IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 September 2024, accepted 5 October 2024, date of publication 10 October 2024, date of current version 3 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3477992

RESEARCH ARTICLE

Exploring Topic Coherence With PCC-LDA and BERT for Contextual Word Generation

SANDEEP KUMAR RACHAMADUGU[1,2], T. P. PUSHPHAVATHI[1], SURBHI BHATIA KHAN[3,4], AND MOHAMMAD ALOJAIL[5]

[1]Department of Computer Science and Engineering, M. S. Ramaiah University of Applied Sciences, Bengaluru, Karnataka 560054, India
[2]Department of Computer Science and Engineering, G. Pulla Reddy Engineering College, Autonomous, JNTUA, Ananthapuramu, Andhra Pradesh 518007, India
[3]School of Science, Engineering and Environment, University of Salford, M5 4WT Manchester, U.K.
[4]Adjunct Research Faculty, Centre for Research Impact and Outcome, Chitkara University, Punjab 140401, India
[5]Management Information System Department, College of Business Administration, King Saud University, Riyadh 11587, Saudi Arabia

Corresponding author: Sandeep Kumar Rachamadugu (sandy.racha@gmail.com)

This work was supported by the Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia under Grant IFKSUOR3-176-6.

ABSTRACT In the field of natural language processing (NLP), topic modeling and word generation are crucial for comprehending and producing texts that resemble human languages. Extracting key phrases is an essential task that aids document summarization, information retrieval, and topic classification. Topic modeling significantly enhances our understanding of the latent structure of textual data. Latent Dirichlet Allocation (LDA) is a popular algorithm for topic modeling, which assumes that every document is a mix of several topics, and each topic will have multiple words. A new model similar to LDA, but a better version called Probabilistic Correlated Clustering Latent Dirichlet Allocation (PCC-LDA) was recently introduced. On the other hand, BERT is an advanced bidirectional pre-trained language model that understands words in a sentence based on the full context to generate more precise and contextually correct words. Topic modeling is a useful way to discover hidden themes or topics within a range of documents aiming to tune better topics from the corpus and enhance topic modeling implementation. The experiments indicated a significant improvement in performance when using this combination approach. Coherence criteria of are utilized to judge whether the words in each topic accord with prior knowledge, which could ensure that topics are interpretable and meaningful. The above results of the topic-level analysis indicate that PCC-LDA consistency topics perform better than LDA and NMF(non-negative matrix factorization Technique) by at least 15.4%,12.9%($k = 5$) and up to nearly 12.5% and 11.8% ($k = 10$) respectively, where k represents the number of topics.

INDEX TERMS BERT, key phrases, LDA, topic coherence, topic modeling.

I. INTRODUCTION

Topic modeling is a popular method for analyzing large document collections.However it is challenging to compare the algorithms of topic models and evaluate their outputs. As existing metrics provide a mixed picture, selecting a suitable measure for outcome assessment is difficult. Several topic models have yielded encouraging results, however the validity of these findings has not yet been properly examined. One study examined commonly used topic modeling techniques and evaluated their effectiveness against established clustering. For any topic modeling, studying the characteristics of coherency is regarded as a complicated task in which semantic coherence plays a vital part together with other evaluation measures for the confirmation of quality. This study restricts topic coherence because the bag-of-words representation from a topic modeling method cannot be validated using only cross-validated coherent and high-quality projected terms for decades [1].

While embedding-based methods and contextual word embeddings, such as BERT, can capture semantic relationships and improve natural language processing performance, they also have computational, memory, data, interpretability, and fine-tuning complexity issues.In order to gain a more

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

in-depth understanding of the context of each word inside a sentence or document, contextual word embeddings such as BERT encode information from the words that surround it. This has the potential to capture the nuances of language. Contextual word embeddings, particularly deep learning based models such as BERT, are frequently criticized for their lack of interpretability and transparency, which makes it difficult to comprehend the process by which individual findings are generated [2]. Rapid advances in deep learning and big data have made furniture topic modeling a mainstream approach for making decisions regarding data. On a large scale, people use different applications such as customer data mining, sentiment analysis and opinion mining.This includes model and parameter selection and,evaluation of performance in terms of topic coherence, scalability,and interpretability of results.

Key phrase extraction is crucial for analyzing customer reviews, understanding feelings, and identifying developing trends. This is a difficult task that requires full understanding of the document's structure, meaning, and challenging language. However, this is important for NLP and information retrieval. The primary goal of the key phrase is to encapsulate the principal concepts included in this study. Topic coherence ensures that the most significant words related to a topic are relevant, interconnected, and aligned with essential phrases in the text. Key terms exhibited a more robust correlation with the documents.

The purpose of key−phrase extraction in the context of scientific articles is to recognize and extract phrases that are significant and representative, with the intention of providing a concise summary of the primary concepts, ideas, and themes presented within the document. Document retrieval, classification, and recommendation systems are examples of jobs that benefit from such information. In addition, the extraction of key phrases can help readers quickly comprehend the primary material of the document, as well as assist them in making decisions, such as whether or not they should read the article. [3]. Automatic key phrase extraction (AKE) recognizes and extracts relevant, descriptive, and emotive phrases or terms from a document or set of documents without human intervention. We extract key phrases to capture the document's core subjects, themes, and vital information, thereby facilitating information retrieval, NLP, and text mining.Problems with current Automatic Key Phrase Extraction (AKPE) methods include low accuracy and performance, redundancy, infrequency of latent Key phrases,scaling to large and growing data [4]. Automatic key phrase extraction(APKE) is limited to small−scale data. This study focuses on how to use word embeddings to make it easier to find keywords in scientific papers. The authors focused on the automatic keyword recognition in scientific articles. Relevant words and phrases were extracted from the documents and,abstract keywords were generated. The authors researched the utilization of word embedding vectors to improve the keyword mining and construction processes. The authors also addressed the limitations of their approach, such as the reliance on domain-wide word embeddings and potential bias in the evaluation datasets. In addition, they emphasize their commitment to future work by exploring the underlying knowledge that appears in word embeddings and always working for better results [5].

A. BACKGROUND AND MOTIVATION

By automatically identifying the most important key phrases in a document, we can improve search engine performance, generate concise summaries, and gain insight into the main themes of text collection.

One of the important methods used to find useful thematic roles or topics from a large document is topic modeling, which has become an essential task using NLP.LDA is a generative probabilistic model for topic modeling. Nevertheless, the conventional LDA method has drawbacks in terms of capturing the intricate semantic connections between words within topics, resulting in less than ideal topic coherence [6]. Some experts have expressed interest in using sophisticated NLP techniques such as neural language models and BERT to enhance multiple NLP tasks, one of which is topic modeling. Some NLP tasks use BERT to render words that are more coherent and relevant by creating contextual representations. This helps to create high-level semantic links between words [7]. LDA and BERT embeddings were employed to analyze and classify the items. We employed the LDA technique to conduct topic modeling to identify latent topics within the text. We used the LDA algorithm for topic modeling to extract latent topics from the text. In addition, we employed BERT embeddings to obtain a dense vector encoding the pre-processed sentence text that would eventually aid in our clustering. In our work, we surveyed state-of-the-art topic modeling techniques (e.g., LDA) to identify suitable methods that would be more effective in improving the accuracy and efficiency of aspect-based opinion mining to provide an automatic way to extract latent topics from large textual data. Given the real-world nature of AB testing during mining, this further stresses the inevitable requirement for correct and prompt evaluation [8]. The motivation behind our concept stems from the need to address the limitations of traditional LDA and further enhance topic coherence by integrating cutting-edge techniques. We recommend the integration of the PCC component with LDA as an enhancement to conventional LDA, which employs probabilistic correlation clustering to enhance the accuracy of topic modeling. Furthermore, our objective was to capitalize on the capabilities of BERT to produce contextual words that more accurately represent the semantic underpinnings of the topics.

By exploring the fusion of PCC-LDA and BERT for word generation, we seek to advance state-of-the-art topic modeling and contribute to the development of more interpretable and coherent topic representations.

Our goal was to provide researchers and practitioners with a robust framework for extracting meaningful insights from

textual data, thereby facilitating a deeper understanding of complex topics across various domains.

Unwinding coherence within topics plays a critical role in interpreting natural languages. Advancements in AI models such as BERT, have improved their performance. AI models, such as BERT have largely resolved the issue of machines not understanding the context and hence lead to lower topic deviation. The entire sentence was considered when predicting a single word in BERT. This method has yielded considerable success in human-machine interactions by making such addresses more natural and committed. This progress was even more enhanced when the topic model was considered which generated more meaningful and coherent topics compared to current neural models and traditional bag-of-words topic models. We employed the LDA technique to conduct topic modeling, to identify latent topics within the text. In addition, we utilized BERT embeddings to generate dense vector representations of the text, thereby enhancing clustering performance. [9].

Planning: In this section, we outline our research questions and search plans. In particular, we asked, ''How were syntactic and semantic contextual words quantified and identified in generating the contextual words?''

Research Issues: We focus our work on the following primary research questions: The following five research questions provide additional information about this topic:

RQ1: Which metrics are the most useful for assessing the coherence of themes generated by PCC-LDA and BERT?

RQ2: What is the difference between the topic coherence of the themes created by BERT and PCC-LDA?

RQ3: How does employing BERT's contextual embeddings affect the word creation quality in comparison with the conventional LDA method?

RQ4: What are the advantages and difficulties of combining BERT with PCC-LDA for improved topic modeling?

RQ5: In terms of topic coherence and contextual word creation, a hybrid model combining PCC-LDA and BERT performs better than either model alone.

The assessment metric section employed and computed the RQ1 and RQ2 metrics to measure the likelihood of a word occurring together within a topic as opposed to each word occurring separately. The impact of BERT is discussed in Section IV for RQ3, and RQ4 limitations are discussed in the conclusion section under future scope. The solution for RQ5 was proven by integrating LDA with the PCC component in a hybrid model.

B. CONTRIBUTION OF THE PAPER

In the field of NLP, this approach substantially contributes to the advancement of word production and topic modeling problems.

- This strategy emphasizes topic coherence in contrast to previous methods that may highlight the importance of individual words. Prioritizing topic coherence enhances the relevance of extracted key phrases to a document's primary issues. This method aims to improve topic coherence in text data by combining BERT for word creation with PCC-LDA for topic modeling.
- NMF gives a semantic presentation of topics, and they give us the ability to interpret them. This, in turn, facilitates the identification of frequent words that serve as the dominant keywords for each topic. Compared to LDA, which incorporates the same spherical topics, NMF only includes semantically similar words in a topic, thereby reducing its coherence and interpretability.
- BERT is a machine learning model that generates contextual phrases to improve subject coherence and relevance. In contrast, the PCC-LDA improves the conventional LDA by including probabilistic correlation clustering.
- This study introduces an innovative approach for extracting key phrases that prioritize topic coherence. It combines BERT and PCC-LDA to improve the coherence of topics. Additionally, this study provides an empirical proof of the efficiency of this method.

C. ORGANIZATION OF THE PAPER

The reminder of this paper is organized as follows: The second section summarizes the literature review, which assesses pertinent research and the LDA and BERT theoretical frameworks. Section III describes the development of the proposed method and its integration with the BERT. Section IV presents the findings of our investigation and analyzes them in relation to the previous research and theoretical frameworks. The 5th section provides a brief synopsis of the study's key points, as well as its consequences and recommendations for further study.

II. LITERATURE REVIEW

A. OVERVIEW OF TOPIC MODELING TECHNIQUES

NLP uses topic modeling to identify the main themes or topics in a collection of text documents. Every document in LDA assumes a combination of topics, with each topic represented by a word distribution. It calculates the likelihood that each word in a document is associated with a specific topic, thereby enabling the detection of the most probable topics within the document collection [10].

This article discusses a range of approaches to topic modeling, focusing on the analysis of documents and short−text data. This study demonstrated the importance of topic models in organizing and summarizing collections of unstructured textual data. In short, this study is crucial as it has pointed out challenges such as noise, scalability, and applications to deal with brief text processing. It also provides evaluation metrics of point wise mutual information(PMI), normalized point wise mutual information (NPMI) and purity to measure the quality of coherence for the model. Topic modeling is widely used in different areas such as industrial applications, bio informatics, recommender systems and financial analysis. This shows how important and flexible topic models draw interesting information from textual data [11].

Topic modeling methods are a very useful framework for extracting the main concepts from an entire text corpus. The LDA and BERT techniques were used for context-based word generation.

A collection of written texts can be analyzed using a technique called topic modeling that exposes previously hidden semantic patterns. Using this method, it is possible to automatically determine the underlying subjects that are discussed in a collection of documents as well as the manner in which these topics are spread throughout the entire collection. Natural language processing, machine learning, and information retrieval are only a few fields that extensively use this technology, and are widely employed in a variety of fields.

The basic concept of topic modeling is to integrate latent themes to represent each document as a collection of words that commonly appear together, with each topic being a probability distribution over this set of words. We aimed to infer this distribution from a collection of studies, assuming that these topics originated from a prior probability distribution. LDA is a generative probabilistic model, which assumes that each word in a document originates from a single topic. The distribution of topics in the collection and the distribution of words within each topic determines the probability of a word. Once trained, the model can identify the most likely themes for a given document and categorize new documents into topics. Topic modeling can be employed to perform other forms of analysis such as content recommendation, trend discovery and exploitation, document summarization,and identification of new documents in an existing database. Even when there are no existing labels or groups, topic modeling is helpful in revealing hidden semantic structures in a corpus of texts. This helps academics learn what a corpus is about, that is it represents significant ideas and themes [12].

B. INTRODUCTION TO LATENT DIRICHLET ALLOCATION

NLP and machine learning use LDA as a popular topic-modeling technique. This generative probabilistic model allows the discovery of hidden topics within a collection of text documents. LDA is a well-known topic modeling method that helps identify the main ideas in a group of text files. Each document is believed to contain a variety of topics and words. LDA determines these topics by examining the frequency with which certain words appear together. To capture the complex relationships between words and topics in texts, we integrated neural network models such as BERT, with LDA. Overall, topic modeling methods are essential for organizing and analyzing large amounts of text data because they automatically group and extract meaningful topics that help us to better understand the content of documents [13].

LDA assumes that each document is a mix of themes and that each word comes from one source. By analyzing the word distribution across topics, LDA can help identify the underlying themes or topics presented in a set of documents. Tasks such as document clustering, topic labeling,and information retrieval can be utilized as powerful tools to uncover the latent structure of textual data [14]. To indicate the latent thematic structure, LDA projects hidden variables into a corpus in the form of themes and topic proportions per article. According to probabilistic sampling criteria, LDA generates articles from an imaginary random process. We only see words in text, and we must determine the hidden structure using statistical inference to answer the following question: Which hidden structure or topic model generates these documents? LDA, a topic modeling technique, enhances the accuracy of identifying topics or groups in scientific abstracts [15].

C. INTRODUCTION TO NON-NEGATIVE MATRIX FACTORIZATION TECHNIQUE

NMF is a topic modeling technique that draws its foundation from linear algebra. It essentially takes a matrix and converts it into two smaller matrices, with the stipulation that all three of those matrices have no negative elements. This, in turn, makes it ideal for datasets where the data are naturally non-negative, such as text data represented by word counts or frequencies. NMF works well in reducing the dimensions of the feature vectors in text data.This method simplifies the handling of high-dimensional data while preserving the essential information. This is very helpful for handling large text corpora for topic modeling [16]. NMF is a non-probabilistic matrix factorization, therefore unlike probabilistic models such as LDA, NMF does not consider uncertainty when performing decomposition. It factorizes a term-document matrix into two non-negative lower-ranking matrices, W and H, using topic model LDA. The non-negativity of topics will give a more human−understandable interpretation as it avoids negative values that could make the topic difficult to understand.

D. LIMITATIONS OF LDA

Researchers have widely investigated and employed natural language processing, but earlier research has revealed significant drawbacks. LDA faces challenges in distinguishing between overlapping topics and capturing textual language. To obtain accurate results, LDA must tune numerous parameters, and its output depends on the quality and quantity of data.

Traditional LDA may struggle to accurately model short documents because it relies on the presence of multiple words to infer topics effectively.

LDA operates without any contextual relevance, therefore it might be difficult to capture the larger context and smaller oscillations in user feedback data. This constraint may slow down the model by understanding user concepts and experiences. LDA, if it cannot correctly understand words or topics can misinterpreted user feedback,which is restricted by the issues highlighted above, thus slowing down its potential to pinpoint user themes and conditions [17].

LDA struggles with short texts because it does not understand the word connections. Another metric may be to determine if the top-level terms of a topic is to closely related to each other. One way to determine the quality (and

conciseness) of a topic coherence is by using coherence, which plays an important role in how LDA works. However, one of the main limitations of LDA is that it looks into bags-of-words for topics and does not consider their actual meaning. Moreover, the proposed method performed better than LDA in terms of the coherence of a topic metric [18].

LDA is not important for text sequences because it addresses this problem. This means that critical information regarding the layout and context of a document is lost. Another problem with LDA is choosing a number of topics, which can be difficult if the corpus is large and diverse and, has many themes or sub-themes. Another issue with LDA is its inability to explain the hierarchical structure of the corpus or demonstrate the relationships between topics. LDA assumes that each word in a document is conditionally independent, whereas a genuine language may not. The order of words in a document may be relevant for certain purposes, however LDA does not. LDA applies the Bag-of-Words (BoW) model, which ignores text syntax and grammar and treats each document as a set of words. Information loss leads to poor topic modeling. Finally, LDA needs to set various hyperparameters, such as the number of topics, which can be challenging and time-consuming.

E. LIMITATIONS OF NMF

NMF is a non-probabilistic linear algebraic algorithm. This method relies on matrix factorization and TF-IDF-transformed data. Occasionally, this method may yield less interpretable topics, because it does not replicate the probabilistic generation of words within a topic using LDA. LDA coherence is heavily dependent on tuning its hyperparameters (number of topics, alpha, and beta). Adjusting the hyperparameters enhances the overall coherence score for the LDA. Applying meticulous tuning can yield superior topics compared to NMF, which might not exhibit as much hyperparameter sensitivity or optimization. Additionally, the non-negative Young model also has an interpretability problem that can limit NMF because it requires all entries in its matrices to be non-negative. This can have consequences for the perceived topic coherence of the NMF-generated topics. We observed that NMF, compared with the LDA, produces more coherent topic word elements in large text corpora with longer documents. This observation highlights the potential challenges that NMF may encounter in maintaining the topic coherence in such scenarios. According to this study, NMF generally resulted in significantly more noisier topics than LDA. This is probably due to the fact that NMF does less statistical modeling than LDA, and therefore it may generate less coherent topics. The matrices are easier to understand because NMF enforces non-negativity constraints. However, the generated topics may not be as coherent and informative as LDA topic models, which means they will not work as well in applications that require a high level of interpretability [19]. Despite being a topic modeling algorithm at its core, the main drawback of NMF is its lack of powerful embedded meaning identification within the corpus. This is attributed in part to the reliance on the Frobenius norm and hence some challenges with the interpretability of findings [20]. This view of coherence metrics provides a complete means to validate semantic coherence for topics created by LDA, and NMF allows us to compare their performance selectively. Unlike probabilistic models such as LDA, NMF does not provide a probabilistic interpretation of topics. The automatically distributed and sparse nature of topics makes it challenging to evaluate them in terms of coherence, a probabilistic measure that provides significant information. Although NMF is typically faster than other methods such as LDA, it can still run into scalability issues with massive amounts of data. This can lead to worse topic coherence when working with longer text corpora [21].

F. INTRODUCTION TO BERT

Google has launched BERT, a new natural English processing system. It includes the code before and after so that you can learn what it does. The more accurate BERT is in understanding language, the better it can stream text into buckets such as answering questions and translating languages. The BERT approach has been proven to work well in many NLP tasks, including some examples of text classification such as determining topics for scientific papers. Many NLP tasks, including finding topics in scientific articles, have demonstrated exceptional efficiency of the BERT method. We trained BERT, which is contextualized language model, and achieved the highest scores on various NLP measures. We trained BERT, a contextualized language model, and achieved the highest scores on a several NLP measures.

This post briefly introduces BERT using Google, a language model. Deep learning models such as BERT learn from a large amount of data and can be used to solve most natural language processing (NLP) tasks such as sentiment classification, named entity recognition, and question answering. BERT undergoes pre-training on a large dataset consisting of various sources, such as Wikipedia and the Book Corpus. Through this process, BERT now has a better understanding of language in general, and can then use that knowledge to easily answer any task-specific question,which is where fine-tuning comes into play. BERT is a neural network design that is leveraged as a computer system analyst for this task.

Further, BERT used a novel training strategy as well that makes it better understand the context and importance of words faster than traditional word embeddings without any contextual information. Finally, this article explains that BERT allows researchers to fine-tune the pre-trained model to improve the performance on a large number of Natural Language Processing (NLP) tasks [22]. In addition, BERT Topics has exhibited a high level of big data analysis capability, making it an optimal choice for analyzing large sets of texts [23].

BERTopic is a topic-based BERT-Optimized Unsupervised Topic Clustering method that combines language embeddings with class-based TF-IDF for better results and provides interpretability of topics. BERTopic might alleviate these

problems of LDA, which performs poorly with short and out-of-domain texts without providing a satisfactory general solution because different domains have disparate terminology for similar ideas [24].

G. PREVIOUS WORK ON IMPROVING THE TOPIC COHERENCE WITH BERT

Researchers have also measured text-topic coherence using BERT embeddings. Researchers can assess text coherence and topicality by comparing BERT embeddings with sentences or paragraphs.According to the previous research, BERT embeddings or fine-tuning the model for specific tasks can improve topic coherence in text production tasks [14].

This demonstrates how advanced language models, such as MPNet, and topic modeling techniques, such as CTM, can improve user feedback data comprehension. This method extracts insights from unstructured text data to improve context-aware topic modeling [17]. This study shows that previous work on improving topic coherence using BERT has shown great promise in capturing lexical-semantic connections between words. We fine-tuned a previously trained BERT model to predict the likelihood of words appearing in a text corpus. This is done by learning detailed descriptions of each word's context, considering both local and global contexts [18]. introducing entity topic models, that use entity information to identify coherent subjects. The drawback of this study was its dependence on entity information, which may not always be accessible or applicable to textual datasets. Additionally, implementing the model in real-world applications can pose challenges owing to its potential complexity [25].

This study introduces an advanced BERTopic framework and algorithm with the objective of enhancing topic coherence and diversity in topic modeling assignments. The limitations of this study include the computational complexity of the enhanced framework and algorithm, which may limit its scalability to large datasets, and the need for extensive parameter tuning for optimal performance [20].

In this study, the authors presented an integrated clustering and BERT framework for improved topic modeling. This method employs clustering techniques and BERT embedding to enhance generated topics. The complexity of merging clustering and BERT frameworks, which may require specific knowledge and skills, and the necessity for considerable computer resources for training and inference, may limit this study [26]. This study found that LDA and NMF work well for topic modeling of long textual data, however BERTopic and Top2Vec work better for shorter texts such as social media content [27]. The goal of using BERT for topic coherence is to improve the quality and applicability of generated topics. Better topic coherence with BERT can be achieved in this study [28] Pre-training contextualized document embeddings can achieve BERT, leading to higher coherence ratings for the generated topics. Coherence is an important statistic that helps determine the importance and specificity of topics and, the semantic similarity of high-scoring terms within a topic cluster measures coherence.

III. METHODOLOGY

A. EXPERIMENT SECTION

Data Collection and Analysis

The dblpv12 dataset can be found in aminer.org/citation. Several publications and citations have been published in literature. It includes a digital library and a bibliography of articles and proceedings. The JSON collection included journals and proceedings. The JSON dataset consists several features. However, the article ID, title, keywords, abstract, and references are crucial to the proposed work. We intended to use this dataset for research purposes only. Many sources exist, including Microsoft Academic Graphs and ACM [29]. Given a collection of documents, the PCC-LDA model first preprocesses the text, tokenizes, and represents documents as bags of words. The following figure is based on a topic modeling study. Figure 1 and 2 shows how the top 30 most relevant terms for the five topics and ten topics were spread out.

Each subject had a unique set of terms that represented the most significant words related to the topic content. The size and frequency of terms for each topic demonstrate their importance and usefulness. This distribution allowed us to learn more about the main ideas and trends of the dataset. This picture helps us see how the different and related areas of focus fit the chosen topic. From 1 and 2 it is clear that as the number of topics increases, the relevance of words in each topic also decreases. A dictionary of words and a document-term matrix with rows for documents and columns for word frequency was created. Table 1 illustrates the document-term matrix, which presents the significance of each term across multiple topics as determined by the topic modeling results.

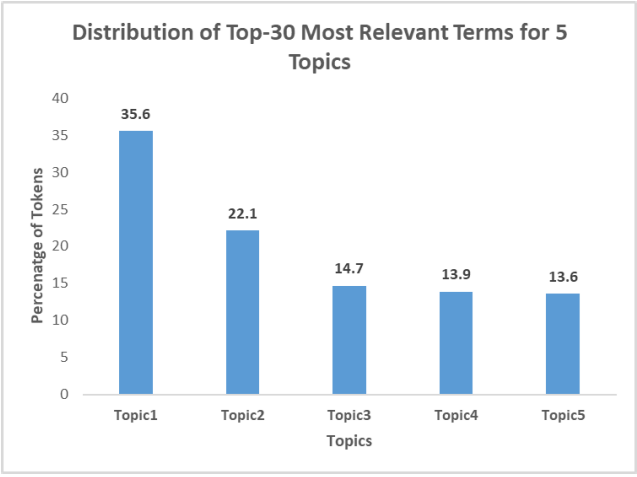

FIGURE 1. Distribution of Top-30 most relevant terms for $k = 5$.

The PCC-LDA model then trained the LDA model using a document-term matrix. It estimates latent topics and their distributions over words in the entire document collection.

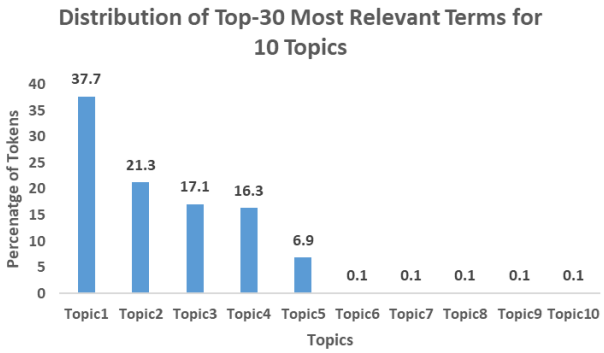

FIGURE 2. Distribution of Top-30 most relevant terms for $k = 10$.

TABLE 1. Document-term matrix.

Term	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
use	0.008	0.004	0.007	0.008	0.014
user	0.006	0	0.004	0	0
paper	0.006	0.006	0.006	0.010	0.008
system	0.006	0.004	0.008	0	0.007
set	0.006	0	0	0	0.005
data	0.005	0.004	0.004	0.005	0.005
design	0.004	0	0	0.005	0
show	0.004	0	0	0	0
result	0.004	0	0	0.005	0
inform	0.004	0	0	0	0
propos	0	0.009	0.006	0	0.004
present	0	0.005	0.007	0.007	0.004
develop	0	0.004	0.004	0	0.005
architectur	0	0.004	0	0	0
project	0	0.004	0	0	0
provid	0	0.004	0	0	0
move	0	0	0.005	0	0
base	0	0	0.006	0	0
imag	0	0	0	0.007	0
approach	0	0	0	0.006	0
improv	0	0	0	0.006	0
gener	0	0	0	0.005	0
method	0	0	0	0	0.006
problem	0	0	0	0	0.005

B. OVERVIEW OF PROPOSED METHOD

By considering the correlation between topics and document clusters, PCC-LDA topic modeling seeks to identify themes or topics within a collection of documents. PCC extends traditional LDA by introducing correlations between topics and document clusters. This assumption implies that topics exhibit correlations within document clusters, rather than being independent. This correlation reflects the tendency for certain topics to co-occur more frequently in specific subsets of documents.

C. CONCEPTUAL FRAMEWORK

Figure 3 depicts the framework for generating the five topics by adding the PCC component to LDA. The framework starts by preprocessing the abstracts using text−processing techniques such as stop removal, stemming, and lemmatization. In the next stage, the LDA and PCC components receive the texts, generate five topics, and compute the relevancy of words between topics using the topic coherence metric.

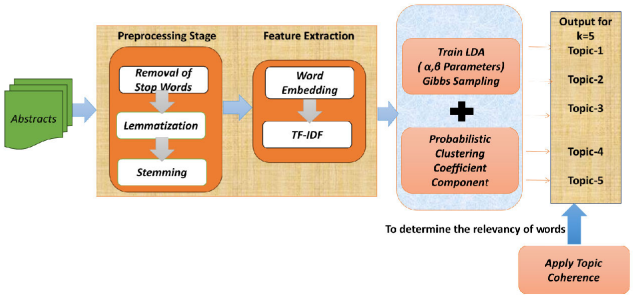

FIGURE 3. PCC−LDA framework.

After training the LDA model, PCC-LDA adds a PCC component to the model. The PCC component represents the correlation between topics and clusters of documents.

It constructs a PCC matrix, where rows represent document clusters, columns represent topics, and each cell represents the strength of the correlation between the topic and cluster.We normalized the PCC matrix to form a probability distribution, ensuring that the sum of the probabilities for each cluster is 1.The LDA model was incorporated into the PCC, allowing it to capture the correlation structure between topics and document clusters.

Finally, the PCC-LDA model extracts topics from each cluster of documents using both the natural topic distributions learned by the LDA and correlations found by the PCC component.The topics were represented as word distributions, and the most probable words for each topic were extracted and displayed. The following is the PCC-LDA algorithm, which represents the Probabilistic Correlated Clustering.This algorithm Probabilistic Correlated Clustering (PCC) and Linear Discriminant Analysis (LDA) to reduce dimensionality and improve classification performance, as detailed below Algorithm 1. The PCC−LDA consists of four stages preprocessing stage, LDA training model, incorporating PCC components, topic extraction

D. EXPERIMENTAL PROCEDURE

Model Training: Similar to a network architect, the PCC-LDA model begins by preprocessing the text, which involves tokenizing and representing documents as bag of words, whereas dictionary $\mathcal{V}$ contained a set of distinct words. Documents $\mathcal{D}$ are constructed, and a matrix $\mathbf{X}$ is created, with each row representing a document and each column representing the frequency of words in that document.

Subsequently the PCC-LDA model used a document term matrix to train the LDA algorithm. The LDA model calculates latent themes and their probabilities of occurrence for all words in the document set. The distribution of topic words, denoted as Φ, and the distribution of documents and topics, denoted as Θ, were randomly initialized. We employ Gibbs sampling, also known as variational inference, to iteratively update the matrices Φ and Θ until convergence is achieved. The distributions obtained as a result were:

$$\Phi_{k,w} = P(w \mid z = k) \quad \text{and} \quad \Theta_{d,k} = P(z = k \mid d)$$

Algorithm 1 PCC-LDA Algorithm

Require: Collection of documents $\mathcal{D}$, Number of topics K, Number of clusters C

Ensure: Most probable words for each topic in each cluster

Model Training:

Tokenize each document $d_i \in \mathcal{D}$ into words.

Construction of dictionary $\mathcal{V}$ of unique words.

Create a document-term matrix $\mathbf{X}$ where X_{ij} is the frequency of word j in document i.

Initialize Φ (topic-word distribution) and Θ (document-topic distribution) randomly.

Gibbs sampling or variational inference is used to estimate the latent topics and their distributions:

$$\Phi_{k,w} = P(w \mid z = k) \quad \text{and} \quad \Theta_{d,k} = P(z = k \mid d) \tag{1}$$

Update Φ and Θ until convergence.

Incorporating PCC Component:

Cluster document $\mathcal{D}$ into C clusters using a clustering algorithm (e.g., k-means) based on Θ.

Construct the PCC matrix $\mathbf{P}$ where P_{cj} is the correlation strength between cluster c and topic j:

$$P_{cj} = \frac{\sum_{d \in c} \Theta_{d,j}}{\sum_{d \in c} \sum_k \Theta_{d,k}} \tag{2}$$

Normalize $\mathbf{P}$ to form a probability distribution:

$$\sum_j P_{cj} = 1 \quad \forall c \tag{3}$$

Initialize an empty list Φ_c

for each cluster $c = 1$ to C **do**

  Initialize Φ_{cluster} as a zero matrix of size $K \times V$

  for each topic $k = 1$ to K **do**

    Adjust the topic distribution:

$$\Phi_{\text{cluster}}[k, :] = P[c, k] \times \Phi[k, :] \tag{4}$$

  end for

  Normalize Φ_{cluster}:

$$\Phi_{\text{cluster}} = \frac{\Phi_{\text{cluster}}}{\sum \Phi_{\text{cluster}}} \tag{5}$$

  Append the normalized Φ_{cluster} to the list Φ_c

end for

Topic Extraction:

for each cluster $c = 1$ to C **do**

  for each topic $k = 1$ to K **do**

    Extract the most probable words for each topic:

$$\text{Top words} = \arg\max_w \Phi_c[k, w] \tag{6}$$

  end for

end for

First, the Φ and Θ parameters in LDA represent hyper-parameters that control the topic-word distribution and the word-topic distribution, respectively. Φ is the document-topic density, with a low value meaning that each document is likely to contain just a few topics and a high value meaning that each document may contain many of them. The topic-word density parameter, known as **Theta**, determines the appropriate number of words to assign to a topic. We explicitly specified the Φ and Θ hyperparameters, where Φ controls the document topic density and Θ controls the topic-word density in the LDA model using the Gensim library for $k = 5$ and $k = 10$ number of topics for λ relevance to a topic.

INCORPORATING PCC COMPONENT

After training the LDA model, the PCC-LDA integrates a PCC component into the model. Documents $\mathcal{D}$ are grouped into C clusters using a clustering technique such as k-means, which is based on Θ. A PCC matrix $\mathbf{P}$ is created, with rows representing document clusters, columns representing topics, and each cell indicating the level of correlation between the topic and cluster:

$$P_{cj} = \frac{\sum_{d \in c} \Theta_{d,j}}{\sum_{d \in c} \sum_k \Theta_{d,k}}$$

The PCC matrix is then normalized to form a probability distribution:

$$\sum_j P_{cj} = 1 \quad \forall c$$

We adjusted the topic distributions for each cluster using the PCC matrix. To find the adjusted topic distribution Φ_{cluster} for each cluster c, multiply the entry in the PCC matrix $P[c, k]$ that goes with it by the topic-word distribution $\Phi[k, :]$. The adjusted distributions are normalized and stored.

$$\Phi_{\text{cluster}}[k, :] = P[c, k] \times \Phi[k, :]$$

$$\Phi_{\text{cluster}} = \frac{\Phi_{\text{cluster}}}{\sum \Phi_{\text{cluster}}}$$

TOPIC EXTRACTION

Using the topic distributions learned by LDA and the correlations gathered by PCC, the PCC-LDA model extracts topics from each group of documents. The most likely terms were extracted for each cluster c and topic k

$$\text{Top words} = \arg\max_w \Phi_c[k, w]$$

PCC-LDA ALGORITHM WITH ADJUSTED TOPIC DISTRIBUTIONS EXPLANATION

- **Initialization**: An empty list Φ_c was initialized to store the adjusted topic distributions for each cluster.
- **Outer Loop (Cluster Loop)**: The outer loop iterates over each cluster c.
- For each cluster, we initialized Φ_{cluster} as a zero matrix of size $K \times V$ to store the adjusted topic distributions.
- **Inner Loop (Topic Loop)**: The inner loop iterates over each topic k.
- For each topic, To modify the topic distribution for each topic, we multiply the appropriate item in the PCC matrix $P[c, k]$ by the topic-word distribution $\Phi[k, :]$.
- **Normalization**: Once the topic distributions for all topics in the cluster have been modified, we normalize

the adjusted topic distribution Φ_{cluster} to verify that the total of each row is equal to one.
- **Storage**: We append the normalized Φ_{cluster} to the list Φ_c.

1) ADVANTAGES OVER TRADITIONAL LDA AND NMF

By implementing PCC-LDA, researchers can uncover topics prevalent in different communities, shedding light on the thematic composition of the dataset and distribution of topics across various document clusters. This integrated approach combines advanced text analysis techniques with network visualization and exploration to facilitate a comprehensive understanding of complex datasets. The main advantages of PCC-LDA over traditional LDA and NMF are correlations for topics with clusters, the flexible capture of complex structures, and enhanced interpretability. In correlations for topics with clusters, PCC-LDA captures complex relationships among the topics and their correlations to document clusters, providing an in-depth examination of whether certain sets of topics are likely to be present within particular communities.LDA can be highly effective in analyzing topics, but it does not provide a model for categorizing documents into specific clusters.NMF provides limited topic extraction and no information about the hierarchy of topics with document clusters.

In capturing the complex structures PCC-LDA captures the complex relationships among topics and their correlations to document clusters, providing an in-depth examination of whether certain sets of topics are likely to be present within particular communities. LDA can be highly effective for analyzing topics, however it does not provide a model for categorizing documents into specific clusters. NMF provides limited topic extraction and no information regarding the hierarchy of topics within document clusters.

In terms of flexible and enhanced interpretability,PCC-LDA establishes correlations between latent topics and clusters, enabling us to determine whether these communities align with classical interpretability. LDA and NMF both extract topics but do not model the relationships between clusters of topics, therefore interpretations are less clear within document communities.

The overall summary indicates that PCC-LDA combines topic modeling and relevance between groups. The slightly improved interpretability and visualization make it ideal for studying larger and more intricate datasets, highlighting the most significant relationships among topics and document communities. LDA and NMF analyze each topic separately, without considering these relationships, therefore, they are less informative about the structure of the dataset.

E. INCORPORATING BERT FOR CONTEXTUAL WORD GENERATION

This study addresses the importance of automatically extracting keywords from unstructured text data and condensing major ideas, themes, concepts, or arguments into words or phrases that make sense. This study discussed the use of contextual word embedding, particularly BERT, to extract keywords. This demonstrates the usefulness of BERT in obtaining semantic and environmental data for keyword extraction [2]. Algorithm 2 explains the use of transformer-based language models to determine how words are used in a certain topic or community and for what contextual word extraction using the Bidirectional Encoder Representations from Transformers(BERT) model. The basic steps are as follows:

Algorithm 2 Algorithm for Identifying Contextual Words With a BERT

Require: Collection of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$
Ensure: Aggregated contextual words $\mathcal{W}$

for each document $d_i \in \mathcal{D}$ **do**
  tokenize d_i using a pre-trained BERT tokenizer.
  Special tokens [CLS] are added at the beginning and [SEP] at the end of the tokenized document.
end for
for each document $d_i \in \mathcal{D}$ **do**
  Obtain the BERT embeddings $\mathbf{E}_i$ for each token in d_i using a pre-trained BERT model.
end for
for each document $d_i \in \mathcal{D}$ **do**
  for each token t_{ij} in document d_i **do**
    compute the change in embeddings:

$$\Delta \mathbf{E}_{ij} = \mathbf{E}_{ij} - \mathbf{E}_{ij-1} \tag{7}$$

  end for
end for
for each document $d_i \in \mathcal{D}$ **do**
  Identify words with the highest embedding changes $\Delta \mathbf{E}_{ij}$ within d_i.
end for
for each document $d_i \in \mathcal{D}$ **do**
  Obtain contextual words $\mathcal{W}_i$ from d_i based on highest $\Delta \mathbf{E}_{ij}$.
end for
Collect and aggregate contextual words across all documents:

$$\mathcal{W} = \bigcup_{i=1}^{N} \mathcal{W}_i \tag{8}$$

Output the aggregated contextual words $\mathcal{W}$.

EXPLANATION

STEP 1: PRE-PROCESS OF THE DOCUMENTS

Every document d_i in set $\mathcal{D}$ was tokenized using a BERT tokenizer trained in advance. Every tokenized document appends tokens [CLS] and [SEP] to its start and end, respectively.

STEP 2: OBTAIN BERT EMBEDDING

Use BERT embeddings $\mathbf{E}_i$ for all tokens in document d_i from a pre-trained BERT model, for every document d_i in set $\mathcal{D}$. These embeddings capture contextual information associated with each token.

STEP 3: COMPUTE EMBEDDING CHANGES

Calculate the change in embeddings for every token t_{ij} in document d_i.

$$\Delta \mathbf{E}_{ij} = \mathbf{E}_{ij} - \mathbf{E}_{ij-1}$$

This metric quantifies the difference in contextual information between adjacent tokens.

STEP 4: GET CONTEXTUAL WORDS

Determine the words that exhibit the most significant alterations in their embeddings. The change in energy, denoted by $\Delta \mathbf{E}_{ij}$, is calculated for each document d_i. These words are regarded as having the most substantial alterations in context.

STEP 5: APPLY THE ALGORITHM TO EACH DOCUMENT

For every document d_i in set $\mathcal{D}$,the contextual words $\mathcal{W}_i$ are extracted by considering the most significant changes in word embeddings $\Delta \mathbf{E}_{ij}$.

STEP 6: COLLECT AND AGGREGATE CONTEXTUAL WORDS

Gather and compile all documents' contextual words:

$$\mathcal{W} = \bigcup_{i=1}^{N} \mathcal{W}_i$$

Consequently, a vocabulary of contextual words is generated, that represents major changes in the context over the entire document set.

STEP 7: OUTPUT THE AGGREGATE CONTEXTUAL WORDS

The most contextually meaningful words across the entire document are represented by aggregated contextual words $\mathcal{W}$, which should be the output. This procedure ensures that the extracted subjects accurately represent the relationship between topics and document clusters, resulting in a more detailed understanding of topic distributions within specific document groups.

F. PREPROCESSING AND FINE-TUNING BERT

Sentence Construction:

We formulate sentences or short contexts for each word in a topic by selecting relevant sentences from documents in which that word appears. These contexts offer significant insight into word usage within a particular topic or community. Encoding Contextual Information with BERT: We use a pre-trained BERT model to encode contextual sentences. BERT generates contextualized word embeddings that capture the nuanced meanings of words in different contexts. The model considers the entire sentence surrounding the word, rather than just the word itself, to understand its usage within a specific context.

Word Embedding Extraction:

After encoding the context sentences with BERT, we extracted embeddings for both the target word and its context from output of the model. These embeddings represent the contextual representation of a word in the specific context.

Comparison and Similarity Calculation:

To identify words with similar contextual usage to the target word within a given topic or community, we compared the extracted word embeddings. We can quantify the similarity between word embeddings and identify words with comparable contextual usage using similarity measures such as cosine similarity.

Iterative Process for Each Topic and Community:

We followed an iterative process to find contextual words for every word in each topic within a community: a. Select a topic from the community. b. Select a word from that topic. c. Form a context sentence around words. d. Use a pre-trained BERT model to encode the context sentence and extract word embeddings. e. For other words related to the same topic and community, repeat Steps a-d.

G. COMBINING BERT WITH PCC-LDA

Using PCC-BERT-based contextual word extraction, researchers can gain insights into the usage of specific words across different topics and communities. This approach enables a deeper understanding of the semantic relationships between words and their contextual usage patterns within a document corpus. Visualization and analysis of the extracted contextual information facilitate the exploration and interpretation of complex textual data structures. The flow of steps that occurred before using the PCC-LDA is shown in Figure 4, and we combined PCC-LDA with BERT to generate contextual words for each topic using the PCC-LDA. This method guarantees that the input to PCC-LDA contains rich semantic information from BERT, which has the potential to enhance the quality of the extracted topics. This method can lead to more coherent and meaningful topics by considering the efficiency of the PCC-LDA with contextual knowledge from BERT. This hybrid approach facilitates a better interpretation of topics, rendering topic modeling a more reliable and insightful process. Key phrase extraction is a fundamental components of NLP, and LDA can significantly improve its performance using advanced methods. One of these is BERT, the next language model that has been previously naturalized. The state-of-the-art, pre-trained language model BERT takes background information and makes accurate representations of text. Furthermore, PCC-LDA can enable even greater improvement in key phrase extraction. In this respect, PCC-LDA incorporates the notion of topic coherence which considers not only the relevancy between key phrases but general topic coherence,if we integrate PCC-LDA into the LDA framework, it is possible to output more meaningful key phrases from a given collection of documents.

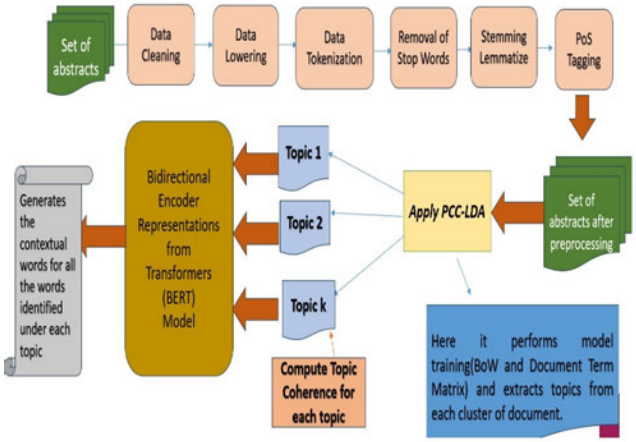

FIGURE 4. Integration of PCC LDA with BERT.

H. EVALUATION METRICS

In this article, the author discusses two popular approaches to determining topic coherence: PMI and NPMI. These assess how similar the most salient words in a topic are to one another. We used them to identify topic models that perform well on an unknown dataset. Strength of Co-occurrence: How strongly two words are associated will be measured with PMI, and it calculates the log likelihood ratio between appearance co-appearance data from a corpus while accounting for chance words. This represents how far apart (on average) two words appear from each other compared with their expected co-occurrence based on individual frequency. The higher the PMI number, the closer the word is to the other. They are used to prepare a list of major words in an area [30].

Topic coherence is important for NLP because it ensures that topics split in text data are related to the main content. Appropriate material is required to obtain accurate consumer insights from social networks. An innovative task is to evaluate the consumer topic coherence. [31].

The concept of topic coherence pertains to the level of semantic co-occurrence frequency among multiple terms within a topic. These techniques seek to evaluate the interpretability and meaningfulness of the obtained topics and to choose the best number of topics [32].

In NLP, topic coherence refers to the quantification of a topic's interpretability or significance as determined by its constituent words. Put simply, this involves evaluating the semantic relationship and coherence of the words comprising a topic in a set of texts. Eliminating irrelevant terms, modifying topic model parameters, or using domain-specific information to assist topic extraction can improve the topic coherence. The goal is to identify subjects that accurately reflect textual themes or concepts to improve comprehension and analysis [33].

$$PMI = \log_2 \left(\frac{p(w_i, w_j)}{p(w_i).p(w_j)} \right) \tag{9}$$

where $p(w_i, w_j)$ is the probability of seeing both words in the same document, and $p(w_i)$ and $p(w_j)$ are the probabilities of seeing each word separately. Equation (2) calculates the topic coherence:

$$TopicCoherence = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} PMI(w_i, w_j) \tag{10}$$

PMI calculates the word co-occurrence probability for specific topics.These steps are frequently used to evaluate the interpretability and coherence of topics acquired using topic modeling methods such as LDA [31].

I. STATISTICAL ANALYSIS FOR VALIDATION

Validation was implemented using statistical analysis testing, in which different degrees of statistics were used to check the accuracy and generalizability of the results. During validation, we determined the extent to which the results were reliable owing to bias and whether they were more likely to occur randomly. A higher score indicated that the result was less likely to be due to random chance. Results: The data were analyzed, and various statistical tests led to important findings in this study.

This study used t-tests to demonstrate the effectiveness of the topic modeling method in the literature analysis. Topic modeling performance evaluation: A T-test was used to determine whether the proportions of topics in the samples differ. We examined whether the topics matched our theme identification intuition and utilized the t-statistic to determine whether they had reasonable test scores. Like many well-known literary subjects, most candidate topics showed favorable test statistics, but they were not statistically significant. Despite having no definition, we used the p-value with the t-test to evaluate statistical significance. Based on subject modeling, we assumed that some bills would strongly represent specific themes. The low p-value of this test demonstrated that the differences between these topics were not random, supporting our hypothesis. Using the same tests, the researchers confirmed the ability of topic modeling to capture literary text aboutness. It claims that topic modeling is promising but not necessarily statistically significant in identifying themes within and between cases, necessitating further validation [34].

IV. RESULTS AND DISCUSSION

We evaluate the performance of a well-established traditional Latent Dirichlet Allocation (LDA) model, non-negative matrix factorization (NMF), and our new probabilistic correlation-based LDA with Probabilistic Correlated Clustering(PCC-LDA) on two large scientific article datasets: DBLPv12 and arXiv. After preprocessing the DBLPv12 dataset, our proposed PCC-LDA model outperformed LDA and NMF in terms of topic coherence. LDA suffers from the intrinsic sparsity of the dataset, NMF has been proven to perform better at finding high-level topics, combining community detection, and PCC-LDA can produce more granular topics and better recommendation relevance. This results in a richer set of content recommendations as well as

more focused topic mining on biospheres. The LDA accuracy suffers from high data sparsity and imprecise topic formation. Compared to Latent Dirichlet Allocation, NMF yielded better results, but it still involves sparsity, similar to LDA. In PCC-LDA, by appropriately clustering documents, it is possible to discover latent topics more efficiently, which eventually leads to better for topic coherence and user profile mapping. Using the arXiv dataset from Figure 5 and 6 shows that, on the arXiv dataset (which is relatively denser and more encouraging for interdisciplinary research), PCC-LDA yields slightly better performance together with a significant consistency improvement. Because of the community-driven nature of PCC-LDA, it identifies articles that have more contextual relevance and hence are exceptionally relevant for cold start recommendation scenarios. LDA and NMF have wide topic diversity and variance in the limitations, but PCC-LDA can effectively address this deficiency. For the arXiv dataset after applying various topic modeling techniques,LDA results are decent in performance but not successful in finding topics associations forever, NMF gives better results, but is not very effective in detecting complex topics and PCC-LDA shows methods with state-of-the-art performance to identify fine-grained topics well and associate them with pseudo-user profiles for better recommendations.

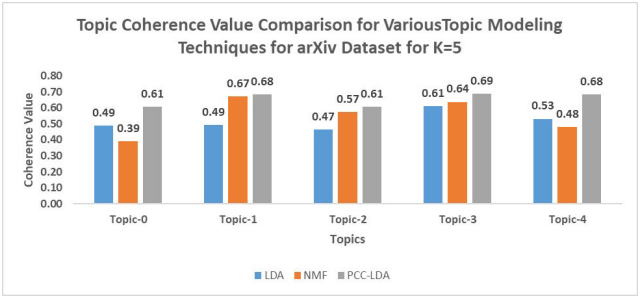

FIGURE 5. Topic coherence for a dataset arXiv for $k = 5$.

A. PERFORMANCE COMPARISON OF PCC-LDA,TRADITIONAL LDA AND NMF

Figure 7, 8 and Figure 9, 10 depict the number of topics and words generated for each topic in the citation network community. We performed coherence analysis on the DBLPv12 dataset, with different numbers of topics to see how our proposed model, PCC-LDA, compared to the traditional LDA model for topic modeling.

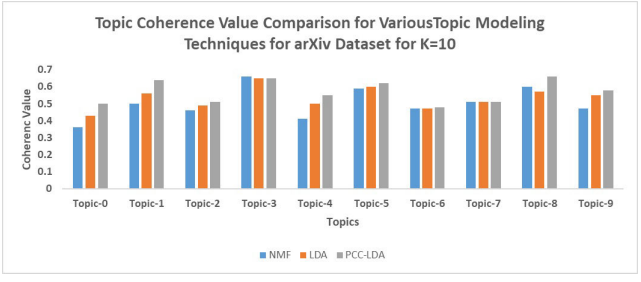

FIGURE 6. Topic coherence for a dataset arXiv for $k = 10$.

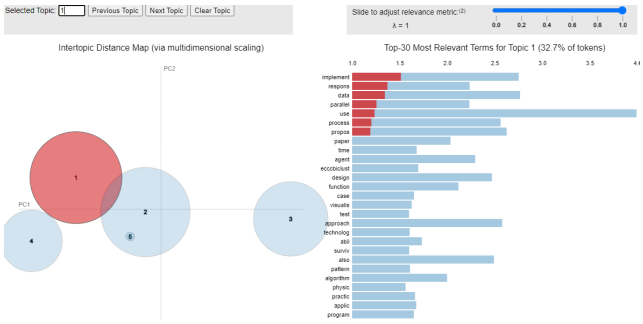

FIGURE 7. LDA topic modeling for community1($k = 5$).

For $k = 10$, the Φ is chosen as 0.1 and Θ chosen as 1.0 with N iterations, where N is the number of iterations used to converge the model.

In Figure 11, 12 and 13, where k = 5, the coherence analysis results indicate the coherence scores for each method (LDA,PCC-LDA and NMF) across the five topics (T1 to T5). For PCC-LDA, the coherence scores range from 0.0029 to 0.0101, whereas for LDA, the scores ranged from 0.0012 to 0.0085 and NMF ranges from 0.002 to 0.0079.

For $k = 5$, the Φ is chosen as 0.1 and Θ chosen as 1.0 with N iterations, where N is the number of iterations used to converge the model.

In Figure 14, 15 and 16 where k = 10, the Coherence Analysis results display the coherence scores for each method (LDA,PCC-LDA and, NMF) across ten topics (T1 to T10). PCC-LDA exhibited coherence scores ranging from 0.0032 to 0.0143, LDA scores ranged from 0.0013 to 0.011, and NMF scores ranged from 0.002 to 0.0087.

This study shows that PCC-LDA is generally better than LDA and NMF, in that it is consistent across a wide range of topics. The coherence scores consistently show that PCC-LDA generates more coherent topics than the traditional LDA model.The DBLPv12 dataset undergoes a coherence analysis with varying topics.

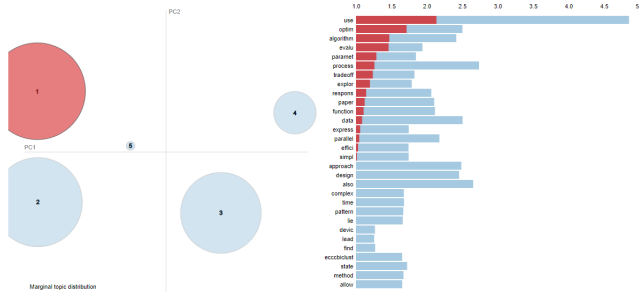

FIGURE 8. PCC−LDA topic modeling for community1($k = 5$).

B. COMPARISON BETWEEN STATISTICAL TESTING AND CONFIDENCE INTERVALS

Table 2 shows a comparison of the three topic modeling algorithms of PCC-LDA, LDA,and NMF which were applied to the dblpv12 dataset. The results and continuous intervals of the statistical tests for each algorithm are presented in

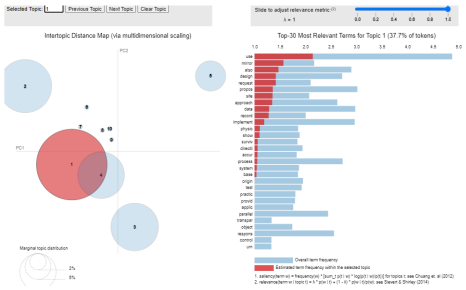

FIGURE 9. LDA topic modeling for community1($k = 10$).

TABLE 2. Results of T-Test, P-Test, and confidence intervals.

Algorithm	T-Test	P-Test	Confidence Intervals(CI)
PCC-LDA	15.425	0.05	95% CI: (0.85, 0.82)
LDA	2.7213	0.0261	95% CI: (0.45, 0.56)
NMF	13.84	0.05	95% CI: (0.85, 0.79)

this table. At 15.425, the PCC-LDA test values differed significantly from those of the null hypothesis. A p-value of 0.05 indicates that the finding narrowly met the statistical significance threshold, right at the point where the null hypothesis is rejected. A high confidence interval (0.85, 0.82) around the coherence score indicates a small range and consistent performance. LDA has a test value of 2.7213, which is lower than those of PCC-LDA and NMF, indicating less statistical evidence against the null hypothesis. The t-statistic suggests an effective size of less than 0.05, resulting in at P-value of.0261 significant. However, PCC-LDA and NMF had broader confidence intervals (0.45, 0.56) than LDA, despite their higher coherence. For NMF, the test value of 13.84 is once again sufficiently high that we must reject the null hypothesis and accept ample to excellent evidence for a difference of 0.05. Thus, as an example, this ability becomes significant in PCC-LDA. Context confidence Intervals are 0.85 and 0.79 (broader than PCC-LDA), which represent coherence but higher variability due to noise. PCC-LDA is a model we use with narrow and high confidence interval (0.82, 0.85), which is an important characteristics to possess intensity and reliability for a topic coherence support. The confidence interval of NMF is also high (0.79, 0.85), slightly wider than the PCC-LDA model alluding to similar levels of variability but still very strong agreement between both topics and reviewers across different bag-of-words representations. Its confidence interval is also much broader (0.45, 0.56), suggesting that its coherence score not only is lower but varies across different runs or sub-samples of the data as well This indicates that LDA is less capable of generating coherent topics compared to PCC-LDA and NMF. After subjecting to a t-test and coherence score testing, the PCC-LDA not only produces overall better results than LDA but also demonstrates its efficiency in this larger datasets. They have some minor differences in coherence (there is more variance among runs for NMF than PCC-LDA, based on the size of its confidence interval) but are both near each other and close to optimal.

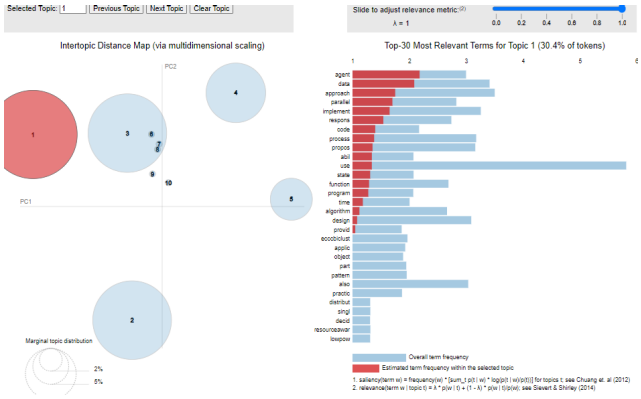

FIGURE 10. PCC-LDA topic modeling for community1($k = 10$).

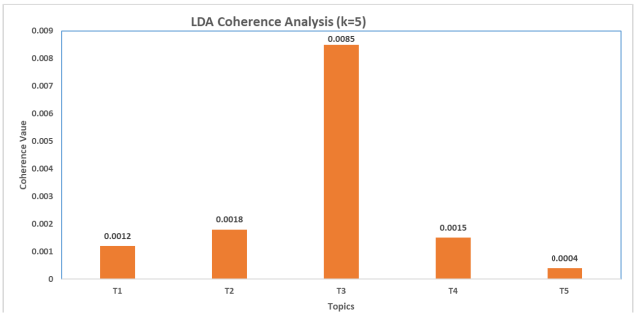

FIGURE 11. LDA topic modeling for community1($k = 5$).

Besides worse statistical power of the results (t-test and confidence intervals), it is also ranked last by coherence score making LDA least effective model in this comparison. In all statistical validation, PCC-LDA and NMF outperform LDA in topic coherence and consistency. PCC-LDA improves performance slightly, indicating that an explicit probability model can statistically improve interpretable subject learning. This validation testing demonstrates that the LDA model is not as good as other models we tested, which is important for researchers who want to employ topic modeling. This demonstrates how relevant the results are to your research and what the model delivers in practice.

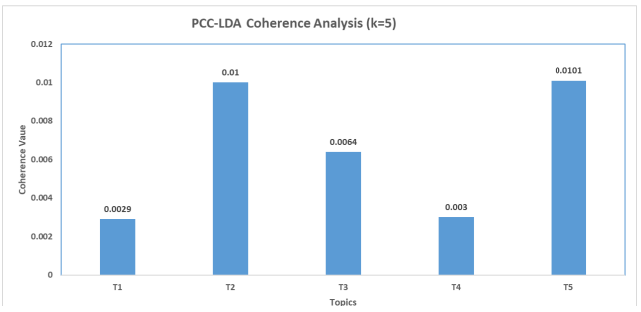

FIGURE 12. PCC–LDA topic modeling for community1($k = 5$).

C. IMPACT OF BERT INTEGRATION

PCC-LDA, on the other hand, contributes to key phrase extraction by prioritizing the subject coherence of the

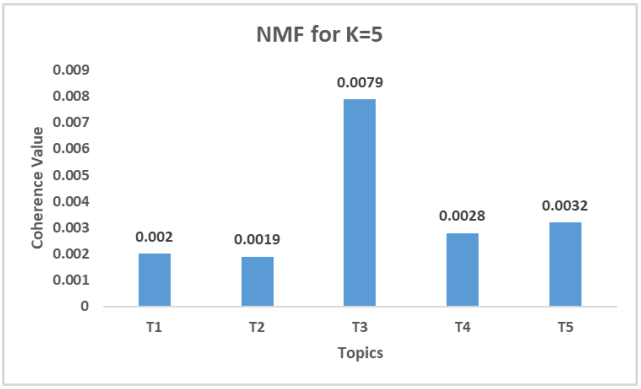

FIGURE 13. Non matrix factorization topic modeling for community1($k = 5$).

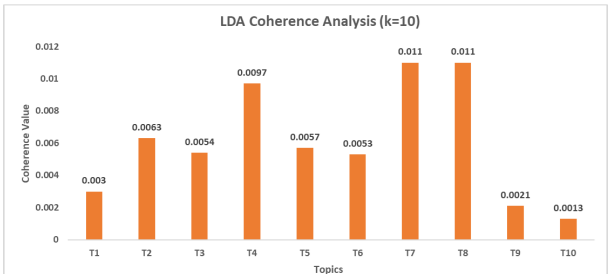

FIGURE 14. LDA topic modeling for community1($k = 10$).

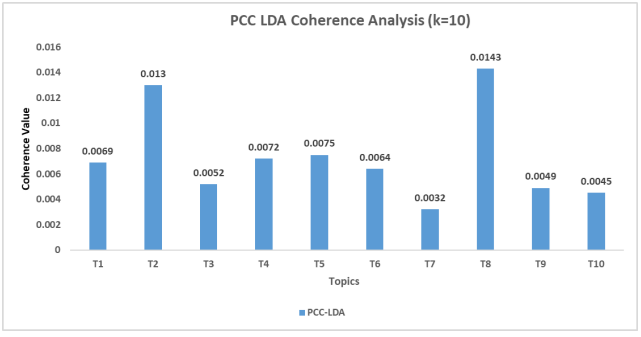

FIGURE 15. PCC−LDA topic modeling for community1($k = 10$).

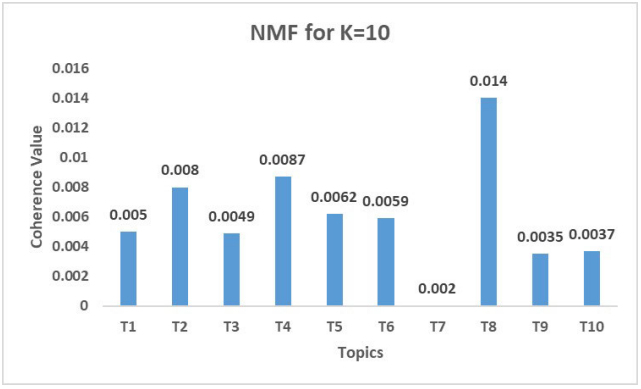

FIGURE 16. Non matrix factorization topic modeling for community 1($k = 10$).

extracted key phrases. By considering the overall coherence of key phrases within a topic, the PCC-LDA ensures that the identified key phrases are not only relevant but also coherent in the broader context of the document.

The BERT model enhances the key phrase extraction process by incorporating contextual information and capturing word associations within a given document. BERT's pre-trained representations facilitate a more precise understanding of the material within a document and aid in the identification of key phrases that effectively capture the most significant information. The sample result, which displays contextual words generated by the PCC-LDA, is as follows:

Enter the Number of Topics: 10 [1091, 1, 1674, 27301, 108390, 121829, 35878] Community 1 - Topic 1: response, algorithm, parallel, bicluster, time, propose, pattern, implement, paper, use

Average Probabilistic Coherence for Topic 1: 0.004313 Context for response in Community 1: cohere approximate paper parallel observe bicluster advance drug either speedup complex exhaust implement understand state tempore program algorithm use test

Context for algorithm in Community 1 function ring approach evaluate problem would optim thu inductor evaluate involve lie algorithm model use efficiency surface topology method underline cohere approximate paper parallel observe bicluster advance drug either speedup complex exhaust implement understand state tempore program algorithm use test.

Community 1 - Topic 2: use, optimum, approach, explore, also, agent, function, evaluate, data, design

Average Probabilistic Coherence for Topic 2: 0.0028867126744974787

Context for use in Community 1:

illustrate accurate object accuracy collect attempt beyond reliable believe interpret case use goal test statue propose physics colour polychrome 2006 explain tracker software site use nearest source user administrator mysql directly resolve base use make package perl http system function ring approach evaluate problem would optim thu inductor evaluate involve lie algorithm model use efficiency surface topology method underline practice agent data focus decide simple program approach autonomic agentu0027 shown preserve state spatial base use remote procedural distribute runtime propose coher approximate paper parallel observe bicluster advance drug either speedup complex exhaust implement understand state tempore program algorithm use test.

Context for approach in Community 1:

illustrate accurate object accuracy collect attempt beyond reliable believe interpret case use goal test statue propose physics colour polychrome 2006 function ring approach evaluate problem would optim thu inductor evaluate involve lie algorithm model use efficiency surface topology method underline practice agent data focus decide simple program approach autonomy agentu0027 shown preserve state spatial base use remote procedural distribute runtime propose.

The combination of PCC-LDA and BERT may provide a more comprehensive understanding of the relationships and themes within scientific articles. The PCC-LDA captures correlated topics and clusters, whereas BERT captures at

semantic nuances. The synergy of PCC-LDA and BERT could result in recommendations that are both localized (within clusters) and context-aware (considering semantic meanings). This may enhance the relevance of recommendations.

D. SUMMARY OF KEY FINDINGS

PCC-LDA improves upon the traditional LDA by integrating probabilistic correlation clustering. BERT generates contextually appropriate phrases, that enhances the coherence and relevance. An empirical evaluation showed that strategy is effective in creating concepts that are coherent and semantically relevant the subjects.

V. CONCLUSION AND FUTURE WORK

This study presents an innovative approach that combines the advantages of BERT and PCC-LDA for key phrase extraction. This method focuses on textual coherence, ensuring that retrieved keywords are closely related to many major topics in the input texts. Using BERT allows for the ability to extract contextually relevant terms that help themes be coherent and accurate helping to generate key phrases that are even more informative. The implications of this research are significant for more accurate documentation summarization, better information retrieval, and enhancements in topic modeling.

A method that adds the probabilistic aspect of PCC to the semantic depth on BERT embeddings generates clusters which are both semantic-rich, cohesive and flexible This hybrid method improves upon topic modeling, document classification and other NLP applications by enabling a comprehensive but detailed view of your textual data. To make the topics coherent, PCC can use BERT embeddings as input which helps provide more semantic value to our topic and is easier for us to understand.

These are the pre-trained embeddings used by BERT, but they might not capture the variance within a particular subject. This problem can be solved using customized embedding or domain adaptation methods. Tuning BERT for key phrase extraction datasets is not easy. Directions for future work include interpretable key phrase extraction, domain-specific adaptation, and generalization to multi-document contexts (text coherence). In conclusion, LDA integrated with BERT is a powerful and efficient method for NLP. It made an effort to combine contextual word conditioning with topic coherence and provided a powerful robust solution. BERT captures intricate contextual word-word interactions. Integration with LDA in BERT increases the accuracy and credibility of topic modeling for longer text documents. This will help to obtain better search results by predicting the correct information. This could be extended to content recommendation, information retrieval systems and any other document clustering system. An interesting approach to explore is combining LDA and BERT because they can leverage most of each other because of their advantages in context-category separation and contextual word representation with BERT better captures topic coherence than classical LDA. We use this combination to distill LDA's topic distribution with BERT's contextual awareness and achieve semantically-rich clusters.

The PCC−LDA with BERT produced higher coherence values than this technique. BERT's extensive contextual embeddings improve topic semantic alignment in PCC LDA, which directly models the topic correlations. This finding emphasizes the importance of using cutting−edge contextual embeddings along with complex probabilistic models such as PCC to obtain the best results in topic modeling tasks. This will help obtain better search results by predicting the correct information using some parametric tests. This could be extended to content recommendation, information retrieval systems, social network analysis, bioinformatics, and market segmentation.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia, for funding this research (IFKSUOR3-176-6).

REFERENCES

[1] M. Rüdiger, D. Antons, A. M. Joshi, and T.-O. Salge, "Topic modeling revisited: New evidence on algorithm performance and quality metrics," *PLoS ONE*, vol. 17, no. 4, Apr. 2022, Art. no. e0266325.
[2] M. Q. Khan, A. Shahid, M. I. Uddin, M. Roman, A. Alharbi, W. Alosaimi, J. Almalki, and S. M. Alshahrani, "Impact analysis of keyword extraction using contextual word embedding," *PeerJ Comput. Sci.*, vol. 8, p. e967, May 2022.
[3] K. Patel and C. Caragea, "Exploiting position and contextual word embeddings for keyphrase extraction from scientific papers," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 1585–1591.
[4] Z. Alami Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: A survey and trends," *J. Intell. Inf. Syst.*, vol. 54, no. 2, pp. 391–424, Apr. 2020.
[5] R. Wang, W. Liu, and C. McDonald, "Using word embeddings to enhance keyword identification for scientific publications," in *Databases Theory and Applications*, M. A. Sharaf, M. A. Cheema, and J. Qi, Eds. Cham, Switzerland: Springer, 2015, pp. 257–268.
[6] A. Srivastav and S. Singh, "Proposed model for context topic identification of English and Hindi news article through LDA approach with NLP technique," *J. Inst. Eng. (India), Ser. B*, vol. 103, no. 2, pp. 591–597, Apr. 2022.
[7] H. Du, S. Thudumu, A. Giardina, R. Vasa, K. Mouzakis, L. Jiang, J. Chisholm, and S. Bista, "Contextual topic discovery using unsupervised keyphrase extraction and hierarchical semantic graph model," *J. Big Data*, vol. 10, no. 1, p. 156, Oct. 2023.
[8] A. F. Pathan and C. Prakash, "Unsupervised aspect extraction algorithm for opinion mining using topic modeling," *Global Transitions Proc.*, vol. 2, no. 2, pp. 492–499, Nov. 2021.
[9] E. Rivadeneira-Pérez and C. Callejas-Hernández, "Leveraging LDA topic modeling and BERT embeddings for thematic unsupervised classification of tourism news in rest-mex competition," in *Proc. IberLEF@ SEPLN*, 2023, pp. 1–8.
[10] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2017, pp. 745–750.
[11] S. Likhitha, B. S. Harish, and H. M. K. Kumar, "A detailed survey on topic modeling for document and short text data," *Int. J. Comput. Appl.*, vol. 178, no. 39, pp. 1–9, Aug. 2019.
[12] K. Ashihara, C. B. El Vaigh, C. Chu, B. Renoust, N. Okubo, N. Takemura, Y. Nakashima, and H. Nagahara, "Improving topic modeling through homophily for legal documents," *Appl. Netw. Sci.*, vol. 5, no. 1, pp. 1–20, Dec. 2020.
[13] G. Papadia, M. Pacella, M. Perrone, and V. Giliberti, "A comparison of different topic modeling methods through a real case study of Italian customer care," *Algorithms*, vol. 16, no. 2, p. 94, Feb. 2023.

[14] Z. Zhou and K. Wakabayashi, "Topic modeling using jointly fine-tuned BERT for phrases and sentences," in *Proc. 14th Data Eng. Inf. Manage.*, 2022, pp.'1–8.
[15] A. Glazkova, "Identifying topics of scientific articles with BERT-based approaches and topic modeling," in *Trends and Applications in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer, May 2021, pp. 98–105.
[16] S. George and S. Vasudevan, "Comparison of LDA and NMF topic modeling techniques for restaurant reviews," *Indian J. Nat. Sci.*, vol. 10, no. 62, pp. 28210–28216, 2020.
[17] T. Saheb, M. Dehghani, and T. Saheb, "Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis," *Sustain. Comput., Informat. Syst.*, vol. 35, Sep. 2022, Art. no. 100699.
[18] S. Avasthi, R. Chauhan, and D. P. Acharjya, "Extracting information and inferences from a large text corpus," *Int. J. Inf. Technol.*, vol. 15, no. 1, pp. 435–445, Jan. 2023.
[19] S. Mifrah, "Topic modeling coherence: A comparative study between LDA and NMF models using COVID'19 corpus," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5756–5761, Aug. 2020.
[20] R. Egger and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts," *Frontiers Sociol.*, vol. 7, May 2022, Art. no. 886498.
[21] S. Si, J. Wang, R. Zhang, Q. Su, and J. Xiao, "Federated non-negative matrix factorization for short texts topic modeling with mutual information," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–7.
[22] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2017, pp. 165–174.
[23] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of topic modelling approaches in the banking context," *Appl. Sci.*, vol. 13, no. 2, p. 797, Jan. 2023.
[24] M. de Groot, M. Aliannejadi, and M. R. Haas, "Experiments on generalizability of BERTopic on multi-domain short text," 2022, *arXiv:2212.08459*.
[25] M. Allahyari and K. Kochut, "Discovering coherent topics with entity topic models," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Oct. 2016, pp. 26–33.
[26] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023.
[27] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 188–197.
[28] H. Zankadi, A. Idrissi, N. Daoudi, and I. Hilal, "Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques," *Educ. Inf. Technol.*, vol. 28, no. 5, pp. 5567–5584, May 2023.
[29] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 990–998.
[30] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," 2020, *arXiv:2004.03974*.
[31] Y. Wang, E. Willis, V. K. Yeruva, D. Ho, and Y. Lee, "A case study of using natural language processing to extract consumer insights from tweets in American cities for public health crises," *BMC Public Health*, vol. 23, no. 1, p. 935, May 2023.
[32] K. Park, S. Park, and J. Joung, "Contextual meaning-based approach to fine-grained online product review analysis for product design," *IEEE Access*, vol. 12, pp. 4225–4238, 2024.
[33] S. J. Blair, Y. Bi, and M. D. Mulvenna, "Aggregated topic models for increasing social media topic coherence," *Appl. Intell.*, vol. 50, no. 1, pp. 138–156, Jan. 2020.
[34] J. Schröter and K. Du, "Validating topic modeling as a method of analyzing sujet and theme," *J. Comput. Literary Stud.*, vol. 1, no. 1, pp. 1–120, 2022.

SANDEEP KUMAR RACHAMADUGU received the B.Tech. and M.Tech. degrees from the Department of Computer Science and Engineering, JNTU Hyderabad. He is currently a Research Scholar with the M. S. Ramaiah University of Applied Sciences, Bengaluru, and an Assistant Professor with the G. Pulla Reddy Engineering College, India. His research interests include explore the applications of social network analysis and recommender systems in the field of natural language processing, with a focus on developing innovative solutions that combine language understanding, network analysis, personalized recommendations, machine learning, deep learning, and data science.

T. P. PUSHPHAVATHI received the B.E. degree in computer science and engineering from SJCE, Mysore, in 1996, the M.Tech. degree in computer network engineering from DSCE, Bangalore, in 2008, and secured the university's second rank, and the Ph.D. degree in computer science and engineering from JAIN University, Bangalore, in 2016. She is an Associate Professor with the Department of Computer Science and Engineering, M.S. Ramaah University of Applied Sciences, Bangalore, India. Her research interests include data mining, science, artificial intelligence, and machine learning.

SURBHI BHATIA KHAN is working with the School of Science, Engineering and Environment, Data Science Department, University of Salford, U.K. She is a Project Management Professional Certified from PMP, USA. She is supervising/cosupervising a number of Ph.D.'s and Research Associates. She has published over 200 articles and books in high indexed outlets with her students and colleagues. Her research interests are machine learning/deep learning, data science in healthcare, and sentiment analysis. She has been awarded with best paper awards, and several other recognitions. She is also serving as Associate and Guest editor in reputed journals. She is also a member in the Women in Data Science Ambassador 2024.

MOHAMMAD ALOJAIL received the doctorate degree in business information systems from RMIT University, Australia. He has published many articles in reputed journals and conferences. He is an Associate Professor of Management Information Systems at the College of Business Administration, King Saud University. He held a few administrative positions and has worked as a consultant in the Education and Training Evaluation Commission. His research interests are information systems, digital transformation, and the IoT.