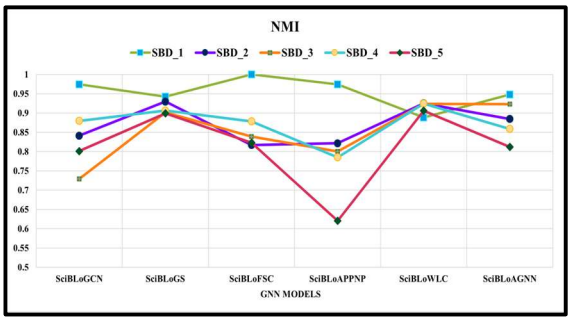
The evaluation metrics are used to evaluate the ability of the model to assign nodes to communities correctly as compared to pseudo-labels. The pseudo-labels are produced by the unsupervised community detection algorithms, which are treated as training and testing targets for the GNN. The similarity measures of partitions are used to compare the level

of agreement between two partitions of the community. They are especially applicable to community detection in cases where ground truth is frequently inaccessible and are compared to baseline partitions. These are measures that evaluate the stability and consistency of identified community structures. Normalized Mutual Information (NMI) is a measure that determines the mutual information between predicted and reference partitions. Adjusted Rand Index (ARI) measures the similarity of two partitions, and it adjusts for chance agreement. The overall performance for detected communities is tabulated in Table 1.

Table 1: Performance of Detected Communities.

Year	Metrics \ Dataset Name		SciBLo GCN	SciBLo GS	SciBLo FSC	SciBLo APPNP	SciBLo WLC	SciBLo AGNN
2006 - 2013	SBD_1	DC	5	5	5	5	6	5
		NMI	0.9742	0.9425	1	0.9742	0.8887	0.9477
		ARI	0.9892	0.9611	1	0.9903	0.9308	0.9798
2014 - 2016	SBD_2	DC	9	9	7	9	8	8
		NMI	0.8417	0.9301	0.8166	0.8218	0.9252	0.8844
		ARI	0.8686	0.9463	0.857	0.8203	0.9462	0.8992
2017	SBD_3	DC	9	8	9	9	8	10
		NMI	0.729	0.9015	0.8387	0.8003	0.9239	0.9233
		ARI	0.7462	0.9312	0.8742	0.8201	0.9529	0.9496
2018	SBD_4	DC	9	8	10	9	12	10
		NMI	0.8794	0.9062	0.8785	0.7851	0.9245	0.8584
		ARI	0.9278	0.939	0.9202	0.8294	0.945	0.8929
2019	SBD_5	DC	7	6	9	6	7	6
		NMI	0.8017	0.8994	0.8232	0.6202	0.9058	0.8118
		ARI	0.8651	0.935	0.8832	0.6806	0.9363	0.8532

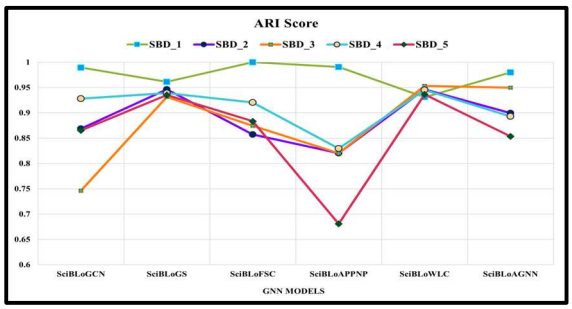
Table 1 shows the performance of six models of Graph Neural Networks on various Scopus-based datasets between the years 2006 and 2019. DC shows the number of Detected Communities, whereas NMI and ARI demonstrate grouping consistency and structural quality of the detected communities. In the case of SBD_1 (2006-2013), 5-6 communities were identified by all models, with SciBLoWLC obtaining 1 in both NMI and ARI, and this aspect shows the outstanding performance of this model on this small dataset. In SBD_2 (2014-2016), the DC falls between 7 and 9, and values of NMI and ARI are the greatest, around 0.9395, which demonstrates the strong performance of community detection on a medium-sized dataset. SBD_3 (2017) showed DC between 8-10, whereas SciBLoGS and Continuous SciBLoAGNN presented the best NMI and ARI above 0.92, which is a good clustering behaviour. In the case of SBD_4 (2018), DC was between 9 and 12, whereas the highest ARI of 0.945 and NMI of 0.9245 were detected in SciBLoAPPNP and SciBLoWLC, indicating trustworthy community detection on a large dataset. DC range between 6 and 9 in SBD_5 (2019) and high values of NMI and ARI of 0.8994 and 0.935 in SciBLoGS, respectively, prove the stability of model effectiveness on a larger dataset. The variation in the size of data among the datasets represents the small, medium and large quantities of publications and keywords that affect the quantity of identified communities and performance of the clustering. The performance of identified communities based on a) NMI score, b) ARI score and c) detected communities count is shown in Figure 11.



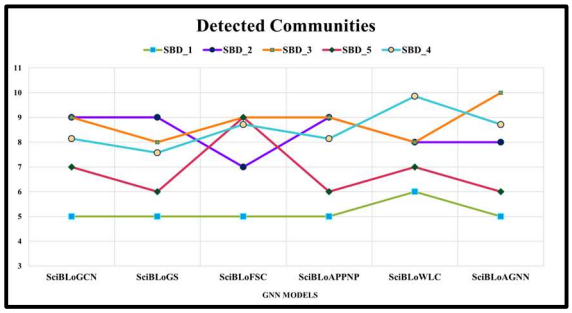
a) NMI Score

Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



b) ARI Score



c) Detected Communities

Figure 11: Performance of Detected Communities

4.2. Hyperparameter and Evaluation Metrics

The hyperparameters manage a particular element of the model architecture or training procedure to enhance performance and guarantee the efficient identification of the community. Table 2 shows the important hyperparameters of the GNN models on five datasets of Scopus. GNN Layers determines the graph convolution layers in the network. All models operate on 2 layers that elicit node relationships but do not overfit on smaller datasets. This value is selected to be at a level of balancing complexity and computation time of the model. Dimensions of Hidden Layers gives the number of neurons in each of the hidden layers. Values of SBD_1 and SBD_2 are 16 and those of SBD_3-SBD_5 are 64; thus, the model will learn more complicated patterns on large datasets. The values are chosen to suit the growing dataset size and network complexity. Dimensions of SciBERT Node Features show the size of the feature vector of each node based on SciBERT embeddings. The fixed dimension of 768 gives a rich semantic representation of the keywords.

Table 2: Hyperparameters Tuning

Hyperparameters	SBD_1	SBD_2	SBD_3	SBD_4	SBD_5
GNN layers	2				
Dimensions of Hidden Layers	16	16	64	64	64
Dimensions of SciBERT Node Features	768				
Optimizer	Adam				
Activation Function	ReLU (Rectified Linear Unit)				
Dropout Rate	0.5				
Learning Rate	0.01				
Number of Epochs	300	300	700	700	700
Temperature	2.0				

Optimizer decides which algorithm to use in changing model weights when training. Stable and efficient gradient updates are done on Adam. The option guarantees rapid convergence and credible training between datasets. Activation Feature stipulates the non-linear operation on layer outputs. ReLU is a non-linear method that is computationally efficient. The purpose is chosen due to its simplicity and ability to work in deep networks. Dropout Rate determines the proportion of neurons that are randomly disabled in training. The rate of 0.5 will minimize overfitting and maintain enough information. The value is decided upon so as to have a balance between regularization and learning capacity. Learning Rate regulates the decrease in weight updates. A balance between stable convergence and learning speed is achieved at a value of 0.01. The rate is chosen to make sure that it is optimized gradually and successfully. The number of Epochs is the number of times the training data has been passed through. The smaller datasets go through 300 epochs, whereas the larger datasets go through 700 epochs to guarantee proper learning. The values are selected so as to provide enough training using the size of the dataset. The Temperature-Scaled SoftMax hyperparameter modifies softness of pseudo-labels by transforming hard one-hot labels into continuous probabilities. The increased temperature (2.0) can be used to allow the GNN architecture to consider a range of community assignments, whereas lower temperature generates stricter labels. Loss Function determines

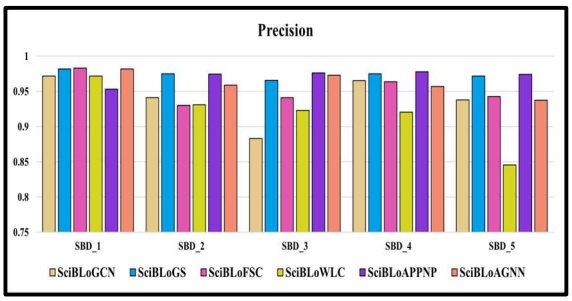
the goal that the model is going to maximize. Negative Log Likelihood loss is useful in managing tasks involving the detection of multi-class communities. The loss has been chosen because it is appropriate for classification and community assignment assignments. The purpose of this tuning is to strike a balance between replicating the traditional algorithms and discovering new structures to enhance the gradient flow and training stability. Loss functions like Negative Log Likelihood Loss (NLL) and Kullback-Leibler (KL) Divergence loss are used to train the model to align its predictions with real patterns on hard labels, and to align the distribution of predictions on soft labels, respectively, and the optimization of the form of its structure is available. Adjustments in the hidden layer size and the number of epochs are dependent on the size of the dataset, as it allows the model to find complex patterns in medium and large datasets.

The performance of these models is based on the comparison of the performance as predicted and actual labels, which can be measured in terms of True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN). These measures give a clear view of the performance of the model. True Positives (TP) in this case indicate those nodes that are properly placed within a community, False Positives (FP) are nodes that are wrongly placed within a community, False Negatives (FN) are those nodes that belong to a community but are not identified and True Negatives (TN) are those nodes that are not misplaced. Precision is used to define the extent of the correctly allocated nodes in relation to all members of a community that are being predicted. Recall indicates the share of the true community members that was identified. The F1-score is a harmonic average of precision and recall, thus giving a balanced score of both. Accuracy represents the general percentage of nodes that are assigned to their communities. The detailed evaluation results are tabulated as performance evaluation metrics in Table 3.

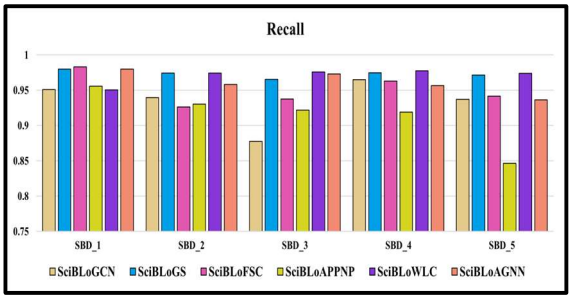
Table 3: Performance Evaluation Metrics

Year	Metrics \ Dataset Name		SciBLo GCN	SciBLo GS	SciBLo FSC	SciBLo APPNP	SciBLo WLC	SciBLo AGNN
2006-2013	SBD_1	Precision	0.9719	0.9819	0.9833	0.9718	0.9534	0.9818
		Recall	0.951	0.9802	0.9833	0.956	0.95045	0.9802
		F1-Score	0.9699	0.9804	0.9826	0.9697	0.95049	0.9798
		Accuracy	0.9703	0.9802	0.9901	0.9703	0.9505	0.9802
2014 - 2016	SBD_2	Precision	0.941	0.975	0.9302	0.931	0.9746	0.9588
		Recall	0.9397	0.9745	0.92646	0.93054	0.97457	0.95812
		F1-Score	0.9397	0.9744	0.927	0.9302	0.9744	0.9579
		Accuracy	0.9397	0.9745	0.9265	0.9305	0.9745	0.9581
2017	SBD_3	Precision	0.8829	0.9656	0.9413	0.923	0.9761	0.973
		Recall	0.8776	0.9654	0.93747	0.92195	0.97605	0.97295
		F1-Score	0.8747	0.9652	0.9382	0.9216	0.976	0.9729
		Accuracy	0.8776	0.9654	0.9375	0.922	0.9761	0.9729
2018	SBD_4	Precision	0.9655	0.9751	0.9637	0.9205	0.9778	0.957
		Recall	0.9652	0.9749	0.96320	0.91892	0.97776	0.95673
		F1-Score	0.965	0.9749	0.9632	0.9181	0.9777	0.9567
		Accuracy	0.9652	0.975	0.9632	0.9189	0.9778	0.9567
2019	SBD_5	Precision	0.9379	0.9717	0.9427	0.8455	0.9742	0.9377
		Recall	0.9371	0.9715	0.94173	0.84642	0.97398	0.93661
		F1-Score	0.9372	0.9716	0.9417	0.8453	0.974	0.9368
		Accuracy	0.9371	0.9715	0.9417	0.8464	0.974	0.9366

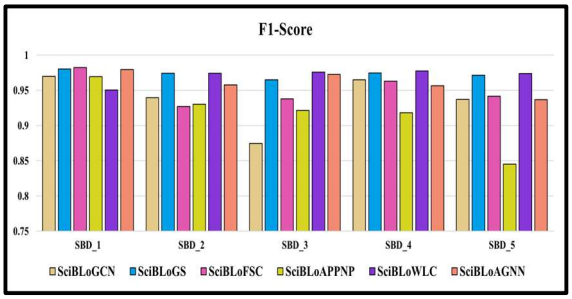
The datasets are chosen to capture the various data scales. SBD_1 is a small dataset that has fewer network structures. SBD_2 and SBD_3 are medium-sized datasets that are of medium complexity. SBD_4 and SBD_5 are bigger and more complicated datasets, which represent complex relationships and difficult community discoveries. In the case of the dataset SBD_1 between 2006 and 2013, SciBLoGCN has given the best results in terms of precision, recall, F1-Score, and accuracy, all greater than 0.99, which can be discussed as highly consistent and reliable predictions. SciBLoGS performed the best in the SBD_2 dataset 2014-2016, with all the evaluation metrics of approximately 0.974-0.975, showing that it can better address the network structure in the period between 2014 and 2016. In the case of SBD_3 data of 2017, SciBLoWLC outperformed the rest of the approaches, demonstrating a precision and accuracy of more than 0.976, which shows that the process is effective in identifying community structures. In 2018, the SciBLoWLC was once again at the top of the results in SBD_4, where the precision and accuracy were nearly 0.978, which is a reliable and correct performance. Lastly, in the SBD_5 dataset of 2019, WLConvContinuous retained its lead with a precision and accuracy of about 0.974, which proved its continued usefulness in diagnosing communities. Generally, SciBLoWLC and SciBLoGS have always scored the best on various datasets, which indicates their strength and effectiveness in community detection activities. The comparison among evaluation metrics is visualized in figure 12 a) to e).



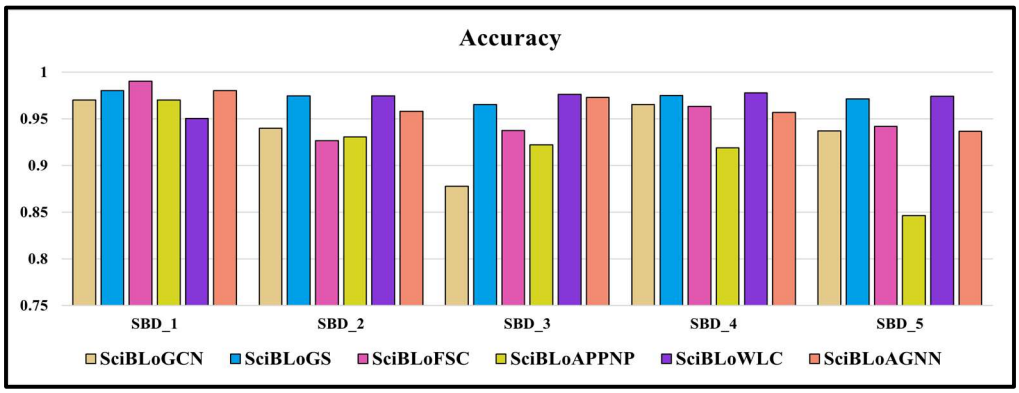
a) Precision



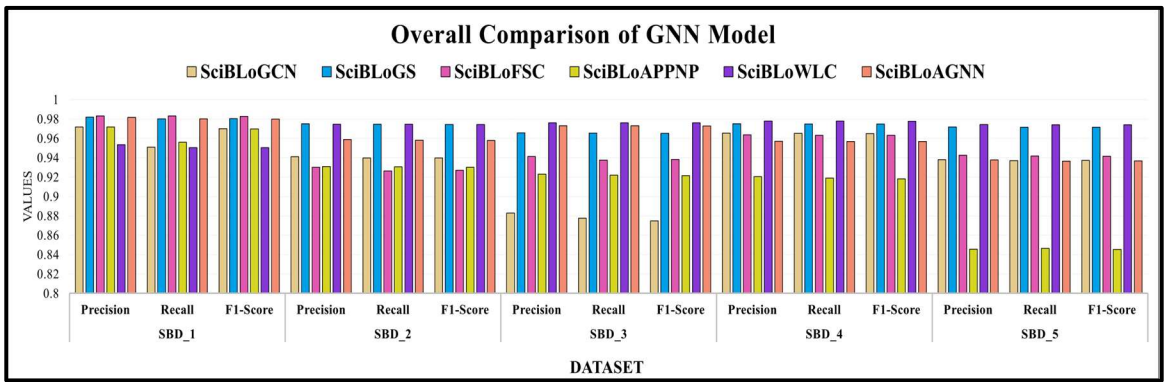
b) Recall



c) F1-Score



d) Accuracy



e) Overall Comparison

Figure 12: Comparison among Evaluation Metrics

4.3. Discussion

Benchmark datasets have fixed ground truth labels to evaluate. Unlike Benchmark datasets, Scopus KCN does not have ground truth because the networks of research keywords in the domain are dynamic and specific to the domain and ground truth is not feasible through manual labeling. Since there are no ground truth community labels available for real-time data, Scopus KCN relies on pseudo-labels. These pseudo-labels enable the GNN to replicate supervised learning and it is able to recover meaningful and data-driven communities that replicate actual research trends and it can also perform performance evaluation like actual benchmark datasets. Pseudo-labels can be used to compare the performance of a model to assess its ability to detect communities where ground truth is unavailable and they also enhance the identification of meaningful communities of keywords. This makes the model evaluation less independent, as the GNN can reproduce the existing structures. The past research has shown that thematic structures are identified by keyword co-occurrence networks through network connectivity. The findings of this work build upon these results by demonstrating that structure is not enough to interpret the community accurately. Communities identified with semantic-enriched node representations are more internally consistent when assessed with an agreement measure. This confirms that community detection using semantic information outperforms when compared with conventional topology-based community detection.

The Louvain traditional community detection approach of identifying communities based on SciBERT Features integrated with Graph Neural Networks models, as experimental results on KCN graphs. The node is expressed by textual features of the key words processed by SciBERT, which enable the GNN to be trained on forming rich feature embeddings that store semantic as well as structural information. In order to determine reference structures, a baseline community detection algorithm, such as Louvain, was used to produce partitions used as pseudo-labels. Training GNNs with pseudo-labels provided by the Louvain algorithm may lead to circular validation since the same procedure determines and measures the communities. To deal with this, the framework strictly separates training and testing datasets in order to provide an unbiased evaluation and valid community detection performance. During the training step, the graph data set is divided at the node level into a training and test set (standard 80:20). The GNN is able to optimize the node representations by reducing the gap between the identified community memberships and the pseudo-labels. Such pseudo-labels give supervision when training the different GNN models, including GCN, GraphSAGE, FeaStConv, APPNP, WLConvContinuous, and AGNN. The training and testing design avoids circular validation by making sure that test nodes are fully independent when the model is being trained. The GNN identifies higher-order relational patterns not based on the Louvain structure by feature representation using SciBERT node embeddings. The independent validation on hidden test data is an assurance that the model is using real community structures and not recreating the available structures. The evolution of SBD_3 into Community 7 Keyword Communities is the gradual narrowing down of the 2017 Scopus Keyword Co-occurrence Network into a narrower thematic cluster of related keywords. This process facilitates community finding by separating closely related sets of keywords, so that specific areas of special research within the larger network can be identified more easily. To build the relationships between indexed

words as they appear in research articles, the original SBD_3 graph of the 2017 Scopus Keyword Co-occurrence Network (KCN) was originally built to showcase the relationship between indexed keywords in their co-occurring context. The community detection process was used to analyze the graph systematically in order to find meaningful groups of most closely related keywords. The identified communities graph showed that there were some distinct clusters of keywords that represented a specific area of research. Based on this step, the KCN was further divided into community-related sets of keywords so that they could be analyzed as thematic patterns. Community 7 was chosen separately as a sample representation to show how KCN are identified as a consistent keyword community with a high internal correlation of its words and the clear focus of research in the context of the larger network. The transformation from the original SBD_3 graph, then Detected Communities Graph using the SciBLoGS model and next separated KCN as Community-wise Keyword Communities into the actually identified Community 7 Keyword Communities shows how the 2017 Scopus KCN was refined over time into the specific thematic patterns, as depicted in Figure 13.

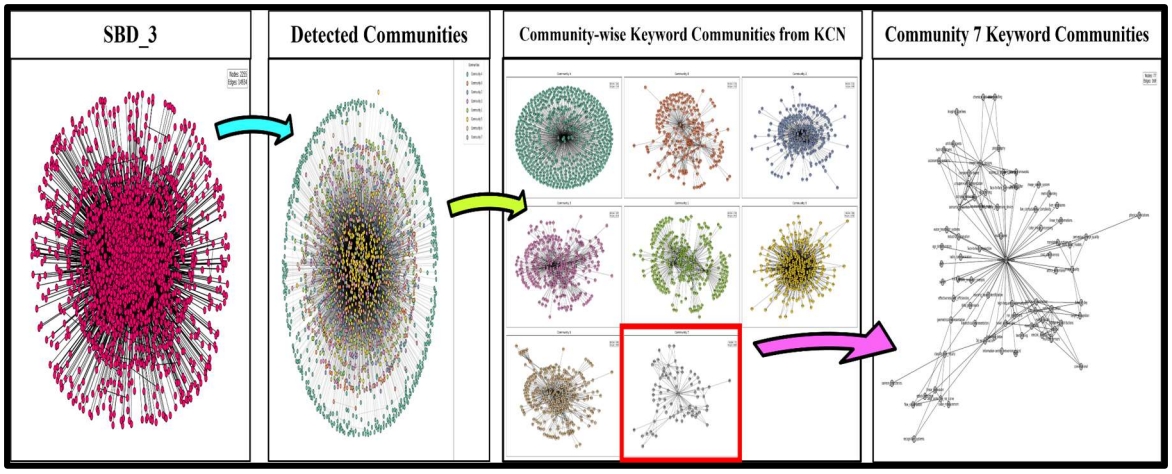


Figure 13: Transformation of SBD_3 into Community 7 Keyword Communities using SciBLoGS

Community 7 Keyword Communities, the point of focus is characterized by a small set of the most powerful keywords, including visual languages, image sensor designs, computer hardware, mobile robots and high-level abstraction. These keywords reflect a high correspondence with the development of intelligent systems and automation, where visual comprehension and sensor networks are important in improving computational efficiency. Availability of computer hardware indicates the basis of constructing effective architectures to accommodate these new technologies, with mobile robots highlighting the applications involving perception, control, and flexibility. High-level abstraction refers to the conceptual frameworks and modelling methods that contribute towards making the complex system designs easier to understand, enhancing functionality and readability. The combination of these keywords shows that there is a unifying and dynamic research direction in the field of intelligent computing and autonomous system development that is directed towards the larger scientific communities. But the identified communities are aimed at grouping similar keywords based on pseudo-labelling, but do not explicitly provide the exact labels for the detected communities based on their research themes. Post-analysis and expert validation still remain necessary to interpret research themes. This constraint has an impact on the direct usability of the identified communities in the automated topic identification and presents the necessity of having theme labelling mechanisms in future work. The main research issue that is addressed in the work revolves around the poor semantic sensitivity of current keyword-based community detection techniques used on bibliometric networks. The majority of the existing methods are based on co-occurrence frequency and network topology that limits meaningful interpretation of research communities. The other research issue is the lack of validation of advanced community detection models on scale on large real world bibliographic data like Scopus. The lack of automatic identification of research themes based on communities identified and the difficulty in combining semantic understanding with scalable graph learning models are open research issues.

Other open problems are associated with the generalisation of community detection models to different times and fluctuating research areas.

5. CONCLUSION

The proposed novel framework of research keyword-based community identification on Scopus bibliometric information shows an innovative method of interpreting recent research trends. This work provides a holistic approach to studying bibliometric networks on a large scale by integrating the traditional community detection methods with more advanced deep learning models. Incorporation of the Louvain algorithm with other Graph Neural Network (GNN) models, with benefits added by the SciBERT-generated node features, has demonstrated promising outcomes in identifying meaningful research communities. This is a hybrid method, which would combine the benefits of both the traditional and modern methods and will enable one to have a more subtle view of how the different domains of research are interconnected.

The results of the experiment demonstrate the existence of effective research communities with references to the structure and interconnections in the keyword co-occurrence network. This framework also gives meaningful information to scholars as they are able to establish relationships among the various research fields and prioritize their research strategies accordingly. The step-wise procedure of the methodology, based on data collection and preprocessing, as well as the implementation of sophisticated GNN models, predetermines a comprehensive and systematic analysis of the bibliometric data. Multiple GNN methods (GCN, GraphSAGE, FeaStConv, APPNP, WLConvContinuous, and AGNN) can be used to make a comparative evaluation and select the most suitable model to be used in this specific application. This novel framework performs well on SciBLoWLC and SciBLoGS models among the other models. This provides a great contribution to the field of bibliometric analysis by presenting a powerful instrument that allows researchers and policymakers to navigate the confusing world of scientific research. This strategy can potentially guide strategic decision-making in academic and research settings by offering a more comprehensive insight into the structures of research communities and how they have changed over time. The framework serves as an important source of knowledge on how the various emerging research areas relate to each other and this helps the researchers to understand the relationship between one topic and another and also the direction that future research should take by developing the automated topic identification and presents the necessity of having theme labelling mechanisms as a future work using LLM. Future research work is to extend the framework to analyze interdisciplinary research by examining connections between traditionally separate fields, such as Cross-domain analysis. Optimize the framework for handling even larger datasets, potentially exploring distributed computing solutions. Develop methods to track the evolution of research communities over time, for identifying emerging trends.

Acknowledgement

We thank PSGR Krishnammal College for Women for the Motivation and Encouragement to make this work a success.

REFERENCES:

[1] C. ; Liu *et al.*, "A Community Detection and Graph-Neural-Network-Based Link Prediction Approach for Scientific Literature," *Mathematics 2024, Vol. 12, Page 369*, vol. 12, no. 3, p. 369, Jan. 2024, doi: 10.3390/MATH12030369.

[2] B. Kamiński, P. Prałat, · François Théberge, and S. Zając, "Predicting properties of nodes via community-aware features," *Social Network Analysis and Mining 2024 14:1*, vol. 14, no. 1, pp. 1–18, Jun. 2024, doi: 10.1007/S13278-024-01281-2.

[3] Y. Malode, A. Aylani, A. Bhardwaj, and D. Hajoary, "Comparative Analysis of Community Detection Algorithms on the SNAP Social Circles Dataset," Feb. 2025, Accessed: Oct. 14, 2025. [Online]. Available: https://arxiv.org/pdf/2502.04341v1

[4] R. Márquez and R. Weber, "Dynamic community detection including node attributes," *Expert Syst Appl*, vol. 223, p. 119791, Aug. 2023, doi: 10.1016/J.ESWA.2023.119791.

[5] Z. Lu, J. Wahlström, and A. Nehorai, "Community Detection in Complex Networks via Clique Conductance," *Sci Rep*, vol. 8, no. 1, pp. 1–16, Dec. 2018, doi: 10.1038/S41598-018-23932-Z.

[6] S. Kumar, A. Mallik, and S. S. Sengar, "Community detection in complex networks using stacked autoencoders and crow search algorithm," *Journal of Supercomputing*, vol. 79, no. 3, pp. 3329–3356, Feb. 2023, doi: 10.1007/S11227-022-04767-Y.

[7] H. Zheng, H. Zhao, and G. Ahmadi, "Towards improving community detection in complex networks using influential nodes," *J Complex*

Netw, vol. 12, no. 1, Dec. 2023, doi: 10.1093/COMNET/CNAE001.

[8] J. H. Patil, P. Potikas, W. B. Andreopoulos, and K. Potika, "Community Detection Using Deep Learning: Combining Variational Graph Autoencoders with Leiden and K-Truss Techniques," *Information 2024, Vol. 15, Page 568*, vol. 15, no. 9, p. 568, Sep. 2024, doi: 10.3390/INFO15090568.

[9] N. Alotaibi and D. Rhouma, "A review on community structures detection in time evolving social networks," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5646–5662, Sep. 2022, doi: 10.1016/J.JKSUCI.2021.08.016.

[10] Y. Malode, A. Aylani, A. Bhardwaj, and D. Hajoary, "Comparative Analysis of Community Detection Algorithms on the SNAP Social Circles Dataset," Feb. 2025, Accessed: Oct. 14, 2025. [Online]. Available: https://arxiv.org/pdf/2502.04341v1

[11] R. Hernández, I. Gutiérrez, and J. Castro, "Social Network Analysis: A Novel Paradigm for Improving Community Detection," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, pp. 1–22, Dec. 2025, doi: 10.1007/S44196-025-00812-9/TABLES/6.

[12] Y. Xian, P. Li, H. Peng, Z. Yu, Y. Xiang, and P. S. Yu, "Community Detection in Large-Scale Complex Networks via Structural Entropy Game," *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28â•fiMay 2, 2025, Sydney, Australia*, vol. 1, Jan. 2025, doi: 10.1145/3696410.3714837.

[13] L. Brahim, M. Mouad, C. Chihab-Eddine, and I. Ali, "A SURVEY ON COMMUNITY DETECTION: APPLICATIONS, ALGORITHMS, AND CHALLENGES," *J Theor Appl Inf Technol*, vol. 30, no. 12, 2024, Accessed: Oct. 14, 2025. [Online]. Available: www.jatit.org

[14] S. Tokala, M. K. Enduri, T. J. Lakshmi, and K. Hajarathaiah, "Evaluating Community Detection Algorithms: A Focus on Effectiveness and Efficiency," *Journal of Scientometric Research*, vol. 14, no. 1, pp. 62–74, Jan. 2025, doi: 10.5530/JSCIRES.20250839.

[15] A. Diboune, H. Slimani, H. Nacer, and K. Beghdad Bey, "A comprehensive survey on community detection methods and applications in complex information networks," *Soc Netw Anal Min*, vol. 14, no. 1, pp. 1–47, Dec. 2024, doi: 10.1007/S13278-024-01246-5/METRICS.

[16] M. El-Moussaoui, M. Hanine, A. Kartit, M. G. Villar, H. Garay, and I. de la Torre Díez, "A systematic review of deep learning methods for community detection in social networks," *Front Artif Intell*, vol. 8, p. 1572645, 2025, doi: 10.3389/FRAI.2025.1572645.

[17] H. Sadiki, M. Ertel, A. Sadqui, and S. Amali, "A NEW ALGORITHM FOR COMMUNITY DETECTION IN COMPLEX SOCIAL NETWORKS," *J Theor Appl Inf Technol*, vol. 15, no. 11, 2024, [Online]. Available: www.jatit.org

[18] D. Jin *et al.*, "A Survey of Community Detection Approaches: From Statistical Modeling to Deep Learning," *IEEE Trans Knowl Data Eng*, vol. 35, no. 2, pp. 1149–1170, Feb. 2023, doi: 10.1109/TKDE.2021.3104155.

[19] M. Adraoui *et al.*, "A novel three-phase expansion algorithm for uncovering communities in social networks using local influence and similarity in embedding space," *Decision Analytics Journal*, vol. 11, p. 100472, Jun. 2024, doi: 10.1016/J.DAJOUR.2024.100472.

[20] J. Li *et al.*, "A Comprehensive Review of Community Detection in Graphs," Jul. 2024, Accessed: Oct. 14, 2025. [Online]. Available: https://arxiv.org/pdf/2309.11798v4

[21] N. R. Smith, P. N. Zivich, L. M. Frerichs, J. Moody, and A. E. Aiello, "A guide for choosing community detection algorithms in social network studies: The Question-Alignment approach," *Am J Prev Med*, vol. 59, no. 4, p. 597, Oct. 2020, doi: 10.1016/J.AMEPRE.2020.04.015.

[22] S. H. H. Anuar, Z. A. Abas, N. M. Yunos, M. F. Mukhtar, T. Setiadi, and A. S. Shibghatullah, "Identifying Communities with Modularity Metric Using Louvain and Leiden Algorithms," *Pertanika J Sci Technol*, vol. 32, no. 3, pp. 1285–1300, Apr. 2024, doi: 10.47836/PJST.32.3.16.

[23] D. Dhanalakshmi and D. G. Rajendran, "AN ENHANCED COMMUNITY DETECTION METHOD USING LABEL PROPAGATION ALGORITHM WITH ANT COLONY OPTIMIZATION TECHNIQUE," *J Theor Appl Inf Technol*, vol. 15, no. 9, 2024, [Online]. Available: www.jatit.org

[24] Y. Zheng, L. Yi, and Z. Wei, "A survey of dynamic graph neural networks," *Front Comput Sci*, vol. 19, no. 6, pp. 1–29, Jun. 2025, doi: 10.1007/S11704-024-3853-2.

[25] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," *J Big Data*,

vol. 11, no. 1, pp. 1–43, Dec. 2024, doi: 10.1186/S40537-023-00876-4.

[26] A. Singh, S. S. Dar, R. Singh, and N. Kumar, "A Hybrid Similarity-Aware Graph Neural Network with Transformer for Node Classification," *Expert Syst Appl*, vol. 279, p. 127292, Jun. 2025, doi: 10.1016/J.ESWA.2025.127292.

[27] M. Wienczkowski, A. Desta, and P. Ugochukwu, "Geometric Properties and Graph-Based Optimization of Neural Networks: Addressing Non-Linearity, Dimensionality, and Scalability," Feb. 2025, Accessed: Oct. 14, 2025. [Online]. Available: https://arxiv.org/pdf/2503.05761v1

[28] X. M. Zhang, L. Liang, L. Liu, and M. J. Tang, "Graph Neural Networks and Their Current Applications in Bioinformatics," *Front Genet*, vol. 12, p. 690049, Jul. 2021, doi: 10.3389/FGENE.2021.690049.

[29] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, Jan. 2020, doi: 10.1016/J.AIOPEN.2021.01.001.

[30] Y. Wang Author and J. Zhao Author, "Research on the application of graph data structure and graph neural network in node classification/clustering tasks," Jul. 2025, Accessed: Oct. 14, 2025. [Online]. Available: https://arxiv.org/pdf/2507.19527v1.

[31] G. Gkarmpounis, C. Vranis, N. Vretos, and P. Daras, "Survey on Graph Neural Networks," *IEEE Access*, vol. 12, pp. 128816–128832, 2024, doi: 10.1109/ACCESS.2024.3456913.

[32] W. Jiang *et al.*, "Graph Neural Networks for Routing Optimization: Challenges and Opportunities," *Sustainability 2024, Vol. 16, Page 9239*, vol. 16, no. 21, p. 9239, Oct. 2024, doi: 10.3390/SU16219239.

[33] J. H. Tanis, C. Giannella, and A. V Mariano, "Introduction to Graph Neural Networks: A Starting Point for Machine Learning Engineers".

[34] Y. Zhou, H. Zheng, X. Huang, S. Hao, D. Li, and J. Zhao, "Graph Neural Networks: Taxonomy, Advances, and Trends," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 1, Jan. 2022, doi: 10.1145/3495161.

[35] S. Job *et al.*, "Towards Causal Classification: A Comprehensive Study on Graph Neural Networks," Jan. 2024, Accessed: Oct. 14, 2025. [Online]. Available: https://arxiv.org/pdf/2401.15444v1.

[36] W. M. Lim, S. Kumar, and N. Donthu, "How to combine and clean bibliometric data and use bibliometric tools synergistically: Guidelines using metaverse research," *J Bus Res*, vol. 182, p. 114760, Sep. 2024, doi: 10.1016/J.JBUSRES.2024.114760.

[37] N. Yazdanjue *et al.*, "A comprehensive bibliometric analysis on social network anonymization: current approaches and future directions," *Knowl Inf Syst*, vol. 67, no. 1, pp. 29–108, Jan. 2025, doi: 10.1007/S10115-024-02289-Y.

[38] M. Nowakowska, "A comprehensive approach to preprocessing data for bibliometric analysis," *Scientometrics*, pp. 1–35, Sep. 2025, doi: 10.1007/S11192-025-05415-X.

[39] W. Lin, X. Wu, Z. Wang, X. Wan, and H. Li, "Topic Network Analysis Based on Co-Occurrence Time Series Clustering," *Mathematics 2022, Vol. 10, Page 2846*, vol. 10, no. 16, p. 2846, Aug. 2022, doi: 10.3390/MATH10162846.

[40] S. Radhakrishnan, S. Erbis, J. A. Isaacs, and S. Kamarthi, "Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature," *PLoS One*, vol. 12, no. 3, p. e0172778, Mar. 2017, doi: 10.1371/JOURNAL.PONE.0172778.

[41] S. Lozano, L. Calzada-Infante, B. Adenso-Díaz, and S. García, "Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature," *Scientometrics*, vol. 120, no. 2, pp. 609–629, Aug. 2019, doi: 10.1007/S11192-019-03132-W/METRICS.

[42] W. Oubaalla and L. Benhlima, "A NOVEL ALGORITHM BASED ON THE NODE TIGHTNESS DEGREE FOR COMMUNITY DETECTION IN LARGE GRAPHS," *J Theor Appl Inf Technol*, vol. 15, no. 15, 2022, [Online]. Available: www.jatit.org

[43] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3615–3620, Mar. 2019, doi: 10.18653/v1/d19-1371.

[44] R. Dwivedi and L. Elluri, "Exploring Generative Artificial Intelligence Research: A Bibliometric Analysis Approach," *IEEE Access*, vol. 12, pp. 119884–119902, 2024, doi: 10.1109/ACCESS.2024.3450629.

[45] M. A. Noor, J. Sheppard, and J. Clark, "Improving Community Detection in Academic Networks by Handling Publication Bias," Jul. 2025, Accessed: Oct. 14, 2025. [Online]. Available: https://arxiv.org/pdf/2507.20449

[46] B. Chigarev, "Keyword Co-Occurrence Analysis Using the FPGrowth Algorithm. An Example of Energies Journal Bibliometric Data for 2023-2024," Jun. 2024, doi: 10.20944/PREPRINTS202406.1380.V1.

[47] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A Bibliometric Review of Large Language Models Research from 2017 to 2023," *ACM Trans Intell Syst Technol*, vol. 15, no. 5, pp. 1–25, Oct. 2024, doi: 10.1145/3664930.

[48] C. Mejia, M. Wu, Y. Zhang, and Y. Kajikawa, "Exploring Topics in Bibliometric Research Through Citation Networks and Semantic Analysis," *Front Res Metr Anal*, vol. 6, p. 742311, 2021, doi: 10.3389/FRMA.2021.742311/FULL.

[49] R. Wu and S. Pourroostaei Ardakani, "Towards Scientific Knowledge Graphs: Dependency Graph Analysis Using Graph Neural Networks for Extracting Scientific Relations," *Electronics 2025, Vol. 14, Page 2276*, vol. 14, no. 11, p. 2276, Jun. 2025, doi: 10.3390/ELECTRONICS14112276.

[50] R. Bhattacharya, N. K. Nagwani, and S. Tripathi, "A community detection model using node embedding approach and graph convolutional network with clustering technique," *Decision Analytics Journal*, vol. 9, p. 100362, Dec. 2023, doi: 10.1016/J.DAJOUR.2023.100362.

[51] W. Fu and S. Akbar, "Expert Profile Identification From Community Detection on Author-Publication-Keyword Graph With Keyword Extraction," *IEEE Access*, vol. 12, pp. 27918–27930, 2024, doi: 10.1109/ACCESS.2024.3368003.

[52] A. Fiallos, K. Jimenes, C. Vaca, and X. Ochoa, "Scientific communities detection and analysis in the bibliographic database: SCOPUS," *2017 4th International Conference on eDemocracy and eGovernment, ICEDEG 2017*, pp. 118–124, Jun. 2017, doi: 10.1109/ICEDEG.2017.7962521.

[53] H. Akkineni, M. Madhu Bala, V. Takellapati, M. Nallamothu, and S. Yadlapati, "MEASURING RESEARCH INTEREST SIMILARITY AMONG AUTHORS USING COMMUNITY DETECTION," *J Theor Appl Inf Technol*, vol. 15, no. 11, 2022, [Online]. Available: www.jatit.org

[54] R. Kiruthika and N. Radha, "Author-Centric Pattern Detection in Scopus Citation Network via Community Structures," pp. 1–15, 2025, doi: 10.1007/978-981-97-7839-3_1.

[55] S. Radhakrishnan, S. Erbis, J. A. Isaacs, and S. Kamarthi, "Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature," *PLoS One*, vol. 12, no. 3, p. e0172778, Mar. 2017, doi: 10.1371/JOURNAL.PONE.0172778.

[56] R. Kiruthika and N. Radha, "Analyzing the Citation Networks Using Community Detection Approaches: A Review," vol. 398, pp. 181–194, 2025, doi: 10.1007/978-981-97-5200-3_13.

[57] A. Aldabobi, A. Sharieh, and R. Jabri, "AN IMPROVED LOUVAIN ALGORITHM BASED ON NODE IMPORTANCE FOR COMMUNITY DETECTION," *J Theor Appl Inf Technol*, vol. 15, no. 23, 2022, [Online]. Available: www.jatit.org

[58] S. Chandrasekharan *et al.*, "Finding scientific communities in citation graphs: Articles and authors," *Quantitative Science Studies*, vol. 2, no. 1, pp. 184–203, Apr. 2021, doi: 10.1162/QSS_A_00095.

[59] H. Z. Ahmed and A. M. N. Alzubaidi, "Enabling the Community Detection with Graph Autoencoder using node features in the Social Networks," *Proceedings - CSCTIT 2022: 5th College of Science International Conference on Recent Trends in Information Technology*, pp. 169–174, 2022, doi: 10.1109/CSCTIT56299.2022.10145756.

[60] J. A. Garrido-Cardenas *et al.*, "The Identification of Scientific Communities and Their Approach to Worldwide Malaria Research," *Int J Environ Res Public Health*, vol. 15, no. 12, Dec. 2018, doi: 10.3390/IJERPH15122703.

[61] Y. Zhang, E. Levina, and J. Zhu, "Community Detection in Networks with Node Features," *Electron J Stat*, vol. 10, no. 2, pp. 3153–3178, Sep. 2015, doi: 10.1214/16-EJS1206.

[62] Z. Wang, G. Shen, D. Mao, X. Wang, Z. Zhang, and D. Quan, "Self-Supervised Community Detection Algorithm Based on Node Feature Convolution," *Proceedings of the 18th IEEE Conference on Industrial Electronics and Applications, ICIEA 2023*, pp. 1214–1219, 2023, doi: 10.1109/ICIEA58696.2023.10241919.

[63] S. Yuan, H. Zeng, and C. Wang, "Community Detection based on Node Relationship Classification," *International Conference on Pattern Recognition Applications and Methods*, vol. 2, pp. 596–601, Feb. 2022, doi: 10.5220/0010850600003122.

[64] G. A. O. Guangliang, W. Liang, M. Yuan, H. Qian, Q. Wang, and C. A. O. Jie, "Triangle-oriented Community Detection considering Node Features and Network Topology," *ACM Transactions on the Web*, vol. 18, no. 1, Jul. 2022, doi: 10.1145/3626190.