International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

The Evaluation of Distributed Topic Models for Recognition of Health-Related Topics in Social Media Through Machine Learning Paradigms

Yerragudipadu Subbarayudu*[1], Alladi Sureshbabu[2]

Submitted: 08/05/2023 **Revised**: 18/07/2023 **Accepted**: 07/08/2023

Abstract: Social media is the most effective technique to obtain enormous amounts of data on tweets from the health field internationally. It is also a well-known source of data for anticipating healthcare-related solutions and looking for health-related phrases. In terms of income and employment, the health care industry has grown to become one of the biggest in the world. People may share their opinions and thoughts on a range of healthcare-related issues on Twitter, which is used by billions of users every day. The research gap in topic modeling related to healthcare topics in social media refers to areas or aspects that have not been extensively explored or adequately addressed by existing studies. Here are a few potential research gaps in this domain. Many studies focus on general healthcare discussions in social media, but there may be specific healthcare topics or subdomains that have not received sufficient attention. Research could focus on exploring topic modeling techniques for niche healthcare areas like mental health, rare diseases, specific treatments, or emerging healthcare technologies. Most topic modeling studies in social media healthcare discussions do not account for the user context and demographics. Research could investigate the influence of user characteristics, such as age, gender, location, or occupation, on the topics discussed, providing a deeper understanding of how different demographics engage with healthcare topics. Social media platforms are highly dynamic, and the popularity and sentiment of healthcare topics can change rapidly. There is a need for research that focuses on analyzing the temporal dynamics of healthcare topics in social media and tracking the evolution of topics over time. Social media platforms not only consist of text-based content but also include visual and audiovisual data. Research could explore topic modeling techniques that can effectively integrate and analyze multimodal data, such as images, videos, or audio, in healthcare-related discussions. While various evaluation metrics exist for topic modeling, they may not capture the unique characteristics and challenges of healthcare-related discussions in social media. Developing domain-specific evaluation metrics or adapting existing metrics to better assess the quality and relevance of topics in healthcare-related social media data is an important research direction. Social media data often raises ethical concerns related to privacy, consent, and data usage. Research gaps exist in exploring ethical guidelines, data anonymization techniques, and best practices for conducting topic modeling research on healthcare topics in social media while ensuring privacy and confidentiality. Addressing these research gaps can contribute to a more comprehensive understanding of healthcare topics in social media discussions and provide valuable insights for healthcare practitioners, policymakers, and researchers. It can help identify emerging healthcare trends, public sentiment, and inform evidence-based decision-making in the healthcare domain. The main objective of this research era is By applying topic modeling methods such as CvLDA and DiCTM to healthcare topics in social media, researchers and practitioners can gain insights into the prevalent themes, concerns, and discussions in the online healthcare domain. It enables the identification of emerging topics, the monitoring of public perceptions and sentiments, the discovery of valuable information for public health interventions, and the understanding of patient experiences and needs in the digital space.

***Keyboards**: Twitter, Hadoop, Machine Learning, LDA, CTM, NMI, TF, TFIDF, DTM, HDFS.*

1. Introduction

Sentiment analysis is a technique used to determine the sentiment or emotion expressed in text data. When applied to healthcare topics, sentiment analysis can help understand the attitudes, opinions, or emotional tone associated with those topics. Topic modeling algorithms, on the other hand, are used to identify latent topics or themes within a collection of documents. Let's explore how sentiment analysis can be combined with topic modeling algorithms in the context of healthcare topics. Topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) or Correlated Topic Model (CTM), are applied to healthcare-related text data to identify the underlying topics or themes within the documents. These algorithms automatically group words that tend to co-occur, allowing for the discovery of latent topics without the need for pre-defined categories.

Sentiment analysis techniques are then applied to the text data to determine the sentiment associated with each document or topic. Sentiment analysis can be performed using machine learning models, rule-based approaches, or lexicon-based methods. The sentiment of a document or topic can be categorized as positive, negative, neutral, or even fine-grained sentiments like happy, sad, angry, etc. Combining Sentiment Analysis with Topic Modeling,

[1] *Research Scholar of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur,51502,AP, India*
ORCID ID : 0000-0003-0349-9831
[2] *Department of Computer Science and Engineering, JNTUA College of Engineering, Anantapur, 515002 AP, India.*
ORCID ID: 0000-0002-1312-9765
** Corresponding Author Email: subbu.jntua@gmail.com*

Sentiment analysis can be applied at different levels within the topic modeling process. Document-level sentiment analysis, Determine the sentiment polarity of each document, indicating whether the overall sentiment expressed in the document is positive, negative, or neutral. This can provide an understanding of the overall sentiment distribution across healthcare topics. Topic-level sentiment analysis, assign sentiment scores to individual topics, indicating the sentiment associated with each topic. This allows for a more granular analysis of sentiment within specific healthcare themes [1].

Sentiment-aware topic modeling, Incorporate sentiment analysis as a factor in the topic modeling process. By considering sentiment information, the topic modeling algorithm can identify topics that are not only distinct in terms of content but also in terms of sentiment expressed. Insights and Applications, by combining sentiment analysis with topic modeling on healthcare topics, several insights can be derived. Identify positive and negative sentiment trends associated with specific healthcare topics, such as public perception of healthcare policies or patient experiences. Discover topics that are strongly associated with positive or negative sentiment, indicating the sentiment polarity of discussions related to those topics [2]. Analyze how sentiment evolves over time for different healthcare topics, allowing for the tracking of sentiment shifts or changes.

Public Opinion Monitoring: Understand public sentiment towards healthcare policies, treatments, or healthcare providers. Brand Reputation Management: Assess the sentiment associated with healthcare organizations or medical products/services. Patient Feedback Analysis: Analyze sentiment in patient reviews or social media posts to gain insights into patient experiences and satisfaction levels. By combining sentiment analysis with topic modeling algorithms, analysts can gain a deeper understanding of sentiment dynamics within healthcare topics. This integration allows for a more nuanced analysis of emotions, opinions, and attitudes expressed in the context of specific healthcare themes, enabling better decision-making, public perception monitoring, and healthcare improvements. A health system is any institution, project, or activity that has as its main objective maintaining, enhancing, or restoring health, according to the World Health institution [3-4].

This includes both activities that improve health directly and those that have an impact on the factors that influence health. Social media is one of the many avenues from which health information may be obtained. Stakeholders can quickly investigate online debates taking place outside of conventional health settings using social media. The most well-known Social Networking site is Twitter, which has more than 320 million active users. On this platform, users are permitted to submit succinct comments with a character limit of 280. In addition to text, tweets can also include photos, videos, and links to other websites. Tweets can be liked, reposted, and replied to by users. Users who are not registered can just view tweets. Twitter has become a resource for the gathering and transmission of information for numerous businesses, including the health industry, due to its broad access to many points of view and large user base. Twitter is already used by health organizations to promote healthy lifestyle choices, identify disease outbreaks, study human behavior, and learn what the public thinks about health-related issues. On Twitter, these organizations also promote health. Department of Health and Human Services of the United States is one such organization that uses Twitter to share crucial health information with the broader public. Along with health organizations, a wide range of other entities, including persons, news sources, businesses, interest groups, and others, also tweet about health-related topics. If left untreated, the severe disease caused by the Ebola virus frequently results in death [5].

2. Related Works

The first Bayesian machine learning models for discovering drugs active against the Ebola virus would be derived from FDA-approved medication screens. (EBOV). Three active compounds were found in vitro thanks to these models: tilorone, pyronaridine, and quinacrine. Infected mice with mice-adapted EBOV responded to one of these medications, tilorone, with 100% in vivo effectiveness. when given at a dose of 30 mg/kg/day intraperitoneally, according to follow-up research. This implied that we could draw conclusions about EBOV inhibition from the published data and apply them to the selection of novel drugs for in vivo testing**.**

They have also identified two other novel compounds, the antimalarial drug arterolane (IC50 = 4.53 M) and the anticancer therapeutic candidate lucanthone (IC50 = 3.27 M), both of which exhibit minimal cytotoxicity and have EBOV inhibitory efficacy in HeLa cells. With the help of quantitative structure-activity relationship data on compounds that are now available and have experimental anti-Ebola activities, they have created a "anti-Ebola" web server. The "DrugRepV" database contained 355 different anti-Ebola drugs, along with their corresponding IC50 values. After extracting the molecular descriptors from the substances, regression-based model construction was performed on them.

William's plot was used to cross-evaluate the developed models' robustness. They believe that this will help the scientific community create efficient Ebola virus inhibitors. To create efficient predictive models, they have implemented three MLTs: RF, SVM, and ANN. SVM, RF, and ANN are three machine learning algorithms that make use of different operational principles. For instance, the

SVM is a nonlinear algorithm, the ANN is a neural network-based approach, and the RF employs the decision tree algorithm family. This work intends to offer ideas on enhancing current research on the Ebola virus (EBOV) by examining the significance of machine learning (ML) methodologies in the prediction of tiny pharmacological inhibitors of EBOV. Anti-EBOV medication forecasting has been done using the Bayesian, Support Vector Machine, and Random Forest approaches, which all provide credible models. We describe how deep learning models could be used to create quick, effective, reliable, and creative algorithms to help with the identification of anti-EBOV medications since their usage for predicting anti-EBOV compounds is underutilized. They also investigated the deep neural network as a potential machine learning strategy to anticipate anti-EBOV medicines.

. There is widespread concern about the lack of FDA-approved alternative medications for the treatment of EBOV [6]. The study recommended employing machine learning techniques to evaluate how well time series projections of Ebola casualties performed. By testing without lag generation, the best results were obtained with the Random Tree Classifier, which had an MAE of 7.85%, RMSE value of 61.14%, and Direction Accuracy of 85.99%. The study may thus conclude that the health authorities are able to guarantee that the appropriate actions are taken to manage the outbreak by using these models to anticipate epidemic spread and establish public health policies [7]. This implies that instead of lowering inflammation directly, host-directed methods for reducing the severity of EVD might improve the control of inflammatory gene expression [8]. Between 2000 and 2016, Uganda saw eight filovirus epidemics, more than any other nation in the world. To aid in epidemic preparation and surveillance, the location of likely filovirus outbreaks in Uganda is predicted using species distribution modelling.

The MaxEnt program, a machine learning modelling technique that makes use of presence-only data, was used to ascertain the relationships between filoviruses and their habitats. From the field and online sources, presence-only information about filovirus epidemics was gathered. To achieve a nominal resolution of 1km x 1km, environmental variables from Africlim were downscaled. Using an average of 100 bootstrap iterations, the final model determined the relative likelihood of filoviruses in the study region. Utilizing Receiver Operating Characteristic (ROC) plots, the model was evaluated. ArcGIS 10.3 mapping software was used to make the maps [9]. This research provides a method for finding cell structures in pictures that have been marked for subviral particles. One may show a relationship between where the cell's structures are located and how the subviral particles are distributed inside an infected cell.

A "Mask-R-CNN" approach that has been presented in this study is used to segment data. An effective and quick recognition of cell structures is made possible by the model, which is a region-based convolutional neural network. It also gives a description of the network architecture. Data verified by professionals is used to test the suggested method. The outcomes indicate that the procedure has a high potential and is appropriate. On unclassified, microscopic photos of infected cells, Mask R-CNN object detection is demonstrated. All cell membranes have been found, as can be shown in (a), despite the mask predictions being off. In (a), the bounding box has been positioned incorrectly, leaving a little area near the left side of the screen unoccupied, preventing the green mask from completely enclosing the cell. Parts of the neighboring cell are covered by the light blue mask, erroneously identifying it. It is likely that a second cell is positioned behind the first one in the center, giving the impression that there are two nuclei in a single cell. Nevertheless, each group was given the appropriate categorization names. It was hard to locate a data collection of this magnitude in the medical sector.

The goal of this project was to create synthetic data for neural network training and to represent a piece using only one delta peak. It was shown that even a little dataset is sufficient for that purpose, and that a simple convolutional network can predict particle locations with almost the same accuracy as the technique developed by Kienzle, Rausch, and colleagues [10].

The results of the KRA are shown alongside the performance of various datasets and square edge lengths. The resulting training dataset grows from 500 to 1000, 2000, and 4000, but only marginally improves the associated detection result. However, the KRA's performance might still be readily met or exceeded. On the other hand, increasing the square edges allows for the inclusion of particle estimates with a wider field of view and lengthens the intersection of pixels between ground truth and estimation, which significantly enhances the overlap measure. When the edge length was extended from two to seven, the OM for the variously sized datasets and the KRA rose by about [11].

The dataset has around 23000 data points that were acquired using the HTS technique. The chemicals are divided into two groups: low-activity class and high-activity class. At the end of the screening 4594 unique molecules have been found. 70% of the dataset is used for training, while the remaining 30% is shared evenly between validation and testing. CPE, cytotox, AlphaLISA are the few algorithms used for this project [12]. This study involved many deep learning and machine learning algorithms to be trained and tested. This thesis' key claim is that SARS analysis can be performed using ML and DL techniques and algorithms. The highest accuracy is given by the Deep learning models

of approximately 99% whereas the machine learning models just lagged a few percentages and recorded a score of 96%. Based on the test results the models were tested [13]. This research mainly focusses on which ML algorithm performs the best on the given covid-19 positive recorded samples. A total of 3 algorithms are used in this study (KNN, SVM, Naïve Bayes). In terms of precision, accuracy score, f1 score SVM recorded the highest and Naïve Bayes gave the least accuracy. 350 examples with multiple features, including infected and non-infected persons from various locations throughout the world, are included in the dataset utilized for this project [14]. When tested with 5-fold cross validation comparison the random forest method resulted in the best accuracy of 75% while with external data most algorithms underperformed with the highest being SVM with 70%. PCA is used to reduce the dimensionality complex of the compounds and make processing much easier and faster [15-16].

The symptoms are classified into 3 levels. The higher the level, the greater the impact on the patient. Fuzzy logic is used to classify the symptoms into 3 different categories. MATLAB is used for computation of several packages [17-18]. A system is developed under this research that will diagnose Lassa fever. VPES programming software is used to structure the process and create a framework. This framework assists both medicinal experts and the patients. The framework analyses the patient's health status using the information it has collected, predicts potential future symptoms, and aids in speedier diagnosis. Few rules are decided which classify whether the patient is having mild, heavy, or critical Lassa fever [22].

Problem Statement:

The main Problem of the research study is to identify the highly discussed diseases in social media platforms like twitter which may lead to pandemic in future. *Limited Coverage of Specific Healthcare Topics*: Many studies focus on general healthcare discussions in social media, but there may be specific healthcare topics or subdomains that have not received sufficient attention. Research could focus on exploring topic modeling techniques for niche healthcare Topic modeling algorithms can be integrated with the Hadoop Distributed File System (HDFS) to analyze healthcare topics in social media at scale. HDFS is a distributed file system that allows for the storage and processing of large volumes of data across multiple machines in a Hadoop cluster. Here's how topic modeling algorithms can leverage HDFS for healthcare topics in social media. ***Data Storage***: HDFS can store healthcare text data obtained from social media platforms. The data is divided into blocks and distributed across multiple nodes in the Hadoop cluster. This distributed storage enables efficient handling of large-scale healthcare datasets, including text data from social media.

areas like mental health, rare diseases, specific treatments, or emerging healthcare technologies. *Incorporation of User Context and Demographics:* Most topic modeling studies in social media healthcare discussions do not account for the user context and demographics. Research could investigate the influence of user characteristics, such as age, gender, location, or occupation, on the topics discussed, providing a deeper understanding of how different demographics engage with healthcare topics. ***Dynamics and Temporal Analysis:*** Social media platforms are highly dynamic, and the popularity and sentiment of healthcare topics can change rapidly. There is a need for research that focuses on analyzing the temporal dynamics of healthcare topics in social media and tracking the evolution of topics over time. *Incorporation of Multimodal Data:* Social media platforms not only consist of text-based content but also include visual and audiovisual data. Research could explore topic modeling techniques that can effectively integrate and analyze multimodal data, such as images, videos, or audio, in healthcare-related discussions.

Domain-Specific Topic Modeling Evaluation: While various evaluation metrics exist for topic modeling, they may not capture the unique characteristics and challenges of healthcare-related discussions in social media. Developing domain-specific evaluation metrics or adapting existing metrics to better assess the quality and relevance of topics in healthcare-related social media data is an important research direction. *Ethical and Privacy Considerations:* Social media data often raises ethical concerns related to privacy, consent, and data usage. Research gaps exist in exploring ethical guidelines, data anonymization techniques, and best practices for conducting topic modeling research on healthcare topics in social media while ensuring privacy and confidentiality. Addressing these research gaps can contribute to a more comprehensive understanding of healthcare topics in social media discussions and provide valuable insights for healthcare practitioners, policymakers, and researchers. It can help identify emerging healthcare trends, public sentiment, and inform evidence-based decision-making in the healthcare domain.

3. Proposed Architecture

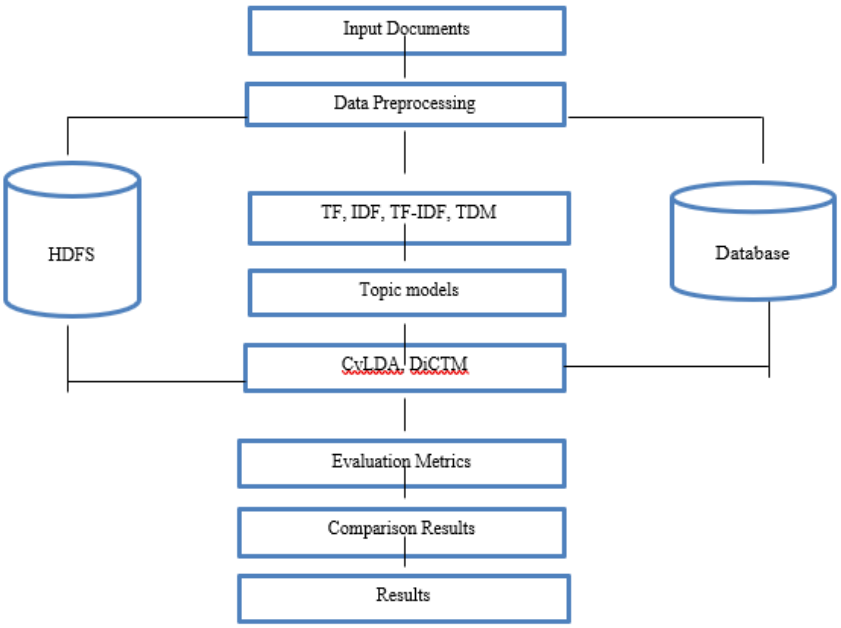


Fig: 1 Control flow of Proposed Architecture

Data Preprocessing: Prior to topic modeling, the healthcare text data stored in HDFS can be preprocessed using MapReduce or other Hadoop processing frameworks. This involves tasks like tokenization, stop word removal, stemming, and creating a document-term matrix or bag-of-words representation. ***Distributed Computation:*** Hadoop's distributed processing framework enables parallel execution of topic modeling algorithms across multiple nodes in the cluster. This allows for faster processing and scalability, especially when dealing with large healthcare datasets. The data is divided into smaller chunks, and each node processes a portion of the data simultaneously.

Hadoop Distributed File System with Topic Models:

A distributed file system called Hadoop Distributed File System (HDFS) is made to process and store massive amounts of data across a cluster of computers called a Hadoop installation. By dividing the data among several cluster nodes, it offers a dependable and scalable approach for processing massive data. A method for finding hidden themes or topics in a group of papers is called topic modelling. Topic modelling can be used in the context of healthcare themes in social media to find pertinent healthcare-related discussions, spot trends, and learn more about what the general public thinks or feels. These approaches can be used to apply topic modelling techniques to healthcare-related social media topics using HDFS.

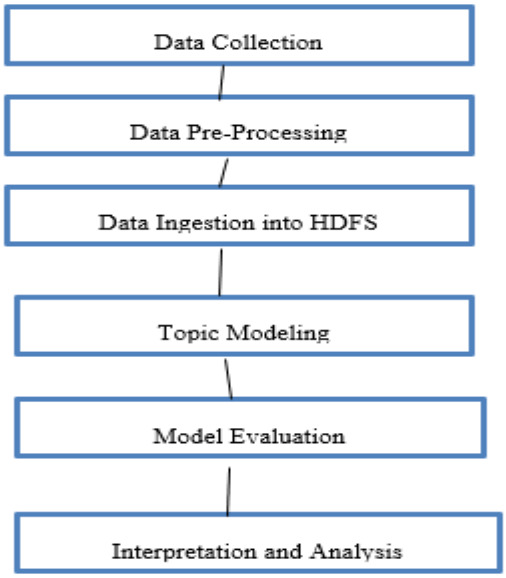


Fig: 2 Working control flow of HDFS with topic Models

In fig (2), ***Data Collection:*** Gather a large collection of social media posts or tweets related to healthcare. This can be done by leveraging APIs provided by social media platforms or using web scraping techniques. ***Data Preprocessing:*** Preprocess the collected data to remove noise, such as irrelevant symbols, URLs, or special characters. You may also want to perform text normalization, including lowercasing, stemming, and removing stop words. This step helps to standardize the text and reduce the dimensionality of the data.

Data Ingestion into HDFS: Load the preprocessed data into HDFS by splitting it into smaller chunks and distributing them across the nodes in the Hadoop cluster. HDFS will automatically handle data replication, fault tolerance, and data distribution across the cluster. *Topic Modeling:* Apply topic modeling algorithms to the data stored in HDFS. One popular algorithm for topic modeling is Cluster Visualized Latent Dirichlet Allocation (CvLDA) and Distributed CTM, which is widely used for discovering latent topics within a collection of documents. You can use libraries like Apache Mahout or Apache Spark MLlib to perform topic modeling tasks in a distributed manner across the Hadoop cluster. **Model Evaluation:** Evaluate the quality and coherence of the generated topics by examining the most representative terms for each topic and assessing their semantic relevance to healthcare. Various metrics like topic coherence or perplexity can be used to measure the quality of the topic model.

Interpretation and Analysis: Once you have generated the topic model, interpret, and analyze the discovered topics in the context of healthcare. Identify the key themes, understand the sentiment or opinion associated with each topic, and extract meaningful insights from the social media data. By leveraging HDFS, you can handle large volumes of social media data efficiently and distribute the topic modeling process across multiple machines in the Hadoop cluster. This enables you to scale your analysis and extract valuable healthcare-related information from social media discussions. Hadoop Distributed File System (HDFS) multi-node clustering involves setting up a cluster of multiple machines to store and process data in a distributed manner. In the context of healthcare topics, this clustering approach allows for efficient storage and analysis of large volumes of healthcare-related data. By utilizing HDFS multi-node clustering, healthcare organizations can store and analyze large volumes of healthcare data efficiently and reliably. The distributed nature of HDFS enables parallel processing, fault tolerance, and scalability, making it well-suited for handling the data-intensive requirements of healthcare topics.

4. Methods and Materials

4.1 Sample Dataset: Here are some sources for datasets on healthcare topics in social media, Twitter API: Twitter provides an API that allows access to public tweets. Researchers can use the Twitter API to collect healthcare-related tweets based on specific keywords, hashtags, or user profiles. *Kaggle:* Kaggle is a popular platform for datasets, and it hosts various datasets related to healthcare topics in social media. You can search for healthcare or social media datasets on Kaggle and filter the results to find relevant datasets. *MIMIC-III:* The Medical Information Mart for Intensive Care (MIMIC) database includes de-identified health data, including text notes from intensive care units. Researchers can utilize text notes to extract healthcare-related discussions from social media platforms. *Social media research platforms:* Some dedicated platforms provide access to social media data for research purposes. Examples include GNIP, Social Studio, or Crimson Hexagon. These platforms often require subscription or licensing. ***Research publications:*** Many research papers related to healthcare topics in social media provide access to their datasets as supplementary material. Checking the publications in this field and accessing the associated datasets can be a valuable resource [21].

4.2 Research collaborations: Collaborating with research institutions or organizations involved in healthcare and social media research can provide access to proprietary datasets or access to data through partnerships. *Academic repositories:* Universities and research institutions may have their repositories hosting datasets related to healthcare topics in social media. Exploring institutional repositories or contacting researchers in the field can help identify available datasets. Remember to review and comply with the terms of use, privacy guidelines, and ethical considerations associated with each dataset source. Additionally, ensure that the datasets are relevant to your specific research goals and adhere to any data anonymization or de-identification protocols. The dataset is Biotext, which includes Medline-collected abstracts of diseases and therapies available in table (1).

weblink:(http://biotext.berkeley.edu/data/dis treat data.html).

Table 1: Data set

Data set Name	***Document Preprocess***	***Terms***	***Unique Terms***
Biotext	40	25921	10267
Twitter	58927	395635	25309

a) Data Preprocessing

Natural language text data is noisy and unstructured. Before text can be fed into a model for additional analysis and learning, it must be translated into a clear and consistent format. Text preprocessing methods can be broad, making them applicable to a wide range of applications, or they can be tailored to achieve a particular objective. An NLP pipeline for categorizing texts may comprise sentence segmentation, word tokenization, lowercasing, stemming or lemmatization, stop word removal, and spelling correction.

b) **Result Aggregation:** The results of the topic modeling algorithms, such as the identified topics, their keywords, and document-topic distributions, can be aggregated from the distributed nodes. This consolidation of results helps in understanding the prevalent healthcare topics discussed in social media and their associations. *Post-Processing and Visualization:* Once the topic modeling results are obtained, further analysis can be performed on the data stored in HDFS. This may include post-processing steps like sentiment analysis, clustering, or association rule mining to gain deeper insights into healthcare topics in social media. Visualization techniques can be employed to present the results in an interpretable manner. By leveraging the capabilities of HDFS, topic modeling algorithms can efficiently process and analyze large volumes of healthcare text data from social media platforms. The distributed nature of HDFS enables parallel computation, scalability, fault tolerance, and storage of the processed data, making it a valuable platform for topic modeling tasks related to healthcare in social media.

4.3 The Bag of Words (BoW) Algorithm:

A common method for transforming text data into a numerical format for additional analysis is the bag-of-words representation. The bag-of-words approach would entail extracting a lexicon of distinctive words from the text data and representing each document as a vector reflecting the presence or frequency of these words in the context of healthcare issues in social media. Here's an illustration of how the bag-of-words metaphor might be used to describe healthcare-related social media topics.

Data Collection: Gather a dataset of social media posts or text data related to healthcare topics in social media (e.g., tweets, forum posts, blog comments). Ensure that the data is relevant to healthcare discussions and contains the necessary information for analysis. ***Text Preprocessing:*** Clean the text data by removing noise, such as URLs, hashtags, punctuation, and special characters. Convert the text to lowercase for consistency. Perform tokenization to split the text into individual words or tokens. Remove stop words, which are common words that do not carry much meaning (e.g., "the," "and" "in").

$$TF(t, d) = \frac{(Number\ of\ occurrences\ of\ term\ t\ in\ document\ d)}{(Total\ number\ of\ terms\ in\ document\ d)}.$$

Eq(1)

4.4Building the Vocabulary: Create a vocabulary by identifying all unique words across the preprocessed documents. Each unique word will become a feature or dimension in the bag-of-words representation. *Bag-of-Words Encoding:* Represent each document as a vector in the bag-of-words format. Assign a value to each dimension of the vector, indicating the presence or frequency of the corresponding word in the document. Common approaches include using binary values (0/1) to indicate presence or term frequency (TF) to represent the number of occurrences of each word in the document. For example, let's say we have the following two social media posts related to healthcare: **Post 1: "I just got vaccinated, and it feels great!" Post 2: "Healthcare access is a major concern in rural areas."** After preprocessing, the vocabulary may consist of the following words: "vaccinated," "feels," "great," "healthcare," "access," "major," "concern," "rural," "areas." The bag-of-words representation for each document would be **Post 1: [1, 1, 1, 0, 0, 0, 0, 0, 0] and Post 2: [0, 0, 0, 1, 1, 1, 1, 1, 1]**

In this representation, each position in the vector corresponds to a specific word in the vocabulary, and a value of 1 indicates the presence of the word in the document. The bag-of-words representation allows for further analysis, such as clustering, classification, or topic modeling, on the healthcare topics in social media data. It provides a numerical format that can be processed by machine learning algorithms to gain insights from the text data. The bag-of-words representation is a simple and intuitive way to convert text data into a numerical format. It involves creating a vocabulary of unique words and representing each document as a vector indicating the presence or frequency of these words. Here are the mathematical equations involved in the bag-of-words representation with a numerical example related to healthcare topics. *Vocabulary Creation:* Let's assume we have a collection of three healthcare-related documents: D1, D2, D3. We extract all the unique words from these documents to create a vocabulary. Let's say our vocabulary consists of six words: {health, care, hospital, patient, medicine, treatment}.

4.5 Term Frequency (TF): Term Frequency measures how often a term appears in a document.

We calculate the term frequency for each word in each document using the following equation (Eq1).

Let's calculate the term frequencies for each word in each document in table (2)

Table 2: Term Frequencies for each word in each document.

Document	*health*	*care*	*hospital*	*patient*	*medicine*	*treatment*
D1	2	1	0	3	1	0
D2	1	2	2	1	0	0
D3	0	0	3	1	2	1

4.7 Document-Term Matrix:

The Document-Term Matrix (DTM) is a matrix that can be used to represent the term frequencies of each word in each document. The columns of the matrix represent the vocabulary words, while the rows represent the documents. The term frequency of each item in the matrix corresponds to the frequency of the word in the relevant document. This is how the DTM for our case would seem, and it is shown in table (3).

Table 3: DTM

Document	**health**	**care**	**hospital**	**patient**	**medicine**	**treatment**
D1	2	1	0	3	1	0
D2	1	2	2	1	0	0
D3	0	0	3	1	2	1

Bag-of-Words Encoding: The bag-of-words encoding allows us to encode each document as a vector. A vector is used to represent each page, and its members are the term frequencies of the vocabulary terms. For our case, the bag-of-words vectors would be.

Table 4: BoW

Document	health	care	hospital	patient	medicine	treatment
D1	2	1	0	3	1	0
D2	1	2	2	1	0	0
D3	0	0	3	1	2	1

D1: [2, 1, 0, 3, 1, 0], D2: [1, 2, 2, 1, 0, 0] and D3: [0, 0, 3, 1, 2, 1]. In this example, we started with three healthcare-related documents and created a vocabulary of six words. We then calculated the term frequencies for each word in each document and represented the documents as bag-of-words vectors. The bag-of-words representation allows us to perform further analysis on the healthcare topics, such as clustering, classification, or topic modeling, using the numerical vectors instead of raw text data.

4.8 Term Frequency-Inverse Document Frequency (TF-IDF):

A numerical statistic called TF-IDF (Term Frequency-Inverse Document Frequency) measures the significance of a phrase in a document within a collection of documents. It is frequently employed in text mining and information retrieval activities. The mathematical formulas for determining TF- IDF are as follows: Frequency of Terms (TF) The frequency at which a term appears in a document is measured. It is determined mathematically in Eq. (2) as the total number of terms in the document (n) divided by the number of times a term (t) appears in the document (d).

$$TF(t, d) = \frac{(Number\ of\ occurrences\ of\ term\ t\ in\ document\ d)}{(Total\ number\ of\ terms\ in\ document\ d)}. \quad Eq(2)$$

4.9 ***Inverse Document Frequency (IDF):*** The inverse document frequency measures the rarity or uniqueness of a term across a collection of documents. It is calculated as the logarithm of the total number of documents (N) divided by the number of documents that contain the term (df(t)). Mathematically,

$$IDF(t) = \log (N / df(t)). \quad Eq(3)$$

4.10 TF-IDF:

The TF-IDF score combines the term frequency and inverse document frequency to determine the importance of a term in a specific document within a collection.

It is calculated as the product of the term frequency (TF) and the inverse document frequency (IDF).

Mathematically,

$$TF\text{-}IDF(t, d) = TF(t, d) * IDF(t). \quad Eq(4)$$

Now, let's illustrate the calculation of TF-IDF with a numerical example. Consider a collection of documents with three documents (D1, D2, D3) and a vocabulary of five terms (T1, T2, T3, T4, T5). The term frequencies (TF) and document frequencies (DF) for each term are available in table (5).

Table 5:TF-IDF

Term	D1	D2	D3	DF
T1	3	0	2	2
T2	1	2	0	2
T3	2	0	1	2
T4	0	4	0	1
T5	1	1	1	3

To calculate the TF-IDF score for each term in each document, Calculate the Term Frequency (TF) for each term in each document. For example, TF (T1, D1) = 3 / (3 + 1 + 2 + 0 + 1) = 0. 375.Calculate the Inverse Document Frequency (IDF) for each term. For example, IDF(T1) = log (3 / 2) = 0.176. Multiply the TF and IDF to obtain the TF-IDF score for each term in each document. For example, TF-IDF (T1, D1) = 0.375 * 0.176 = 0. 066.Repeat these steps for each term and document to calculate the TF-IDF scores for the entire collection. The TF-IDF scores provide a numerical representation of the importance of each term in each document, where higher scores indicate greater relevance. This information can be used for various tasks such as keyword extraction, document ranking, text classification, and information retrieval.

NLTK (Natural Language Toolkit) is a popular Python library that provides various tools and resources for natural language processing tasks. While NLTK itself does not directly implement topic modeling algorithms, it offers several components and functionalities that can be used in conjunction with topic modeling techniques for healthcare topics. Here are some ways NLTK can be utilized. ***Text Preprocessing:*** NLTK provides a range of preprocessing functions for cleaning and transforming text data. This includes tokenization, stop word removal, stemming, lemmatization, and handling of special characters. These preprocessing steps are crucial in preparing textual data for topic modeling. ***Part-of-Speech Tagging:*** NLTK's part-of-speech tagging capabilities allow for the identification of different grammatical elements in text, such as nouns, verbs, adjectives, and adverbs. This information can be useful in filtering or enhancing the topic modeling process by considering specific parts of speech that are more relevant to healthcare topics.

Named Entity Recognition: NLTK offers named entity recognition functionality to identify and classify named entities in text, such as medical terms, drugs, diseases, and healthcare organizations. Incorporating named entity recognition can help improve the accuracy and specificity of topic modeling results by focusing on healthcare-specific entities. ***Corpus Management:*** NLTK provides utilities for managing corpora, which are collections of text documents. This can be helpful in organizing and processing healthcare-related text data from social media sources or other healthcare-specific datasets. ***Language Modeling:*** NLTK offers tools for language modeling, including n-gram models and smoothing techniques. Language modeling can complement topic modeling by capturing the statistical relationships between words and improving the coherence of topics extracted from healthcare texts. ***Sentiment Analysis:*** NLTK includes sentiment analysis capabilities that can be used to analyze the sentiment or polarity of healthcare-related social media posts. Integrating sentiment analysis with topic modeling can provide additional insights into the emotional or subjective aspects of healthcare discussions on social media. While NLTK itself may not provide topic modeling algorithms out of the box, it serves as a valuable toolkit for text preprocessing, linguistic analysis, and general NLP tasks that can enhance the topic modeling process for healthcare topics. Researchers and practitioners can leverage NLTK's functionalities in combination with dedicated topic modeling libraries, such as Gensim or scikit-learn, to develop comprehensive topic modeling pipelines for healthcare text data.

4.11 Document-Term Matrix:

The Document-Term Matrix (DTM) is a crucial representation used to analyze healthcare topics in social

media. Let's explore how DTM works in the context of healthcare discussions. *Data Collection:* Gather a collection of social media posts or tweets related to healthcare topics from platforms like Twitter. These posts may cover a wide range of subjects, such as diseases, treatments, healthcare policies, public health campaigns, or personal experiences. ***Preprocessing:*** Perform preprocessing steps to clean and prepare the text data for analysis. Steps may include removing punctuation, converting text to lowercase, removing stop words (common words like "and" "the," etc.), and handling special characters or URLs. ***Tokenization:*** Tokenize the preprocessed text by splitting it into individual words or terms.

For example, the sentence "I had a flu shot today" would be tokenized into ["I," "had," "a," "flu," "shot," "today"]. *Vocabulary Creation:* Create a vocabulary of unique terms from the tokenized text. Each term represents a word that appears in the social media posts. The vocabulary helps create the columns of the DTM. *Construction of DTM:* Construct the DTM, where each row represents a social media post or tweet, and each column corresponds to a term from the vocabulary. The entries in the matrix represent the frequency, occurrence, or weighting of each term in each document. DTM can be binary (indicating presence or absence of a term), frequency-based (count of term occurrences), or TF-IDF weighted (reflecting term importance). *DTM Representation:* The DTM serves as a numerical representation of social media data, enabling quantitative analysis and modeling. Each entry in the matrix reflects the importance or relevance of a term within a specific document.

The DTM captures the distribution of terms across the documents and allows for various computational techniques to be applied. *Analysis and Interpretation:* With the DTM constructed, various analyses can be performed on healthcare topics in social media. Topic Modeling: Techniques like Latent Dirichlet Allocation (LDA) or Correlated Topic Modeling on healthcare be applied to identify latent topics within the documents. Sentiment Analysis: Sentiment analysis algorithms can assess the polarity (positive, negative, neutral) of the text to gauge public opinion on healthcare topics. Keyword Extraction: Statistical methods or natural language processing techniques can identify frequently occurring or important keywords related to healthcare.

Clustering or Classification: Machine learning algorithms can group or classify documents based on their content, helping identify distinct themes or categories within healthcare discussions. By utilizing the Document-Term Matrix (DTM) representation, analysts can gain insights into the patterns, trends, and themes related to healthcare topics in social media discussions. The DTM serves as a foundation for various computational and analytical techniques, enabling researchers to explore, understand, and extract valuable information from the vast amount of textual data available in social media platforms.

Assume that the corpus comprises the following three Documents:

Document 1: He is suffering from heart disease.

Document 2: These diseases usually move quickly through a population.

Document 3: The disease is usually more serious in adults.

A document-word matrix, or DTM as it is often known, may represent any corpus, or group of documents. It is well knowledge that the first stage of text data processing entails text cleaning, preprocessing, and tokenization. Get the following document word matrix after preprocessing the documents, which includes, the words are designated by the letters W1 through W4, while the three sheets are marked by the letters D1, D2, and D3. As a result, the matrix is shaped as follows: 3 * 4 (three rows by four columns):

Table:6 Document Word Matrix

	W1	*W2*	*W3*	*W4*
D1	0	1	1	0
D2	0	0	1	1
D3	1	1	1	0

Most of the corpus is now represented using the previously mentioned document-word matrix, where each row represents a document, and each column represents a token or word.

4.12 Cluster Visualized Latent Dirichlet Allocation (CvLDA):

Cluster Visualized Latent Dirichlet Allocation (LDA) is a topic modeling algorithm that can be applied to healthcare topics in social media, such as Twitter, to uncover latent themes and patterns within the discussions. LDA is

particularly useful for analyzing large volumes of text data and identifying the underlying topics without prior labeling or knowledge. Here's how LDA is related to healthcare topics on Twitter. *Data Collection:* Collect a dataset of healthcare-related tweets from Twitter using appropriate keywords and filters. This dataset should consist of tweets discussing various healthcare topics like diseases, treatments, healthcare providers, public health issues, and more. *Preprocessing*: Clean the collected tweets by removing noise, such as URLs, hashtags, user mentions, and irrelevant characters. Tokenize the tweets into individual words and perform other preprocessing tasks like stemming, stop-word removal, and lowercasing. First, decide on the number of topics you want to extract from the Twitter dataset. This will depend on the specific healthcare domain and the granularity of topics you are interested in. Second, Apply the LDA algorithm to the preprocessed tweets to estimate the topic-word distributions and document-topic distributions. This involves running the algorithm iterations to learn the model parameters. *Topic Interpretation:* Analyze the learned LDA model to interpret the discovered topics and their associated words. Each

topic will represent a specific healthcare theme discussed on Twitter, such as diabetes, mental health, vaccination, or healthcare policies. By examining the top words in each topic, you can understand the content and focus of the discussions related to each topic. *Topic Visualization and Analysis:* Visualize the results of the LDA model to gain insights into the healthcare topics being discussed on Twitter. Techniques like word clouds, topic networks, or topic distributions can help you understand the prevalence and relationships between different healthcare themes.

Track the trends and patterns of healthcare topics on Twitter over time. This can assist in monitoring public opinion, identifying emerging health concerns, or evaluating the impact of healthcare campaigns or policies. Analyze the sentiment associated with each healthcare topic to understand public sentiment or opinion towards specific health issues, treatments, or healthcare providers. Content Recommendation: Utilize the LDA model to improve content recommendation systems in the healthcare domain on social media platforms. By associating tweets with relevant topics, you can enhance the accuracy of recommending healthcare information or resources to users. Applying LDA to healthcare topics on Twitter enables a deeper understanding of the discussions, prevalent themes, sentiment, and emerging trends in the healthcare domain. It can provide valuable insights to healthcare professionals, policymakers, researchers, and enable more targeted and informed decision-making.

Algorithm: *CvLDA*

1. Notations:

a. D→no. of documents

b. N→no.of words in a document

c. K→no. of topics

d. w→Word index in the vocabulary

e. d→Document index

f. z→ Topic index

2. Topic Proportions:

a. θ_d→Topic proportions for document d

b. θ_d ~ Dirichlet(α), where α is a hyperparameter

3. Word Distribution:

a. β_k→ Word distribution for topic k

b. β_k ~ Dirichlet(η), where η is a hyperparameter

4. Topic Assignment:

a. z_dn→Topic assignment for word n in document d

b. z_dn ~ Multinomial(θ_d), representing the probability of word n belonging to each topic in document d

5. Word Generation:

a. w_dn→Observed word in document d at position n

b. w_dn ~ Multinomial(β_{z_dn}), indicating the probability of each word being generated from the topic assigned to

it

6. Latent Dirichlet Allocation Equation:

a. $P(w\_dn = w \mid \theta\_d, \beta, z\_dn) = \sum\_k P(w\_dn = w \mid z\_dn = k, \beta\_k) * P(z\_dn = k \mid \theta\_d)$

b. This equation calculates the probability of observing word w in position n of document d, given the topic proportions θ_d, word distributions β, and topic assignment z_dn.

7. Joint Probability of the Model:

a. $P(w, \theta, \beta, z \mid \alpha, \eta) = \prod\_d P(\theta\_d \mid \alpha) * (\prod\_k P(\beta\_k \mid \eta)) * (\prod\_n P(z\_dn \mid \theta\_d)) * P(w\_dn = w \mid \beta\_\{z\_dn\})$

b. This equation represents the joint probability of the observed words, topic proportions, word distributions, and topic assignments, given the hyperparameters α and η.

In order to classify or categorize the text in a document and the words per subject, Latent Dirichlet Allocation (LDA), a method and strategy for topic modelling, employs models based on Dirichlet distributions and processes. The LDA is based on two important hypotheses that-Documents have a variety of themes, and topics contain a variety of tokens (or words). Additionally, the probability distribution is used to create the words derived from these topics. In contrast to the probability density (or distribution) of words in texts, topics in documents have a different distribution or probability density. The two fundamental presumptions outlined above are first applied to the corpus via LDA.

4.13 Distributed Correlated Topic Model (DiCTM)

The Distributed Correlated Topic Model (CTM) can be applied to health care topics in social media, such as Twitter, to extract meaningful and correlated themes from the discussions and posts. By analyzing the content shared on Twitter, the CTM can help identify and understand the interrelationships between different health care topics being discussed by users. Here's how DiCTM can be related to health care topics on Twitter. Data Collection: Collect a large dataset of health-related tweets from Twitter using appropriate keywords and filters. This dataset will consist of tweets related to various health care topics, such as diseases, treatments, symptoms, medications, healthcare policies, and more. *Preprocessing:* Clean the collected tweets by removing noise, such as irrelevant characters, URLs, hashtags, and user mentions. Tokenize the tweets into individual words and perform other preprocessing tasks like stemming, stop-word removal, and lowercasing.

Determine the Number of Topics: Decide the number of topics you want to extract from the Twitter dataset. This will depend on the specific health care domain and the granularity of topics you are interested in. *Model Training:* Apply the CTM algorithm to the preprocessed tweets to estimate the topic proportions, topic correlations, and word distributions. This involves running the E-step and M-step iterations of the CTM algorithm to learn the model parameters. *Topic Interpretation:* Analyze the learned model to interpret the discovered topics and their correlations. Each topic will represent a specific health care theme discussed on Twitter, and the correlations between topics will capture the relationships and dependencies among these themes. *Topic Visualization and Analysis:* Visualize the results of the DiCTM to gain insights into the health care topics being discussed on Twitter. You can use techniques such as word clouds, topic networks, or topic distributions to understand the prevalence and relationships between different health care themes.

In Fig (3),The block diagram for the Distributed Correlated Topic Model typically consists of the following components. *Input Data:* This represents the collection of documents or textual data on which the CTM will be applied. It could be a dataset of healthcare-related topics in social media, for example. **Text Preprocessing:** This module involves the preprocessing steps such as tokenization, stop word removal, stemming, and other techniques to clean and prepare the textual data for further analysis. *Topic Model Training:* This component represents the training phase of the CTM. It involves applying the CTM algorithm on the preprocessed text data to estimate the topic proportions and correlations between topics.

Topic Modeling Output: This module generates the output of the CTM algorithm, which includes the estimated topics, their associated probabilities, and the topic correlations. This information helps in understanding the underlying themes and relationships between topics in the dataset. Post-processing and Analysis: After obtaining the topic modeling output, this module involves further analysis and interpretation of the results. It may include tasks such as topic labeling, topic coherence evaluation, topic visualization, and extracting insights from the modeled topics. *Evaluation and Validation:* This component assesses the quality and effectiveness of the CTM by evaluating various metrics, such as coherence scores, perplexity, or other domain-specific evaluation measures. It helps in validating the performance of the DiCTM on the healthcare topics in social media.

Trend Analysis: Identify the trends and patterns of health care topics on Twitter over time. This can help in

monitoring public opinion, tracking emerging health issues, or assessing the impact of health campaigns or policies. Sentiment Analysis: Analyze the sentiment associated with each health care topic to understand the public sentiment or opinion towards specific health issues, treatments, or healthcare providers. Information Retrieval: Utilize the DiCTM to improve information retrieval in the health care domain on social media platforms. By associating tweets with relevant topics, you can enhance the accuracy of retrieving health-related information from Twitter. Applying the DiCTM to health care topics on Twitter allows for a comprehensive analysis of the discussions, correlations between topics, and emerging trends. It can assist in understanding public health concerns, tracking public sentiment, and providing valuable insights to healthcare professionals, policymakers, and researchers.

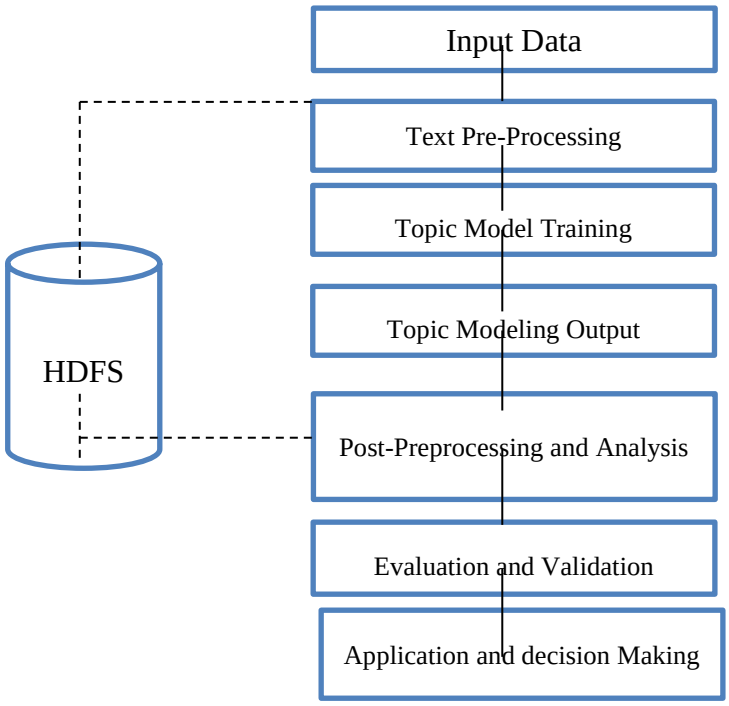


Fig:3 Control Flow of Distributed CTM

Application and Decision-making: This final module involves applying the DiCTM results to practical applications in healthcare. It could include tasks such as topic-based recommendation systems, trend analysis, opinion mining, or other applications that leverage the knowledge extracted from the topic modeling process.

Algorithm: *DiCTM*

1. *Notations:*

a. *D*→no. *of documents*

b. *N*→no. *of words in a document*

c. *K*→no. *of topics*

d. *w*→*Word index in the vocabulary*

e. *d*→*Document index*

f. *z*→*Topic index*

2. *Topic Proportions:*

a. *θ_d*→ *Topic proportions for document d*

b. *θ_d ~ Dirichlet(α), where α is a hyperparameter*

3. *Word Distribution:*

a. *β_k→ Word distribution for topic k*

b. *$\beta_k \sim Dirichlet(\eta)$, where η is a hyperparameter*

4. *Correlation Matrix:*

a. *ρ→ Correlation matrix of size K x K, capturing the correlations between topics*

b. *ρ_{kl}: Correlation coefficient between topic k and topic l*

5. *Topic Assignment:*

a. *z_{dn}→ Topic assignment for word n in document d*

b. *$z_{dn} \sim Multinomial(\theta_d)$, representing the probability of word n belonging to each topic in document d*

6. *Word Generation:*

a. *w_{dn}→ Observed word in document d at position n*

b. *$w_{dn} \sim Multinomial(\beta_{z_{dn}})$, indicating the probability of each word being generated from the topic assigned to it*

7. *Correlated Topic Model Equation:*

a. *$P(w_{dn} = w \mid \theta_d, \beta, z_{dn}) = \sum_k \theta_{dk} * \beta_{kw}$ This equation computes the probability of observing word w in position n of document d, given the topic proportions θ_d, word distribution β, and topic assignment z_{dn}.*

8. *Joint Probability of the Model:*

a. *$P(w, \theta, \beta, z \mid \alpha, \eta, \rho) = \prod_d P(\theta_d \mid \alpha) * (\prod_k P(\beta_k \mid \eta)) * (\prod_n P(z_{dn} \mid \theta_d)) * P(w_{dn} = w \mid \beta_{z_{dn}})$ This equation represents the joint probability of the observed words, topic proportions, word distributions, and topic assignments, given the hyperparameters α, η, and the correlation matrix ρ.*

9. *Inference:*

a. *The inference process in DiCTM involves estimating the posterior distribution of the latent variables (topic proportions θ, word distributions β, and topic assignments z) given the observed words w and the model parameters α, η, and ρ. This is typically done using variational inference or Gibbs sampling.*

The equations mentioned above provide the foundations for the Correlated Topic Model and its probabilistic modeling of correlations between topics in a document collection. These equations are utilized in the estimation and inference steps to uncover the latent thematic structure and topic correlations within the data.

DiCTM stands for Distributed Correlated Topic Model, which is a probabilistic topic model that extends the Latent Dirichlet Allocation (LDA) algorithm by incorporating correlations among topics. While LDA assumes that topics are independent of each other, DiCTM relaxes this assumption and allows for correlations between topics. The Correlated Topic Model addresses the limitations of LDA by introducing a correlation matrix that models the pairwise relationships between topics. This correlation matrix captures the dependencies or associations between different topics in the document collection. The DiCTM assumes that the topic proportions within each document are drawn from a multivariate Gaussian distribution, where the meaning is determined by the document-specific topic correlations.

4.14 Evaluations Metrics:

Normalized Mutual Information (NMI) is a measure of the similarity or mutual dependence between two sets of variables. In the context of topic modeling applied to healthcare topics in social media, NMI can be used to assess the agreement between the discovered topics and predefined topic labels or ground truth information. Here's how NMI can be applied to evaluate topic modeling results in healthcare social media analysis.

Topic Modeling: Apply a topic modeling algorithm, such as Latent Dirichlet Allocation (LDA), Distributed Correlated Topic Model (CTM), or any other relevant approach, to extract topics from healthcare-related social media data. The output of the topic modeling process is a set of topics, each represented by a distribution of words. *Predefined Topic Labels:* Establish a set of predefined topic labels or ground truth information that reflect the expected topics related to healthcare in social media. These labels can be created through expert knowledge, domain-specific guidelines, or manual annotation of a subset of the data. *Document-Topic Assignment:* Assign each document in the social media dataset to the most relevant topic based on the topic

modeling results. This assignment is typically done by considering the probabilities of the document belonging to different topics, as inferred from the topic modeling algorithm. *NMI Calculation:* Calculate the Normalized Mutual Information between the predefined topic labels and the assigned topics for the documents.

NMI measures the similarity between the two sets of labels, considering both the relative overlap and the size of the sets. NMI ranges between 0 (no similarity) and 1 (perfect agreement). *Interpretation*: A higher NMI value indicates a higher level of agreement between the discovered topics and the predefined labels, suggesting that the topic modeling algorithm has successfully captured the expected topics in the healthcare social media data. A lower NMI value indicates less agreement, implying that the discovered topics may not align well with the predefined labels or that the predefined labels themselves may have limitations. By using NMI, researchers can quantitatively evaluate the performance of their topic modeling approach in relation to predefined topic labels or ground truth information. This evaluation provides insights into the effectiveness and alignment of the topics discovered through topic modeling with the expected healthcare topics in social media.

Normalized Mutual Information (NMI) is a measure used to assess the similarity or agreement between two sets of labels. It quantifies the mutual dependence or information shared between the sets. Here is the mathematical equation for calculating NMI:

$$NMI = (2 * I(X, Y)) / (H(X) + H(Y)) \quad Eq(5)$$

Where NMI represents Normalized Mutual Information. I(X, Y) is the Mutual Information between the two sets of labels X and Y. H(X) and H(Y) are the entropies of X and Y, respectively. The Mutual Information (I(X, Y)) between two sets of labels is calculated using the following equation:

$$I(X, Y) = \sum\sum P(x, y) * \log(P(x, y) / (P(x) * P(y))) \quad Eq(6)$$

Where: P (x, y) is the joint probability of observing the pair (x, y) in the two sets of labels. P(x) and P(y) are the marginal probabilities of observing labels x and y, respectively. The entropy (H(X) or H(Y)) of a set of labels is calculated as:

$$H(X) = -\sum P(x) * \log(P(x)) \quad Eq(7)$$

Where: P(x) is the probability of observing label x in the set. Now, let's illustrate the calculation of NMI with a numerical example: Suppose we have two sets of labels X and Y, and we want to calculate their NMI. Here are the contingency tables representing the co-occurrence frequencies of the labels.

Table:7 NMI

	Y1	**Y2**	**Y3**
X1	10	5	2
X2	3	8	4

to calculate NMI, first need to Calculate the joint probabilities. **P(x, y):** P(X1, Y1) = 10 / (10 + 5 + 2 + 3 + 8 + 4) = 0.370 , P(X1, Y2) = 5 / (10 + 5 + 2 + 3 + 8 + 4) = 0.185,P(X1, Y3) = 2 / (10 + 5 + 2 + 3 + 8 + 4) = 0.074,P(X2, Y1) = 3 / (10 + 5 + 2 + 3 + 8 + 4) = 0.111,P(X2, Y2) = 8 / (10 + 5 + 2 + 3 + 8 + 4) = 0.296,P(X2, Y3) = 4 / (10 + 5 + 2 + 3 + 8 + 4) = 0.148. Calculate the marginal probabilities **P(x) and P(y):**P(X1) = (10 + 5 + 2) / (10 + 5 + 2 + 3 + 8 + 4) = 0.481, P(X2) = (3 + 8 + 4) / (10 + 5 + 2 + 3 + 8 + 4) = 0.519, P(Y1) = (10 + 3) / (10 + 5 + 2 + 3 + 8 + 4) = 0.407, P(Y2)=(5+8)/ (10 + 5 + 2 + 3 + 8 + 4)=0.375, P(Y2)=(2+4)/ (10 + 5 + 2 + 3 + 8 + 4)=0.187.

4.15 Cosine Similarity

Cosine similarity is a measure that quantifies the similarity between two vectors in a vector space. In the context of topic modeling applied to healthcare topics in social media, cosine similarity can be used to assess the similarity between topics based on their word distributions. Here's how cosine similarity can be applied. *Topic Modeling:* Apply a topic modeling algorithm, such as Latent Dirichlet Allocation (LDA), Distributed Correlated Topic Model (DiCTM), or any other relevant approach, to extract topics from healthcare-related social media data. The output of the topic modeling process is a set of topics, each represented by a distribution of words.

Word Vector Representation: Represent each topic as a vector in a high-dimensional space, where each dimension corresponds to a unique word in the vocabulary. The value of each dimension represents the weight or probability of the corresponding word in the topic's word distribution. **Cosine Similarity** Calculation: Calculate the cosine similarity between pairs of topic vectors. The cosine similarity between two topic vectors A and B is computed as the cosine of the angle between the vectors, and it ranges from -1 to 1.

$$\textbf{\textit{cosine similarity (A, B)}} = \frac{(\boldsymbol{A} \cdot \boldsymbol{B})}{(||\boldsymbol{A}|| * ||\boldsymbol{B}||)} \quad Eq\ (8)$$

where (A · B) represents the dot product of vectors A and B, and ||A|| and ||B|| represent their respective Euclidean norms [19].

Interpretation: A cosine similarity of 1 indicates that two topics have identical word distributions, meaning they are highly similar. A cosine similarity of 0 suggests that two topics are dissimilar and have no overlap in their word distributions. A cosine similarity of -1 indicates that two topics have completely opposite word distributions. By calculating cosine similarity between topics, researchers can identify similar or related topics within the healthcare social media data. This information can be useful for grouping related topics, identifying topic clusters, or identifying topics that are semantically similar in terms of the words they contain. It provides a quantitative measure to assess the similarity between topics and can aid in understanding the relationships between different healthcare topics in social media discussions.

Precision, recall, and F1 score are commonly used evaluation metrics in topic modeling tasks related to healthcare topics in social media. These metrics help assess the performance and effectiveness of the topic modeling algorithms. Here's an explanation of the mathematical equations for precision, recall, and F1 score: Precision: Precision measures the proportion of correctly identified topics (true positives) out of all the topics identified by the model. It provides an indication of how reliable the model is in identifying relevant topics.

$$\textbf{Precision} = \frac{\textbf{True Positives}}{(\textbf{True Positives} + \textbf{False Positives})} \qquad Eq\ (9)$$

True Positives (TP): The number of topics correctly identified as positive (relevant topics). False Positives (FP): The number of topics incorrectly identified as positive (non-relevant topics identified as relevant). A high precision score indicates a low false positive rate, meaning the model is accurately identifying relevant topics without many false alarms. Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of correctly identified topics (true positives) out of all the actual positive topics in the dataset. It provides an indication of how well the model captures all relevant topics.

$$\textbf{Recall} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Negatives}} \qquad Eq\ (10)$$

False Negatives (FN): The number of topics that are positive (relevant topics) but are incorrectly identified as negative (non-relevant topics). A high recall score indicates a low false negative rate, meaning the model is effectively capturing most of the relevant topics without missing many. F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that considers both precision and recall. The F1 score is often used when there is an imbalance between the positive and negative topics or when both precision and recall are equally important.

$$\textbf{F1 Score} = \frac{2 * (\textbf{Precision} * \textbf{Recall})}{(\textbf{Precision} + \textbf{Recall})} \qquad Eq\ (11)$$

The F1 score ranges between 0 and 1, with a higher value indicating better performance in terms of both precision and recall. These mathematical equations allow us to quantitatively evaluate the performance of topic modeling algorithms in capturing relevant healthcare topics in social media. By calculating precision, recall, and F1 score, we can assess the model's ability to accurately identify relevant topics while minimizing false positives and false negatives.

4.16 Comparison results:

In this section, To compare the Distributed Correlated Topic Model (DiCTM) and Cluster Visualized Latent Dirichlet Allocation (CvLDA) models in the context of healthcare topics on social media Twitter. Here's a comparison analysis, CvLDA assumes that documents are generated based on a mixture of topics, where each topic is a distribution over words. It treats documents as independent, making it suitable for capturing broad topical themes. Distributed CTM extends CvLDA by introducing correlations between topics. It assumes that documents are generated based on a mixture of correlated topics, allowing for capturing more nuanced relationships between topics. ***Topic Dependencies: Cv***LDA assumes topics are independent of each other, making it difficult to model correlations or dependencies between topics. It treats each topic as a separate entity without considering their interrelationships. Distributed CTM explicitly models the correlations between topics. It allows for capturing topic dependencies and provides a more accurate representation of the relationships between different healthcare topics on Twitter. ***Interpretability: Cv***LDA provides interpretable topics based on word distributions. It assigns each word a probability of belonging to a particular topic, allowing for easy interpretation of the main themes in the data. Distributed CTM also provides interpretable topics like CvLDA. However, it considers topic correlations, which can provide additional insights into how different healthcare topics relate to each other in social media discussions.

Table 8: Results with models

Method	Accuracy (%)	Precision	Recall	F1 Score	Topics
LSA[21]	57.52	0.67	0.72	0.69	50
LDA[21]	60.95	0.69	0.74	0.71	50
DiBERT	85.78	0.78	0.79	0.73	50
DiXLnet	91.12	0.88	0.81	0.76	50
CvLDA	93.31	0.89	0.83	0.79	50
DiCTM	95.43	0.98	0.91	0.88	50
LSA[21]	56.19	0.67	0.68	0.67	100
LDA[21]	58.85	0.69	0.70	0.69	100
DiBERT	75.67	0.79	0.76	0.72	100
DiXLnet	86.34	0.87	0.85	0.83	100
CvLDA	88.21	0.88	0.87	0.87	100
DiCTM	91.34	0.90	0.91	0.89	100
LSA[21]	62.67	0.71	0.75	0.73	150
LDA[21]	59.23	0.70	0.68	0.69	150
DiBERT	72.77	0.73	0.78	0.79	150
DiXLnet	86.44	0.88	0.82	0.82	150
CvLDA	90.12	0.91	0.90	0.89	150
DiCTM	93.67	0.92	0.92	0.91	150
LSA[21]	60.00	0.70	0.70	0.70	200
LDA[21]	63.42	0.70	0.78	0.74	200
DiBERT	73.23	0.78	0.79	0.76	200
DiXLnet	87.98	0.83	0.84	0.83	200
CvLDA	89.98	0.85	0.87	0.85	200
DiCTM	**90.32**	**0.92**	**0.91**	**0.86**	**200**

Granularity: LDA is well-suited for capturing broad topical themes and can be used to identify major healthcare topics on Twitter, such as diseases, treatments, or public health issues. DiCTM is more effective in capturing fine-grained correlations and dependencies between healthcare topics. It can uncover more nuanced relationships, such as the association between specific diseases and their related symptoms or medications. Suppose we have a Twitter dataset focused on healthcare discussions. Using CvLDA, we might identify topics like "COVID-19," "vaccination," "healthcare policies," and "mental health." These topics represent the major themes in the Twitter data without considering their interdependencies. With Distributed CTM, we can identify not only the major themes but also the correlations between them. For example, we might discover that the "COVID-19" topic is highly correlated with the "vaccination" and "public health" topics, indicating the strong association between these topics in healthcare discussions on Twitter. In summary, while CvLDA provides a straightforward approach for topic modeling in healthcare on Twitter, DiCTM offers a more advanced modeling technique that accounts for topic correlations. DiCTM can provide a deeper understanding of the relationships between healthcare topics, allowing for more nuanced insights and analysis. The choice between CvLDA and DiCTM depends on the specific research objectives and the level of granularity and detail required in modeling the healthcare topics on Twitter.

Table:6 Comparison Result Analysis with 50-topics

Method	Accuracy(a)(%)	Precision(p)	Recall(R)	F1 Score	Topics(T)
LSA[21]	57.52	0.6667	0.7221	0.6933	50
LDA[21]	60.95	0.6938	0.7356	0.7141	50
DiBERT	85.78	0.78	0.79	0.73	50
DiXLnet	91.12	0.88	0.81	0.76	50
CvLDA	**93.31**	**0.89**	**0.83**	**0.79**	**50**
DiCTM	**95.43**	**0.98**	**0.91**	**0.88**	**50**

As compared to current models, Distributed CTM achieves the highest accuracy (%), precision, recall, and F1 scores with 50 clusters of topics relevant to health (Table 6 categorizes the precision and recovery or recall values of the suggested models). The model functions most well when a cosine metric from the Twitter corpus of Euclid is utilized. Because of its superior accuracy, distributed CTM that uses a cosine similarity estimate beats other models.

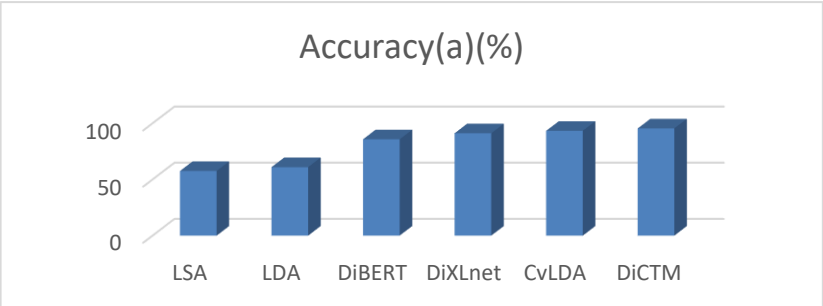


Fig:4 Model Comparison with 50 Topics

With this accuracy, a corpus of tweets from 50 to 200 data participants was made available. The Distributed CTM model performs better than the other models in terms of accuracy. The cosine distance similarity indicators clearly indicated the highest level of topical identification of all sorts of tweets in this context.

Table:7 Comparison Result Analysis with 100-Topics

Method	Accuracy(a)(%)	Precision(p)	Recall(R)	F1 Score	Topics(T)
LSA[21]	56.19	0.6676	0.6791	0.6733	100
LDA[21]	58.85	0.6854	0.7011	0.6932	100
DiBERT	75.67	0.79	0.76	0.72	100
DiXLnet	86.34	0.87	0.85	0.83	100
CvLDA	**88.21**	**0.88**	**0.87**	**0.87**	**100**
DiCTM	**91.34**	**0.90**	**0.91**	**0.89**	**100**

Distributed CTM received high-quality accuracy (%), precision, recall, and F1 score using 100 health-related topic cluster when compared to current models, as shown in table 7. The precision and recovery or recall values from the suggested models are also categorized in this table.

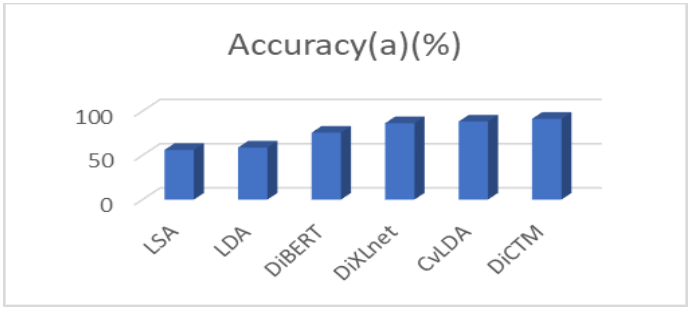


Fig:5 Model Comparison with 100 topics

The model functions most well when a cosine metric from Euclid's Twitter corpus is employed. Distributed CTM utilizing a cosine similarity metric surpasses other models in terms of accuracy. Because of this precision, a corpus of tweets from 50 to 200 data participants was produced. It has been shown that the Distributed CTM model beats the other models in terms of accuracy. The maximum level of acknowledgment of subjects connected to all forms of tweets was effectively proven in this context by the cosine distance similarity indicators.

Table:8 Comparison Result Analysis with 150-Topics

Method	Accuracy(a)(%)	Precision(p)	Recall(R)	F1 Score	Topics(T)
LSA[21]	62.67	0.705	0.753	0.7285	150
LDA[21]	59.23	0.699	0.679	0.689	150
DiBERT	72.77	0.73	0.78	0.79	150
DiXLnet	86.44	0.88	0.82	0.82	150
CvLDA	**90.12**	**0.91**	**0.90**	**0.89**	**150**
DiCTM	**93.67**	**0.92**	**0.92**	**0.91**	**150**

As compared to current models, Distributed CTM achieves the highest accuracy (%), precision, recall (R), and F1 scores using 150 health-related topic clusters. The precision and recovery or recall values of the suggested models are categorized in table 8. The model functions most well when a cosine metric from Euclid's Twitter corpus is employed.

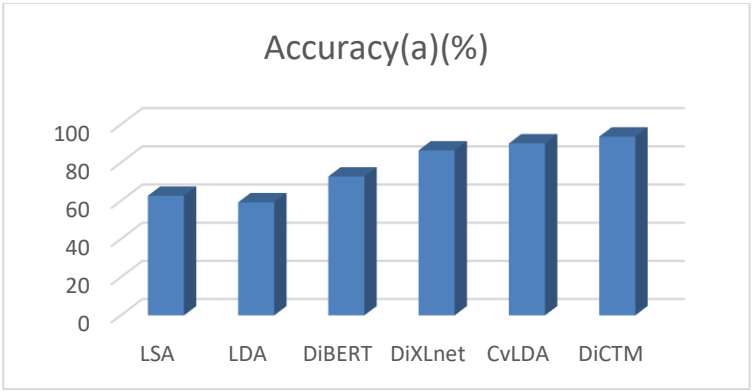


Fig:5 Model Comparison with 150 topics

Distributed CTM utilizing a cosine similarity metric surpasses other models in terms of accuracy. Because of this precision, a corpus of tweets from 50 to 200 data participants was produced. It has been shown that the Distributed CTM model beats the other models in terms of accuracy. The cosine distance similarity indicators in this instance were able to successfully exhibit the highest level of recognition of pertinent themes to all different forms of twitter postings.

Table:9 Comparison Result Analysis with 200-Topics

Method	Accuracy(a)(%)	Precision(p)	Recall(R)	F1 Score	Topics(T)
LSA[21]	60.00	0.698	0.702	0.7	200
LDA[21]	63.42	0.7039	0.776	0.7384	200
DiBERT	73.23	0.78	0.79	0.76	200
DiXLnet	87.98	0.83	0.84	0.83	200
CvLDA	**89.98**	**0.85**	**0.87**	**0.85**	**200**
DiCTM	**90.32**	**0.92**	**0.91**	**0.86**	**200**

Using 200 topic clusters relating to health, Distributed CTM outperforms previous models in terms of accuracy (%), precision, recall, and F1 scores. The model works best when a cosine metric from the Twitter corpus of Euclid is used.

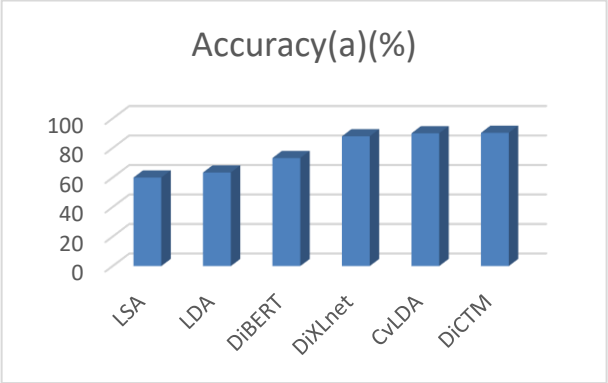
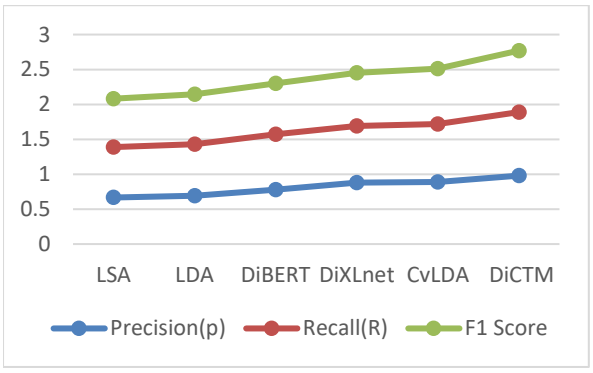
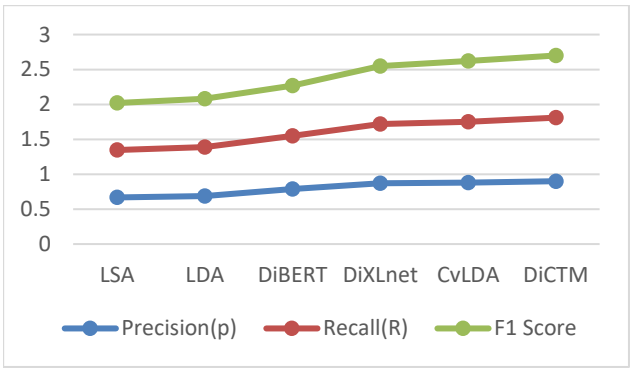


Fig: 6 Model Comparison with 200 topics

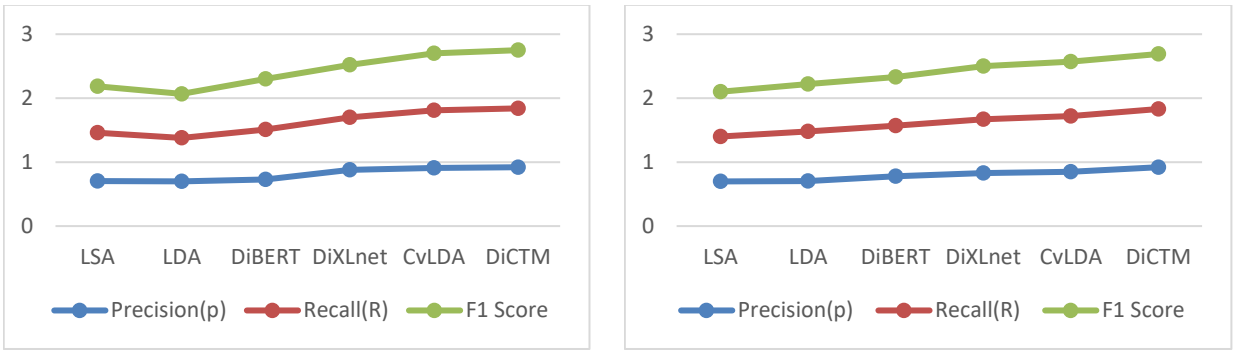
Using a cosine similarity metric, distributed CTM outperforms other models in terms of accuracy. This accuracy allowed for the creation of a corpus of tweets from 50 to 200 data subjects. In terms of accuracy, the Distributed CTM model performs better than the competition. In this manner, the cosine distance similarity indicators effectively illustrated a greater degree of awareness of pertinent themes, covering all varieties of Twitter remarks.



50 Topics



100 Topics



150 topics **200 topics**

Fig:7 Comparison Analysis on proposed models with 200 topics

The identified topics and their correlations provide valuable insights into public perceptions, concerns, and interests related to healthcare on Twitter. Analysis can inform strategic decision-making, public health campaigns, or targeted interventions to address specific healthcare issues or improve healthcare communication. Overall, the conclusion of a DiCTM analysis on healthcare topics in social media Twitter should summarize the key findings, highlight the main themes and their correlations, discuss the implications for healthcare stakeholders, and acknowledge any limitations or areas for further research. It should emphasize the value of DiCTM in uncovering insights from social media discussions and its potential to inform healthcare strategies and decision-making. The future scope of topic modeling techniques on healthcare topics in social media is promising, with several potential avenues for further exploration and development. Here are some future directions included.

Here's a comparison analysis, CvLDA assumes that documents are generated based on a mixture of topics, where each topic is a distribution over words. It treats documents as independent, making it suitable for capturing broad topical themes. Distributed CTM extends CvLDA by introducing correlations between topics. It assumes that documents are generated based on a mixture of correlated topics, allowing for capturing more nuanced relationships between topics. DiCTM can provide a deeper understanding of the relationships between healthcare topics, allowing for more nuanced insights and analysis. The choice between CvLDA and DiCTM depends on the specific research objectives and the level of granularity and detail required in modeling the healthcare topics on Twitter.

5. Conclusion and Future Enhancements:

In this section, the Distributed Correlated Topic Model (DiCTM) analysis on healthcare topics in social media Twitter would depend on the specific findings and insights derived from the model. However, here are some general points that could be included in the conclusion. The DiCTM analysis successfully identified and extracted latent topics from the healthcare-related tweets in the Twitter dataset. The identified topics represent the main themes and discussions related to healthcare in social media. The DiCTM captured the correlations and dependencies between different healthcare topics discussed on Twitter. The model revealed how certain topics tend to co-occur or have similar patterns of discussion, indicating relationships and associations among healthcare themes. The analysis provided interpretable topics with associated word distributions, enabling the understanding of the key terms and concepts within each healthcare theme. The identified topics shed light on the prevalent discussions and concerns in the realm of healthcare on Twitter.

By examining the temporal aspects of the DiCTM analysis, insights into the evolution of healthcare topics over time were gained. The analysis revealed any changes or shifts in the prominence or sentiment of different healthcare themes on Twitter. Integrating sentiment analysis with DiCTM provided an understanding of the sentiment polarity associated with each healthcare topic. The sentiment analysis component revealed positive, negative, or neutral sentiment trends within the discussions related to different healthcare themes. The findings from the DiCTM analysis can have implications for healthcare organizations, policymakers, and researchers.

The identified topics and their correlations provide valuable insights into public perceptions, concerns, and interests related to healthcare on Twitter. Analysis can inform strategic decision-making, public health campaigns, or targeted interventions to address specific healthcare issues or improve healthcare communication. Overall, the conclusion of a DiCTM analysis on healthcare topics in social media Twitter should summarize the key findings, highlight the main themes and their correlations, discuss the implications for healthcare stakeholders, and acknowledge any limitations or areas for further research. It should emphasize the value of DiCTM in uncovering insights from social media discussions and its potential to inform healthcare strategies and decision-making. The future scope of topic modeling techniques on healthcare topics in social media is promising, with several potential avenues for further exploration and development. Here are some future

directions included.

Improved Topic Interpretability: Enhancing the interpretability of topics generated by topic modeling techniques can provide more meaningful insights for healthcare professionals and stakeholders. Future research can focus on developing techniques that produce topics with clearer and more coherent interpretations, allowing for easier understanding and actionable insights.

Incorporation of Contextual Information: Contextual information, such as user profiles, network structures, temporal dynamics, or geographical data, can enhance the accuracy and relevance of topic modeling in healthcare social media analysis. Future studies can explore techniques that effectively incorporate these contextual factors into topic modeling algorithms, providing a more comprehensive understanding of healthcare discussions. *Integration of Multimodal Data:* Social media platforms host diverse types of data, including text, images, videos, and audio. Future research can explore innovative approaches to incorporate and analyze multimodal data within topic modeling frameworks. Integrating multiple modalities can provide richer insights into healthcare topics, such as analyzing image content or sentiment in videos related to healthcare discussions.

Fine-grained Sentiment Analysis: Sentiment analysis is an essential component of healthcare social media analysis, but it can be further refined to capture fine-grained sentiment. Future studies can explore sentiment analysis techniques that go beyond basic polarity detection and capture more nuanced emotions and opinions expressed in healthcare discussions. *Topic Evolution and Trend Analysis:* Understanding how healthcare topics evolve and identifying emerging trends is crucial for staying up to date with the ever-changing healthcare landscape. Future research can focus on developing dynamic topic modeling techniques that capture the temporal aspects of healthcare discussions in social media and identify shifting trends over time.

Personalized Topic Modeling: Different users may have unique interests and perspectives within healthcare discussions. Future studies can explore personalized topic modeling techniques that adapt to individual user preferences, enabling tailored healthcare topic recommendations or personalized content delivery. *Ethical and Privacy Considerations:* As the use of social media data in healthcare research continues to grow, addressing ethical considerations and ensuring privacy protection is crucial. Future studies can focus on developing ethical guidelines, privacy-preserving techniques, and frameworks that prioritize data privacy and address potential biases or ethical concerns. *Real-time Monitoring and Early Detection:* Social media platforms provide a wealth of real-time data that can be leveraged for monitoring public health trends and detecting early warning signs. Future research can explore topic modeling techniques that enable real-time monitoring and early detection of emerging healthcare issues, epidemics, or adverse events based on social media discussions. By exploring these future directions, topic modeling techniques can advance the understanding of healthcare topics in social media, enabling more accurate, timely, and actionable insights for healthcare professionals, policymakers, and researchers.

References:

[1] kanksha Rajput, Manoj Kumar,"Anti-Ebola: an initiative to predict Ebola virus inhibitors through machine learning","Mol Divers

[2] 2022 Jun;26(3):1635-1644. doi: 10.1007/s11030-021-10291-7. Epub 2021 Aug 6."

[3] Samuel K. Kwofie, Joseph Adams, Emmanuel Broni, Kweku S. Enninful, Clement Agoni, Mahmoud E. S. Soliman, Michael D. Wilson. "Artificial Intelligence, Machine Learning, and Big Data for Ebola Virus Drug Discovery" , Pharmaceuticals, 2023

[4] Manu Anantpadma,†#∇ Thomas Lane,‡# Kimberley M. Zorn,‡ Mary A. Lingerfelt,‡ Alex M. Clark,§ Joel S. Freundlich, Robert A. Davey, Peter B. Madrid, and Sean Ekinscorresponding author"Ebola Virus Bayesian Machine Learning Models Enable New in Vitro Leads ", ACS Omega. 2019 Jan 31; 4(1): 2353–2361. Published online 2019 Jan 30. doi: 10.1021/acsomega.8b02948

[5] Mujahed I. Mustafa, Shaza W. Shantier. "Next generation multi epitope based peptide vaccine against Marburg Virus disease combined with molecular docking studies" , Informatics in Medicine Unlocked, 2022.

[6] Victor O. Gawriljuk, Phyo Phyo Kyaw Zin, Ana C. Puhl, Kimberley M. Zorn et al. "Machine Learning Models Identify Inhibitors of SARSCoV-2" , Journal of Chemical Information and Modeling, 2021

[7] Songyuan Geng, Qiling Luo, Kun Liu, Yunchao Li, Yuchen Hou, Wujian Long. "Research status and prospect of machine learning in construction 3D printing" , Case Studies in Construction Materials, 2023

[8] Fritz Heinrich Obermeyer, Martin Jankowiak, Nikolaos Barkas, Stephen F. Schaffner et al. "Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness" , Cold Spring Harbor Laboratory, 2022.

[9] Jayashree Jayashree, Shivaprakash T., Venugopal K.R.. "MUNPE:Multi-view Uncorrelated

Neighborhood Preserving Embedding for unsupervised feature extraction" , Institute of Electrical and Electronics Engineers (IEEE), 2023.

[10] Roh, S.M.; Eun, B.W.; Seo, J.Y. Does coronavirus disease 2019 affect body mass index of children and adolescents who visited a growth clinic in South Korea?: A single-center study. Ann. Pediatr. Endocrinol. Metab. 2022, 27, 52–59.

[11] Barrett, C.E.; Koyama, A.K.; Alvarez, P.; Chow, W.; Lundeen, E.A.; Perrine, C.G.; Pavkov, M.E.; Rolka, D.B.; Wiltz, J.L.; Bull-Otterson, L.; et al. Risk for newly diagnosed diabetes >30 days after SARS-CoV-2 infection among persons aged <18 years—United States, March 1, 2020–June 28, 2021. Morb. Mortal. Wkly. Rep. 2022, 71, 59–65.

[12] Kim, Y.; Park, S.; Oh, K.; Choi, H.; Jeong, E.K. Changes in the management of hypertension, diabetes mellitus, and hypercholes terolemia in Korean adults before and during the coronavirus disease 2019 pandemic: Data from the 2010–2020 Korea National Health and Nutrition Examination Survey. Epidemiol. Health 2023, e2023014, Online ahead of print.

[13] Korea Disease Control and Prevention Agency. Weekly Updates for Countries with Major Outbreaks. 2022. Available online: http://ncov.mohw.go.kr/bdBoardList_Real.do?brdId=1&brdGubun=11&ncvContSeq=&contSeq=&board_id=&gubun= (accessed on 25 June 2022).

[14] Korean Diabetes Association. A Statement from the Korean Diabetes Association Regarding the COVID-19 Vaccine. 2021. Available online: https://www.diabetes.or.kr/popup/2021/pop20210126.html (accessed on 25 June 2022).

[15] World Health organization. Coronavirus Disease 2019 (COVID-19): Situation Report, 51; World Health Organization: Geneva, Switzerland, 2020; pp. 1–9. Available online: https://apps.who.int/iris/handle/10665/331475 (accessed on 21 September 2022).

[16] Ko, Y.S.; Lee, S.B.; Cha, M.J.; Kim, S.D.; Lee, J.H.; Han, J.Y.; Song, M. Topic modeling insomnia social media corpus using BERTopic and building automatic deep learning classification model. J. Korean Soc. Inf. Manag. 2022, 39, 111–129.

[17] Hossain, M.M.; Tasnim, S.; Sultana, A.; Faizah, F.; Mazumder, H.; Zou, L.; McKyer, E.L.J.; Ahmed, H.U.; Ma, P. Epidemiology of mental health problems in COVID-19: A review. F1000Research 2020, 9, 636.

[18] Rossi, R.; Socci, V.; Talevi, D.; Mensi, S.; Niolu, C.; Pacitti, F.; Di Marco, A.; Rossi, A.; Siracusano, A.; Di Lorenzo, G. COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy. Front. Psychiatry 2020, 11, 790.

[19] De Santis, E.; Martino, A.; Rizzi, A. An infoveillance system for detecting and tracking relevant topics from Italian tweets during the COVID-19 event. IEEE Access 2020, 8, 132527–132538.

[20] Wang, T.; Lu, K.; Chow, K.P.; Zhu, Q. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. IEEE Access 2020, 8, 138162–138169.

[21] Gourisaria, M.K.; Jee, G.; Harshvardhan, G.M.; Singh, V.; Singh, P.K.; Workneh, T.C. Data science appositeness in diabetes mellitus diagnosis for healthcare systems of developing nations. IET Commun. 2022, 16, 532–547.

[22] JUNAID RASHID,SYED MUHAMMAD ADNAN SHAH,AUN IRTAZA TOQEER MAHMOOD,MUHAMMAD WASIF NISAR,, MUHAMMAD SHAFIQ,AND AKBER GARDEZI,"Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clusterin","Digital Object Identifier 10.1109/ACCESS.2019.2944973".IEEE Access.

[23] Singh, V.; Gourisaria, M.K.; Gm, H.; Rautaray, S.S.; Pandey, M.; Sahni, M.; Leon-Castro, E.; Espinoza-Audelo, L.F. Diagnosis of Intracranial Tumors via the Selective CNN Data Modeling Technique. Appl. Sci. 2022, 12, 2900.

[24] Priya, S. ., & Suganthi, P. . (2023). Enlightening Network Lifetime based on Dynamic Time Orient Energy Optimization in Wireless Sensor Network. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 149–155. https://doi.org/10.17762/ijritcc.v11i4s.6321

[25] Dhablia, A. (2021). Integrated Sentimental Analysis with Machine Learning Model to Evaluate the Review of Viewers. Machine Learning Applications in Engineering Education and Management, 1(2), 07–12. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/12

[26] Earshia V., D. ., & M., S. . (2023). Interpolation of Low-Resolution Images for Improved Accuracy Using an ANN Quadratic Interpolator . International Journal on Recent and Innovation Trends in

Computing and Communication, 11(4s), 135–140. https://doi.org/10.17762/ijritcc.v11i4s.6319

[27] Wiling, B. (2021). Locust Genetic Image Processing Classification Model-Based Brain Tumor Classification in MRI Images for Early Diagnosis. Machine Learning Applications in Engineering Education and Management, 1(1), 19–23. Retrieved from http://yashikajournals.com/index.php/mlaeem/article/view/6