FedExLSA: Design and Development of algorithms on Federated Data Exploration of Topic Prediction Using Latent Semantic Analysis

Original Scientific Paper

Saranya M*
SRM Institute of Science and Technology, Kattankulathur
Research Scholar, Department of Computing Technologies, School of Computing
Kattankulathur, Chennai, India
sm2317@srmist.edu.in

Amutha B
SRM Institute of Science and Technology, Kattankulathur
Professor, Department of Computing Technologies, School of Computing
Kattankulathur, Chennai, India
amuthab@srmist.edu.in

*Corresponding author

Abstract *– Every government in the world has multiple departments that must function and operate to address the various inquiries raised by the population. The government's diverse range of websites offers citizens a platform to submit inquiries, thereby facilitating the fulfilling of their requirements. Comprehending the subjects addressed in People Query is essential for government services. Unstructured query data is analyzed using extracting information from text techniques such as allocation of Latent diffuser (LDA) and analysis of hidden semantics (LSA). LSA outperforms other methods in terms of performance because of its minimal complexity and quick installation process. Research on decentralized learning techniques for natural language processing (NLP) is necessary due to concerns about limited data availability and privacy. Federated learning (FL) employs methods that enable different users to collectively train an integrated broad model while maintaining their information regionally stored and accessible. Nevertheless, the current body of literature lacks a thorough examination and evaluation of FL techniques. Data federation is an approach to data integration that allows the government to access and query data from multiple diverse sources as if they were a single, unified repository. Functioning as a form of data virtualization, it facilitates the creation of a comprehensive representation of data, thereby enhancing operational efficiency and the accuracy of decision-making. FedEx utilizes Federated Learning to apply topic modelling techniques to common NLP tasks. The proposed structure integrates the FL Methodology with Latent Semantic Analysis to deliver outcomes for intelligent data analysis and management.*

Keywords*:* *Natural Language Processing, Topic Modeling, Federated Learning, Latent Semantic analysis, Text Mining*

Received: March 27, 2024; Received in revised form: June 7, 2024; Accepted: June 7, 2024

1. INTRODUCTION

Open governments and easy-to-contact are usually the most effective. The government is one of several that have recognized the significance of this. This is demonstrated by the creation of the Government Website, which addresses the concerns and desires of the general public. Facebook, email, and a personal visit to the government center’s office are some other ways to get in touch with the website. Researchers examine the Internet for possible data sources. Researchers can find data on websites in various ways. A key component is the website's focus on text messaging and its restriction on character counts. One more thing about the website is that it has an API that makes it accessible from any location in the globe. In addition, the Separate Website program has a large user base throughout several states. Public complaints in the government are also received by numerous regional offices through online platforms. By evaluating all the searches, the government might access the most recent data on the website information that the users themselves have contributed. If the government is serious about improving its performance, it should hear the recommendations, comments, and opinions of its citizens. A great deal of information was retrieved from these individuals during the interrogation. It would take a very long time to read all of this in sequence. However, the government needs answers to these questions to move forward; therefore, it would be a great shame if they were disregarded. The results of this study could help the government do its job better. With public support, new policies can be sustained. Besides the accelerated procedure, the government also can deal with any concerns that might arise. Take the licensing procedure as an example. If someone has a problem with it, the government should try to fix it by making it more clear and transparent.

This investigation utilizes textual data. When text documents are grouped, overlapping data results. A substantial amount of ground could potentially be explored with a single inquiry. Consequently, this research employed topic modelling methodologies. The objective of this study is to investigate the application of latent semantic analysis (LSA) and latent Drichlet allocation (LDA) to topic modelling. A set of methods known as topic modelling is employed to uncover concealed subjects within a query [1]. There are two perspectives from which to examine topics: probabilistic and linear. Linear topic modelling is surpassed by positivity topic modelling. This is illustrated by the "Latent Semantic Indexing" linear topic model. The technique in question is referred to as "Latent Semantic Analysis" (LSA). Probabilistic topic modelling is illustrated by the works of Latent Drichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) [2]. The LSA model yields results that do not account for the correlation between the query and the corpus. Using LDA, one could investigate the interrelationships between the documents in the corpus. Numerous scholars have implemented LDA on datasets other than Query data. By applying the LDA method for topic modelling, [2] was able to ascertain the content that was discussed in People Queries about two distinct industries.

One way to train ML models [3] collaboratively without revealing any local data is by using Federated Learning (FL). It often requires several clients to collaborate with one or more servers that mediate the setting of agreements, privacy assurances, and the aggregation of node updates. Many researchers are looking into FL's possible application in topic modelling because of its privacy-preserving and decentralized data-leveraging features. Many scholars have focused on developing LDA-like or federated frameworks [4], whereas others have proposed federated general-purpose topic models [5]. FedLSA, also known as Federated Latent Semantic Analysis, contributes significantly to the large field of data science and natural language processing (NLP) research by delivering innovative solutions to pressing challenges. FedLSA is a system that allows for the analysis of distributed text data while adhering to tight privacy constraints, which is especially significant in an era where privacy is becoming increasingly crucial. Furthermore, as the number of edge devices and Internet of Things (IoT) technologies grows, its decentralized learning method enables collaborative analysis across several data sources without centralizing the data. This not only ensures the ability to handle large volumes of text data but also encourages the creation of applications that can be utilized in various fields, including healthcare, finance, and social media analysis. FedLSA marks a fundamental shift in data science and NLP research. It encourages collaboration and information sharing through federated analysis while maintaining data privacy. This technique user in a new era of collaborative, privacy-preserving analysis. However, no research has been conducted on federated latent semantic analysis implementations. Finding the best topic modelling technique for data derived from citizen queries in government is the goal of this research. Topic prediction often requires a large training dataset, which is not always readily available.

Below is a summary of the work's contributions:

- Our proposed framework is Fed LSA, which stands for federation. With its help, a large number of users could train a topic model with LSA and SVD.
- Fed LSA to perform even better, recommend combining it with machine learning. The goal of this method is to increase the degree to which text features resemble their abstract counterparts.
- Using the DigiLocker NAD, IHMCL, Kerala Startup Mission, and KILA datasets all of which are publicly available online, test our methods to make sure it works and also investigate its performance in various federated settings.

In this, Federated Latent Semantic Analysis (FedLSA) provides a strong framework for studying latent themes in text data from many government departments while protecting data privacy and confidentiality. FedLSA uses secure communication and aggregation techniques to allow for collaborative data analysis without centralizing sensitive information. The benefits of greater privacy, cooperative insights, and scalability make FedLSA an interesting technique for governmental data analysis, despite potential downsides like as communication costs and aggregation problems.

2. LITERATURE SURVEY

Text mining is the practice of using appropriate analysis to extract valuable information from a database of documents. Text mining uses query data, which is essentially unstructured text data. Data sources can be mined for important information using extraction, which involves discovering and analyzing interesting patterns. Using preprocessing procedures, text miners can convert the query's unstructured data into an intermediate form that is more specifically arranged [6]. The primary objective of preprocessing is to enhance accuracy. There are several People Queries collections on sites like the web. In particular, there is a growing need for automated methods that can read, evaluate, and summarize large document sets. Various topic modelling techniques are used in numerous applications. Each topic modelling technique differs from both centralized and federated learning. Table 1 displays the comparison.

Table 1. Comparison of the various topic models

Technique	Context	Advantages	Disadvantages
LDA	Centralized	Topics that could be easily understood and are commonly utilized	Requires significant computational resources and highly influenced by hyper parameters.
GD	Centralized	Efficiently computes straightforward and positive outcomes, capable of processing matrices with many zero elements	Requires the specification of the number of Topics and the convergence of local minima.
LSA	Centralized	The concept is straightforward and does not require a specific Topic number.	Assumes linear correlations, more difficult to analyze
RmsProp	Centralized	Documents the progression of a Logic	Requires data with time stamps due to its high level of intricacy.
Federated LDA	Federated	Ensures confidentiality, decentralized analysis	Overhead in communication and intricate aggregation
Federated GD	Federated	Ensures confidentiality, manages decentralized information	Secure aggregation is necessary to address convergence difficulties.
Federated LSA	Federated	Efficient and expeditious data privacy	There are difficulties in communication and aggregation.
Federated RmsProp	Federated	Effectively manages differences in local datasets between clients.	Requires a high level of intricacy and necessitates synchronization.

Each topic modelling approach has a unique set of tradeoffs. Centralized approaches such as LDA, GD, LSA, and RmsProp are effective, but they might be constrained by computing resources and data privacy issues. Federated techniques alleviate privacy concerns and enable collaborative study of distant datasets; however, they add complexity to communication and aggregation.

As proposed by [7], the objective of Federated Topic Modelling (FL) is to train a model effectively using decentralized data from multiple clients. To reduce communication costs without sacrificing performance, numerous algorithms have been developed, including Fed Avg [8] and FedRmsProp [9]. In general, these algorithms operate in a two-step process: initially, they train the model using local data while simultaneously synchronizing the server with the latest model weights. Each client transmits the modified weights to the server for aggregation once training is complete. These potent FL algorithms are compatible with our Federal LSA architecture. Federated topic models have received limited attention from researchers. The federated topic modelling approach, as illustrated in reference [10], showcases the implementation of novel methodologies including heterogeneous model integration and topic-wise normalization. They implemented an innovative local differential privacy (LDP) technique to federate the LDA. The focus of this study is topic modelling via LDA. The focus of this study, on the other hand, is topic modelling in federated environments via LSA. Recently, there has been much buzz around latent topic modelling, which is an unsupervised approach to topic discovery in large document collections. An example of such a model is the LDA [11]. Using statistical (Bayesian) topic models, Latent Drichlet Allocation (LDA) is a well-liked approach for text mining. A generative model of writing is what LDA achieves. Consequently, it strives to generate a document that is relevant to the given subject. This approach can also process other types of data. Various methods, such as latent Drichlet co-clustering, topic modelling, author-topic analysis, and temporal text mining, use topics to express queries; each topic is a discrete probability distribution that specifies the likelihood of each word appearing in that topic. These subject probabilities can be used to describe a document. In this sense, a "Query" is just a "bag of words" sorted just by Topic and word count. As shown graphically in Fig. 1 and the LDA Mathematical Notation shown in Table 2. The LDA is an example of a generative probabilistic model, the next step for LDA to produce a specific corpus is to follow the following procedure:

1. Choose a distribution, $\theta t \sim D(y)$, over words for any subject t, where t is in the interval {1… t}.
2. For every query Ds, where q is an integer from 1 to q, Select a distribution where topics are $\delta q \sim D(\alpha)$. As if they were random variables obeying Drichlet distributions with parameters x and y, respectively, the word distributions for themes and the document topic distributions were considered. The likelihood of the corpus in Equation 1 is given by a set of Q Queries denoted as $D = \{D1, D2, ..., DQ\}$.

$$p(D|x,y) = \prod_{q=1}^{q} \int p(\alpha|\beta) \prod_{W=1}^{W} \sum_{a=1}^{t} (p(aqw|\delta q)p(\theta a|y))d\alpha q, d\alpha a \tag{1}$$

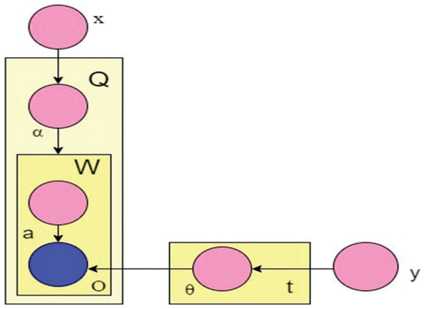


Fig.1. Graphical Representation of the LDA Model. The blue shade represents the Observed Model. Pink Shaded represent the Latent Variable.

The corpus was subjected to multiple runs of the LDA algorithm with varying numbers of topics. They started each experiment with the two hyper parameters set to $x = 0.1$ and $y = 0.01$

Singular Value Decomposition (SVD) is also utilized by Latent Semantic Analysis (LSA) to reorganize information. Using a matrix-based technique, SVD reorganizes and calculates all contractions of vector space. In addition, compute the reductions in vector space and arrange them in descending order of importance. The meaning of the text can be inferred using the most significant assumption if the LSA [12] assumption phase does not use the least important assumption. Finding words with a similar vector is one approach to finding words with many similarities. An important initial stage in LSA is to gather a large amount of relevant content and arrange it according to topics. In the second step, create a matrix that shows how often each word and document appears. Kindly provide the cell names (e.g., "document a," "terms b") and dimensional values (m for terms and n for documents) for each entry so that there is no room for misunderstanding.

Table 2. Table of Mathematical Notation in Latent Drichlet Allocation

Notation	Description
Q	Number of Queries in Corpus
N	Number of topics
W	Number of words in one Query
X	hyper Topic-specific parameter (if symmetric scalar vector t)
Y	pre-word distribution hyper parameter (if symmetric scalar vector)
α	Topic Mixture ratio QXT Matrix (one row per Q Query)
θ	Word distribution TXS Matrix (S is the size of Vocabulary) topic t with θt
a	topics generating word Q vectors(one per w words)value of at is in (1…t)

Running the calculations and making adjustments to each cell is the third stage. To conclude, SVD is going to be an enormous assistance in calculating all the diminutions and creating the three matrices. The SVD operational principle was discussed in Section III.

This research has connections to three different fields: federated learning, similarity information, and LSA-based [13] topic modelling. a) One method for training models proposed by [14] is FL, which stands for Federated Topic Modeling. Its objective is to facilitate the effective utilization of distributed data among several People. Since its inception, numerous algorithms have been developed to reduce communication costs without compromising performance. These include FedAvg's and FedRmsProp. There are usually two steps to these algorithms in a typical implementation: first, the clients use local data to train the model while synchronizing the most recent model weights with the server. After training is completed, the clients send the updated weights back to the server for aggregated data. These cutting-edge FL algorithms are compatible with the proposed Fed LSA design. Few academics have focused on federated topic modelling. As an illustration, federated topic modelling has been described in [15]. It integrates innovative approaches such as topic-wise Semantic analysis, private Metropolis-Hastings, and heterogeneous model integration. Constructed federated LDA using a novel local differential privacy method. The major focus of this study is LDA-based topic modelling. In contrast, this study delves into LSA-based topic modelling in federated settings.

3. METHODOLOGY

Federated Latent Semantic Analysis (FLSA) is an approach that ensures the preservation of privacy while analyzing vast quantities of text data distributed across multiple clients, including institutions and smartphones. The methodology commences by assigning unique local datasets to each client and initializing local models with shared initial parameters. At the local level, individual clients perform data preprocessing, compute Term Frequency-Inverse Document Frequency (TF-IDF) vectors, and employ Singular Value Decomposition (SVD) to extract latent semantic structures. Based on their LSA results, clients subsequently generate local model updates and implement privacy-preserving strategies. Local updates are transmitted to a centralized server, which securely aggregates them while maintaining the confidentiality of individual data. By merging local modifications, the server modifies the global model and returns the updated model to the clients. The updated global parameters are subsequently incorporated into the local models by the clients, thereby enhancing the local LSA tasks. The process is iterative, consisting of multiple iterations of local processing and global aggregation. This iterative approach progressively enhances the global model, thereby improving the accuracy and generalizability of the latent semantic structures, all the while safeguarding data privacy. Several FedLSA stages are illustrated in Fig. 2.

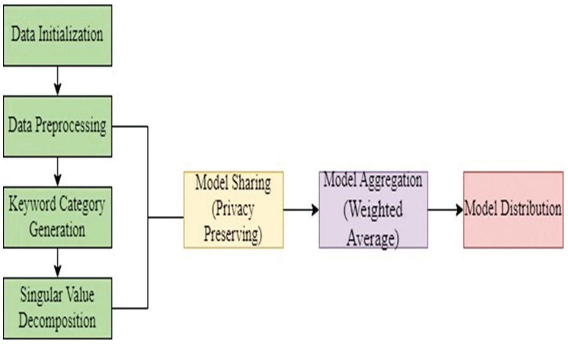

Data Initialization
Data Preprocessing
Keyword Category Generation
Singular Value Decomposition
Model Sharing (Privacy Preserving)
Model Aggregation (Weighted Average)
Model Distribution

Fig 2. Stages of Federated Latent Semantic Analysis

By utilizing Latent Semantic Analysis (LSA), the dimensionality of a document representation is diminished. In a word vector, LSA employs a vector comprising latent semantic concepts. A large word-document matrix is subjected to singular value decomposition (SVD) by LSA [14] to reduce the dimensionality of the data.

Three matrices comprise a massive term-document matrix: one for documents, one for singular values, and one for concepts and terms. To reduce the dimension of the word document matrix, singular value decomposition (SVD) is abstained from in this instance. This methodology is founded upon two fundamental assumptions: (1) the number of subjects addressed in each document and (2) the vocabulary size corresponding to each subject as determined by Equations (2) and (3), respectively. Let us consider two variables: the number of subjects (T) and the size of the vocabulary (V). Within the given context, the notation ^{3} (t, d) signifies the occurrence of topic t in document d, whereas γ (w, t) denotes the creation of term w by topic t. The following is one possible configuration for the two assumptions:

$$r\,(t|dr) = \mu_{(t,d} \; \textstyle\sum_{w \in T}^{t} \; \mu_{(t,dr)} = 1 \tag{2}$$

$$r\,(w|t) = \beta_{(w,t)} \; \textstyle\sum_{w \in V}^{v} \; \beta_{(w,t)} = 1 \tag{3}$$

Additionally, the topic of the paper and each word are generated using Singular Value Decomposition. Finding two semantic vector matrices A and B for a matrix M such that $M \sim AB$ and two matrices W and H such that $M \approx WX$ are the three main goals of LSA. Cutting down on the following L (φ) loss concerning W and X is an easy way to do it.

$$L\,(M, W, X) = (M_{i,j} - M\hat{}_{i,j})^{2} \tag{4}$$

It is possible to express each document using a count (column) vector overlaid on top of the bag-of-word representation. Concerning the i-th client, M_i represents the count feature matrix for documents. The union of all matrices with $i = 0,...,X$ allows for the decomposition of this matrix.

3.1. FEDERATED LSA

The FedLSA factorization procedure for client-distributed matrices is illustrated in Fig. 3. Use the GD algorithm to minimize loss within the federated learning architecture. However, as shown in Section IV studies, using FedAvg's approach to optimize the loss on each client alone results in poor topic models. Below are the factors that FedLSA follows. Imagine a network of X client devices, where the i-th device's data distribution function (ddi) might vary for different values of i. In distributed learning environments, X clients are usually trained using a single global model. Finally, under the assumption of a Federated Semantic Analysis (FedLSA) architecture with layers = $0,...,L$, all clients share the set of weights $\varphi = \{W\}\ L{=}0$. Mastering the art of limiting the average loss for every client could help achieve the global goal. This is the general principle behind many federated learning approaches.

For instance, fedLSA aims to minimize the following goals in Equation (5):

$$Min\ L\,(\varphi) = \textstyle\sum_{i=0}^{X} Ni\ Li(\varphi) \tag{5}$$

The Weight of each device i, represented as $Ni > 0$, and the number of Peoples, X, are input into the local objective function, $Li\,(\varphi) := Pxi{\sim}ddi\,[li\,(xi;\varphi)]$.

Unfortunately, statistical heterogeneity means that there is no silver bullet when it comes to fitting the global model to individual clients. This, in turn, impacts the degree to which a client's local distribution resembles the population distribution. People who share fewer attributes might view this strategy as unfair. In comparison, X local models = $\{W'\ i\}\ L = 0$ are learned, where each model is trained using only ddi.

The data distribution of each client i determines the set of weights χi to the maximum extent possible. Considering that each client typically has limited data that may not be sufficient to train a comprehensive model without over fitting, the total number of parameters that must be learned across all clients increases as X decreases. Using shared learning problems or comparable client data distributions, simultaneously learn X distinct models.

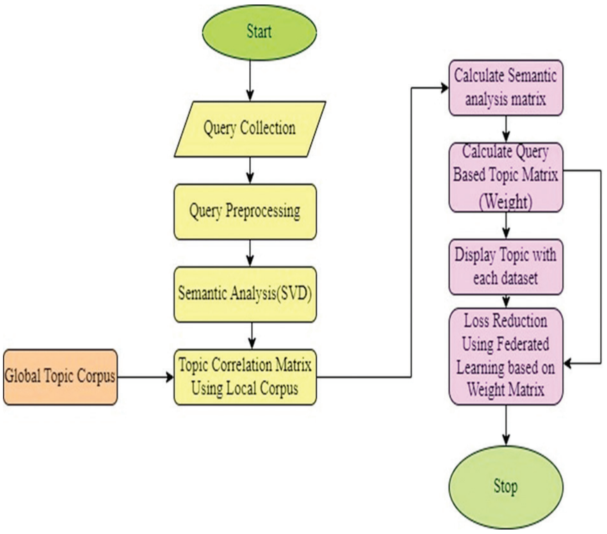

Start
Query Collection
Query Preprocessing
Semantic Analysis(SVD)
Global Topic Corpus
Topic Correlation Matrix Using Local Corpus
Calculate Semantic analysis matrix
Calculate Query Based Topic Matrix (Weight)
Display Topic with each dataset
Loss Reduction Using Federated Learning based on Weight Matrix
Stop

Fig. 3. Architecture of Client distributed Matrices using Federated LSA Model

Aiming for a compromise between X distinct local models and one global model is the optimal approach to data utilization. Ensure that all models utilize the same vocabulary of combined components, but have each client modify their model by their specific distribution in their local area. By factorizing the subsequent equations (1) and (2) using layer-wise decomposition, they construct each weight matrix.

Algorithm1. Federated Latent Semantic Analysis for X Number of Clients Communication

Server Side Execution

```
Server Executes:
Initialize $W$ (0) and $\varphi$
for each round $r$= 0, 1,2 . . . do
$Qr$ ←( group of $X$ clients)
for each client $i \in Qr$ in parallel do
($W(r+1)$
$i$, $\varphi$ ($r$+1)
) ← Update ($i$, $W(r)$, $\varphi$ ($r$))
```

```
end for
```

Client Side Execution

```
Client Update ($i$, $W$, $\varphi$): //Run on client $i$
for each local updation from 0 to $U$ do
For each $Wx\in\varphi$
$\varphi M$ Compute the Similarity Measure
loss $L$ is defined in Eq. (3)
return ($W$, $\varphi$) to serve
end for each
end
```

Algorithm 1 displays the entire pseudo code of the proposed FedLSA framework. This method describes execution on the server and client sides. Existing Federated learning algorithms like FedAvg's and FedRmsProp are compatible with this architecture, which is dubbed FedLSA.

4. EXPERIMENTAL RESULTS

Demonstrate the efficacy of our algorithms by running them on many publicly available datasets and comparing their results with those of the state-of-the-art FedLSA. Python Tensor Flow is used to implement all the models.

4.1. DATASETS

In the tests, the four real-world text datasets represent People's Queries related to different topics. Tesz, which stands for "Questions and Answers in Various People Queries," is associated with the following four datasets.

a. Digi locker NAD: This dataset is a subset of Dig locker, which contains user queries about many domains, including education and various schemes. Around 1,250 authentic and encrypted queries written in English make it up.

b. IHMCL: Transportation-related queries, such as Fas tags, are a component of the Government Highway Management, which includes this data set. It has 800 queries written in English and authorized by users.

c. Kerala Startup Mission: One thousand queries for the Kerala State's Medical and Educational Schemes are contained in this collection. The Government of Kerala was responsible for its upkeep.

d. KILA: This is a database of 2000 questions about various forms of education (seminars, conferences, workshops, etc.) that have been posted on People. The Kerala government made this dataset available.

Table 3 shows the fundamental statistics of the datasets. "Files" indicates the total number of queries in the Records dataset, "terminology" indicates the total number of terms in the dataset, and "types" indicates the total number of categories in the dataset. A dataset's "Record length" is its mean Record length. After removing stop words and tokens using text preprocessing techniques, the figures were calculated. For our experiments, they used 70% of the datasets for training and 30% for testing.

Table 3. Data Preprocessing

Dataset	Files(Queries)	Terminology	Types(Fields)
DigiLockerNAD	1200	2300	3
IHMCL	800	1280	4
Kerala Startup Mission	1000	1600	2
KILA	2000	3460	5

4.2. COHERENCE METRICS

For query-based data, three separate types of coherence metrics should be used: PMI, LSA, and Word Embedding (WE) [11]. Here, we will go over the 9 metrics that come out of these measurements. It begins by outlining the existing PMI- and LSA-based metrics for theme consistency assessment and then provides a novel Word Embedding-based statistic. The top $n = 20$ words ($\{w1, w2,..., w10\}$) chosen based on their probabilities ($p\ (w|z)$) in the collection μ could represent a topic t inside this subject. One could determine the coherence of a topic by averaging the semantic similarity of the word pairs related to it (Equation (6).

$$Topic\ Coherence(x) = 1/\textstyle\sum_{x=1}^{y-1} a = x + 1CS\ (w_a, w_b) \tag{6}$$

Demonstrated that the pair-word PMI might represent the coherence [16] of topics identified in both the standard and Query corpora. For additional accuracy, they could use Equation (6) to determine how similar wa and wb are. This is accomplished by pulling co-occurrence statistics from a backdrop corpus that contains DigiLocker NAD, IHMCL, KILA, and the Kerala Startup Mission. Equation (1) is also used for this objective.

Note that to precalculate the PMIs of word pairs, certain extra datasets are required. Finding out how similar two-word pairings are in meaning is another usage of LSA [17]. Words are represented by dense vectors in the reduced LSA [18] space (Vxi) to apply LSA. To obtain this vector from a background corpus, Singular Value Decomposition is employed. The degree of similarity between two words can be assessed using the LSA metric [19], which uses a cosine function to calculate the distance between the word vectors. The following is substituted into Equation (6) within Equation (7) to accomplish this:

$$C(w_a,wb) = PMI(w_a,wb) = logp(w_a,wb)/p(w_a)Xp(w_b) \tag{7}$$

As indicated before, word embedding is more accurate than LSA. Table 5 displays the results of the coherence score for several topic modelling techniques. The topics generated by federated learning approaches, such as FedLDA, FedAvg, FedGD, and FedRmsProp algorithms, are of high quality. These algorithms provide instances of topic terms on the dataset. Table 5 demonstrates that FedLDA tends to produce repetitive subjects.

While the topics of FedAvg and FedRmsProp appear diversified, they are less informative and lack coherence. The Topic Quality of FedLSA is superior to those of other methods. Fig. 4 displays the coherence score graphically. Not long ago, this investigation was conducted. Application of Word Embedding vectors Vwa, which are obtained using a Word Embedding model that has been pre-trained on a large text dataset. The Topic Coherence Notation is displayed in Table 4.

Table 4. Table of Notations for Topic Coherence

Notation	Description
CS	Coherence Score
PMI	Point Wise Mutual Information
W_a, w_b	Two Different Kind of Words
Vxi	Vector Space
$p(w_a), p(w_b)$	Probability of a Particular word

According to Equation (8), the cosine similarity of two words' word vectors is larger when the words have a comparable semantic meaning.

$$CS\ (wa, wb) = cosine\ (Va, Vb) \tag{8}$$

Table 5. Coherence Score of Topic Modeling Methods

Technique	DigiLockerNAD	IHMCL	Kerala Startup Mission	KILA
LDA	0.524	0.410	0.523	0.467
LSA	0.536	0.456	0.510	0.426
LDAGD	0.517	0.345	0.444	0.356
LSAGD	0.578	0.444	0.543	0.432
FedLDA	0.657	0.334	0.437	0.543
FedLSA	0.432	0.523	0.324	0.326
FedGD	0.523	0.432	0.324	0.467
FedAvg	0.434	0.433	0.456	0.345
FedRMSprop	0.343	0.435	0.439	0.346

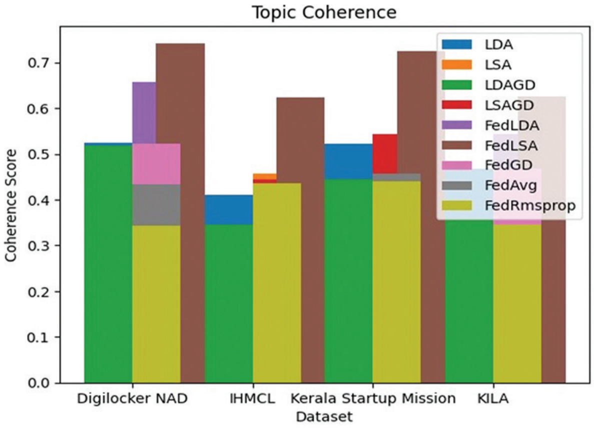


Fig. 4. Topic Coherence of Various Topic Modeling Techniques

They examined the similarity between WE, PMI, and LSA measures and human evaluations using the methodologies described in Section III. They also examined whether the WE-based metric could capture the coherence of People Queries topics.

4.3. COMPARISON METHODS

From centralized to federated, these are the topic modelling strategies they used in our studies.LDA and LSA are examples of traditional topic models used with centralized text data.

- Centralized GD-based Methods: To confirm that Semantic analysis is effective for the centralized LSA topic modelling, they incorporate GD-based LSA [9] methods (LSA+GD) into our trials. The main idea of LSA+GD is to maximize LSA's least square loss using mini-batch GD.
- Three federated topic modelling approaches are put into practice by us: FedAvg, FedGD, and FedRmsProp; FedLSA, which is based on variational inference; and FedRmsProp.

4.4. EXPERIMENTAL SETUP

In this experiment, they constructed datasets for several clients according to the methods described in [20]. For the sake of precision, it is assumed that the training samples for each client are chosen at random with class labels based on a categorical distribution over l classes, where v is a vector with elements ($vi > 0$, $i \in [1, l]$ and $Pvi = 1$). They pull $v \sim Dir$ (µq) from a Drichlet distribution, where q is the label distribution of a specific dataset and β controls the degree of client identity, to generate a set of clients that are not identical. Every client has the same distribution relative to q as β gets closer to infinity. On the other hand, as β gets closer to zero, each client only saves instances from one label. To conduct our experiments, they manipulated the heterogeneity of the client data using β and generated different FL settings by changing the client number N. Throughout the experiment, they allocated N to the set {10, 20, 30, 40}. They divided the overall sample size by the People number N to obtain the number of documents (Queries) given to each client. This ensured that our results would be similar. They then create a test and training set using the aggregated topic weight vectors from all documents. They then used the findings to calculate the accuracy and macro F1 score using a Logistic Regression (LR) classifier. Federated topic modelling is the next step in this procedure. By adjusting the value of n to 10, 20, 30, and 40 for all datasets, thorough results were obtained. This federated topic modelling method uses a constant participant fraction of $P = 1$ throughout all iterations. With each cycle, they tweaked the local batch size from the set {20, 40, 60, and 80} and the local GD training epoch count from the set {10, 20, 30, and 40}. When FedAvg and FedRmsProp are run by default, the hyper parameters are set to [21].

5. RESULT AND DISCUSSION

Following this, discuss more about the topics produced by the top model, which is the optimized LSA model that performed best across all of these criteria in terms of the topic Coherence Score. Not only are the five resulting subjects easily distinct but they are also thoroughly relevant and cohesive. Look at Table 6 for the subject keywords for every single idea. Additionally, it delves further into each subject: • DigiLocker Non-Disclosure Agreement: This covers matters about the realm of education and student conduct. The fact that users provide comments on elements about education suggests a strong connection with the Dig Locker app, and this app in particular. User comments regarding app performance and troubleshooting in the education domain are the focus of this section.

- IHMCL People Support: This section addresses Transportation-Related Questions, Concerns, and management, service-related problems, fast tags, and Toll Information. The significance of dependable and trouble-free transportation services for individuals is showcased. Regarding the Kerala Startup Mission, this section addresses public comments on health and education initiatives. Surprisingly, it is highly related to the knowledge and expertise acquired by the Kerala Government's various schemes. Disagreements over the app's usability and user interface are widely discussed.
- KILA: This category houses all of the questions that attendees of educational events like seminars, conferences, and workshops may have.

Table 6. Keywords related to each Topic based on the Dataset

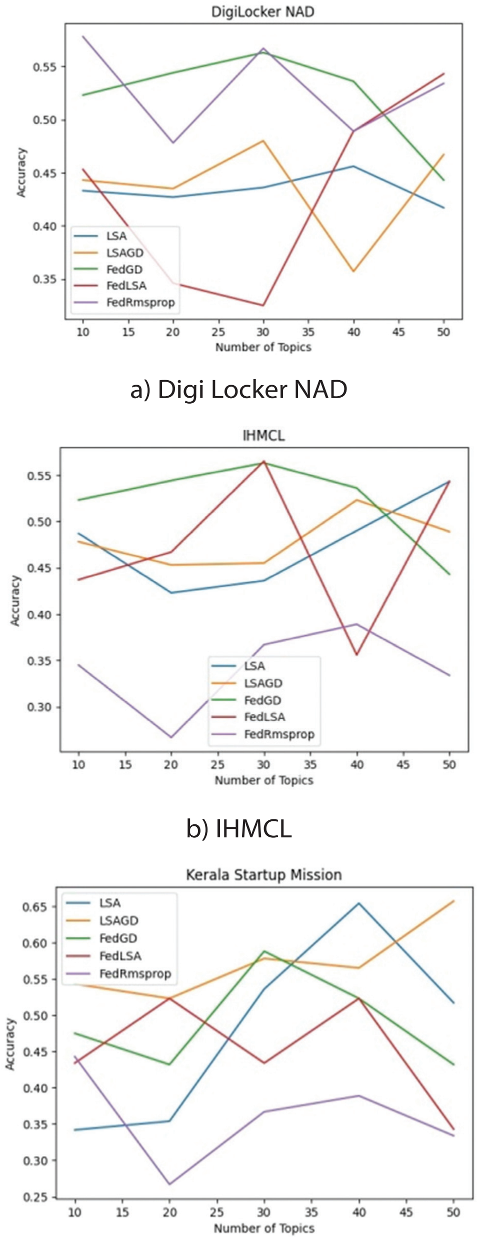
Topic	Topic Keyword
1	['Student',' Certificate', 'Digital’‘new', 'download']
2	['People', 'Transport', 'Renew', 'FasTag', 'car']
3	['Citizen', 'Medisep', 'Scheme',' Scholarship']
4	['Education', 'Conference', 'Seminar',' workshop']

Tables 5 and 7 show the results of text classification metrics and coherence scores for different topic modelling methods across the four datasets and the graphic representation shown in Fig. 5. On four separate topics (X=10, 20, 30, and 40), they averaged the provided coherence scores, F1 scores, and accuracy values. The setting for the FL environment is $\varphi = 1.5$ and $X = 10$.

These are the key points to remember. First, these papers are typically shorter than 20 words in length, and when they compare classical LDA with LSA, They see that LSA performs better on all datasets. Therefore, when it comes to People Query data, LSA typically performs better than LDA. 2) Modeling LSA topics in federated and centralized environments. The coherence score and classification both reveal this. As an example, Table 5 shows that across all four datasets, LSA+GD produces consistently higher F1 scores than LSA+GD in centralized learning, with the 15% gap being most pronounced on KILA. The efficacy of LSA-based topic modelling is demonstrated. 3) On all four datasets, FedLSA techniques (including FedAvg, FedGD, and FedRMSProp) outperform FedLDA among the federated topic models. Both the F1 Score and the Accuracy display this. Table 7 shows that across all four datasets used for centralized learning, LSA+GD consistently outperform LSA in terms of F1 scores.

Table 7. Evaluation Metrics for Four Different Datasets with Four Topic

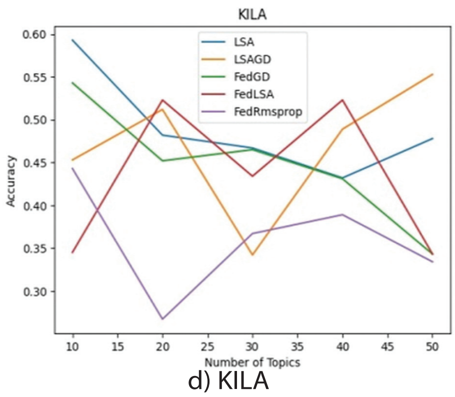
Dataset Metrics	DigiLockerNAD F1score Acc	IHMCL F1score Acc	Kerala StartupMission F1score Acc	KILA F1score Acc
LDA	0.424 0.433	0.487 0.423	0.342 0.354	0.468 0.456
LSA	0.436 0.456	0.490 0.543	0.536 0.654	0.593 0.482
LDAGD	0.417 0.443	0.478 0.453	0.517 0.543	0.467 0.432
LSAGD	0.435 0.480	0.455 0.523	0.578 0.565	0.478 0.453
FedLDA	0.357 0.467	0.489 0.437	0.657 0.475	0.512 0.342
FedLSA	0.523 0.554	0.467 0.565	0.432 0.588	0.489 0.553
FedGD	0.563 0.356	0.356 0.543	0.523 0.432	0.543 0.454
FedAvg	0.443 0.453	0.498 0.453	0.434 0.523	0.465 0.431
FedRMSprop	0.346 0.325	0.489 0.465	0.343 0.443	0.343 0.345



a) Digi Locker NAD

b) IHMCL

c) Kerala Startup Mission



d) KILA

Fig. 5. The Performance of Various Federated Latent Semantic Analysis Techniques

Comparison of Overall Performance to that of Related Articles Using locally stored documents, this research presents FedLSA, a framework for federated topic modelling algorithms based on LSA that generate high-quality topics. To mitigate the impact of client-side data heterogeneity on performance. Decentralized short text analysis and short document content mining are just two of the many potential uses for our FedLSA algorithms in light of the rising tide of privacy concerns.

6. CONCLUSION

Fed LSA is a framework that is introduced in this article to support federated topic modelling approaches that are based on LSA. Whether the documents are stored locally or not, these approaches could still produce high-quality topics. They provide the FedLSA design to fix performance problems brought on by data heterogeneity on the client side. Semantic analysis further optimizes the relationship between topic weights and the amount of input text. This elucidates the possible benefits of LSA for subject modelling. In light of the growing number of privacy concerns, our FedLSA algorithms have numerous potential uses, one of which is the distributed analysis of People Query documents.

7. REFERENCES

[1] G. Masson, N. Sneddon, R. Alghamdi, K. Alfalqi, "A survey of topic modelling in text mining", International Journal of Advanced Computer Science and Applications, Vol. 6, No. 1, 2015.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Drichlet allocation", Journal of Machine Learning Research, Vol. 3, 2003, pp. 993-1022.

[3] A. Fallah, A. Mokhtari, A. Ozdaglar, "Personalized federated learning: A meta-learning approach", arXiv:2002.07948, 2020.

[4] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, F. Wang, "Federated learning for healthcare informatics", Journal of Healthcare Informatics Research, Vol. 5, No. 1, 2021, pp. 1-19.

[5] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, "Automatic evaluation of topic coherence", Proceedings of Human language technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA, 2-4 June 2010, pp. 100-108

[6] C. C. Aggarwal, C. Zhai, "A survey of text classification algorithms", Mining text data, Springer 2012, pp. 163-222.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data", Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, Fl, USA, 2017, pp. 1273- 1282.

[8] D. Jiang, Y. Song, Y. Tong, X. Wu, W. Zhao, Q. Xu, Q. Yang, "Federated topic modeling", Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3-7 November 2019, pp. 1071-1080.

[9] Y. Wang, Y. Tong, D. Shi, "Federated latent drichlet allocation: A local differential privacy based framework", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 4, 2020, pp. 6283-6290.

[10] S. Si, J. Wang, R. Zhang, Q. Su, J. Xiao, "Federated Non-negative Matrix Factorization for Short Texts Topic Modeling with Mutual Information", Proceedings of the International Joint Conference on Neural Networks, Padua, Italy, 18-23 July 2002.

[11] Y. Li, T. Yang, "Word embedding for understanding natural language: A survey," Guide to Big Data Applications, Springer, 2018, pp. 83-104.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, Vol. 41, No. 6, 1990, pp. 391-407.

[13] T. Williams, J. Betak, "A Comparison of LSA and LDA for the Analysis of Railroad Accident Text", Procedia Computer Science, Vol. 130, 2018, pp. 98-102.

[14] J. Stremmel, A. Singh, "Pretraining federated text models for next word prediction", arXiv:2005.04828, 2020.

[15] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, Q. Yang, “Secure boost: A lossless federated learning framework”, arXiv:1901.08755, 2019.

[16] A. Fang, C. Macdonald, I. Ounis, P. Habel, “Using word embedding to evaluate the coherence of topics from twitter data”, Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy, 17-21 July 2016, pp. 1057-1060.

[17] R. Barzilay, M. Lapata, “Modeling local coherence: An entitybased approach”, Computational Linguistics, Vol. 34, No. 1, 2008, pp. 1-34.

[18] S. Zhou, X. Xu, Y. Liu, R. Chang, Y. Xiao, ”Text Similarity Measurement of Semantic Cognition Based on Word Vector Distance Decentralization With Clustering Analysis”, IEEE Access, Vol. 7, 2019, pp. 107247-107258.

[19] J. Hoblos, “Experimenting with Latent Semantic Analysis and Latent Drichlet Allocation on Automated Essay Grading”, Proceedings of the Seventh International Conference on Social Networks Analysis, Management and Security, Paris, France, 14-16 December 2020.

[20] T.-M. H. Hsu, H. Qi, M. Brown, “Measuring the effects of nonidentical data distribution for federated visual classification”, arXiv:1909.06335, 2019.

[21] R. Alghamdi, K. Alfalqi “A Survey of Topic Modeling in Text Mining”, International Journal of Advanced Computer Science and Applications, Vol. 6, No. 1, 2015.