# Topic Modeling: A Comprehensive Review

**Article** *in* ICST Transactions on Scalable Information Systems · July 2018

DOI: 10.4108/eai.13-7-2018.159623



**2 authors:**

Pooja Kherwa
Maharaja Surajmal Institute Of Technology
**26** PUBLICATIONS **354** CITATIONS
SEE PROFILE

Poonam Bansal
Maharaja Surajmal Institute Of Technology
**36** PUBLICATIONS **437** CITATIONS
SEE PROFILE

EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

# Topic Modeling: A Comprehensive Review

Pooja Kherwa[1,]*, Poonam Bansal[2]

[1,2] Maharaja Surajmal Institute of Technology, C-4 Janak Puri. GGSIPU. New Delhi-110058.

## Abstract

Topic modelling is the new revolution in text mining. It is a statistical technique for revealing the underlying semantic structure in large collection of documents. After analysing approximately 300 research articles on topic modeling, a comprehensive survey on topic modelling has been presented in this paper. It includes classification hierarchy, Topic modelling methods, Posterior Inference techniques, different evolution models of latent Dirichlet allocation (LDA) and its applications in different areas of technology including Scientific Literature, Bioinformatics, Software Engineering and analysing social network is presented. Quantitative evaluation of topic modeling techniques is also presented in detail for better understanding the concept of topic modeling. At the end paper is concluded with detailed discussion on challenges of topic modelling, which will definitely give researchers an insight for good research.

**Keywords**. Topic Modeling, Latent Dirichlet Allocation, Latent Semantic Analysis, Inference. Dimension reduction.

Received on 25 February 2019, accepted on 03 July 2019, published on 24 July 2019



doi: 10.4108/eai.13-7-2018.159623

*Corresponding author. Email: poona281280@gmail.com

## 1. Introduction

Nowadays, when majorities of literature in every sphere of life, become digitized and stored electronically in the databases, through digital libraries or social networking databases. Researcher needs powerful automatic tools to read this data and understand underlying themes of data like human beings. Topic modeling is a technique comes with group of algorithms that reveal, discover and annotate thematic structure in collection of documents [1]. Vector space model is the foundation for many advanced information retrieval techniques [2][3] and topic models [4]. The VSM (Vector space model) was the first simple algebraic model directly based on the term document matrix for extracting semantic information from word usage [5]. A basic vector space model can break the text into character unigrams or bigrams. It is based on Bag of words (BOW) approach, where the exact ordering of term in a document is ignored but the frequency of occurrence of each term is retained an important factor [6][7]. VSM has spread its wings in variety of applications including machine learning, search engines, natural language processing, machine translation that involve measuring the similarity of concepts, context between words, phrases, and documents [3][9]. For modeling a retrieval system like in search engine or digital collection, all words in a document are not equally important. So, we assign each term in a document, a weight that depends on the number of occurrences of the term in a document. The concept of weighting is to give more weight to unexpected events and less weight to frequently occurring events. The most popular way to formalize the idea for term document matrices is (term frequency*inverse document frequency) [tf*idf], from family of weighting function [8][10]. Topic modeling algorithms provides technique from multiple perspectives to find hidden semantics in document collection and cluster the themes as topics. The distributional model like vector space, Latent semantic analysis, Probabilistic latent semantic model and Latent Dirichlet allocation can be used to derive word meaning representation based on the analysis of statistics [2]. In last two decades, there have been several papers and research results devoted to various topic modeling techniques including LSA (latent Semantic analysis), Plsa (Probabilistic latent semantic analysis) and LDA (Latent Dirichlet Allocation). Researchers from various fields mathematicians, statisticians, computer scientists, biologists, and neuroscientists, have explored the topic modeling from diverse perspectives. Major researchers in the topic modeling are David M Blei, Michael I. Jordan, Thomas L. Griffiths, Hanna M. Wallach, Jordan Boyd Graber, Mark Johnson, D. Mimno, Jonathan Chang, and Katherine Heller In these David M Blei is the most dominant researcher with continuous research publication since 2003 to 2017 with around 200 publications in various reputed journals and international conferences. Jordan Boyd Graber has written first paper in topic modeling in 2007 with David M. Blei on use of topic modeling for word sense disambiguates. In 2008 presented the concept of multilingual topic model and syntactic topic model and complete his PhD on linguistic extension of topic models in 2010.In 2011 discovered the concept of interactive topic modeling with Brianna Sati off. Jordan contributed majorly in sentiment analysis, multilingual topic modeling and topic modeling on speech – large scale topic modeling and efficient tree-based topic modeling and online topic modeling. David Mimno come with the idea of detecting author influence in collection through author topic modeling and more on correlation among topics and articles in corpus using Pachinko Allocation models.

Since topic modeling is a very powerful technique, has spread its wings in multiple domains from Natural language processing to scientific literature, Software Engineering, Bioinformatics, Humanities etc. In Last two decades, hundreds of papers have been published, but in the literature of topic modeling, there is no any paper that explore the whole literature in a single comprehensive manner. Some significance contribution by core researcher are only on the latent Dirichlet allocation(LDA)[11][12].The concept of dimension reduction and semantic theme detection is prevalent in the literature of from early 1999.So in this paper we get motivate to provide a comprehensive review on topic modeling which includes a brief classification hierarchy as well as explanation of each method with algorithms like latent semantic analysis, probabilistic latent semantic analysis, non-negative matrix factorization and then finally the latent Dirichlet Allocation(LDA) is presented in great detail.

So, in totality this paper makes these contributions:

- Classification hierarchy of different topic modeling methods is designed. And topic modeling methods are briefly explained.
- In this paper scholarly article from 2003-2018 related to latent Dirichlet allocation-based topic modeling are explored fully and taxonomy is designed for first time.
- Application of topic modeling application in various domains like scientific literature, Software Engineering, Bioinformatics, Sentiment analysis are discussed in detail.
- Challenges of topic modeling for exploring the topic modeling from multiple perspective are given in detail.
- Quantitative evaluation with standard metric on two data set are also done on both probabilistic and non-probabilistic (Latent Dirichlet allocation, Latent Semantic analysis) topic modelling method.
- In the end some topic modeling tools and packages are also explained.

## 2. Topic Modeling Classification

Topic model can be considered as a methodology to present the huge volume of data generated due to advancement in computer & web technology in low dimension and to present the hidden concepts, prominent feature or latent variables of data, depending on the application context, should be identified efficiently. Dimension reduction was initially seen through algebraic perspective, decomposing the original matrix into factor matrix. So, our topic modeling classification strategy in this survey described as

- Probabilistic Model
- Non-Probabilistic topic model (Algebraic Model).

Non probabilistic approaches are matrix factorization algebraic approaches and come in to existence in early 1990 with the concept of Latent Semantic Analysis [13] and Non-Negative Matrix Factorization [14][15][16][17].Both LSA and NNMF works on Bag of words approach, where the corpus is converted into term document matrix and order of terms is totally neglected, only terms count in document matters. Probabilistic model came into existence to improve the algebraic model like latent semantic analysis by adding the probability sense using generative model approaches [1].

The next level in classification tree is based on Supervised and unsupervised approach to topic modeling, in this hierarchy, probabilistic topic model like Plsa and Latent Dirichlet Allocation falls. Initially both the Plsa and LDA was fully unsupervised approaches, but later on many researchers works in Latent Dirichlet Allocation (LDA) model with supervised approach for model learning, Plsa is explored with semi Supervised fashion in a very limited domains of Application. Non probabilistic models have not found significant contribution in supervised fashions, and considered as outside the of scope of paper.

The last level in classification hierarchy is considering the sequence of words during topic modelling. Till 2006, the complete topic modeling approaches based on BOW (Bag of Words, in 2006 Hanna M. Wallach introduced the importance of incorporating sequence of words in topic modeling through n-gram statistics, and a Hierarchical Dirichlet Bigram model with better accuracy than BOW approaches is presented [18]. Although researchers are exploring Bigram and N-gram based topic modeling perspective, till today the most dominant approach is still based on Bag of Words. where the corpus is converted into term document matrix and order of terms is totally neglected, only terms count in document matters.

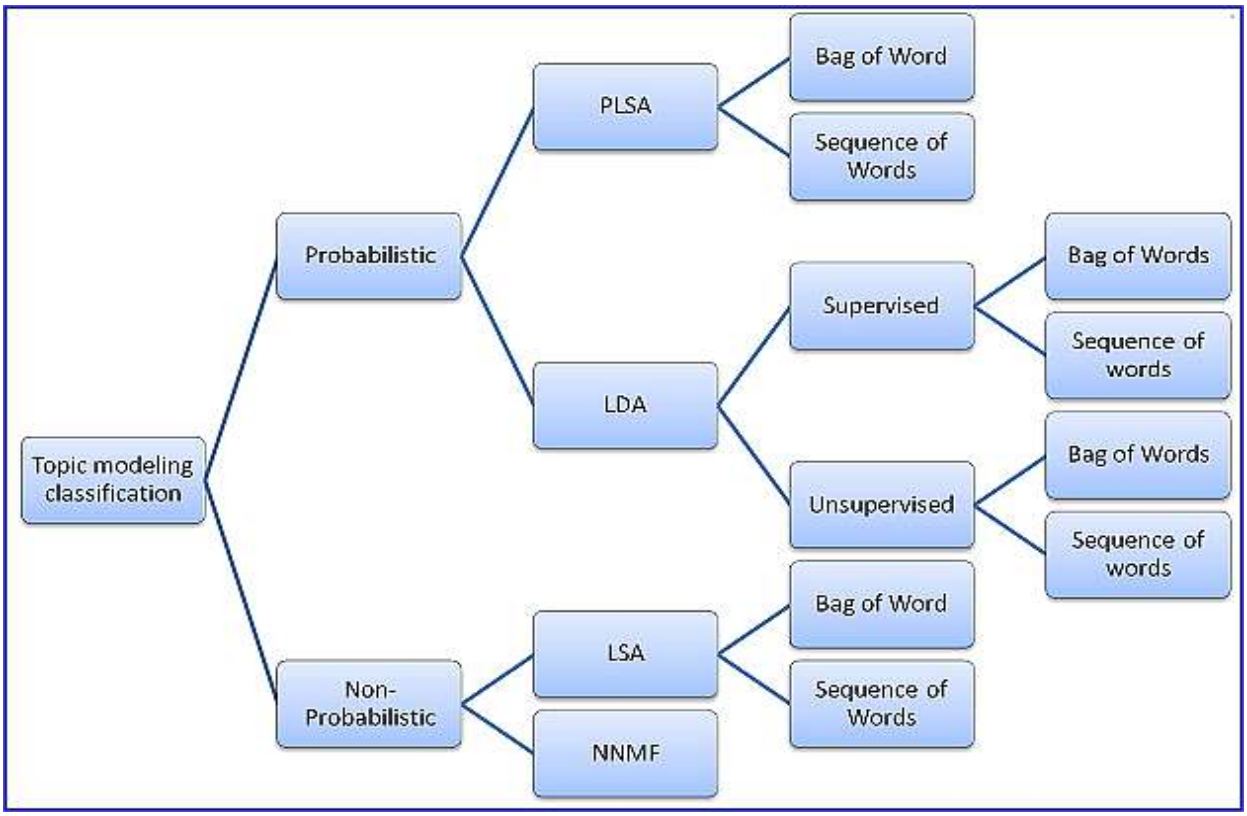


**Figure 1**. Topic Modeling Classification Hierarchy

## 3. Topic Modeling Methods

The most popular topic modeling algorithms that contributed every sphere of text analysis in multiple domains includes Latent semantic analysis, Non-Negative Matrix Factorization, Probabilistic Latent semantic analysis, and Latent Dirichlet Allocation

### 3.1 Latent Semantic Analysis

Latent Semantic analysis is an algebraic method based on single value decomposition (SVD) and presenting the semantic space of documents in which more semantic relationship, contextual usage come closer. Unlike human child a machine learns everything from language. In 1997 Landauer and Dumais proposed latent semantic analysis technique [13]. They used Singular value decomposition (SVD) that had been applied to many different areas including information retrieval, natural language processing and modeling of human language knowledge or at least the basis of human knowledge [18][19][20].
The theoretical foundation of LSA is distributional hypotheses [42], which states that terms with similar meaning also occur very closer in their contextual usage. All the semantic relation between texts is inferred directly from given corpus text. It uses vector representation of text to compute the similarity between texts to find similar words in text, to semantically organize text into semantic clusters. Latent Semantic Analysis has its application in text mining including automatic Grading of Essay Intelligent tutoring, Information retrieval, Social network Analysis, and Text Summarization.

#### 3.1.1 Algorithm for LSA

1. For text corpus a term by document matrix A is created where M is the terms or words in a document and N is the number of documents in corpus.
2. Each entry is weighted with function that associates importance of each word in the document and in the whole corpus. A weighting function should give a low weight to a high frequency types that occur in many documents and high weight to types that occur in some documents but not all [19].
3. Apply SVD (Singular Value Decomposition) to matrix A, which reveals the correlation between terms in documents. SVD decomposes the matrix A into three matrix such that
$A=U\sum V^T$
Where U is an orthogonal matrix ($U^TU$) and V is an orthogonal matrix ($V^TV$) and $\sum$ is a diagonal matrix ($\sum$=diagonal ($\sigma_1$, $\sigma 2$.....$\sigma_K$) with the remaining matrix cell all zeros [20]
4. The matrix A is truncated as $A_k$ ,the truncated Ax is the best or closest rank k approximation to the original matrix A[21][22][23].An experiment is conducted to project the high dimensional space into low dimension with single value decomposition(SVD) and showing the correlation of terms in semantic space with different ranks[20].

### 3.2 Non-Negative Matrix factorization (NNMF)

It is a new dimension reduction technique in which the problem of negative numbers present in dataset are addressed by placing non-negativity constraints on the data model. The concept of NNMF used by Pattero in 1994 for environmental data [24]. In 1997 again used in unsupervised learning by Lee & Seung [25]. NMF aims to address the problem of negative component in data models, because in many applications, the negative components contradict physical reality. So NMF is a representation of data confined to nonnegative vectors.
The idea is decomposing a nonnegative matrix B into non negative factors V & H,
$B=VH+C$, $V\geq 0, H\geq 0$.
This decomposition is approximate in nature and C is noise or error matrix.
This concept is called non-negative matrix factorization or also called PMF (Positive matrix factorization). NNMF also supports sparsity constraints in many application [26][27][28][29]. In microarray dataset many machine learning approaches like classification and clustering have been applied to cancer identification using molecular gene expression data set [34][39][76]. A comparative experiment on filtered set of genes for cancer classification using discrimination methods [41] was done and abnormalities in cell behaviour are highlighted with valid proofs In text mining ,NNMF and clustering based text summarization approach is also presented with better performance than clustering[40].NNMF has been used in many applications including segmentation, dimension reduction, pattern recognition, image processing, language modeling etc.

### 3.3 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis(PLSA) is a technique of dimension reduction in text mining based on bag of words (BOW) for detecting semantic co-occurrence of terms using probabilistic framework in a corpus. It was developed by the Hoffman [30].
The first statistical model for revealing semantic co-occurrence in document term matrix of corpus was Aspect model [31]. It is based on the concept that each word generate from single topic and different word in a document may be generated from different topics. Each document is represented as a list of mixing proportion for these mixture components and thereby reduced to a probability distribution on a fixed set of topics.

#### 3.3.1 Algorithm for PLSA

The model is defined with three parameters $\phi$, $\theta$ and $\rho$, where $\phi_z$ is the distribution of words in latent topics z given by p(w|z), $\theta_d$ is the distribution of topics in document d given by p(z|d) and pd is the probability of choosing document d given by p(d).The words are generated as

1.Choose document d~ρ.
2.Choose topic z~ $\theta_d$.
3.Choose word w ~ $\phi_z$.

$$P(d,w) = P(d)\sum_z \theta_d(z)\varphi_z(w) \quad (1)$$

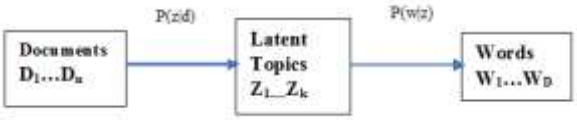


**Figure 2**. Probabilistic Latent Semantic Analysis (Plsa)

The probability of the observed word document pair(d,w) can be obtained by marginalization over latent topics.

## 3.4 Latent Dirichlet Allocation

### 3.4.1 LDA Generative Process

Latent Dirichlet Allocation(LDA) is an approach ,based on definneti theorem [32] to capture significant inter as well as intra document statistical structure via mixing distributional assumes that document arise from multiple topics, a topic is defined as distribution over a vocabulary .Corpus is associated with predefined number of topics k, and each document in corpus contain these topics with different proportion. Topic Modeling aims to learn these topics from data or corpus. LDA is based on hidden variable model, prevalent in machine learning from decades. Generative model does not consider order of words in producing documents, so purely based on bag of words (BOW) approach. Topic modeling is to automatically discover the topics from a collection of documents, so in a data set, the K topic distribution must be learned through statistical inference. This algorithm defined generative process as a joint probability distribution over both observed and hidden variables. The process of learning the topic distribution is described through plate notation given in figure.

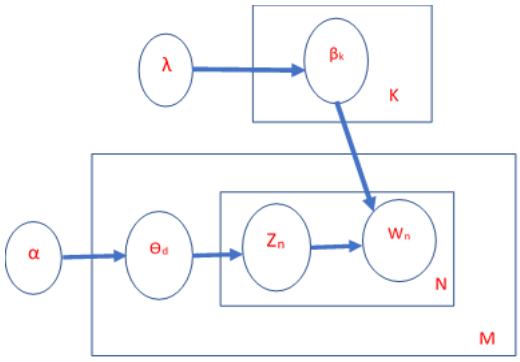


**Figure 3**. Latent Dirichlet Allocation (LDA)Plate Notation.

### 3.4.2 LDA Algorithm

For each document

(a) Distribution over topics $\theta_{d\sim}$Dir(α). where Dir(.)is a draw from a uniform Dirichlet distribution with scaling parameter α.
(b) For each word in the document

1.Draw a specific topic $z_{d,n\sim}$ multi($\theta_d$)
2. Draw a word $w_{d,n\sim}\beta_{z,d,n}$.

The central inferential problem for LDA is determining the posterior distribution of latent variables given the document.

$$P(\theta, z \mid w, \alpha, \beta) = \frac{P(\theta, z, w \mid \alpha, \beta)}{P(w \mid \alpha, \beta)} \quad (2)$$

The aim of Topic modeling is to automatically discover the topics from a collection of documents. The documents are observed, while the topic structure, the topics, per document topic distribution and per document per word topic assignment is hidden structure. The utility of topic modelling is come from the property that the inferred hidden structure resembles the thematic structure of the collection. This interpretable hidden structure annotates each document in the collection and theses annotation can be used in information retrieval, classification, and corpus exploration

## 4. Posterior Inference for Latent Dirichlet Allocation

Various inference techniques that provide approximately solution for marginal probabilities for observed data under various topic models includes variational inference [4][35], collapsed variational inference [36] Expectation propagation [37]and Gibbs sampling [32]. Each approximation technique has its advantages and disadvantages. Variational Inference provides a deterministic methodology for approximating likelihood and posteriors [38]. Variational inference is based on reformulating the problem of computing the posterior distribution as an optimization problem, perturbing that problem and finding solutions to the perturbed problem. Variational inference generally performs worse than sampling methods due to an approximate bias, but runs faster. In addition, being faster than sampling methods and is particularly well suited for large scale problems

## 4.1 Mean Field Variational Method

Mean field methods are based on optimizing Kullback-Leibler (K-L) divergence with respect to a so-called variational distribution. In contrast to the true posterior, the mean field variational distribution for LDA is one where the variables are independent of each other, and each governed by a different variational parameter:

## 4.2 Gibbs Sampling

Gibbs sampling is a typical Markov Chain Monte Carlo

method was originally proposed for image restoration [42].

## 5. Quantitative Analysis for Topic Modeling Methods

For Quantitative analysis of topic modeling methods, we choose two topic modeling techniques: one is chosen from the algebraic method and other is from Probabilistic method. Algebraic topic modeling technique based on matrix factorization is latent semantic analysis (LSA) and probabilistic topic modeling technique is Latent Dirichlet Allocation (LDA). The term topic modeling in machine learning community is given by David Blie et al in 2003, by using latent Dirichlet allocation algorithm for semantic themes detection in large collection of text, since then LDA is considered as synonymous for topic modeling in machine learning community. All the evolutionary algorithms of topic modeling like correlated topic model, Dynamic topic model, Author topic model, Multilingual Topic Model, Syntactic topic model and many more discussed in taxonomy designed in section 5, are either extension of LDA or based on LDA. So, till today LDA is the most widely used approach for topic modeling.

### 5.1 Data Set Chosen

For the empirical evaluation we choose two data set. Movies Review: The labeled dataset consists of 5000 IMDB movie reviews, specially selected for sentiment analysis. The sentiment of the reviews is binary, meaning an IMDB rating < 5 results in a sentiment score of 0, and a rating >=7 has a sentiment score of 1. No individual movie has more than 30 review
National Institutes of Health (NIH): This dataset holds information on research grants awarded by the National Institutes of Health (NIH)in 2014. It includes both 'projects' and 'abstracts' files. The dataset is data-Frame of 100 randomly-sampled grants' abstracts and metadata

### 5.2 Pre-processing

For pre-processing of data set, a pre-processor function is designed which includes to remove all the numeric digits from dataset and also to convert all capital letters characters into lowercase for further processing. In next step a word tokenizer is also constructed and applied to the pre-processed input. This tokenizer will split the document at individual word level and consider that word as token for further processing.

A vocabulary-based document term matrix is created, we collect unique terms from all the documents and mark each of them with a unique id using the create vocabulary function. In typical topic modeling, we begin by constructing a document term matrix (DTM)from input documents. This is the first step to vectorizing text by creating a map from words to a vector space.

### 5.3 Parameter Setting

For topic modeling evaluation we set algorithm parameters for both latent semantic algorithm and latent Dirichlet allocation. Although both the algorithms provide automatic insight in large dataset in the form of semantic themes, the way of modeling is totally different in the algorithms. Latent Semantic Analysis learns latent topics by performing matrix decomposition on the term document matrix. Latent Dirichlet allocation (LDA) is a generative probabilistic model that assumes a Dirichlet prior over the latent topics[1].In this experiment, we choose K =10 (number of topics in LDA and rate of factorization for SVD in LSA)in Movie reviews dataset and K=10 for NIH(According to data size).

### 5.4 Evaluation Measure

In the literature perplexity and log-likelihood is the most widely used approach for evaluating topic quality of topic model[99], but our algebraic topic modeling approach latent semantic analysis is a matrix factorization method based on single value decomposition, so no use of expectation maximization algorithms in learning topics, So we choose the evaluation metrics that are applicable to both topic modeling methods, so for these models, we choose coherence as the most prominent quality measures.

Probabilistic Coherence: Probabilistic coherence measures how associated words are in a topic, controlling for statistical independence. For example, suppose you have a corpus of articles from the sports section of a newspaper. A topic with the words {sport, sports, ball, fan, athlete} would look great. But we actually know that it's a terrible topic because the words are so frequent in this corpus as to be meaningless. In other words, they are highly correlated with each other but they are statistically-independent of each other[100].For each pair of words {a, b}in the top M words in a topic, probabilistic coherence calculates P(b|a) - P(b), where {a} is more probable than {b} in the topic. Suppose the top 4 words in a topic are {a, b, c, d}. Then, we calculate
P(a|b) - P(b), P(a|c) - P(c), P(a|d) - P(d)
P(b|c) - P(c), P(b|d) - P(d)
P(c|d) - P(d)
And all 6 differences are averaged together, giving the probabilistic coherence measure.

Table 1. Coherence Analysis for both data set with different number of topics

| **Movie Review Dataset(K=10)** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LDA** | | | | | | | | | |
| T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| 0.050 | 0.095 | 0.075 | 0.051 | 0.045 | 0.089 | 0.067 | 0.072 | 0.071 | 0.091 |
| **LSA** | | | | | | | | | |
| T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| 0.059 | 0.068 | 0.032 | 0.14 | -0.002 | 0.121 | 0.035 | 0.051 | 0.057 | 0.057 |
| **National institutes of Health (NIH)(K=10)** | | | | | | | | | |
| **LDA** | | | | | | | | | |
| T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| 0.084 | 0.039 | 0.077 | 0.11 | 0.28 | 0.07 | 0.12 | 0.055 | 0.16 | 0.12 |
| **LSA** | | | | | | | | | |
| T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
| 0.0018 | 0.14 | 0.25 | 0.39 | 0.71 | 0.58 | 0.048 | 0.54 | 0.32 | 0.18 |
| **National institutes of Health (NIH)(K=5)** | | | | | | | | | |
| **LDA** | | | | | | | | | |
| T1 | T2 | T3 | T4 | T5 | | | | | |
| 0.15 | 0.087 | 0.093 | 0.17 | 0.14 | | | | | |
| **LSA** | | | | | | | | | |
| T1 | T2 | T3 | T4 | T5 | | | | | |
| 0.0018 | 0.14 | 0.25 | 0.39 | 0.71 | | | | | |

Table 2. Top terms from both the topic modeling algorithm in Movie Review Dataset

| **LDA K=10** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| T1 | **T2** | **T3** | **T4** | **T5** | **T6** | **T7** | **T8** | **T9** | **T10** |
| Show | Role | characters | love | Comedy | M | man | horror | Version | war |
| Years | Play | films | life | Best | Why | where | match | Original | us |
| Saw | character | character | family | Funny | Your | action | night | production | life |
| Now | John | work | father | Fun | Know | gets | here | Book | our |
| Tv | performance | director | woman | Always | Thing | scene | films | Director | those |
| **LSA K=10** | | | | | | | | | |
| T1 | **T2** | **T3** | **T4** | **T5** | **T6** | **T7** | **T8** | **T9** | **T10** |
| character | films | films | mary | Films | Show | man | life | Years | man |
| Films | characters | man | lou | Show | Mary | films | zomb | M | where |
| Man | actors | life | horror | Race | Lou | where | page | say | page |
| Where | these | match | prom | Horror | Prom | guy | flic | books | here |
| Life | character | role | night | Actors | night | gray | bett | something | bettie |

## 5.5 Results & Discussion

In this empirical evaluation, the probabilistic coherence for both data set, for both topic modeling methods with specified number of topics is calculated at individual topic. Coherence score for all the ten topics are shown in table, In Movie Review data set, it is clearly visible that Latent Dirichlet Allocation model gives more coherent topics than latent semantic analysis, except only topic 1. In the case of NIH sample dataset, where we consider only 100 sample, and number of topics taken is 10, here in table shown that latent semantic analysis gives more coherent topics. But in small dataset with only 100 samples, selected number of topics 10 is quite large, for good results we take k (the number of topics 5, and then probabilistic coherence, and it is clearly visible in table that LDA produce more coherent number of topics than Latent semantic analysis.

In analysis of top terms in both the topic modeling algorithms, it is visible through top terms analysis that both the topic modeling algorithms try to give a clear human like true picture of dataset. but when we compare thoroughly other than statistical measures already discussed, top terms of all topics in latent Dirichlet allocation are clearly showing very crisp topics, clearly Separated and also cohesive to tell nature of individual topic like topic one is about tv show, topic2 is about john's performance in play, topic 3 is about films character, topic 4 is about family, topic 5 is about fun etc. In latent semantic topics are quite clear, but not as crisp

as in latent Dirichlet allocation , like topic 1 and topic two are not well separated also topic 4 and 5 .So in conclusion it can be stated that latent semantic analysis can be used in many natural language processing applications like text summarization, word ambiguity detection in information retrieval, but for efficient topic modeling like presenting semantic themes in large collection of huge data set, Latent Dirichlet allocation based on probabilistic theory give acceptable results.

## 6. Evolution of LDA Based topic models

Evolution in LDA started with the aim for best language model incorporating the interest of modeler (as per domain requirement) or application specific and presenting the hidden semantic of dataset as real as given by human. For example: User may be interested in finding correlation among different topics discussed in the corpus, like proceedings of conferences from many decades and many domains, so practitioner may be interested in particular time research interest, author contents and how domains are changing with time, how vocabulary in domains are changing in time like machine learning terminology is changing with time, from keyword “machine learning” to “deep learning” and this is true for all domains. These evolutions are generated in literature of topic modeling by relaxing the basic assumption of latent Dirichlet allocation as well as deleting unnecessary parameters or adding important ones.

### 6.1 Supervised Topic Models

Supervised topic model is developed to use topic modeling in predictive analysis. Each document is presented with a response variable [44].STM has provision for single –label as well as multi-label annotation in classification and regression problems. A supervised updating of LDA as Slda [44], Med-LDA [50]and Discriminative LDA(Disc-LDA) [49] are developed for single-label category of classification. For multi label classification, the generative model is labelled-LDA (L-Lda) [48]. It is a technology for annotation or labelling of images in big scenes and rating a document based on review on webpage, and in a scientific collection of articles, annotate the document with no. of downloads in a certain time. Recent work on supervised topic modeling includes Prior-LDA and Dependency-LDA, for multilabel classification. The advantage of STM is that prediction of label is done at the word level, instead of traditional classification approaches where it is done at document level and labelling of all document is completed simultaneously, again an important factor for more accurate classification. For a regression perspective, other than Slda for most relevant approaches are the Dirichlet multinomial regression [51]and the inverse regression topic model [52].

### 6.2 Correlated Topic Model

Latent Dirichlet allocation is capable enough for considering an effective tool for the statistics analysis of document collections. LDA consider that words of each document arise from multiple topics and each individual topic is a distribution over the vocabulary. In real life data, all the topics
in collection are correlated in some way. For e.g. a document about sports is more likely being about health than finance. So, the correlated topic model algorithm is developed in this topic proportion exhibit correlation through logistic normal distribution [53]. Correlated Topic Model (CTM) with more flexible distribution in topic proportions find a covariance structure among the components [45]. This covariance structure laid the foundation for revealing a more realistic model of latent topic structure to find correlation among the topics. In JSTOR archive of journal science data [12], it is demonstrated that CTM provides a natural way of visualizing and exploring collection of documents. Posterior inference is big challenge in CTM, so mean field variational methods are used to efficiently obtain an approximation of this posterior distribution.

### 6.3. Pachinko Allocation Topic Model

Pachinko Allocation Topic Model (PAM) is an advanced algorithm to explore correlation in complicated unstructured documents through directed acyclic graph (DAG) structure [54]. It is powerful enough to learn nested, arbitrary and sparse topic correlation in corpus. PAM exhibit both word correlation in vocabulary as well as inter topic correlations [55]. It can be simple hierarchy or directed acyclic graph with cross connected edges or with missing nodes. In PAM each leaf node is associated with vocabulary word and each interior node corresponds to topic, with Dirichlet distribution over its children topic and these children topics are traditional topics in latent Dirichlet allocation (LDA) model, thus representing a mixture over topics. Although PAM provides a powerful way to extract inter-topic correlation in large collection of text. It has the same traditional problem –How to determine optimal number of topics. So significant contribution includes nonparametric Bayesian prior to automatically determine number of topics [65].

### 6.4. Hierarchical Topic Model

Hierarchical Topic model is an algorithm that relaxes the dominant theory of Latent Dirichlet allocation (LDA) that each document could only have limited topics. In LDA mixture distribution is taken at predefined k number of topics. In an advanced CTM LDA is extended to consider – find relationship between different topics, but number of topics is fixed unable to draw level of abstraction of a topic in hierarchy. So Hierarchical Topic Modeling is an approach

in TM to find more absolute topics and organize the topics according to a hierarchy in which more abstract topic appear near the root of hierarchy and more concrete are appear near the leaves [68]. The Hirarchial-lda generate a mixture distribution on topics using a nested CRP (Chinese Restaurants Process) [56].

## 6.5. Spherical Topic Models

Spherical Admixture Model is a topic model that integrates the features of mixture of von Mises-Fisher distributions (movMF) [57] with LDA, for modeling different aspects of data [58]. In both topic modeling and dimensionality reduction application Spherical Admixture Model was found to produce more relevant topic feature than either the movMF spherical mixture model or LDA. Spherical Admixture Model outperform because of its three attribute-Cosine distance, negative terms, and word absence /presence. Other model for modeling word absence is multivariate Bernoulli likelihood [59]. Spherical Admixture Model uses Mises fisher likelihood to model data more accurately than multinomial model [60]. Spherical Admixture Model uses variational mean field as inference technique.

## 6.6. Author topic model

Author topic model is generative model approach for author topic model that simulates the contents of document and interest of authors [61]. It is an extension of LDA to include authorship information. Each author is associated with a multinomial distribution over topics and each topic is associated with multinomial distribution over vocabulary. Before Author topic model other approaches were based on Stylometric features [62] like stylistic features of text corpus including frequency of certain words, stop words, sentence length, and diversity of author's vocabulary used for finding authors attributes in text collections.

## 6.7 Multilingual Topic Model

Multilingual topic model (MTM) is the methodology to detect semantic themes or topics in corpus, where corpus vocabulary is not restricted to only one language. The world wide web is contributing to knowledge in multiple languages emerges from different continents [46]. MTM has its root deep in 2004, when first latent semantic analysis model for cross lingual languages are developed [46]. PolyLDA is based on the concept of latent Dirichlet allocation and extends it for extracting semantic themes across multiple languages parallelly. This model take input from multiple documents consists of multiple languages and discovers T topics in each language parallelly [45]. PolyLDA implemented for corpus of German, Greek, English, Danish, and Spanish etc. MuTo (Multilingual Topic model for unaligned text) assumes that similar themes appear in multiple languages and discover the topics at vocabulary level in corpus [47].

## 6.8 Dynamic Topic Model

DTM is a topic model that bound the exchangeability assumption of LDA unlike to CTM at corpus level. It is assumed in CTM that terms are exchangeable within document as well as in corpus. But in reality, this assumption is not appropriate, any domain's literature evolve with time, and also the vocabulary used to represent the study area, for example, Nanotechnology in 1950 is totally different from that was in 1990.The topics of collection evolve over time. So DTM is concerned with detecting the topics in corpus with their dynamics as well. The DTM reveal the evolution of topics in corpora by dividing the corpora in time slice. For e.g. by Year, by Decades etc. Each slice is modelled with k component topic model and each time slice t evolve from t-1 slice. In this logistic normal distribution is used and variational methods are used for approximate posterior inference. In DTM the Hellinger distance is used for similarity score. This similarity metrics, has the potential to provide a time-based notion of document similarity. The two articles about physics might be looked similar even one uses vocabulary of 1950 and other of 2017.A dynamic topic model of learning analytics Iterative-Dynamic Topic Model(Iterative-DTM) that can accommodate the evolution of all the unbounded number of topics, topics can die or be born at any epoch and the representation of each topic evolve according to a Markovian dynamics [63].

## 6.9 Syntactic Topic Model

Topic modeling based on BOW (Bag of Word) approach where words are exchangeable in corpus, and their probability is stable in permutation [65]. These models have proved useful for quickly analysing very large corpus and successful for classification and information retrieval [64]. BOW based topic model is sufficient where abstract statistical footprint of themes of document is sufficient for success, but fails where finer grained, qualities of language are required. So Syntactic topic model, a nonparametric Bayesian topic model is developed to infer both syntactically and semantically coherent topics. Previous work on considering syntactic feature includes semantic space model [65] and linear sequence models [66] [67] These models have significant work in the domain of word sense disambiguation application in natural language processing.

## 7. Applications of Topic modeling in various research Area of Technology

Topic modeling has spread its wings in various area of research in last two decades since its inception like Scientific Research, Bioinformatics, Social network, Software Engineering and in many other domains.

### 7.1 Scientific Research

The first generative model for topic modeling known as latent Dirichlet Allocation (LDA), In this David M.Blei explore the scientific literature of 16000 documents from a subset of the TREC AP corpus(Harman 1992)with 100 topics[1].In 2009 the extended model named correlated topic model and dynamic topic model was applied on dataset of JSTOR archives to the article published from 1980-2002 [12].17000 articles from science magazine with 100 topics and Yale law journal with 20 topics was done by David M. Blei [4].
Thomas L Griffiths and Mark Steyvers use topic modeling for PNAS abstracts dataset and explored research trends in published articles and distinguish them as hot and cold topics. Topics discovered by LDA using Gibbs sampling inference pick out meaningful structure of science and reveal some of the relationship between scientific papers in different disciplines [68]. Nicholas analysed area of interdisciplinary research funded by science foundation using topic modeling algorithms [69]. This helps science foundation administrators to better understand the content and context of funding portfolios in order to help promote future science funding plans. A topic model is presented for simultaneously modeling papers, authors, and publication venues known as Author conference topic model [73]. It includes data of 14134 authors 10716 papers and 1434 conferences from Arnet Miner. In one more advanced topic model where traditional collaborative filtering and probabilistic topic modeling are combined and provides a system to recommend scientific article to users of online community [70][71]72]. One another perspective of topic modeling of research fields by Michael Paul, in this LDA is used to classify research papers based on topic & language. In this various insightful statistics and correlation within and across three research field: linguistics, Computational linguistics, Education were found [85].

### 7.2 Bioinformatics

Topic mode is a useful method as compare to traditional means like classification & clustering in bioinformatics and it enhances practioner's ability to interpret biological information [87]. In 2006 a study of statistical modeling of biomedical corpora by David Blei demonstrate an LDA model can be employed to infer hidden factors permeating a biomedical text corpus to synthesize and organize information about complex biological phenomena [88]. The Probabilistic Latent Semantic Analysis (PLSA) in 2010 employed on expression microarray dataset for extraction of bio-Clusters, this model simultaneously groups' genes and samples. An Ensemble topic model for sharing health care data and predicting disease risk was developed by Andrew et al. In this approach, model is developed to allow sharing of beneficial information while staying within the bounds of data privacy [84].
Genome level composition of DNA sequences are analysed using probabilistic models [74]. John Barnett and Tommy Jakkola use topic modeling for Gene expression analysis. Gene expression profiling provides a window into the inner working of the cell, as exhibited through messenger RNA levels [75]. Drugs impact different pathways within the cell and the active pathway may be different within cell types. Understanding more precisely which pathway are affected by drugs in specific cell types open up new possibilities for more targeted drugs or combination drug therapies.

### 7.3 Social Network Analysis

The large amount of data available on social web platform opens new opportunities for mining information about real world. Because of its popularity and wide spread usage, it can be used to infer important aspects about the users of those services and about the things happening in their surroundings. Young Chul Cha et al described Social network analysis using topic model, in this paper, applied LDA to analyse the relationship graph in a large social network [77]. Different from the usual approaches that extract topics from textual data such as Bio and tweets, this approach relies purely on social network graph consisting of follow edges. This approach is especially useful when only linkage data is available. In 2009 Justin Grimmer presented a Bayesian hierarchical topic model for political analysis for measuring expressed agendas in senate press release [78]. It addresses the issues to efficiently measure the priorities political actors emphasize in statements. Debashis Naskar et al, developed SentLDA –a generative model to identify the sentiments of the users in the social network through their conversation [79]. An Author recipient topic model for social network analysis is developed [81] which learns topic distribution based on the direction sensitive messages sent between entities. A Multi attribute latent Dirichlet allocation (MA-LDA) is developed, in which the time and hash tag attributes of micro blogs are incorporated into the LDA model [80]. This method provides strong evidence of the importance of the temporal factor in extracting hot topics.

### 7.4 Software Engineering

With the evolution of software industry, a large volume of data is produced in various software repositories, among those data, it is estimated that between 80% and 85% of the data in software repositories is unstructured. For example, source code, documentation, test cases, bug repositories etc. Mining theses unstructured repositories can uncover interesting and actionable information to support various software Engineering tasks such as program comprehension, location, and traceability link recovery [83].

The underlying assumption for using topic modeling on software repositories is that software arty-facts share the same textual characteristics as text in natural language. So, configuration of LDA for software Engineering tasks using the same parameter used for natural language is quite easy.
So LDA-Ga, A genetic approach for software engineering task like traceability link discovery was developed to identify a near optimal configuration of LDA. For LDA parameters more specifically LDA-GA uses the coherence of documents pertaining to the same topic to derive the evolution of GA [82]. Again, a detailed analysis of software evolution on the source code history of well-known system Hot Draw and J-Edit was performed using topic modeling.

## 8. Challenges of Topic modeling

In last two decades although topic modelling has explored in various domains, including text mining in scientific literature, Software engineering, bioinformatics and social network analysis and many information retrieval research areas. And thousands of papers have written and models has been developed. According to our studies, some domains where topic modelling can be used are still unexplored. And certain aspect through which topic modelling can be understood in more significant way should be developed. So as per our study, some issues require further research, and require more contribution, to fully explore the topic modeling's vision of distant reading make possible. These areas of further exploration are followings.

### 8.1 Visualization

Topic models represents an abstract modeling which is totally abstract for modeler and for end users in many ways [47]. To understand the document collection, the topics in a topic model that is the distribution over words with the maximum probability in a topic explain-what is a topic all about. So, the most common way to understand topics is through visualization. The most common output of topic modeling algorithms is top terms of each individual topic using word list. Another most popular approach is word cloud. The topic model visualization (TMVE)is a tool that represents each topic with the most related topics as well as most related documents. This is calculated through $\theta_d$ _the percentage of topics proportion in each document.
Visualization tools are developed like termite [85] Termite is a compact, intuitive interactive visualization of the topics in a topic model, but by only including terms that rank high in saliency or distinctiveness, which are global properties of terms. It is restricted to providing a global view of the model rather than allowing a user to deeply inspect individual topic by visualizing a potentially different set of terms for every single topic.
Topic Nets [86] uses dimension reduction on topic composition to plot documents in 2D space. But does not show topic or document compositions. LDAvis [87] directly addresses the model checking problem by aiding topic interpretation through a relevance method for ranking term within topics for display to end users. It provides a global view of the topics, while at the same time allowing for a deep inspection of the terms most highly associated with individual topic. Terms are ranked by their probability under a topic is suboptimal. Uncertainty in topic models arises because of Multi-Modality methods, Different text-pre-processing methods and multiple human viewpoint & perspective on judgment of topical quality.

### 8.2 Interpretation of Topic Modeling

Visualization can help in great details about understanding topic model output and issues in modeling. Evaluating topic model for quality is a step to interpretability by product. The most widely used metric for quality of topic model was held-out-likelihood [96] Chang et al., shows that held out likelihood quality metric not suitable for providing ease of interpretability in probabilistic models. Automatic measurement of topic quality is good candidate for quality checks and interpretability [97][98]. Another work emphasis on finding the suitable topic model for specific application and also finding the relationship between multiple models for better interpretable by humans.

### 8.3 Memory Efficient Topic Modeling

Topic models provide a useful tool in analysing complicated text collection, but their computation complexity has hindered their use in large scale and real time application. Although there has been enough work on modeling and improving training time estimation in topic models has been done [88][89]. There has been little attention paid to the practically important problem of inferring topic distribution given existing models. LDA is effective Topic modeling used in classification, feature selection, and information retrieval, however using LDA on large dataset take hours even days due to the heavy computation and intensive memory access.
So, GPU (Graphical Processing Unit) is used to accelerate LDA training on single machine. In this GLDA achieve a speedup of 15X over the original CPU based LDA for large dataset efficient training of LDA on a GPU by Mean for Mode estimation [90]. Mean for Mode estimation is a variant of un-collapsed Gibbs sampling that uses the mean of Dirichlet distribution to make point estimate, can be implemented efficiently on GPU's. The algorithm exposes a lot of parallelism and has good statistical performance.

### 8.4 Stability

Another important challenge in topic modelling is stability. Stability can be defined in topic modelling as that topic modelling algorithm applied to the same data set with same parameters in multiple runs at multiple times the output not necessarily be same. Stability analysis in topic modelling has shown that the traditional topic modelling outcome is unstable both when the model is retained on same input

documents and when it is updated with new documents. So how, we can produce a definitive topic model that is both stable and accurate. A very little work done on stability analysis in Negative matrix factorization algorithms and including NNMF. Most authors on TM don't address this issue and instead simply utilize a single random initialization and present the result of TM as being definitive. Another challenge in TM is the identification of coherent topics in noisy data such as tweets [91].

The most common strategy for reducing instability in unsupervised learning is to use ensemble clustering techniques, which is based on the idea that combining large and diverse sets of clustering can produce more stable and accurate solution. [92] [93] [94], Works regarding the optimality and consistency of solutions produced by clustering and Bio clustering algorithms. In TM, some initial work for stability using ensembles done in matrix factorization. This include using a hierarchical scheme to combine multiple factorization, in the study of Protein networks [95]. However, these studies have not investigated the extend of the problems introduced by instability in the context of Topic Modelling.

## 9.Topic Modeling Toolkit

Two decades of research in topic modelling, produced several toolkits for the development of broad range of application of topic models. Theses toolkits are used in various domains including Cognitive sciences, Social web analysis, Bioinformatics, Natural language processing, Software Engineering and in many more.

Gensim: It is a free library designed to automatically extract semantic topics from documents very efficiently in python. The algorithms implemented in Gensim are LSA, LDA and Random Projections. These algorithms are Unsupervised in nature, means semantic patterns are statistically found in documents without any metadata, only need a corpus of plain text documents [107].

Stanford Topic Modeling ToolBox (TMT): This toolbox was written at the Stanford NLP group by Daniel Ramage &Evan Rosen in September 2009[108]. It was in written in Scala language using linear algebra library. TMT can be used only to detect semantic pattern in textual dataset. LDA and its two enhanced version Labelled LDA and PLDA are implemented in Stanford TMT.

MALLET: The MALLET topic model implementation is an extremely fast and highly scalable tool with Gibbs sampling, topic hyper parameter optimization and modules for inferring topics for new documents given trained models [106]. MALLET includes Latent Dirichlet Allocation, Pachinko Allocation, and Hierarchical LDA. For optimized implementation of Topic modelling algorithm, it depends on numerical optimization.

R Packages: R language has a rich set of packages or special library for efficient topic modelling includes lsa [101], lsafun[102],lda[103],topicmodels[105],textmineR[100],text2 vec[104].

## 10. Conclusion

From last one and half decades, topic modeling using latent Dirichlet allocation (LDA) is very popular in Machine learning and natural language processing community for handling large amount of unstructured data and annotating these data with themes and topic. In this paper we provide a comprehensive survey on topic models using a new classification hierarchy, Different Topic model methods like LSA, NMF, PLSA, and LDA are also explained in detail. Quantative evaluation of Topic modelling techniques is also presented with detail analysis, Different evolutionary model of Latent Dirichlet allocation are also presented first time, Different application area of topic model like Scientific research, Bioinformatics, Social network analysis and software engineering are explained. Although topic modeling have been explored in many areas of research to full extent, but according to our studies, some domains where topic modelling can be used are still unexplored like speech processing, image processing and detecting themes in audio and video data set. And certain aspect through which topic modelling can be understood in more significant way should be developed. In this paper various challenging aspects of topic modelling includes visualization of topics, Interpretation of topics as human, Memory efficient topic Models, Stability of topic modelling are also identified and discussed. Although topic modelling come a long distance in two decades, but certain questions are still unresolved. These includes:

- How to select optimal number of topics. some approaches like perplexity, log-likelihood of held out documents, Harmonic mean, Cross validation are some approaches but results are still ambiguous.
- How to choose priors, the choice of priors has direct effect on probability of held out documents and the quality of inferred topics. Which priors are best like Symmetric or Asymmetric and which is most suitable for problem in hand? No rigorous study is available on the choice of priors.
- Which inference technique is best suitable for problem domain, like Variational inference and Gibbs sampling has dominated the topic modelling from last two decades, but which is best is still not answerable?

So, with the advancement in topic modelling, various extended version of LDA and based on related theory come into existence and used in last two decades in various applications including text mining, biomedical research, sentiment analysis, speech technology, image processing, collaborative filtering, disability survey data, population genetics and the modelling of sequential data and user etc.

## References

[1] Blei, D. M. (2012) Probabilistic topic models. *Communications of the ACM*, 55(4): pp.77_84.

[2] Crain, S. P., Zhou, K., Yang, S. H., & Zha, H. (2012) Dimensionality reduction and topic modeling: From latent semantic indexing to latent Dirichlet allocation and beyond. *In Mining text data. Springer, Boston*, pp.129_161.

[3] Ghorab, M. R., Zhou, D., O'Connor, A., & Wade, V. (2013) Personalised information retrieval: Survey and classification, *User Modeling and User-Adapted Interaction*, 23(4): pp.381_443.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003) Latent Dirichlet allocation, *Journal of machine Learning research*, 3(Jan): pp.993_1022.

[5] Salton, G., Wong, A., & Yang, C. S. (1975) A vector space model for automatic indexing, *Communications of the ACM*, 18(11): pp. 613_620.

[6] Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrices, vector spaces, and information retrieval, *SIAM review*, 41(2): pp.335_362.

[7] Berry, M. W., Dumais, S. T., & O, Brien, G. W. (1995) Using linear algebra for intelligent information retrieval, *SIAM review*, 37(4): pp.573_595.

[8] Turney, P. D., & Pantel, P. (2010) From frequency to meaning: Vector space models of semantics, Journal of artificial intelligence research, 37: pp.141_188.

[9] Nakov, P., & Hearst, M. A, (2008) Solving relational similarity problems using the web as a corpus, Proceedings of ACL-08: HLT, pp.452_460.

[10] Sparck Jones, K, "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation, 28(1), pp.11_21,2012.

[11] Jelisavčić, V., Furlan, B., Protić, J, & Milutinović. (2012) Topic models and advanced algorithms for profiling of knowledge in scientific papers, In MIPRO, Proceedings of the 35th International Convention. pp.1030_1035.

[12] Alghamdi, R., & Alfalqi, K. (2015) A survey of topic modeling in text mining, Int. J. Adv. Computer. Sci. Appl. (IJACSA), 6(1).

[13] Deerwester, S., Dumais, S. T., Furnas, G. W, Landauer, T. K., & Harshman, R. (1990) Indexing by latent semantic analysis, Journal of the American society for information science, 41(6): 391.

[14] Paatero, P., & Tapper, U (1994) Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5(2): pp.111_126.

[15] Paatero, P. (1997) Least squares formulation of robust non-negative factor analysis, Chemometrics and intelligent laboratory systems, 37(1): pp.23_35.

[16] Lee, D. D., & Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization, Nature, 401: pp.6755_67788.

[17] Lee, D. D., & Seung, H. S. (2001) Algorithms for non-negative matrix factorization, In Advances in neural information processing systems, pp.556_562.

[18] Berry, M. W., & Martin, D. (2005) Principle component analysis for information retrieval, In E. Kontoghiorghes (Series Ed.), Statistics: A series of textbooks and monographs: Handbook of parallel computing and statistics, pp. 399_413,

[19] Buckley, C., Allan, J., & Salton, G. (1994) Automatic routing and ad-hoc retrieval using SMART: TREC 2. NIST SPECIAL PUBLICATION SP: pp.45_45.

[20] Kherwa, P., & Bansal, P. (2017) Latent Semantic Analysis: An Approach to Understand Semantic of Text, In International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp.870_874.

[21] Berry, M. W. (1992) Large-scale sparse singular value computations, The International Journal of Supercomputing Applications, 6(1), pp.13_49,

[22] Golub, G. C, Van Loan. (1989) Matrix Computations, John Hopkins U. P, Baltimore.

[23] Berry, M. W., & Fierro, R. D. (1999) Low-rank orthogonal decompositions for information retrieval applications, Numerical linear algebra with applications, 3(4): pp.301_327.

[24] Paatero, P., & Tapper, U. (1994) Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics,5(2): pp.111_126.

[25] Lee, D. D., & Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization. Nature, 401: pp.6755_ 6788.

[26] Cai, D., He, X., Wu, X., & Han, J. (2008) Non-negative matrix factorization on manifold, In Data Mining, ICDM'08. Eighth IEEE International Conference on Data Mining. pp. 63_72.

[27] Kreutz-Delgado, K., Murray, J. F., Rao, B. D., Engan, K., Lee, T. W., & Sejnowski, T. J. (2003) Dictionary learning algorithms for sparse representation, *Neural computation*, 15(2): pp.349_396.

[28] Zdunek, R., & Cichocki, A. (2007) Nonnegative matrix factorization with constrained second-order optimization, *Signal Processing*, 87(8): pp.1904_1916.

[29] Cichocki, A., & Zdunek, R. (2007) Regularized alternating least squares algorithms for non-negative matrix/tensor factorization, *In International Symposium on Neural Networks*: pp.793_802, J. Springer, Berlin, Heidelberg.

[30] Hofmann, T. (1999) Probabilistic latent semantic analysis, *In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp.289_296.Morgan Kaufmann Publishers Inc..

[31] Hofmann, T. (2017) Probabilistic latent semantic indexing", *In ACM SIGIR Forum* Vol.51, No. 2: pp.211_218.

[32] De Finetti, B. (2017) Theory of probability: a critical introductory treatment, Vol. 6, John Wiley & Sons.

[33] Kintsch, W. (2006) Latent semantic analysis: *A road to meaning. Laurence Erlbaum*.

[34] Lee, D.D.and Seung, H.S. (1999) Learning the parts of the objects by non-negative matrix factorization, *Nature*, 401: pp.788-791.

[35] J.-P. Brunet, P. Tamayo, T. R. Golun, and J. P. Mesirov (2004) Metagenes and molecular pattern

discovery using matrix factorization, *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 12. pp.4164–4169

[36] Xing, E. P., Jordan, M. I., & Russell, S. (2002) A generalized mean field algorithm for variational inference in exponential families, *In Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*.Pp.583_591. Morgan Kaufmann Publishers Inc.

[37] Teh, Y. W., Newman, D., & Welling, M. (2007) A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation, *In Advances in neural information processing systems*, pp. 1353_1360.

[38] Minka, T., & Lafferty, J. (2002) Expectation-propagation for the generative aspect model, *In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp.352_359. Morgan Kaufmann Publishers Inc.

[39] Wainwright, M. J., & Jordan, M. I. (2005) A Variational Principle for Graphical Models.

[40] K. Bryan, P. Cunningham, and N. Bolshunov. (2006) Application of simulated annealing to the bio clustering of gene expression data. *IEEE Trans. Inf. Technol. Biomed*. vol. 10, no. 3: pp.519–525.

[41] Park, S., Cha, B., & An, D. (2010) Automatic multi-document summarization based on clustering and nonnegative matrix factorization, *IETE Technical Review*, 27(2): pp.167_178.

[42] S. Dudoit, J. Fridlyand, and T. P (2002) Speed: Comparison of discrimination methods for the classification of tumor using gene expression data, *J. Amer. Stat. Assoc*. vol. 97: pp.77_87.

[43] Geman, S. (1984) Gibbs distribution, and the Bayesian restoration of images, *IEEE Proc. Pattern Analysis and Machine Intelligence*, 6: pp.774_ 778.

[44] McAuliffe, J. D., & Blei, D. M. (2008) Supervised topic models, *In Advances in neural information processing systems*, pp.121_128.

[45] Blei, D. M., & Lafferty, J. D. (2007) A correlated topic model of science", The Annals of Applied Statistics, pp.17_35.

[46] Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009) Polylingual topic models, *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Volume2: pp.880_889.

[47] Boyd-Graber, J., & Blei, D. M. (2009) Multilingual topic models for unaligned text, *In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 75_82. AUAI Press.

[48] Liu, X., Duh, K., & Matsumoto, Y. (2015) Multilingual Topic Models for Bilingual Dictionary Extraction, *ACM Transactions on Asian and Low-Resource Language Information Processing*, 14(3), 11.

[49] Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing:* Volume 1, pp.248_256.Association for Computational Linguistics.

[50] Lacoste-Julien, S., Sha, F., & Jordan, M. I. (2009) Disc-LDA: Discriminative learning for dimensionality reduction and classification, *In Advances in neural information processing systems*.pp.897_904.

[51] Zhu, J., Ahmed, A., & Xing, E. P. (2012) Med-LDA: Maximum margin supervised topic models, *Journal of Machine Learning Research*, 13: pp.2237_2278.

[52] Mimno, D. (2008) Topic models conditioned on arbitrary features with Dirichlet-multinomial regression, *In Proceedings of 24th Conference on Uncertainty in Artificial 1 Intelligence*.

[53] Rabinovich, M., & Blei, D. (2014) The inverse regression topic model, *In International Conference on Machine Learning*, pp. 199_207.

[54] Aitchison, J. (1982) The statistical analysis of compositional data, *Journal of the Royal Statistical Society*, Series B, (Methodological): pp.139_177.

[55] Li, W., & McCallum, A. (2006) Pachinko allocation: DAG-structured mixture models of topic correlations, *In Proceedings of the 23rd international conference on Machine learning*, pp. 577_584, ACM

[56] Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (2004) Hierarchical topic models and the nested Chinese restaurant process, In *Advances in neural information processing systems*, pp.17_24.

[57] Wang, C., & Blei, D. M (2009) Variational inference for the nested Chinese restaurant process, *In Advances in Neural Information Processing Systems*, pp.1990_1998.

[58] Banerjee, A., Dhillon, I. S., Ghosh, J., & Sra, S. (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions, *Journal of Machine Learning Research*, 6(Sep): pp.1345_1382.

[59] Reisinger, J., Waters, A., Silverthorn, B., & Mooney, R. J. (2010) Spherical topic models, *In Proceedings of the 27th International conference on machine learning* (ICML-10): pp. 903_910.

[60] McCallum, A., & Nigam, K. (1998) A comparison of event models for naive Bayes text classification, *In AAAI-98 workshop on learning for text categorization*, Vol. 752, No. 1: pp.41_48.

[61] Banerjee, A., Dhillon, I. S., Ghosh, J., & Sra, S. (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions, *Journal of Machine Learning Research*, 6, pp.1345_1382.

[62] Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P (2004) The author-topic model for authors and documents, *In Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487_494.

[63] Holmes, D. I. (1998) The evolution of stylometry in humanities scholarship, *Literary and linguistic computing*, 13(3): pp.111_117.

[64] Girolami, M., & Kabán, A. (2004) Simplicial mixtures of Markov chains: Distributed modelling of dynamic user profiles, *In Advances in neural information processing systems*, pp. 9_16.

[65] Steyvers, M., & Griffiths, T. (2007) Probabilistic topic models, *Handbook of latent semantic analysis*, 427(7), pp.424_440.

[66] Padó, S., & Lapata, M. (2007) Dependency-based construction of semantic space models, *Computational Linguistics*, 33(2): pp.161_199.

[67] Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005) Integrating topics and syntax, *In Advances in neural information processing systems*, pp.537_544.

[68] Gruber, A., Weiss, Y., & Rosen-Zvi, M. (2007) Hidden topic markov models, *In Artificial intelligence and statistics*. pp. 163_170.

[69] Griffiths, T. L., & Steyvers, M. (2004) Finding scientific topics, *Proceedings of the National academy of Sciences*, 101, suppl 1: pp.5228_5235.

[70] Aletras, N., & Stevenson, M. (2014) Measuring the similarity between automatically generated topics, *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers. pp. 22_27.

[71] Wang, C., & Blei, D. M. (2011) Collaborative topic modeling for recommending scientific articles, *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 448_456.ACM.

[72] Paul, M., & Girju, R. (2009) Topic modeling of research fields: An interdisciplinary perspective, *In Proceedings of the International Conference RANLP-2009*, pp.337_342.

[73] Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016) An overview of topic modeling and its current applications in bioinformatics, *Springer Plus*, 5(1),1608.

[74] Wang, J., Hu, X., Tu, X., & He, T. (2012) Author-conference topic-connection model for academic network search, *Proceedings of the 21st ACM international conference on Information and knowledge management*. Pp.2179-2183.

[75] Zhao, W., Zou, W., & Chen, J. J. (2014) Topic modeling for cluster analysis of large biological and medical datasets, *In BMC bioinformatics* Vol. 15, No. 11: p. S11,

[76] Barnett, J., & Jaakkola, T. Research Abstracts-2007

[77] Bindra, A. (2012) Social-lda Scalable topic modeling in social networks, *University of Washington.*

[78] Cha, Y., & Cho, J. (2012) Social-network analysis using topic models, *In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 565_574.

[79] Grimmer, J. (2009) A Bayesian Hierarchical Topic Model for Political Texts, *Supplemental Appendix*.

[80] Naskar, D., Mokaddem, S., Rebollo, M., & Onaindia, E. (2016) Sentiment Analysis in Social Networks through Topic modeling, 2016. In LREC.

[81] Wang, J., Li, L., Tan, F., Zhu, Y., & Feng, W. (2015) Detecting hotspot information using multi-attribute-based topic mode, l. PloS one, 10(10.

[82] Kuang, X., Chae, H. S., Hughes, B., & Natriello, G. An LDA Topic Model and Social Network Analysis of a School Blogging Platform.

[83] Panichella, A., Dit, B., Oliveto, R., Di Penta, M., Poshynanyk, D., & De Lucia, A. (2013) How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms, *In Proceedings of Software Engineering (ICSE-2013) 35th International Conference on Software Engineering*. pp.522_531.

[84] Hindle, A., Godfrey, M. W., & Holt, R. C. (2009) What's hot and what's not: Windowed developer topic analysis, *In Software Maintenance, ICSM-2009.IEEE International Conference on Software Maintenance*, pp. 339_348.

[85] Yao, L., Mimno, D., & McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 937-946. ACM.

[86] Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models, *In Proceedings of the international working conference on advanced visual interfaces* pp. 74-77. ACM.

[87] Gretarsson, B., O'donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. (2012). Topic nets: Visual analysis of large text corpora with topic modelling, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2), 23.

[88] Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *In Proceedings of the workshop on interactive language learning, visualization, and interfaces*.pp. 63-70.

[89] Newman, D., Smyth, P., Welling, M., & Asuncion, A. U. (2008). Distributed inference for latent Dirichlet allocation. *In Advances in neural information processing systems*. pp. 1081-1088.

[90] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed Gibbs sampling for latent Dirichlet allocation, *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.pp. 569-577.

[91] Tristan, J. B., Tassarotti, J., & Steele, G. (2015). Efficient training of lda on a GPU by mean-for-mode estimation. *In International Conference on Machine Learning* pp. 59-68.

[92] Iwata, T., Yamada, T., & Ueda, N. (2009). Modeling social annotation data with content relevance using a topic model. *In Advances in Neural Information Processing Systems*. pp. 835-843.

[93] Bertoni, A., & Valentini, G. (2005, July). Random projections for assessing gene expression cluster stability. *In Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, 2005. (Vol. 1, pp.149-154.

[94] Pio, G., Ceci, M., Malerba, D., & D'Elia, D. (2015). ComiRNet: a web-based system for the analysis of

miRNA-gene regulatory networks. *BMC bioinformatics*, 16(9), S7.
[95] Strehl, A., & Ghosh, J. (2002). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec), pp.583-617.
[96] Greene, D., Cagney, G., Krogan, N., & Cunningham, P. (2008). Ensemble non-negative matrix factorization methods for clustering protein–protein interactions. *Bioinformatics*, 24(15), pp.1722-1728.
[97] Wallach, H. M., Murray, I., Salakhutdinov, R., Mimno, D. (2009, June). Evaluation methods for topic models. *In Proceedings of the 26th annual international conference on machine learning*.pp. 1105-1112.
[98] Mimno, D., & Blei, D. (2011) Bayesian checking for topic models, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 227-237. Association for Computational Linguistics.
[99] Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.pp. 530-539.
[100] Thomas W. Jones. (2019) TextmineR: Function for Text Mining & TopicModeling.R Package. https://cran.rproject.org/web/packages/textmineR/index.html
[101] Fridolin Wild.(2015)Latent Semantic Analysis(lsa): The R project for statistical package. https://cran.r-project.org/web/packages/lsa/index.html.
[102] Fritz Guenth.Applied Latent Semantic Analysis (LSA) Functions: The R project for statistical package. https://cran.rproject.org/web/packages/LSAfun/LSAfun.pdf
[103] Bettina Grun, Kurt Hornik.(2018)topic models: An R Package for Fitting Topic Models. https://cran.r-project.org/web/packages/topicmodels/index.html
[104] Dmitriy Selivanov(2018) . text2vec: Modern Text Mining Framework for R. https://cran.r-project.org/web/packages/text2vec/vignettes/text-vectorization.html
[105] Jonathan Chang.(2015)Latent Dirichlet Allocation: Collapsed Gibbs Sampling Methods for TopicModelshttps://cran.rproject.org/web/packages/lda/index.html
[106] McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.
[107] Radim Rehuk. Gensim: Topic modelling for humans .https://radimrehurek.com/gensim/
[108] https://nlp.stanford.edu/software/tmt/tmt-0.4/

Pooja Kherwa is an assistant professor of Maharaja Surajmal Institute of Technology, New Delhi. She received her M. Tech in information Technology from Guru Govind Singh Indraprastha University; Dwarka, New Delhi in 2010.Currently she is pursuing her PhD from Guru Govind Singh Indraprastha University, Dwarka- New Delhi. Her research interest includes Topic Modeling, Sentiment Analysis, Machine Learning.

Dr. Poonam Bansal is a Professor of Maharaja Surajmal Institute of Technology. Also working as Deputy Director of Institute. She has received her PhD from Guru Govind Singh Indraprastha University, Dwarka, New Delhi in 2010.Her area of interest includes Speech recognition, Data Mining, Machine learning