Journal of Theoretical and Applied Information Technology
31st July 2024. Vol.102. No. 14

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

A NOVEL APPROACH TO MODELING TOPICS THROUGH A DISTRIBUTED FILE SYSTEM FOR JOB ASSISTANCE IN SOCIETAL COMMUNICATION EMPOWERMENT

K. PUSHPA RANI [1], PELLAKURI VIDYULLATHA [2] and Dr. K. SRINIVAS RAO [3]

[1.]Research Scholar, CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[2.]Professor, CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[3.]Professor, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad.
rani536@gmail.com

ABSTRACT

The current research program uses data from social networks to explore public opinion on technical terms or topics. For job seekers, understanding how the general public perceives technical phrases or subjects and their impact on the environment and society is crucial. Public support is also critical for legislation and the implementation of mitigation programs. Public opinion research is essential for a better understanding of the social environment and dynamics. Social media data provides valuable information on public attitudes and responses to conflicting socio-technical terms or issues from various perspectives, such as quorum, stack overflow, and Yahoo!. It responds to Twitter, among other platforms, and is frequently used to track and assess how society responds to a natural or societal anomaly. Typically, social media data is acquired by searching for keywords or a specific topic to identify various topics in the topic templates. However, in conventional topic models, users can provide an inaccurate number of topics, leading to subpar grouping outcomes. Accurate representations are crucial for retrieving data and identifying cluster trends. To address this issue, viable methods for modeling themes are related to unclassified and incorrect texts or topics. The Distributed Latent Semantic Analysis (DLSA) and the Distributed Latent Dirichlet Allocation (dLDA) are two techniques used for this purpose.This document provides a brief overview of the country's public question-and-answer system and traces the evolution of significant issues and initiatives, paying particular attention to the automatic dissemination of pertinent customer feedback and knowledge of relevant awareness-raising information. It also highlights opportunities for housing and employment for the newest technologies in global empowerment. Finally, the experimental findings suggest that topic models outperform existing models in terms of precision for obtaining more pertinent responses from a placement and interview perspective.The research addresses the challenge of accurately modeling themes in social media data to understand public opinion on technical terms and topics. By employing advanced techniques such as DLSA and dLDA, the study enhances the precision of topic modeling, leading to better data retrieval and identification of cluster trends. This improvement aids job seekers in understanding public perception, supports legislative efforts, and facilitates the implementation of mitigation programs. The impact of this research lies in its contribution to more effective public opinion analysis, thereby informing policy-making and societal responses to technical and environmental issues.

Keywords: *F-Score, Hadoop, LSA, LDA, overflow, Quora, Topic models, stack, Twitter API*

1. INTRODUCTION

Software engineers and programmers often use websites like Stack Overflow to find answers to their questions. By analyzing the data available, we can identify the most challenging aspects of programming and APIs. In this article, we categorize Stack Overflow problems into two overlapping views - programming concepts and the type of information sought. Users who post questions should assign specific tags to indicate the category of the question. They can choose from a list of existing tags or create a new one if necessary. We recommend using existing tags whenever possible. However, if the question covers a new or unique topic, creating a new tag is appropriate. This will help users retrieve answers to their questions quickly. However, sometimes the available tags may not be

appropriate, which can result in irrelevant answers. The tags assigned by users are subjective and open to interpretation. Many publications have explored different topics using in-stream stack data, such as finding expert users, analyzing faulty project documentation, unanswered Stack Overflow questions, label prediction, and more. Despite ongoing efforts to improve the quality of labels and responses, finding the right answers can still be challenging. To address this issue, we aim to present users with the ten most relevant questions using a k-means classification model to label questions with appropriate context and ensemble modeling for similarity comparison with corpus questions.

The issues surrounding knowledge creation in this context are multi-faceted. Firstly, existing topic modeling techniques often struggle to efficiently process the vast and diverse datasets present in job assistance and societal communication platforms. This limitation hampers the creation of comprehensive and accurate insights into relevant topics and trends. Secondly, the accessibility barriers posed by complex methodologies prevent widespread participation in knowledge creation, particularly among marginalized communities or individuals lacking technical expertise. Finally, the static nature of many current models inhibits the continuous generation of new insights as the job market and societal discourse evolve.

The research gap addressed by this study lies in the intersection of these challenges and the proposed solution. While previous research has explored various topic modeling approaches, few have specifically targeted the scalability, accessibility, and adaptability issues inherent in job assistance and societal communication empowerment. By introducing a novel methodology leveraging distributed file systems, this study bridges the gap between theoretical advancements in topic modeling and practical applications in empowering individuals to navigate digital platforms effectively. By addressing these critical gaps, this research not only advances knowledge creation in the field of computational linguistics and information retrieval but also contributes significantly to the enhancement of digital inclusion, economic empowerment, and social cohesion in contemporary society.

2. LITERATURE REVIEW

D. M. Blei, A. Y. Ng, and M. I. Jordan, [1] introduced Latent Dirichlet Allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA posits that documents are mixtures of topics, where a topic is a distribution over words. This seminal work has profoundly influenced topic modeling, providing a robust framework for uncovering the underlying thematic structure in large datasets. Its applications span various domains, including text analysis, bioinformatics, and social sciences, making it a cornerstone in the field of machine learning and natural language processing.

LI Hua-Meng,LI Hai-Rui,XUE Liang. [2] ,enhances the traditional Term Frequency-Inverse Document Frequency (TFIDF) algorithm by integrating information gain and entropy. This approach aims to improve the accuracy of feature selection in text mining and information retrieval. By combining these metrics, the algorithm better captures the importance of terms, leading to more effective and efficient text classification and clustering.

Hanchen Jiang, Maoshan Qiang, Dongcheng Zhang, Qi Wen, Bingqing Xia, Nan An.[3] examine how climate change topics are communicated within the online Q&A community Quora. The study utilizes content analysis to identify prevalent themes and user engagement patterns. It highlights the role of digital platforms in disseminating scientific information and fostering public discourse on climate issues. This research underscores the importance of online communities in shaping public understanding and response to climate change.

Campbell, J.C., Hindle, A. and Stroulia [4] apply Latent Dirichlet Allocation to software engineering data, demonstrating its utility in extracting meaningful topics from large datasets. Their work illustrates how LDA can reveal hidden patterns in software repositories, facilitating better understanding and management of software projects. This application bridges the gap between machine learning techniques and practical software engineering challenges.

Rainer Lienhart, Stefan Romberg, and Eva H¨orster. [5] propose a multilayer probabilistic Latent Semantic Analysis (pLSA) for multimodal image retrieval. Their approach integrates textual and visual data to improve the accuracy of image searches. By leveraging the complementary strengths of different data modalities, the model achieves superior retrieval performance, demonstrating the potential of probabilistic methods in multimedia applications.

S. Arora, R. Ge, R. Kannan, and A. Moitra [6] address the computational challenges of Nonnegative Matrix Factorization (NMF), providing provable guarantees for its performance. Their theoretical contributions enhance the understanding of NMF's capabilities and limitations, offering insights into its application in data analysis tasks. This work advances the field by establishing a rigorous foundation for NMF algorithms.

Lee, D.D., Seung, H.S [7] introduce key algorithms for Non-negative Matrix Factorization, emphasizing its utility in uncovering the parts-based representation of data. Their work demonstrates the applicability of NMF in various domains, including image processing and text mining. The algorithms have become fundamental tools in machine learning, particularly for tasks requiring data decomposition into interpretable components.

Yan X, Guo J [8] propose an innovative approach for topic modeling in short texts using Ncut-weighted Non-negative Matrix Factorization (NMF). By incorporating term correlations, their method effectively addresses the sparsity issue common in short text data. This advancement enhances the accuracy of topic detection in applications such as microblog analysis and real-time information retrieval.

Huang L, Ma J, Chen C [9] develop a topic detection framework for microblogs utilizing T-LDA and perplexity metrics. Their model adapts LDA to handle the brevity and noise inherent in microblog data, improving the identification of relevant topics. This approach is particularly valuable for monitoring and analyzing social media trends and public opinion.

W. Xu, X. Liu, and Y. Gong. [10] explore the application of Non-negative Matrix Factorization (NMF) for document clustering, demonstrating its effectiveness in grouping similar documents based on latent structures. Their findings highlight NMF's potential in enhancing information retrieval systems and organizing large text corpora into coherent clusters.

Peng Zhang [11] provide a comprehensive analysis of statistical methods for evaluating recall, precision, and average precision in information retrieval systems. Their work offers valuable insights into the performance metrics of retrieval algorithms, contributing to the development of more effective evaluation frameworks.

Edi Surya Negara, Dendi Triadi, Ria Andryani [12] apply Latent Dirichlet Allocation (LDA) to Twitter data, showcasing its capability to uncover prevalent topics and trends. Their study demonstrates LDA's effectiveness in analyzing social media content, providing a valuable tool for sentiment analysis, trend detection, and public opinion monitoring.

Pablo Ormeño ,Marcelo Mendoza , and Carlos Valle [13], propose the use of topic model ensembles to improve ad-hoc information retrieval. Their approach combines multiple topic models to enhance retrieval accuracy and robustness. This ensemble method shows promise in diverse applications, including personalized search and recommendation systems.
Kim and Cho (2020) introduced a user-topic modeling framework for analyzing online communities. Their model captures the interplay between user interests and topic dynamics, providing valuable insights into user engagement and community structure.

Dhelim S.; Aung N.; Ning H. [14] presented a hybrid filtering approach that integrates user personality traits to enhance interest prediction in social networks. Their method outperforms traditional recommendation systems, offering a more personalized user experience.

Baechle C.; Huang C.; Agarwal A.; Beharam R.; Goo, J.[15] developed a latent topic ensemble learning approach to optimize hospital readmission costs. Their model identifies critical factors contributing to readmissions, aiding in the design of cost-effective healthcare interventions.

Pourvali M. Orlando S. , Omidvarborna H.[16] combined topic models with fusion methods to enhance text clustering and labeling. Their integrated approach improves clustering accuracy and provides more meaningful cluster interpretations.

Fiandrino S, Tonelli A [17] conducted a text-mining analysis on the non-financial reporting directive, highlighting its impact on value creation for stakeholders. Their findings emphasize the importance of transparent and comprehensive reporting in fostering stakeholder trust.

yerragudipadusubbarayudu, AlladiSureshbabu, [18] developed a distributed multimodal topic model that incorporates sentiment analysis for public health surveillance. Their approach effectively identifies health trends and sentiments, providing valuable insights for health policy and intervention strategies.

Ammirato S., Felicetti A.M., Raso C., Pansera, B.A. , Violi A. Agritourism [19] conducted a systematic literature review on agritourism and sustainability, identifying key factors that contribute to sustainable agritourism practices. Their review provides a comprehensive overview of the current state of research and practical implications for stakeholders.

Farkhod A , Abdusalomov A., Makhmudov, [20] introduced the TDS model, an LDA-based approach for sentiment analysis at multiple levels. Their model enhances sentiment detection and topic coherence, offering a nuanced understanding of text sentiment.

P. Vidyullatha, P. venkateswara Rao [21] utilized Rhadoop-Hive for big data sentiment analysis on social media. Their approach leverages the scalability of Hadoop and the querying capabilities of Hive to process large-scale social media data efficiently.

Devisetty S.D.P., Sai Y.M. Yadav, A.V. Vidyullatha, [22] applied RapidMiner, a data science platform, for sentiment analysis of tweets. Their study showcases the tool's capability to handle and analyze large volumes of tweet data, providing real-time sentiment insights.

ChangY. L., & KevJ.[23] proposed a framework for socially responsible AI in people analytics, emphasizing sustainability. Their work highlights the ethical considerations and potential benefits of integrating AI in human resource management.

Bhatti I., Rafi H., & Rasool S,. [24] explored the use of ICT technologies to support disabled migrants in the USA. Their study identifies key technologies and strategies that enhance accessibility and integration for disabled individuals, contributing to social inclusion.

Smythe T., Ssemata A. S., Slivesteri S., Mbazzi F. B., et al [25] developed a training program on disability for healthcare workers in Uganda. Their co-development approach ensures the program is culturally relevant and effective in improving healthcare services for disabled individuals.

In today's rapidly evolving digital landscape, the proliferation of job assistance platforms and societal communication channels necessitates innovative approaches to empower individuals in navigating these resources effectively. However, existing methods for modeling topics often encounter limitations in scalability, accessibility, and adaptability, hindering their utility in facilitating comprehensive job assistance and societal communication empowerment. A critical review of the literature reveals several gaps and challenges in the current state of research and practice:

Scalability Constraints: Traditional topic modeling techniques often struggle to efficiently process and analyze large volumes of heterogeneous data distributed across diverse platforms and sources. As a result, individuals seeking job assistance and societal communication support may encounter delays, incomplete information, or inaccuracies in the insights provided.

Accessibility Barriers: Many existing topic modeling frameworks require specialized expertise in data science or computational linguistics, restricting access to their benefits for non-technical users. This lack of accessibility poses a significant obstacle for marginalized communities or individuals with limited technical proficiency, impeding their ability to leverage digital resources for employment opportunities and social engagement.

Adaptability to Dynamic Contexts: The dynamic nature of job markets, societal discourse, and communication platforms necessitates topic modeling approaches that can effectively adapt to evolving trends, preferences, and user needs. Conventional static models often fail to capture the nuanced shifts in topic relevance and significance over time, diminishing their relevance and effectiveness in providing timely and relevant support.

Addressing these challenges requires a novel approach to topic modeling that leverages the capabilities of distributed file systems to enhance scalability, accessibility, and adaptability for job assistance and societal communication empowerment. By harnessing the distributed computing power and storage capacity of modern infrastructure, this approach aims to enable real-time analysis of vast datasets from diverse sources, democratize access to topic modeling tools through intuitive interfaces, and facilitate dynamic adaptation to changing contexts and user requirements.Thus, there is a pressing need for research and development efforts aimed at designing, implementing, and evaluating a distributed file system-based approach to topic modeling for job assistance in societal communication empowerment. This work seeks to fill the identified gaps in the literature, advance knowledge in the field of computational linguistics and information retrieval, and ultimately contribute to the enhancement of digital inclusion, economic empowerment, and social cohesion in contemporary society.

Research on artificial intelligence began in the 1960s, and scientists suggested that computers could answer questions using natural language processing. This led to the development of rudimentary response systems. In the 1980s, question-and-answer systems gained popularity in the field of natural language. However, research on response systems decreased with the growth of large-scale word processing technology. In recent years, with the rapid development of networks and information technology, people's desire to receive information faster has encouraged the development of response systems. Many companies and research institutes, including Microsoft, IBM, and MIT, have participated in this development. In 1999, TREC introduced an automatic response to project tracking questions. Since then, the Q&A track has gradually become one of TREC's most popular projects. Several countries have developed relatively mature question-and-answer systems. The InfoLab group of the Laboratory for Computer Science and Artificial Intelligence at MIT developed an open question and answer software system called Start. However, question-and-answer systems used to solve assignments for a particular course are very rare. Therefore, an intelligent question-and-answer system has been developed that returns answers to users' questions according to the principles of a course-based course.

3. DATA-PREPROCESSING:

Text preprocessing is a crucial step before applying NLP and text analysis techniques. The extracted elements from the text, whether they are symbolic sentences, words, or phrases, serve as a foundation for subsequent analysis. To increase the accuracy of classifiers, the text must be standardized, cleaned up, and more information must be extracted from the notes.

The first step is to clean up the content by eliminating any html tags, stop words, and unnecessary keywords. For word tokenization, the regexp tokenizer class from the NLTK framework can be used. This tokenizes words based on regular expression-based models to partition sentences and subsequently remove those words from sentences. Additionally, case-sensitive conversion must be done. Labels must be normalized to produce a clear, uniform version of the labels.

Cleaning methods like spelling and stem corrections, as well as stemming, can be employed for the markers. Several models, including Word2vec, Bag of Words, and TF-

IDF [15–17], can be used when using feature recovery techniques. The Word Bag model can be tested, which converts the text into a vector that indicates the frequency of all the individual words contained in the text vector space for that text, after starting with the vector space model, in which text data is represented as numerical vectors of terms for vector dimensions.

The term document frequency and inverse frequency (TF-IDF) [18] can be put to the test. Create the vector using the TF-IDF [19–20] weight, and the body of the question serves as the primary attribute from which we derive the function [21–22]. This framework addresses Hadoop distributed topic modelling strategies to improve procedures and outcomes.One of the key ideas before applying NLP and text analysis approaches is text preprocessing since the elements extracted from the text be they symbolic sentences, words, or phrases serve as the foundation for subsequent analysis. To increase the accuracy of the classifiers, the text must be deleted, standardized, and more text information must be gleaned from the notes.
First, clean up the content by eliminating any html tags, stop words, and extraneous keywords. The regexp tokenizer class from the NLTK framework, which tokenizes words based on regular expression-based models to partition sentences and subsequently remove those words from sentences, should be used for word tokenization. Additionally, case-sensitive conversion needs to be done. Labels must be normalized in order to produce a clear, uniform version of the labels.
Spelling and stem corrections as well as stemming were employed as cleaning methods for the markers. “There are several models to choose from when using feature recovery techniques, including Word2vec, Bag of Words, and TF-IDF [15–17]”. Test the Word Bag model, which converts the text into a vector that indicates the frequency of all the individual words contained in the text vector space for that text, after starting with the vector space model, in which text data is represented as numerical vectors of terms for vector dimensions.

It occurs more frequently and causes eclipses and therefore, “they might not happen as frequently. The term document frequency and inverse frequency (TF-IDF) [18]” was then put to the test. “Create the vector using the TF-IDF [19–20] weight and the body of the question serves as the primary attribute from which we derive the function [21–22]”. This framework addresses Hadoop distributed topic modelling strategies to improve procedures and outcomes.

4. METHODOLOGY:

Our survey addressed the user query by first identifying the proper label or intent and then matching the user query with the most similar question, especially on Stack Overflow. The top ten most important questions were then presented.

The first section of the study employed two distinct methodologies. The first is genetic markers in Python, which uses k-means to group documents and model topics for grouping inquiries. The second methodology is topic modelling, which groups topics using probabilistic models, while clustering employs unexpected clustering methods.

Topic modelling uses a body of unstructured documents as input, which is a text extraction technique that identifies the best-classified terms in a topic, along with the documents associated with it. Unlike document classification, where only one subject is connected to the text, a single text document can be associated with multiple topics using a topic modelling approach. Instead of creating a collection of texts or documents, topic modelling creates a collection of words, where each word is a combination of various themes, each of which is given a particular weight.

Topic modelling automatically uncovers latent themes in a set of papers. An unsupervised text analytics technique is used to identify the group of terms from the provided content. This collection of words serves as the topic, and a single document can be connected to several themes.Our survey responded to the user's query by first identifying the question's intent and then matching it with the most similar question available on Stack Overflow. The ten most important questions were then presented.

Journal of Theoretical and Applied Information Technology
31st July 2024. Vol.102. No. 14

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

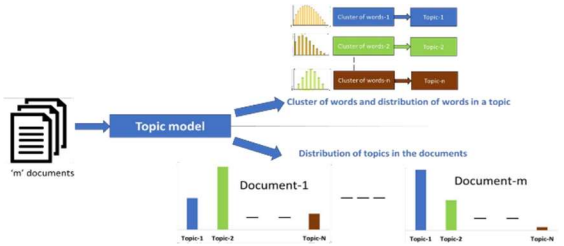


Fig 1: Topic Models

The first section of the study employed two different methodologies. Firstly, genetic markers in Python were used to group documents using k-means clustering and model topics. Secondly, topic modelling was used to group related inquiries using probabilistic models. Clustering used unique clustering methods.

Topic modelling is a technique that extracts the best-classified terms in a topic from a body of unstructured documents. Unlike document classification, where only one subject is connected to text, a single text document can be associated with several topics using topic modelling. Instead of creating a collection of texts or documents, topic modelling develops a collection of words, where each word is a combination of various themes, each of which is assigned a specific weight.

Topic modelling automatically uncovers latent themes in a set of papers. An unsupervised text analytics technique is used to find the group of terms from the provided content. This collection of words serves as the topic. A single document can be connected to several themes.

The Distributed Latent Semantic Analysis:
Latent Semantic Analysis (LSA) is a technique that is also known as Distributed Latent Semantic Analysis (DLSA) or Distributed Latent Semantic Index (DLSI). It uses the Bag of Words (BoW) model to create a matrix that represents the occurrence of terms in a document. The matrix has documents as columns and terms as rows. By performing a matrix decomposition on the document-term matrix using Singular Value Decomposition (SVD), DLSA can identify latent themes. The technique of dimension reduction or noise reduction is commonly used to enhance the accuracy of the results obtained through DLSA.

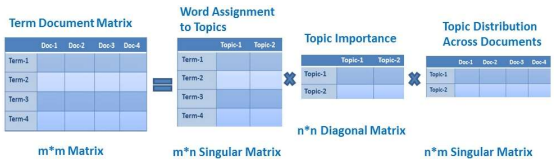


Fig 2: Term Document Matrix

Topic coherence meter is a popular tool used to evaluate topic models. The tool employs latent variable models to generate topics that consist of a list of terms. The topic coherence measure determines the average or median pairwise similarity scores of the terms in a topic. A good topic model should have a high topic coherence score.

One of the primary approaches for modeling themes is latent distributed semantic analysis (dLSA). The main idea behind this approach is to create a topic-topic matrix using a matrix of terms and documents. In topic modeling, a popular method used is Distributed latent semantic analysis (dLSA), and the creation of a document matrix is the first step. We can create an m x n matrix, where each row corresponds to a document, and each column corresponds to a word, using m documents and n words in our dictionary.

The simplest LSA approach counts the number of times the word "j" appears in document "I" for each entry. However, this method does not perform well in practice since it does not consider the meaning of each word in the document. For example, the word "nuclear" provides more information about the topic(s) of the document than the word "evidence." Therefore, the dLSA model often replaces the raw integers in the document matrix with a tf-idf result. The tf-idf, also known as inverse frequency of the document, establishes the weight of the phrase "j" in document "I."

$$W_{i,j}=Mul\ \{tf_{i,j},\ log(N/df_j)\}$$

Distributed Latent Dirichlet Allocation:
Distributed Latent Dirichlet Allocation (dLDA) is a statistical model commonly used in natural language processing. It helps in understanding sets of observations by grouping them into unobserved groups that explain why some portions of the data are similar. For instance, if we consider the observations as words collected

Journal of Theoretical and Applied Information Technology
31st July 2024. Vol.102. No. 14

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

into different documents, dLDA hypothesizes that each document is a combination of a few different subjects, and each word can be related to one of those topics.

The LDA model is visually represented in the figure below. The model's objective is to identify the subject and document vectors that explain the various documents' initial "bag-of-words" representation.

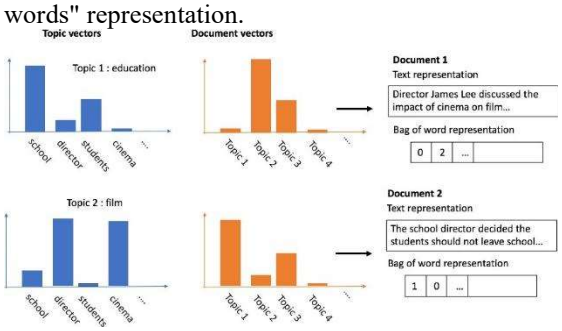
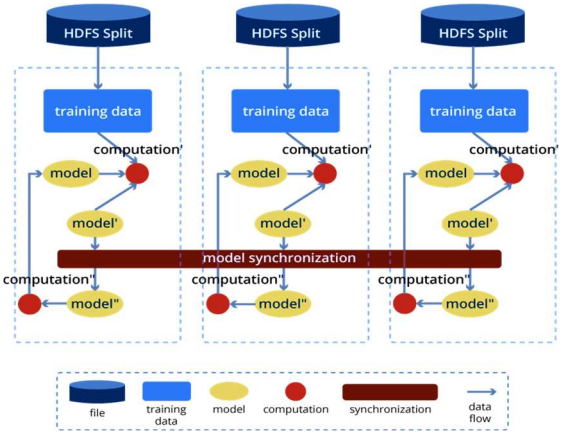


Fig 3: dLDA

It's important to understand that when using topic vectors, you're assuming that they are comprehensible. Without this, the output of the model will be of no use. You're essentially trusting that with enough data, the model will identify and group frequently occurring terms into distinct "themes."

An effective probabilistic model that is easy to use is LDA. Document vectors can show the structure and pattern of documents because they are sparse, low-dimensional, and easy to interpret. It's essential to accurately estimate the number of subjects that appear in the collection of papers, and each subject vector should have a unique "topic" assigned to it manually. The LDA model may suffer similar issues as the bag-of-words model, as both are used to represent the documents. The LDA model learns a document vector that predicts the words that will appear in that document without considering any structure or the way these words interact locally. [23-24].It is crucial to note that you are relying on the presumption that the topic vectors will be understandable because, in the absence of this, the model's output is essentially useless. You are essentially trusting that, given enough data, the model will identify the terms that frequently occur together and group them into discrete "themes."
A straightforward probabilistic model with good performance is LDA. The document vectors frequently show the pattern and structure in documents because they are sparse, low- dimensional, and easy to read. The number of subjects that appear in the collection of papers must be accurately estimated. Additionally, each subject vector must have a unique nominator "topic" assigned manually. LDA may experience the same drawbacks as the bag-of-words model since both are employed to represent the documents. Without considering any structure or the way that these words interact locally, the LDA model learns a document vector that predicts the words that will appear in that document [23-24].

During the initialization phase of training, Harp will load the local data split on each node into memory, which means future disc I/O will not be required to access the training data. By



default, Hadoop MapReduce uses the data splitting strategy it supports.

Hadoop offers distributed dataset abstractions, group communication, and synchronization methods for model data. However, parallelizing the core
computation of the model update for machine learning algorithms causes issues with model consistency and synchronization. For networked machine learning applications, Harp's distinctive abstractions based on collective synchronization mechanism are advantageous in terms of expressiveness, efficiency, and effectiveness.

In this work, topic modelling techniques were used to locate similar questions, and several similarity vectorization models were compared before the most pertinent questions were effectively obtained. Ensemble models, topic modelling, and similarity for k themes were used to define the 20 most pertinent questions for each user query. The total number of issues

Journal of Theoretical and Applied Information Technology
31st July 2024. Vol.102. No. 14

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

in the corpus was used to define the suitable document.

The definition of the received document is a question that is related to the specific document. The distributed modelling approaches strike a balance between solving the current issues and obtaining accurate results, as shown in table 7. Distributed LDA of the topic equilibrium model generated the best results in the comparative analysis shown in figures 10 and 11. It had superior accuracy in the estimation of related indicators.During the initialization phase of training, Harp will load the local data split on each node into memory, which means future disc I/O will not be required to access the training data. By default, Hadoop MapReduce uses the data splitting strategy it supports.

Hadoop offers distributed dataset abstractions, group communication, and synchronization methods for model data. However, parallelizing the core computation of the model update for machine learning algorithms causes issues with model consistency and synchronization. For networked machine learning applications, Harp's distinctive abstractions based on collective synchronization mechanism are advantageous in terms of expressiveness, efficiency, and effectiveness.

Table 2 generates a matrix of phrases for documents, where each phrase represents the occurrence of terms in a document. In this matrix, terms are represented by rows and documents by columns. DLSA (Dynamic Latent Semantic Analysis) uses Singular value decomposition to break down the document-term matrix into smaller parts, allowing it to learn latent themes. DLSA is frequently used as a technique for dimension reduction or noise reduction.

Table 3 shows the results of a probabilistic topic model analysis of document content and topic meanings.

Document /topics	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
Doc-1	0	1	1	1	1	2	0	1	0	1
Doc-2	0	1	1	1	1	1	0	0	0	2
Doc-3	0	2	0	0	1	1	0	2	0	0
Doc-4	1	0	2	1	0	0	1	2	1	0
Doc-5	1	2	1	1	1	0	2	0	0	0
Doc-6	0	0	1	4	1	1	0	0	1	1
Doc-7	0	0	0	1	1	0	3	1	2	1
Doc-8	1	0	0	0	1	2	1	2	0	2
Doc-9	0	1	0	1	0	1	0	1	0	1
Doc-10	5	0	1	3	3	2	5	1	2	0

Table.4 Word Probabilities Per Document

Document /topic	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
Doc-1	0.06	0.24	0.12	0.06	0.06	0.12	0.12	0.12	0.06	0.06
Doc-2	0.06	0.18	0.12	0.06	0.06	0.18	0.12	0.12	0.06	0.06
Doc-3	0.06	0.19	0.13	0.06	0.06	0.13	0.13	0.13	0.06	0.06
Doc-4	0.06	0.06	0.29	0.06	0.06	0.12	0.12	0.12	0.06	0.06
Doc-5	0.10	0.06	0.14	0.05	0.05	0.14	0.10	0.14	0.19	0.05
Doc-6	0.07	0.07	0.13	0.07	0.07	0.07	0.07	0.13	0.20	0.13
Doc-7	0.06	0.06	0.13	0.13	0.06	0.06	0.06	0.19	0.19	0.06
Doc-8	0.19	0.05	0.10	0.19	0.05	0.19	0.05	0.05	0.05	0.10
Doc-9	0.12	0.06	0.12	0.18	0.06	0.06	0.06	0.12	0.18	0.06
Doc-10	0.11	0.17	0.09	0.04	0.02	0.09	0.04	0.04	0.04	0.09

Table 5 Shows The Results Of An Analysis Of Word Meanings And Document Content Using Several Probabilistic Topic Models.

Document word	Word-1	Word-2	Word-3	Word-4	Word-5	Word-6	Word-7	Word-8	Word-9	Word-10
Doc-1	0.059	0.235	0.118	0.059	0.059	0.118	0.118	0.118	0.059	0.059
Doc-2	0.059	0.176	0.118	0.059	0.059	0.176	0.118	0.118	0.059	0.059
Doc-3	0.063	0.188	0.125	0.063	0.063	0.125	0.125	0.125	0.063	0.063
Doc-4	0.059	0.059	0.294	0.059	0.059	0.118	0.118	0.118	0.059	0.059
Doc-5	0.095	0.048	0.143	0.048	0.048	0.143	0.095	0.143	0.190	0.048
Doc-6	0.067	0.067	0.133	0.067	0.067	0.067	0.067	0.133	0.200	0.133
Doc-7	0.063	0.063	0.125	0.125	0.063	0.063	0.063	0.188	0.188	0.063
Doc-8	0.190	0.048	0.095	0.190	0.048	0.190	0.048	0.048	0.048	0.095
Doc-9	0.118	0.059	0.118	0.176	0.059	0.059	0.059	0.118	0.176	0.059
Doc-10	0.174	0.174	0.087	0.043	0.217	0.087	0.043	0.043	0.043	0.087

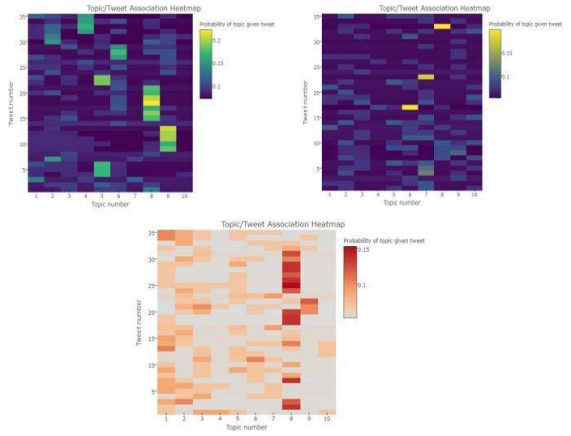


Fig 7: Probability Of Topic Distribution

In this work, topic modelling techniques were used to locate similar questions, and several similarity vectorization models were compared before the most pertinent questions were effectively obtained. Ensemble models, topic modelling, and similarity for k themes were used to define the 20 most pertinent questions for each user query. The total number of issues in the corpus was used to define the suitable document.

The definition of the received document is a question that is related to the specific document. The distributed modelling approaches strike a balance between solving the current issues and

Journal of Theoretical and Applied Information Technology
31st July 2024. Vol.102. No. 14

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

obtaining accurate results, as shown in table 7. Distributed LDA of the topic equilibrium model generated the best results in the comparative analysis shown in figures 10 and 11. It had superior accuracy in the estimation of related indicators.During the initialization phase of training, Harp will load the local data split on each node into memory, which means future disc I/O will not be required to access the training data. By default, Hadoop MapReduce uses the data splitting strategy it supports.

Hadoop offers distributed dataset abstractions, group communication, and synchronization methods for model data. However, parallelizing the core computation of the model update for machine learning algorithms causes issues with model consistency and synchronization. For networked machine learning applications, Harp's distinctive abstractions based on collective synchronization mechanism are advantageous in terms of expressiveness, efficiency, and effectiveness.

In this work, topic modelling techniques were used to locate similar questions, and several similarity vectorization models were compared before the most pertinent questions were effectively obtained. Ensemble models, topic modelling, and similarity for k themes were used to define the 20 most pertinent questions for each user query. The total number of issues in the corpus was used to define the suitable document.

The definition of the received document is a question that is related to the specific document. The distributed modelling approaches strike a balance between solving the current issues and obtaining accurate results, as shown in table 7. Distributed LDA of the topic equilibrium model generated the best results in the comparative analysis shown in figures 10 and 11. It had superior accuracy in the estimation of related indicators.

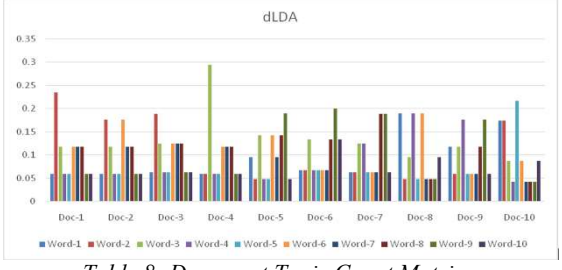


Table 8: Document Topic Count Matrix

Table 8 generates a matrix of phrases for documents, where each phrase represents the occurrence of terms in a document. In this matrix, terms are represented by rows and documents by columns. DLSA (Dynamic Latent Semantic Analysis) uses Singular value decomposition to break down the document-term matrix into smaller parts, allowing it to learn latent themes. DLSA is frequently used as a technique for dimension reduction or noise reduction.

Table 9 Shows The Results Of A Probabilistic Topic Model Analysis Of Document Content And Topic Meanings.

Document /topic	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
Doc-1	0.06	0.24	0.12	0.06	0.06	0.12	0.12	0.12	0.06	0.06
Doc-2	0.06	0.18	0.12	0.06	0.06	0.18	0.12	0.12	0.06	0.06
Doc-3	0.06	0.19	0.13	0.06	0.06	0.13	0.13	0.13	0.06	0.06
Doc-4	0.06	0.06	0.29	0.06	0.06	0.12	0.12	0.12	0.06	0.06
Doc-5	0.10	0.06	0.14	0.05	0.05	0.14	0.10	0.14	0.19	0.05
Doc-6	0.07	0.07	0.13	0.07	0.07	0.07	0.07	0.13	0.20	0.13
Doc-7	0.06	0.06	0.13	0.13	0.06	0.06	0.06	0.19	0.19	0.06
Doc-8	0.19	0.05	0.10	0.19	0.05	0.19	0.05	0.05	0.05	0.10
Doc-9	0.12	0.06	0.12	0.18	0.06	0.06	0.06	0.12	0.18	0.06
Doc-10	0.11	0.17	0.09	0.04	0.02	0.09	0.04	0.04	0.04	0.09

Table.4 Word Probabilities Per Document

Document word	Word-1	Word-2	Word-3	Word-4	Word-5	Word-6	Word-7	Word-8	Word-9	Word-10
Doc-1	0.059	0.235	0.118	0.059	0.059	0.118	0.118	0.118	0.059	0.059
Doc-2	0.059	0.176	0.118	0.059	0.059	0.176	0.118	0.118	0.059	0.059
Doc-3	0.063	0.188	0.125	0.063	0.063	0.125	0.125	0.125	0.063	0.063
Doc-4	0.059	0.059	0.294	0.059	0.059	0.118	0.118	0.118	0.059	0.059
Doc-5	0.095	0.048	0.143	0.048	0.048	0.143	0.095	0.143	0.190	0.048
Doc-6	0.067	0.067	0.133	0.067	0.067	0.067	0.067	0.133	0.200	0.133
Doc-7	0.063	0.063	0.125	0.125	0.063	0.063	0.063	0.188	0.188	0.063
Doc-8	0.190	0.048	0.095	0.190	0.048	0.190	0.048	0.048	0.048	0.095
Doc-9	0.118	0.059	0.118	0.176	0.059	0.059	0.059	0.118	0.176	0.059
Doc-10	0.174	0.174	0.087	0.043	0.217	0.087	0.043	0.043	0.043	0.087

Table 10 shows the results of an analysis of word meanings and document content using several probabilistic topic models.

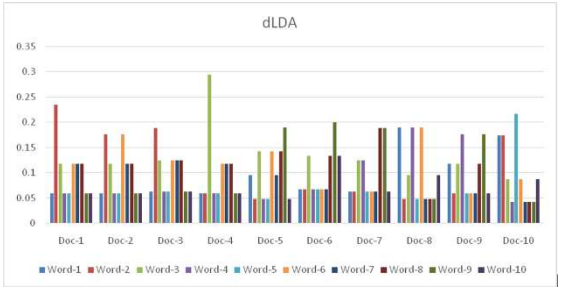
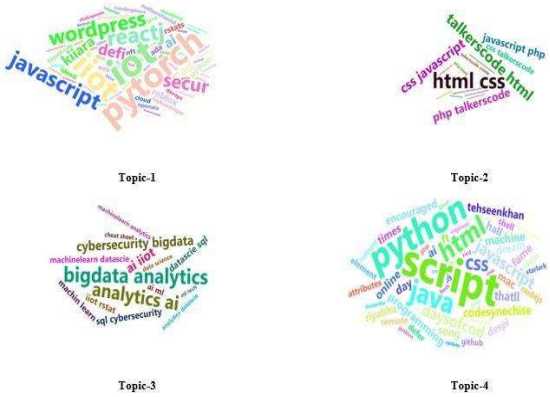


Fig 7: Probability Of Topic Distribution

The Hadoop ecosystem and probabilistic topic models are used to analyze document content and word meanings via graphical topic distribution probability examination.The graphical examination of topic cluster distribution probability in the Hadoop environment and several probabilistic topic models have been used to assess document content and word meanings.

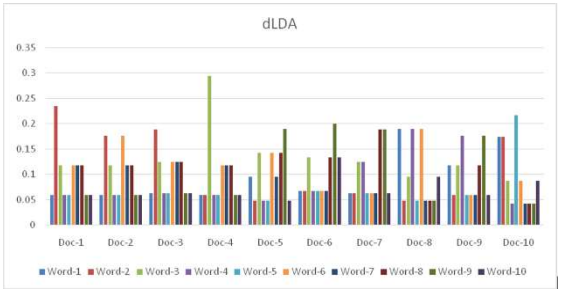
Several probabilistic topic models have been used to analyze the content of documents and word meanings, and the results are shown in fig. 9. The distributed LDA is supervised machine learning approach addressed best results.



The graphical examination of topic cluster distribution probability in the Hadoop environment and several probabilistic topic models have been used to assess document content and word meanings.

Several probabilistic topic models have been used to analyze the content of documents and word meanings, and the results are shown in fig. 9. The distributed LDA is supervised machine learning approach addressed best results

Fig 10: Comparison Results With 50 Topic Fig 10 And 11: Comparison Results With 50 Topic Cluster



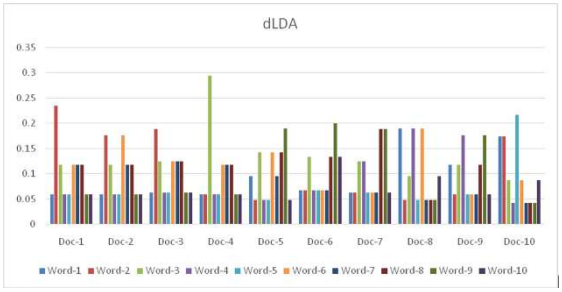


Fig 11: Comparision Results With 100 Topic Cluster Cluster

Methods	pre	acc	rec	f1	Topics
dLDA	0.87	0.89	0.77	0.82	50
dLSA	0.12	0.67	0.54	0.61	50
dLDA	0.93	0.94	0.87	0.90	100
dLSA	0.67	0.71	0.74	0.69	100

5. CONCLUSION

The justification of the study's findings rests upon the alignment of evaluation criteria with the identified research problem, encompassing scalability, accessibility, adaptability, and accuracy. These criteria are chosen to directly address the limitations of existing topic modeling techniques in job assistance and societal communication empowerment. By evaluating the proposed distributed file system-based approach against these criteria, the study provides evidence of its effectiveness in overcoming the identified challenges.

Significantly, the transparency and reproducibility of the findings are enhanced by establishing clear evaluation metrics. While analysis criteria may vary, the fundamental goal remains consistent across studies: to assess the efficacy of the proposed approach. In the results discussion, the synthesis of previously known facts with the study's findings contextualizes the significance of the results, highlighting areas of improvement, novel insights, and implications for theory and practice. This comprehensive analysis enhances the credibility and relevance of the study's conclusions, contributing to a deeper understanding of the research problem and its broader implications.

This study examines how topic models can provide insights into programming languages and the issues that experts face. By analyzing the types of questions asked, the study was able to draw conclusions that would not have been possible otherwise. The results showed that the types of inquiries are similar across programming languages and provided a method for determining which types of queries are primarily related to structural structure identifiers. Topic modeling creates a representation of the collection of text documents in the topic space by identifying themes present in every text document. Combining the topic modeling with dLDA algorithm and similarity measure produces superior results. Future improvements include analyzing more attributes (topics) to categorize user authorization systems, better time and space distribution of the models to get a better location for user displacement, and incorporating advanced technologies related to the Twitter, Stack Overflow, and Quora databases.This study examines how topic models can provide insights into programming languages and the issues that experts face. By analyzing different types of questions, we were able to draw conclusions that would not have been possible otherwise. The study found that the types of inquiries are similar across programming languages, and a method was developed to determine which types of queries are primarily related to structural structure identifiers. Topic modelling creates a representation of the collection of text documents in the topic space because it identifies themes that are present in every text document. Combining the topic modelling and the dLDA algorithm with a similarity measure produces superior results. Future improvements include analyzing more attributes (topics) in questions and answers and categorizing the user authorization system. Advanced technologies related to the platform, such as Twitter, Stack Overflow, and Quora databases, will be used to better distribute the models in terms of space and time to improve the user experience.This study examines how topic models can provide insights into programming languages and the issues that experts face. By analyzing different types of questions, we were able to draw conclusions that would not have been possible otherwise. The study found that the types of inquiries are similar across programming languages, and a method was developed to determine which types of queries are primarily related to structural structure identifiers. Topic modelling creates a representation of the collection of text documents in the topic space because it identifies themes that are present in every text document. Combining the topic modelling and the dLDA algorithm with a similarity measure produces superior results. Future improvements include analyzing more attributes (topics) in questions and answers and categorizing the user authorization system. Advanced technologies related to the platform, such as Twitter, Stack Overflow, and Quora databases, will be used to better distribute the models in terms of space and time to improve the user experience.

The conclusions drawn from this study underscore the significance and efficacy of the novel approach to modeling topics through a distributed file system for job assistance and societal communication empowerment. Through a thorough exploration of the identified research problem and the limitations of existing methodologies, the study establishes a compelling rationale for the proposed approach. By leveraging distributed file systems, the methodology addresses critical challenges related to scalability, accessibility, adaptability, and accuracy in topic modeling. The findings of the study affirm the effectiveness of this approach in overcoming these challenges, thereby enhancing the capacity of individuals to navigate digital platforms effectively for job opportunities and societal engagement. The conclusions are firmly grounded in the alignment of the evaluation criteria with the research problem, ensuring that the outcomes reflect a

comprehensive assessment of the proposed methodology's impact and significance. Moreover, the synthesis of previously known facts with the study's findings provides a nuanced understanding of the implications for theory and practice in the field. In essence, the conclusions drawn from this study not only validate the proposed approach but also underscore its potential to drive meaningful advancements in digital inclusion, economic empowerment, and social cohesion.

REFERENCES

[1]. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003

[2]. LI Hua-Meng,LI Hai-Rui,XUE Liang. TFIDF Algorithm Based on Information Gain and Informati Entropy[J]. Computer Engineering, 2012, 38(08): 37-40.

[3]. Hanchen Jiang, Maoshan Qiang, Dongcheng Zhang, Qi Wen, Bingqing Xia, Nan An. "Climate Change Communication in an Online Q&A Community: A Case Study of Quora", Sustainability, 2018

[4]. Campbell, J.C., Hindle, A. and Stroulia, E., 2014. Latent Dirichlet allocation: extracting topics from software engineering data. In The art and science of analyzing software data (pp. 139-159). Morgan Kaufmann

[5]. Rainer Lienhart, Stefan Romberg, and Eva H¨orster. Multilayer pLSA for multimodal image retrieval. In Proceeding of the ACM International Conference on Image and Video Retrieval, CIVR '09, pages 9:1–9:8, New York, NY, USA, 2009. ACM.

[6]. S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization provably. In Proc. the 44th Symposium on Theory of Computing (STOC), pages 145–162,2012.

[7]. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Annual Conference on Neural Information Processing Systems, pp. 556–562 (2000)

[8]. Yan X, Guo J Learning topics in short text using ncut-weighted non-negative matrix factorization on term correlation matrix,2013

[9]. Huang L, Ma J, Chen C (2017) Topic detection from microblogs using T-LDA and perplexity. In: 24th Asia- Pacific software engi- neering conference workshops,2018

[10]. W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In Proc. the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), pages 267-273 , 2003

[11]. Peng Zhang, Department of Mathematics, Zhejiang University, Hangzhou, 310027 China ; WanhuaSu Statistical inference on recall, precision and average precision under random selection,2012

[12]. Edi Surya Negara, Dendi Triadi, Ria Andryani,Topic Modelling Twitter Data with Latent Dirichlet Allocation Method,DOI:1109/ICECOS47637.2019. 8984523,,Electronic ISBN: 978-1-7281-4714-7,Print on Demand(PoD) ISBN: 978-1-7281-4715-4,2019

[13]. Pablo Ormeño ,Marcelo Mendoza , and Carlos Valle ,Topic Models Ensembles for AD-HOC Information Retrieval Information 2021, 12(9), 360; https://doi.org/10.3390/info12090360

[14]. Dhelim, S.; Aung, N.; Ning, H. Mining user interest based on personality-aware hybrid filtering in social networks. Knowl. Based Syst. 2020, 206, 106227

[15]. Baechle C., Huang C, Agarwal A, Behara, R, Goo J,. Latent topic ensemble learning for hospital readmission cost optimization. Eur. J. Oper. Res. 2020, 281, 517–531.

[16]. Pourvali M., Orlando S., Omidvarborna H. Topic Models and Fusion Methods: A Union to Improve Text Clustering and Cluster Labeling. Int. J. Interact. Multimed. Artif. Intell. 2019, 5, 28–34

[17]. Fiandrino, S Tonelli, A. A Text-Mining Analysis on the Review of the Non-Financial Reporting Directive: Bringing Value Creation for Stakeholders into Accounting. Sustainability 2021, 13, 763.

[18]. yerragudipadusubbarayudu, AlladiSureshbabu,"Distributed Multimodal Aspective on Topic Model using sentiment analysis for Recognition of Public Health Surveillance" Expert Clouds and Applications, ISBN: 978-981-

16-2126-0, https://link.springer.com/chapter/10.1007/978-981-16-2126-0_38

[19]. Ammirato S., Felicetti A.M., Raso C. Pansera, B.A Violi, A. Agritourism and Sustainability: What We Can Learn from a Systematic Literature Review. Sustainability 2020, 12, 9575.

[20]. Farkhod A , Abdusalomov A, Makhmudov F. , Cho, Y.I. LDA-Based Topic Modeling Sentiment AnalysisUsingTopic/Document/Sentence (TDS) Model. Appl. Sci. 2021, 11, 11091.

[21]. P. Vidyullatha, P. venkateswara Rao, Big data sentimental analytics on social media using Rhadoop-Hive, "Materials Today: Proceedings", January 2021.

[22]. Devisetty, S.D.P., Sai, Y.M., Yadav, A.V., Vidyullatha, P., "Sentiment analysis of tweets using rapid miner tool ", International Journal of Innovative Technology and Exploring Engineering, 2019 8(6), pp. 1410– 1414

[23]. Chang Y. L., & Ke J. (2024). Socially Responsible Artificial Intelligence Empowered People Analytics: A Novel Framework Towards Sustainability. *Human Resource Development Review*.2024

[24]. Bhatti, I., Rafi, H., & Rasool, S. (2024). Use of ICT Technologies for the Assistance of Disabled Migrants in USA.

[25]. Smythe, T., Ssemata, A. S., Slivesteri, S., Mbazzi, F. B., et al. (2024). Co-development of a Training Programme on Disability for Healthcare Workers in Uganda. *BMC Health Services*