I.J. Intelligent Systems and Applications, **2019, 2, 70-82**
Published Online February 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2019.02.08

A Trend Analysis of Machine Learning Research with Topic Models and Mann-Kendall Test

Deepak Sharma[1]
[1]Department of Computer Engineering, Netaji Subash Institute of Technology,
Sector-3, Dwarka, New Delhi, 110078, India
E-mail: deepak.btg@gmail.com

Bijendra Kumar[1], **Satish Chand**[2]
[1]Department of Computer Engineering, Netaji Subash Institute of Technology,
Sector-3, Dwarka, New Delhi, 110078, India
[2]School of Computer & Systems Sciences, Jawaharlal Nehru University,
New Delhi, 110067, India
E-mail: bizender@gmail.com, schand20@gmail.com

Received: 06 March 2018; Accepted: 20 May 2018; Published: 08 February 2019

***Abstract*—**This paper aims to systematically examine the literature of machine learning for the period of 1968~2017 to identify and analyze the research trends. A list of journals from well-established publishers ScienceDirect, Springer, JMLR, IEEE (approximately 23,365 journal articles) related to machine learning is used to prepare a content collection. To the best of our information, it is the first effort to comprehend the trend analysis in machine learning research with topic models: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and LDA with Coherent Model (LDA_CM). The LDA_CM topic model gives the highest topic coherence amongst all topic models under consideration. This study provides a scientific ground that helps to overcome the subjectivity of collective opinion. The Mann-Kendall test is used to understand the trend of the topics. Our findings provide indicative of paradigmatic shifts in research methodology of significant patterns of topical prominence and the evolving research areas. It is used to highlight the evolution regarding the previous and recent trends in research topics in the area of machine learning. Understanding such an intellectual structure and future trends will assist the researchers to adopt the divergent developments of this research in one place. This paper analyzes the overall trends of the machine learning research since 1968, based on the latent topics identified in the period of 2007~2017 that may be helpful to the researchers exploring the recommended areas and publish their research articles.

***Index Terms*—**Latent Semantic Analysis, Latent Dirichlet Allocation, Coherence Model, Text Mining, Data Mining, Machine Learning, Trend Analysis.

I. Introduction

The name of Machine Learning firstly devised by Arthur Samuel, who was acknowledged for the checkers-playing program to improve game by game and studying which moves makeup winning strategies and incorporating those moves into the program. It is a subset of artificial intelligence area, and its methodology has unconventional in performance with the primary concerns of the field. The machine learning field has reincarnated many times in past and is known for its existence for many decades. The development of this area has been presented very well in [1-2]. For decades, there have always been challenges to researchers in artificial intelligence to build machines that can mimic the human intellect. The machine learning algorithms have motivated the researchers to empower a computer to autonomously drive cars, write and publish sport match reports, communicate with human beings and find the suspected terrorist. These algorithms are used unconventionally to obtain knowledge from the data. In machine learning, the computers don’t require to be explicitly programmed, but they can improve and change their algorithms by themselves. The machine learning systems automatically learn the program from data, which is a challenging task to make them manually. In the last couple of decades, the use of machine learning has spread rapidly in various disciplines [3]. Therefore, the algorithms in machine learning field are also known as algorithm about algorithms. In particular, the popularity of machine learning motivates us to understand the research trends in this field since the existing machine learning techniques have been applied to large-scale data processing environments or have extended to various application areas such as fraud detection, the stock market, weather forecasting, etc. Also, the algorithms changed according to newly emerged technology. So understanding the machine learning research themes of the past five decades will help to study the current machine learning trends and applies it to practical applications.

Copyright © 2019 MECS

The trends of machine learning research can be examined algorithmically or manually. The manual process offers an intuition into the articles, nonetheless it is not ever free from biases as researchers keep on persuaded in the direction of more cited articles [4]. The Natural language processing (NLP) offers a robust procedure that derives unseen trends from an extensive set of documents. In contrary manual tagging, is very extensive and needs expertise in the documents of subject-matter proficient, whereas the algorithmic based analysis is an automatic process [5, 6, 7] known as topic modeling. The detailed understanding of the topic modeling is available in some past surveys that include [8, 9, 10, 11, 12]. In survey paper [8, 11], presents a classification of directed probabilistic topic models and explains a broader view on graphical models. It may consider as an enormous initial point for the venture in the field of topic modeling. In [9] and [10] a preliminary discussion about topic modeling is presented. In paper [12] discusses the classification of probabilistic topic modeling algorithms and the journey of topic modeling techniques from its initial model till the advance models using deep learning.

II. Related Work

The topic model feeds a corpus, detects the topics and enhances the semantic meaning to the vocabulary. Both topic analysis and clustering methods can be utilized with topic modeling. Nonetheless, topic analysis is more suitable as compared to clustering for detection of trends in research articles of the dataset [13]. In a topic analysis, a document is distributed to a combination of topics, whereas in clustering, every article is prescribed to join exactly one cluster.

Topic analysis and labeling have been united to find the underlying topics and their trends in the text corpus. Latent Semantic Analysis (LSA) [14] and Latent Dirichlet Allocation (LDA) [15] are the two important topic modeling methods. Both of these methods have applied for trend analysis of Machine learning. Moreover, in addition to this, best LDA model has been evaluated with Coherence model (LDA_CM) for trend analysis. Amongst all the topic models, LDA_CM shows highest topic coherence amongst others topic models used in this work. The trends evaluated based on LSA, LDA, and LDA_CM topic model techniques. Also, finally Mann-Kendall test applied on topic models to understand the nature of trends as increasing, decreasing or no_trend for our work.

Similar work on various research areas has been performed using LSA, as in [16] LSA applied to understand the trend analysis of behavioral operation in supply chain management. In [17], LDA employed to understand the research trends and topics in software effort estimation. In [18], proposed a method for topic identification in web documents using web design features. In [19], opinion mining performed on online product reviews with LDA topic clusters using feature ontology tree and sentiwordnet. Also, in [20] Mann-Kendall test is used to understand the nature of trends in rainfall of different regions in India. Here, as per our knowledge first time, we are applying LDA_CM to compare the trends with LSA and LDA for trend analysis in Machine Learning along with Mann-Kendall test to understand the nature of trends.

This paper aims to analyze and understand the trends in Machine Learning research published in well-respected mainstream journals in past five decades, i.e., 1968~2017. The journals included as Journal of machine learning research (JMLR), Springer machine learning (Sp-ML), ScienceDirect pattern recognition (ScD-PR), ScienceDirect neural networks (ScD-NN), IEEE transactions on pattern analysis and machine intelligence (IEEE-PAMI), and IEEE transactions on neural networks (IEEE-NN), are used as primary data source in this work. The title and abstract of the published articles pre-processed before applying the data mining techniques. We have applied LSA, LDA and LDA_CM techniques to study the title and abstract of the articles and use the Mann-Kendall test to find the nature of trends in machine learning research.

Realizing this structure of scholarly work and future trends will help machine learning researchers adopt the distributed developments of this research in one place. This analysis will also be useful for planning particular issues in academia in projecting research directions, setting up themes for conferences in this field and publish an article in journals. The significant contribution of this work is summarized as follows:

i. The primary motivation of this work was to intellectualize the evolution of research topics in machine learning over a period of five decades, i.e., 1968~2017. This work allowed us to visualize and examine the development of research topics over the time.
ii. Evaluate the best topic model using topic coherence and finding the evolving trends with topic models such as LSA, LDA, and LDA_CM.
iii. This work proposed the overall trend analysis of machine learning researchers since 1968~2017 based on the latent topics identified in the period of 2007~2017.

This paper divided into six sections. Section III discusses the process of data collection and preprocessing steps. Section IV is an introductory discussion on the methodological analysis of topic modeling techniques, i.e., LSA, LDA, LDA_CM, and Mann Kendall test to get acquainted with the method. Section V discusses trend analysis setup for finding trends in machine learning. Results of this work addressed in Section VI. Finally, Section VII concludes our work.

III. Data Collection and Preprocessing

In this section, discusses the process of data preparation, description of the corpus, and data preprocessing corpus before feeding for training to topic *models.*

A. Data Preparation

The research data collected from various well-known journals published with high-quality research articles in machine learning. We include established journals like Journal of machine learning research (JMLR), Springer machine learning (Sp-ML), ScienceDirect pattern recognition (ScD-PR), ScienceDirect neural networks (ScD-NN), IEEE transactions on pattern analysis and machine intelligence (IEEE-PAMI) and IEEE transactions on neural networks (IEEE-NN).

Table 1. Number of Articles included in this study

S.No.	Journal Name	Duration	#Years	#Articles Published
1	IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE-PAMI)	1979~2017	39	5323
2	IEEE Transactions on Neural Network (IEEE-NN)	1990~2017	28	4349
3	Journal of Machine Learning Research (JMLR)	2000~2017	18	1755
4	Science Direct- Pattern Recognition (ScD-PR)	1968~2017	50	7314
5	Science Direct Neural Network (ScD-NN)	1988~2017	30	3294
6	Springer- Machine Learning (Sp-ML)	1986~2017	32	1330
Total				23365

Titles and abstracts of research papers considered from the electronic library of the mentioned journal articles. Only journal articles included for our work. The results are drawn from the time span of 50 years, i.e., from 1968 to 2017. Table 1 lists the number of articles included in our work according to the journals. The number of articles published in 2007~2017 and 1994~2006 is quite high compared with other decades. Each dataset considered as a separate corpus.

B. Description of Corpus

The corpus prepared by collecting articles. The corpus divided into four datasets. Table 2 lists the number of articles included in our work according to the corpus set. The first dataset (a.k.a. Dataset1), second dataset (a.k.a. Dataset2), third dataset (a.k.a. Dataset3), and fourth dataset (a.k.a. Dataset4) ranges from 1968~1980, 1981~1993, 1994~2006, and 2007~2017 respectively.

Table 2. Number of Articles in four datasets

S.No.	Corpus Set	Duration	#Years	#Articles Published
1	Dataset1	1968~1980	13	466
2	Dataset2	1981~1993	13	3113
3	Dataset3	1994~2006	13	8552
4	Dataset4	2007~2017	11	11234
Total				23365

C. Data Preprocessing

Preprocessing phase comprises the exclusion of noisy characters/ words from the text corpus. The following phases have followed for preprocessing the articles:

Lexical analysis. In this phase, abstracts and titles of the research articles are tokenized into tokens. Then, produced tokens transformed into lowercase letters for each document. The exclusion of punctuations, apostrophe, commas, quotation marks, exclamation points, question marks, and hyphen performed. Then, numeric values were eliminated to get only the textual tokens.

Stop-word removal. The nltk python package [21] included the standard English words as stop words and the phrases used to develop the literature dataset are removed.

Stemming. Porter Stemmer algorithm [22] was utilized to stem the tokens for each document and converted the inflected words to their base stem for preparing a useful literature dataset.

Converting documents to sparse vectors. The text files in corpus contain titles and abstracts of articles. Bag-of-words document representations used for converting documents to vectors. In this representation, each article represented by one vector where each vector element depicts a pair of word-wordcount. The mapping between the words and their word count is called a dictionary. Finally, sparse vectors are created counts merely the number of occurrences of each distinct word, converts the word to its integer word_id.

Transforming vectors to TfIdf vector. Converting articles from one vector representation into another serves two purposes; firstly, to bring out hidden structure in the corpus to discover relationships between words and describe the documents more semantically. Secondly, to make the document representation more compact. In this step, convert the sparse vectors of the corpus to TfIdf vectors using Eq.(1). As a formula for TfIdf weights of term *i* in document *j* in a corpus of D documents.

$$\text{weight}\{\text{i,j}\} = frequency_{\{i,j\}} * \log_2\left(\frac{D}{document_{freq_{\{i\}}}}\right) \quad (1)$$

Thus, the above steps used to transform corpus into vector representation for training topic models.

IV. Methodological Analysis

In this section, briefly discusses the LSA, LDA, LDA_CM and Mann-Kendall test as they used in our work. The topic modeling techniques are used to build an information retrieval system to get the similar documents from a given content collection. The Mann-Kendall test is

Copyright © 2019 MECS

utilized to find the nature of trends in the documents retrieved using the topic models.

A. Latent Semantic Analysis

The Latent Semantic Analysis introduced in the late 1980s for analyzing the relationship between the documents and terms [14]. Initially, it was used to improve the performance of the information retrieval systems for library indexing and search engine [23-27]. Gradually, it was adopted by researchers working in the area of psychology [28], and now, it is being used in many other areas such as artificial intelligence, cognitive sciences, education, information systems, etc. The basic idea of LSA is to extract the hidden knowledge from a set of text documents. It processes the text, also called document, from a set of files, also called corpus, and identifies the keywords, which are also called terms. Further, it helps to find the latent factors, also called topics, from these extracted terms. The mathematical foundation of LSA is discussed in [29-30]. A review of limitations and solution techniques of several challenges in LSA are presented in [13]. Each of the LSA concepts can be viewed as a function that accepts an *m*-dimensional vector as input and calculates a linear combination (weighted sum) of its coordinates, a scalar. With *k* LSA concepts, each input vector (i.e., document) is represented by a vector in $\mathbb{R}^{k}$. This transformation from the *TfIdf* vector space into the *k* latent ideas are realized through multiplication by a suitable matrix, which places LSA into the category of linear models. Finding the suitable projection matrix is done using the Singular Value Decomposition (SVD) algorithm. The SVD is applied to the *TfIdf* vector of the corpus that gives three matrices as shown in Eq. (2).

$$SVD(TfIdf) = U.S.V^{T} \tag{2}$$

It decomposes the *TfIdf* matrix of corpus into three matrices: *U*, *S*, *V*; where U refers to term eigenvectors; *V* relates to document eigenvectors; *S* refers to a diagonal matrix of singular values.

B. Latent Dirichlet Allocation

The LDA is applied to the corpus to facilitate retrieving and querying a large corpus of data to identify the latent ideas that describe the corpus as a whole [15]. In LDA, a document considered as a mixture of latent topics, and each term in the document related with one of these topics. Using the hidden clues, the topic model connects words having similar meaning and differentiates the words having different meaning [4, 31]. So, the latent topics signify multiple observed entities that have similar patterns identified from the corpus. The LDA is applied to pre-processed corpus data as discussed in [15, 32, 33]. It produces topic models based on the three input parameters, namely, some topics, hyper-parameters α, and β, and the number of iterations needed for the model to converge. The parameter α is the magnitude of the Dirichlet prior over the topic distribution of a document. This parameter is considered as some "pseudowords," divided evenly between all topics present in every document, no matter how the other words are allocated to topics. The parameter β is per-word-weight of the Dirichlet prior over topic-word distributions. The magnitude of the distribution (the sum over all words) is ascertained by the number of words in the vocabulary. The α and β hyper-parameters are smoothing parameters that change the distribution over the topics and words respectively, and initializing these parameters correctly can result in high-quality topic distribution. The value of α has been kept as *50/T*, where *T* is a number of topics; and β is fixed as 0.01 for all topic solutions.

C. Latent Dirichlet Allocation with Coherence Model

One of the problems with LDA is that if it is trained on a large number of topics, the topics get "lost" among the numbers. Since the LDA model is a probabilistic model, it provides different topics each time. To control the quality of the topic model to be produced, we can see what the interpretability of the best topic is and keep evaluating the topic model until the defined threshold is crossed. The coherence model is used to maintain a model for topic coherence. The topic coherence is a measure used to evaluate the topic models that automatically generate topics from a collection of documents, using the latent variable models [34]. It is an alternative measure of human interpretability of topics. The value of threshold (τ) has been kept as 0.75 for evaluating the best topic model. The algorithm for the LDA_CM model is given in Algorithm1.

```
Algorithm1
Input: Corpus as a corpus, input text for training as train_texts,
dictionary(train_texts) as a dictionary, the threshold as τ.
Output: Final evaluated topic model as lmodel.
topTopics ← an empty dictionary
while topTopics[0][1] less than τ do
   lmodel←LdaModel (corpus, dictionary)
   coherence_values← an empty dictionary
   for n, topic in lmodel.show_topics (num_topics=-1,
         formatted=False) do
            topic← list of top words in topics
            cmodel←CoherenceModel(topics=[topic], texts=train_texts,
                      dictionary=dictionary, window_size=10)
            coherence_values[n] ←cmodel.get_coherence ()
   end for
   topTopics←sort coherence_values in descending order
end while
return l
```

D. Mann-Kendall Test

In this work, the Mann-Kendall (MK) test is used to identify the significant trends of the similar documents retrieved (as discussed in 2.1) that generate a time series of the selected latent factor. This is a rank non-parametric test developed by Mann [35] and Kendall [36]. It is superior for detecting linear or non-linear trends as discussed in [37-38]. The adopted method used in our work is described as follows.

In this test, the null (H_0) and alternative hypotheses (H_1) are equal to the non-existence and existence of a trend in the time series of the observational data, respectively. The associated equations for computing the

Copyright © 2019 MECS

MK test statistic S and the standardized test statistic Z_{MK} are as given by Eq. (3) to Eq. (6).

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(X_j - X_i) \tag{3}$$

$$sign(X_j - X_i) = \begin{Bmatrix} +1 if (X_j - X_i) > 0 \\ 0 if (X_j - X_i) = 0 \\ -1 if (X_j - X_i) < 0 \end{Bmatrix} \tag{4}$$

$$Var(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p}^{q} t_p (t_p - 1)(2t_p + 5) \right] \tag{5}$$

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} if S < 0 \\ 0 if S = 0 \\ \frac{S-1}{\sqrt{Var(S)}} if S > 0 \end{cases} \tag{6}$$

where X_i and X_j are sequential data values of the time series in the years i and j, n is the length of time series; t_p is a number of ties for p^{th} value, and q is a number of tied values.

The positive values of Z_{MK} indicate increasing trends, and negative Z_{MK} values indicate decreasing trends in a time series. For $|Z_{MK}| > Z_{1-\alpha/2}$ the null hypothesis is rejected, and a significant trend exists in the time series. $Z_{1-\alpha/2}$ is the critical value of Z from the standard normal table. For *0.05* significant level, the value of $Z_{1-\alpha/2}$ is *1.96*. In next section, we discuss the trend analysis setup required for our work.

V. Trend Analysis Setup

In this section, discusses the process required for trend analysis. Next, evaluate the optimal number of topics, the python library is used for implementing the trend analysis and comparing the topic models utilizing the topic coherence. Identifying and labeling the latent factors, and finally plot the trend analysis graph.

A. Trend Analysis Process

We use the topic models on a machine learning corpus for the trend analysis as shown in Fig. 1. Initially, the corpus is prepared including the title and abstract of the articles, then preprocessing is done as discussed in Section III. In next step, the topic models are applied to the preprocessed data to identify the latent factors and labeled them as topics using the Nominal Group Technique (NGT). Next, the queried latent factor is transformed into query vector. The cosine similarity between the query vector and the trained topic model latent space is computed. The number of similar articles is mined whose similarity score is above 0.75. In next step, we have calculated the normalized frequency of the publications for each year in datasets. Then, the Mann-Kendall test is applied to identify the trends as increasing trend, decreasing trend or no_trend for the respective latent topic. Finally, the trend analysis graphs are plotted for the topic models.

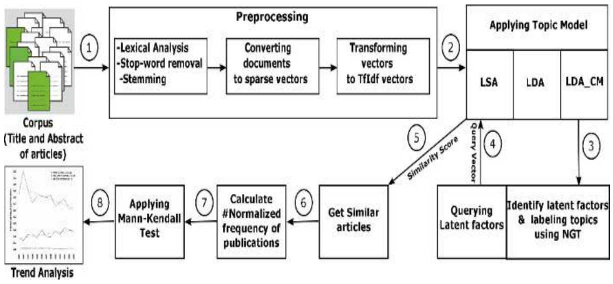


Fig.1. Trend Analysis Process

B. Evaluating Optimal Number of Topics

In an unstructured set of documents, where the numbers of appropriate trends are unknown in advance, and it is a difficult task to identify the optimal number of topics. The coarse topic model is generated if the number of topics is insufficient, whereas an excessive number of topics can result in a complex model, thus, making interpretation difficult [39]. There is no traditional measure to defend the optimal number of solutions. However, the topic coherence is applied to find the optimal range of topic solutions. The maximum coherence score leads to an optimal number of topics for the dataset. Fig.2., shows the optimal numbers of topics are 81, 59, 77 and 69 for each dataset. Based on these heuristics and findings of the study [40], the optimal number of topic solutions for identifying the search trends are chosen as eight as an optimal low-level solution as discussed in [41].

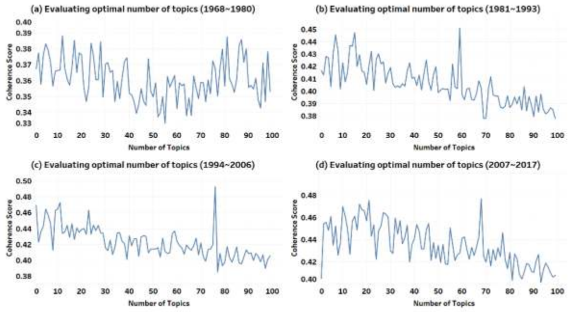


Fig.2(a-d). Evaluating the optimal number of topics for each dataset

C. Document Retrieval Using Gensim

The gensim package is a robust python library/framework for natural language processing. It is based on the idea of handling on substantial unstructured text corpora, document after document, in a memory-independent fashion. Also, it implements the Vector Space Model (VSM) algorithms [42] as LSI, LDA, and, Coherent Model. For experimental purpose, the Gensim as a pure python library is used for implementing the trend analysis and created by the idea of document streaming [43]. The document streaming has two objectives: the first one represents indexing of digital document and similarity search and the second one describes memory-efficient, fast and scalable algorithms for topic models for the unsupervised learning and semantic analysis of the plain text in digital collections. It requires the open source NumPy for n-dimensional array manipulation and SciPy for numerical integration and optimization. The advantages of Gensim are fast processing of large datasets and memory independence because the term-by-document matrix does not have to be stored in memory. Also, it enables the direct application of the topic models on a term-by-document matrix with term frequency weightings.

D. Topic Coherence

Topic coherence essentially measures the human interpretability of a topic model. Traditionally the perplexity has been used to evaluate the topic models; However, it does not correlate with human annotations at times. The topic coherence is another way to evaluate the topic models with a much higher guarantee on human interpretability [34]. In this work, the evaluation of the topic models: LSA, LDA, and LDA_CM have been performed regarding their corresponding coherence. The LDA_CM topic model produces the highest topic coherence as compared to LSA and LDA in each dataset as shown in Fig. 3. The highest topic coherence captures the excellent model with coherent topics.

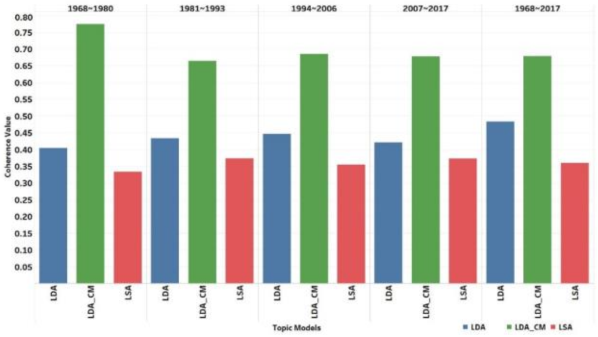


Fig.3. Comparing topic coherence for topic models

E. Identify and Labeling Latent Factors

The latent factors are identified by applying the topic models on the preprocessed corpus, and *TfIdf* vectors are used as input to train the topic models which in turn generate the latent factors. Eight latent factors are identified in each of four datasets for next step. A query model is built where these latent factors are used as a query to train the topic models and return the similar documents based on the query. While preparing the corpus, a unique name for each document in the corpus (such as 'SNo_ML_JournalName_Year_Month.txt,' e.g., '0001_ML_IEEEPAMI_1979_01.txt') format to create the document file names. The query result is arranged to get the count for similar articles published each year depending on the latent factor and then the topic trends are depicted for each dataset. The prime focus of this section is to understand the process of identifying latent factors. Based on the expert opinion, 32 topics (latent factors) are selected for trend analysis. Some of these topics are found redundant in this study, and these redundant topics help to analyze the trends in four datasets. To label, latent factors are highly subjective. The Nominal Group Technique (NGT) has been followed to accumulate expert opinion for labeling the latent factors [44]. Two senior researchers and two peer researchers working in machine learning are consulted for expert opinion. The latent terms in each topic have been selected through voting. The few redundant solutions have been kept to foresee the trends in the different period concerning the dataset of the corpus. Table 3 represents total thirty-two topics of machine learning research for our study.

F. Plotting the Trend Analysis Graph

The similar articles have retrieved by evaluating the cosine similarity between the latent factor as query vector and trained topic models latent space. These numbers of related articles are counted year wise for each dataset. Then, the normalized frequency of publication (*#nfp$_y$*) for each latent factor is calculated by dividing the count of similar articles for a year (*#sa$_y$*) to the total number of articles published in the same year (*#ta$_y$*) as shown in Eq. (7). E.g., the number of similar articles retrieved for the year 1968 is 3, and a total number of articles published in 1968 is 10, then the normalized frequency of publication is 0.3.

$$\#nfp_y = \frac{\#sa_y}{\#ta_y} \tag{7}$$

Where *y* is the year of publication. Similarly, *#nfp$_y$* for each latent factor is calculated for each year in four datasets and plot the trend graph for each latent factor between *#nfp$_y$* and the period for each dataset. Next section discusses the result of trend analysis in machine learning research.

Table 3. Labeling latent factors as topics in this study

TopicId	Topic Label	Key Latent Factors
1.a	Classification accuracy and efficient algorithm	accuraci algorithm classif comput cost degree effici error freedom object
1.b	Clustering techniques	biclust cluster correl density distribut fuzzy inform clos nearest neighbor
1.c	Edge detection and texture feature	boundari detect edg extract feature imag line object shape texture
1.d	Factor Analysis	analysi compon dimens exploratori factor independ linear discrimin princip reduct
1.e	Feature Engineering	combin correl differ encod extract featur linear select space vector
1.f	Genetic, EM and gradient descent algorithm	algorithm converg cost descent expect genet gradient maxim mutat paramet
1.g	Maximum likelihood estimation and local minima	coeffici convex distribut estim function likelihood local maxim minima paramet
1.h	Pattern classification and recognition	algorithm classif knowledge label learn match model pattern recognit regular
2.a	Artificial neural network	activ artifici back hidden multilay network neuron perceptron propag threshold
2.b	Character and object recognition	charact extract featur global imag object optic process recognit segment
2.c	Computer vision and simulation	comput detect diffract edg environ model simul synthet system vision
2.d	Hidden markov model and markov random field	chain distribut factor field graphi hidden markov model random undirect
2.e	Image analysis	3D analysi databas imag process recognit represent retriev segment sequenc
2.f	Signal processing and speech recognition	frequenc intens process recognit represent signal speech transform vector window
2.g	Supervised, unsupervised and reinforcement learning	action algorithm associ classif data label learn reinforce supervis unsupervis
2.h	Support vector machine and kernel method	binary boundari classif hyperplan kernel linear machin method support vector
3.a	Artificial neural network	activ artifici back hidden multilay network neuron perceptron propag threshold
3.b	Bayesian network	bayesian condit distribut graphic independ knowledg model network prior probabilist
3.c	Character and object recognition	charact detect extract featur global imag object optic recognit segment
3.d	Clustering techniques	clos cluster correl density distribut fuzzy inform k-mean nearest neighbor
3.e	Digital image processing	color digit feature imag inform mutual process project segment self-organ
3.f	Face detection, recognition, and expression	analysi detect eigen express face frame pattern recognit templat vector
3.g	Linear and non-linear regression analysis	analysi curv fit function linear model polynomi predict regress transform
3.h	Support vector machine and kernel method	binary boundari classif hyperplan kernel linear machin method support vector
4.a	Bayesian Network	bayesian condit distribut graphic independ knowledg model network prior probabilist
4.b	Deep learning and artificial neural network	artifici contrast convolut diverg multilay neuron perceptron propag recur restrict
4.c	Computer vision and image analysis	analysi detect edg extract feature imag object recognit shape vision
4.d	Support vector machine and kernel method	binary boundari classif hyperplan kernel linear machin method support vector
4.e	High dimensional data	data dimens distribut high larg manifold project quantiz reduct scale
4.f	Linear and non-linear regression analysis	analysi curv fit function linear model polynomi predict regress transform
4.g	Statistical learning and pattern recognition	estim learn model nonparametr pattern probabl recognit space statist theory
4.h	Supervised, unsupervised and reinforcement learning	action algorithm associ classif data label learn reinforce supervis unsupervis

VI. Results and Discussion

The result presented in forms of tables and graphs to enhance the readability. Firstly, the outcome of LSA, LDA and LDA_CM topic models identified as the eight topics (or latent factors) for each dataset. Then, representation of the trends in topics of machine learning research is discussed based on the latent factors for each dataset. Finally, representing the trends of latest latent factors identified in dataset4 over last five decades. The new remarks present in this work are an estimate of a scholarly system view of machine learning.

A. *Trend Analysis of Topics in Each Dataset*

In this section, we discuss the trends analysis of topics in each dataset. Fig. 4 to Fig. 11 depict the topics trend of each dataset for their respective period. The results provide a holistic view of a machine learning researcher who has just started venturing into this area. The title and abstract of research articles have been considered for generating the trends that summarize the full content and thus provide a complete understanding of the topic discussed in a particular research article. Tableau has been used as data visualization tool for preparing the graphs in this work. The eight factors from each of the four different datasets have been considered to analyze the machine learning topics. The trends generated through each topic model for the latent factors have been compared. The graphs represent each topic with its trend as increasing trend, decreasing trend, or no trend. The nature of patterns has been evaluated using the Mann-

Kendall test as discussed in Section IV.D. A few expected comments include the decline of research areas which were prime foci in the early decades.

Fig. 4 and Fig. 5 depict the topics trend of dataset1 for the period of 1968~1980. During this time, the initial growth of the topics was slow, and there was a significant increase in research in later years. The term classification accuracy has always been a concern for researchers. The LDA and LDA_CM show no trends, whereas the LSA shows that the research has increased to improve the efficiency of the algorithms in later years. Similarly, for clustering and edge detection topics, the LDA and LDA_CM show no trends, whereas the LSA shows the increase in trends in the following years. The contribution by a researcher in factor analysis has been throughout the time interval. The LDA shows the decreasing trends, whereas the LDA_CM and LSA show no trends for the factor analysis. Also, the LSA has been observed to be overfitting when the retrieved irrelevant articles result from the normalized frequency of publication beyond one. No trends are seen by the topic models for feature engineering topic. The LDA and LDA_CM show no trend for algorithms and maximum-likelihood topics, whereas the LSA shows an increase in trends, a gradual rise as a result of significant improvement in this area. Finally, the pattern classification and recognition topic trend have decreased drastically for LSA, and no trends shown by LDA and LDA_CM topic models.

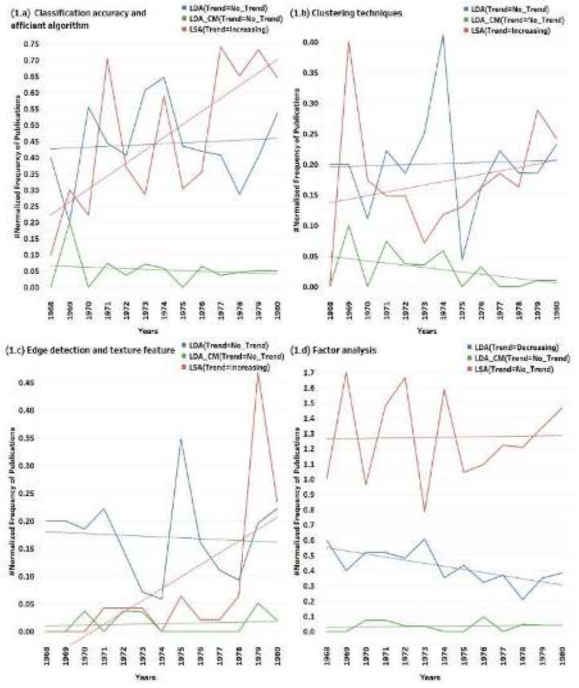


Fig.4. Trend analysis of topic for the period 1968~1980 (1.a to 1.d)

Fig. 6 and Fig. 7 depict the topics trend of dataset2 for the period of 1981~1993. The artificial neural network was well known since 1950, whereas the trend has rapidly increased for all topic models after 1987 due to the advent of backpropagation learning algorithm. Only the LDA has shown the increasing trend for character and object recognition topic. Similarly, the LDA_CM has showed the growing trend in computer vision topic. No trends have been observed by topic models for image analysis topic in this time interval. The LSA has shown the increasing trend for hidden Markov and supervised learning topics, whereas no trend has been demonstrated by the LDA and LDA_CM. Each topic model has shown different trends for signal processing and support vector topics. The LDA shows an increasing trend, LSA shows a decreasing trend and no trend has been demonstrated by LDA_CM.

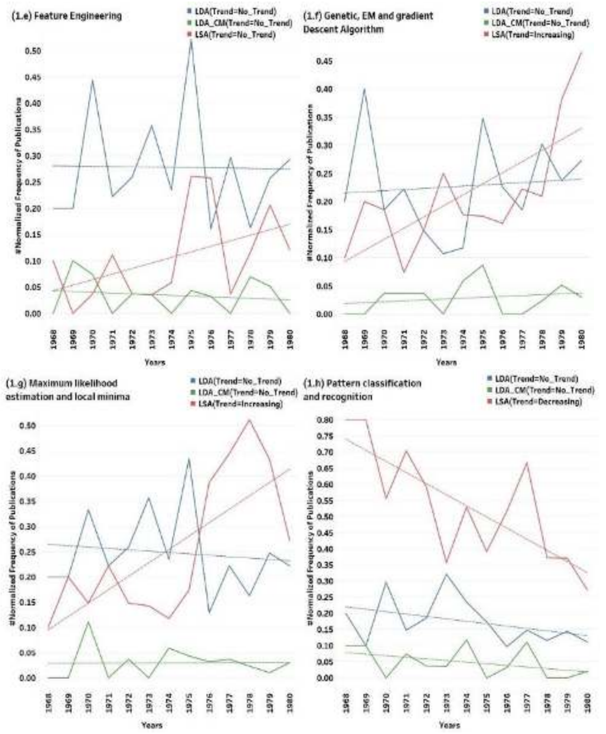


Fig.5. Trend analysis of topic for the period 1968~1980 (1.e to 1.h)

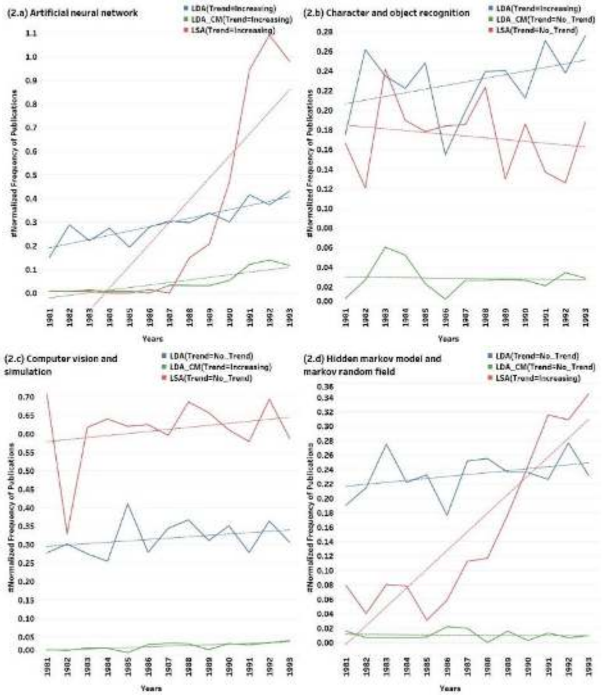


Fig.6. Trend analysis of topic for the period 1981~1993 (2.a to 2.d)

Copyright © 2019 MECS

The trend of the Bayesian network has shown an increase by LDA and LSA due to the probabilistic approach adopted by the researchers that have helped them to uncover the solutions to research problems. Only the LSA has shown the increasing trend for digital image processing topic. The face recognition topic trend was also shown to increase significantly by LDA_CM and LSA due to high potential in numerous commercial and government applications. The image preprocessing can extensively improve the reliability of optical inspection. Multiple filter techniques that intensify or reduce specific image information allow easier or much faster evaluation. Image pre-processing increases the chances of correct matching and decreases the processing time. Similarly, the LDA_CM and LSA show the increasing trend for the clustering techniques.

The trend of linear and non-linear regression analysis topics has gradually increased due to its suite well for predictive modeling which uncovers the relationship between two or more variables. The trend of support vector machine has rapidly grown during this time interval; thus, resulting in a better alternative to the artificial neural networks for researchers.

Fig. 10 and Fig. 11 depict the topics trend of dataset4 for the period of 2007~2017. Topic models for the Bayesian network have shown no trend during this time interval due to the significantly fewer usage of probabilistic models by the researchers. Similarly, the high dimensional data topics have shown no trends. Surprisingly, no trend has demonstrated for computer vision topic in spite of very highly demandable research topic. The LSA has observed the overfitting for the retrieved irrelevant articles resulted in the normalized frequency of publication beyond one of these topics. Inevitably, these trends motivate the researchers to dive into the area of computer vision and high dimensional data. Also, the researchers have designed parallel algorithms that can efficiently execute and extract the potential information from these high dimensional data.

The trend of an artificial network has shown significantly increase by the LDA_CM during this period due to the deep learning models as a consequence of the availability of highly computation machines using the Graphical Processing Units (GPUs). The LDA_CM has shown the decreasing trend for the support vector machine and linear regression topics in this time interval. No trend has demonstrated by the topic models for the statistical learning topic. Earlier, the statistical model enabled the researchers to look at a broad set of data and condense into the meaningful information that has become the heart of most of the machine learning. The machine learning model and inferences have been derived using a statistical approach. Now, due to the advent of deep learning models, no prior assumptions are required about the underlying relationships between the variables. Lastly, only the LSA has shown the increasing trend for the supervised, unsupervised or reinforcement learning topics.

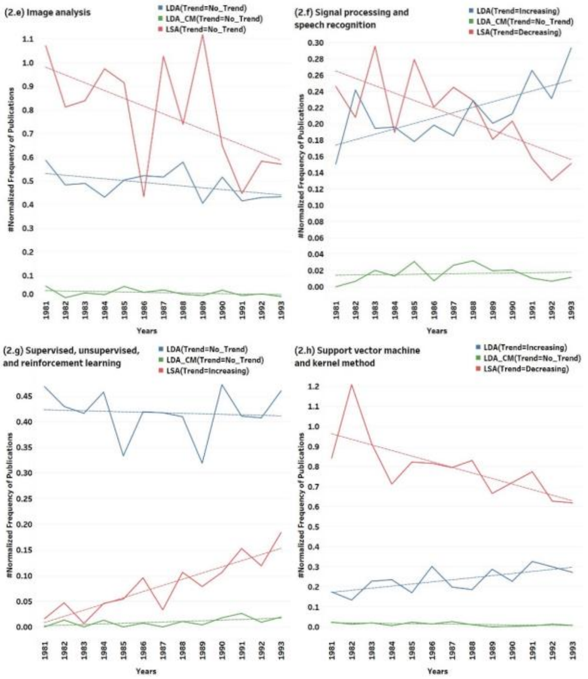


Fig.7. Trend analysis of topic for the period 1981~1993 (2.e to 2.h)

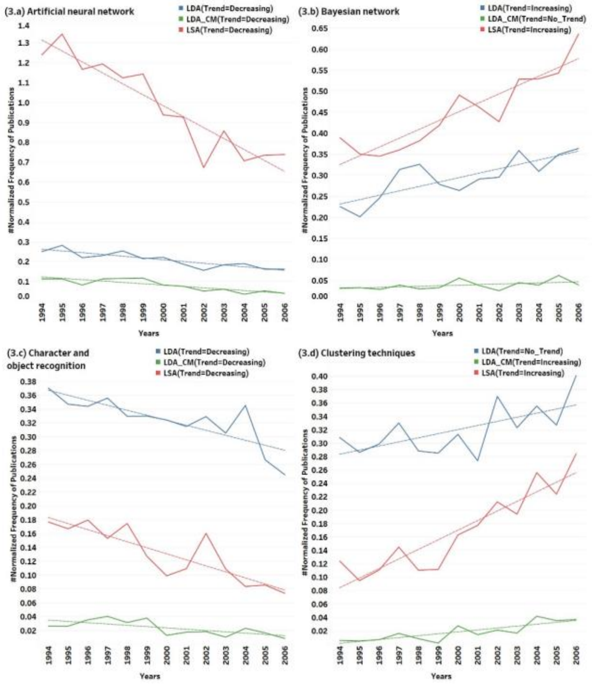


Fig.8. Trend analysis of topic for the period 1994~2006 (3.a to 3.d)

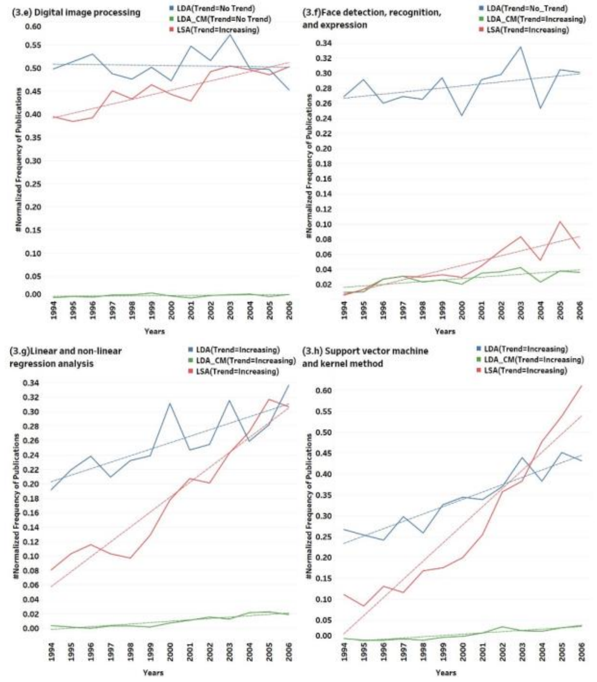


Fig.9. Trend analysis of topic for the period 1994~2006 (3.e to 3.h)

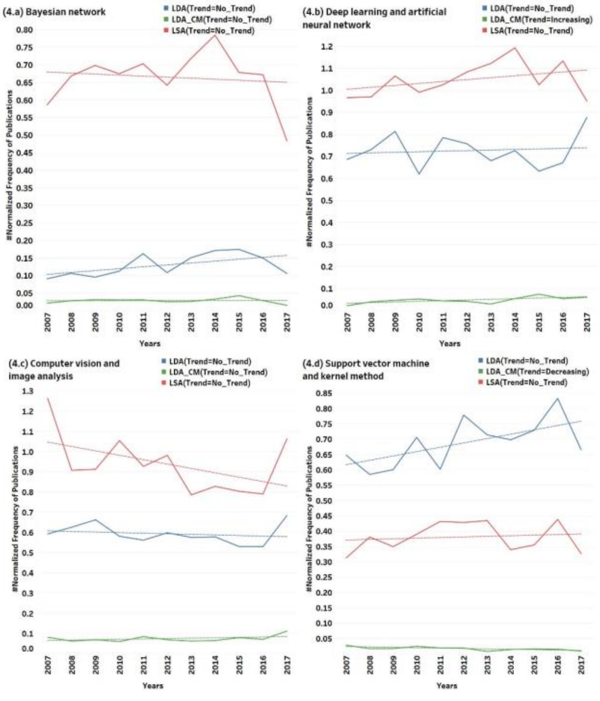
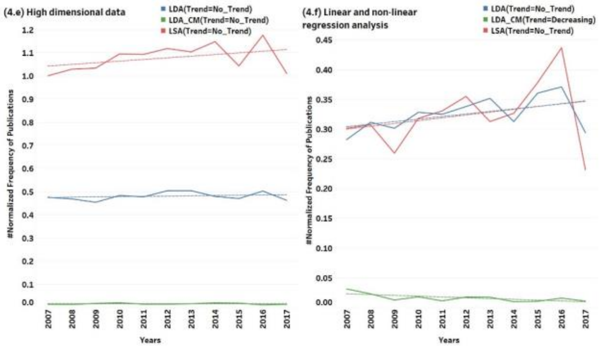


Fig.10. Trend analysis of topic for the period 2007~2017 (4.a to 4.d)



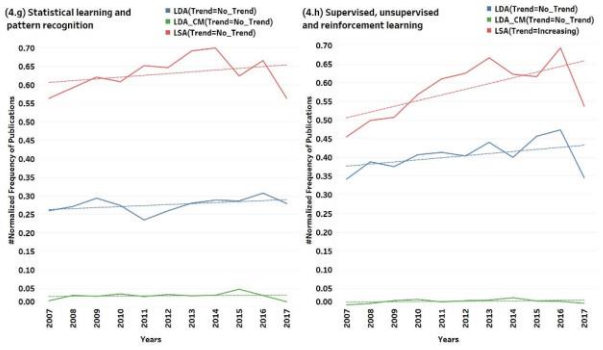


Fig.11. Trend analysis of topic for the period 2007~2017 (4.e to 4.h)

B. Overall Trend Analysis of Latest Topics Over Five Decades

In this subsection, we have discussed the assembled analysis of trends over five decades. Fig. 12 and Fig. 13 have depicted the overall topics trend based on the latest topics identified from the dataset4 for the period of 1968~2017. This kind of analysis helps the future researchers to understand the aggregate trends of the most recent topics. The trend of the Bayesian network has shown a gradual increase in all the topic models, and the similar trend has found for statistical learning. Both of these topics have become a vital tool in Bayesian methods and probabilistic modeling that has given rise to the probabilistic graphical models. The trends of deep learning and the artificial neural network are showing a rapid increase and have demonstrated low demand mostly in early two decades, and then it has gradually increased in rest of the period. Currently, the trends in deep learning have exhibited it as the cutting edge technology for most of the research problems.

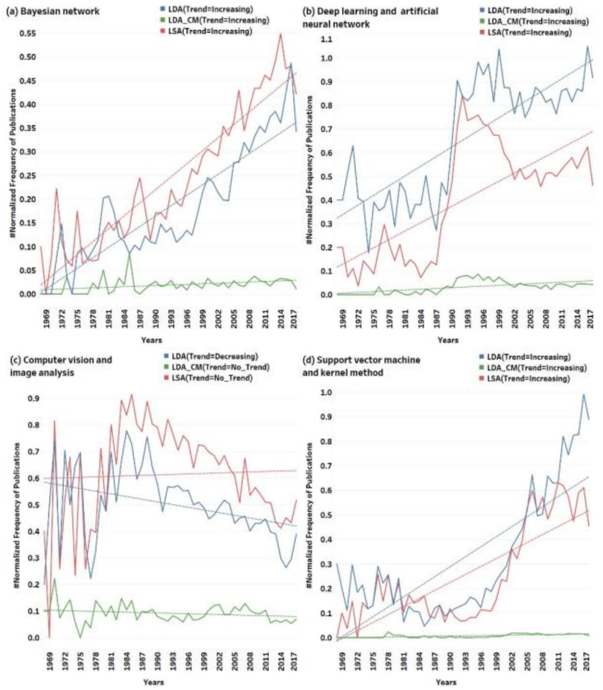


Fig.12. Trend analysis of topic for the period 1968~2017 (4.a to 4.d)

Also, the LDA has shown the decreasing trend in computer vision, whereas the LDA_CM and LSA have shown no trend. The contribution in this field is gradually increasing with time, but this creates a vital opportunity to research community to explore this area. Also, the combination of deep learning with computer vision is another exciting area to study by the future researchers. Similarly, the trend in support vector is increasing rapidly, but it is evident from the graph that the support vector had outperformed when the artificial neural network topic was facing a lot of challenges. The trend of regression analysis has been a slow start in first two decades and increased gradually later due to its application in the business domain. Regression analysis is considered the best tool for predictive modeling in business or the stock market. High dimensionality or large-scale data has always been a concern for researchers due to the availability of the structured as well as unstructured data. Its trend is increasing, and newer techniques have developed using the parallel processing. Lastly, the trends in supervised learning have grown rapidly during the second half of the period due to more and more availability of datasets and its applicability to applied research problems.

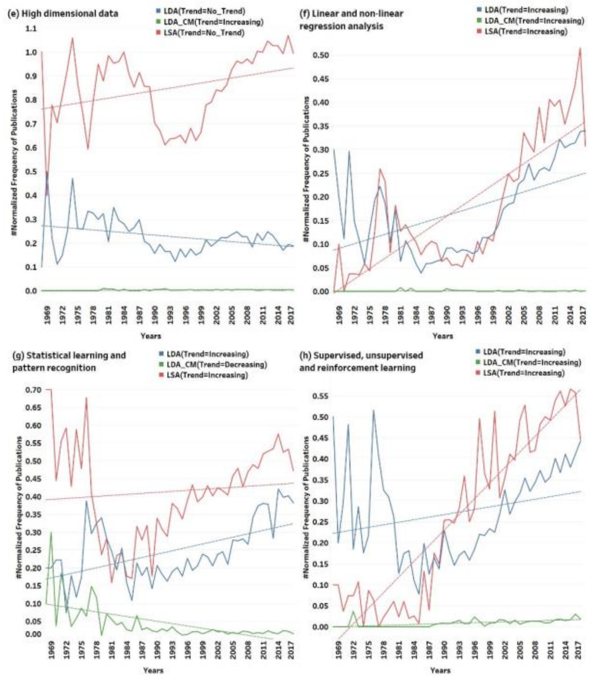


Fig.13. Trend analysis of topic for the period 1968~2017 (4.e to 4.h)

In summary, the trends of topics identified in each dataset have analyzed. Also, the overall trends of the latest topics for the period of 2007~2017 examined over five decades. This analysis can motivate the future researchers to understand the trends of the machine learning topics and give them the opportunity to explore further.

VII. Conclusion

In this work, we have carried out trend analysis on the machine learning research done over last five decades. A long time horizon (i.e., 50 years) can be considered a highly expanded domain of study. The topic modeling has used the knowledge extraction methodology. The documents are extracted by using the latent factors as a query to the topic models. Using this text analysis method, we combined the statistical methods and expert judgment to extract the knowledge in the form of significant latent factors. These latent factors may be considered as the footprints of development in this particular domain across the journey from traditional machine learning to modern machine learning, i.e., computer vision, deep learning, data analytics, etc. This work is more oriented towards analyzing the trends in machine learning research based on the four datasets of the corpus and on latest latent factors. From the results we can infer that the overall trends of deep learning showed gradual increase since its inception. Also, the overall trend of computer vision depicted no trends by topic models. In summary, we can see that the machine learning research will open a wide range of opportunities for future researchers and data scientists. This work provides an approach for identifying the rise and fall of research trends in machine learning. The future research aims at building a web-based application where the interested researchers who are newly venturing into this field can run the model to understand the effectiveness of the trend analysis.

References

[1] Carbonell, Jaime G., Ryszard S. Michalski, and Tom M. Mitchell, "Machine learning: a historical and methodological analysis," AI Magazine 4.3 (1983): 69.

[2] Marr, Bernard, "A Short History of Machine Learning—Every Manager Should Read," Forbes. http://tinyurl.com/gslvr6k (2016).

[3] Domingos, Pedro, "A few useful things to know about machine learning," Communications of the ACM 55.10 (2012): 78-87.

[4] Yalcinkaya, Mehmet, and Vishal Singh, "Patterns and trends in building information modeling (BIM) research: a latent semantic analysis," Automation in Construction 59 (2015): 68-80.

[5] Campbell, Joshua Charles, Abram Hindle, and Eleni Stroulia, "Latent Dirichlet allocation: extracting topics from software engineering data," The art and science of analyzing software data. 2016. 139-159.

[6] Canini, Kevin, Lei Shi, and Thomas Griffiths, "Online inference of topics with latent Dirichlet allocation," Artificial Intelligence and Statistics. 2009.

[7] Saini, Shubham, Bhavesh Kasliwal, and Shraey Bhatia, "Language identification using g-lda," International Journal of Research in Engineering and Technology (2013).

[8] Daud, Ali, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. "Knowledge discovery through directed probabilistic topic models: a survey." Frontiers of computer science in China 4, no. 2 (2010): 280-301.

[9] Blei, David M. "Probabilistic topic models." Communications of the ACM 55, no. 4 (2012): 77-84.

[10] Steyvers, Mark, and Tom Griffiths. "Probabilistic topic models." Handbook of latent semantic analysis 427, no. 7 (2007): 424-440.

[11] Jelisavcic, V., Furlan, B., Protic, J., & Milutinovic, V. M.,

"Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers", 35th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'2012), 1030–1035.

[12] Sharma, Deepak, Bijendra Kumar, and Satish Chand. "A Survey on Journey of Topic Modeling Techniques from SVD to Deep Learning." International Journal of Modern Education and Computer Science 9.7 (2017): 50.

[13] Evangelopoulos, Nicholas, Xiaoni Zhang, and Victor R. Prybutok, "Latent semantic analysis: five methodological recommendations," European Journal of Information Systems21.1 (2012): 70-86.

[14] Deerwester, Scott, et al., "Indexing by latent semantic analysis," Journal of the American society for information science 41.6 (1990): 391.

[15] Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research3.Jan (2003): 993-1022.

[16] Kundu, Anirban, et al., "A journey from normative to behavioral operations in supply chain management: A review using Latent Semantic Analysis," Expert Systems with Applications42.2 (2015): 796-809.

[17] Sehra, Sumeet Kaur, et al., "Research patterns and trends in software effort estimation," Information and Software Technology 91 (2017): 1-21.

[18] Taghandiki, Kazem, Ahmad Zaeri, and Amirreza Shirani. "A Supervised Approach for Automatic Web Documents Topic Extraction Using Well-Known Web Design Features." International Journal of Modern Education and Computer Science 8.11 (2016): 20.

[19] Santosh, D. Teja, et al. "Opinion mining of online product reviews from traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet." IJEME 6 (2016): 1-11.

[20] Mondal, Arun, Sananda Kundu, and Anirban Mukhopadhyay, "Rainfall trend analysis by Mann-Kendall test: A case study of north-eastern part of Cuttack district, Orissa," International Journal of Geology, Earth and Environmental Sciences 2.1 (2012): 70-78.

[21] Bird, Steven, and Edward Loper, "NLTK: the natural language toolkit," Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004.

[22] Porter, Martin F, "An algorithm for suffix stripping," Program14.3 (1980): 130-137.

[23] Cios, Krzysztof J., et al., "Data mining: a knowledge discovery approach," Springer Science & Business Media, 2007.

[24] Dumais, Susan T., "Latent semantic analysis," Annual review of information science and technology 38.1 (2004): 188-230.

[25] Dumais, Susan T, "LSA and information retrieval: Getting back to basics," Handbook of latent semantic analysis 293 (2007): 322.

[26] Han, Jiawei, Jian Pei, and Micheline Kamber, "Data mining: concepts and techniques," Elsevier, 2011.

[27] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze, "Introduction to information retrieval," Vol. 1. No. 1. Cambridge: Cambridge university press, 2008.

[28] Landauer, Thomas K, "LSA as a theory of meaning," Handbook of latent semantic analysis 6 (2007): 3-34.

[29] Martin, Dian I., and Michael W. Berry, "Mathematical foundations behind latent semantic analysis," Handbook of latent semantic analysis (2007): 35-56.

[30] Valle-Lisboa, Juan C., and Eduardo Mizraji, "The uncovering of hidden structures by latent semantic analysis," Information sciences 177.19 (2007): 4122-4147.

[31] Steyvers, Mark, and Tom Griffiths, "Probabilistic topic models," Handbook of latent semantic analysis 427.7 (2007): 424-440.

[32] Mavridis, Themistoklis, and Andreas L. Symeonidis, "Semantic analysis of web documents for the generation of optimal content," Engineering Applications of Artificial Intelligence 35 (2014): 114-130.

[33] Alghamdi, Rubayyi, and Khalid Alfalqi, "A survey of topic modeling in text mining," Int. J. Adv. Comput. Sci. Appl. (IJACSA) 6.1 (2015).

[34] Röder, Michael, Andreas Both, and Alexander Hinneburg, "Exploring the space of topic coherence measures," Proceedings of the eighth ACM international conference on Web search and data mining. ACM, 2015.

[35] Mann, Henry B, "Nonparametric tests against trend," Econometrica: Journal of the Econometric Society (1945): 245-259.

[36] Mg, Kendall, "Rank correlation methods," London: Charles Griffin 35 (1975).

[37] Hisdal, Hege, et al., "Have streamflow droughts in Europe become more severe or frequent?," International Journal of Climatology 21.3 (2001): 317-333.

[38] Wu, Hong, et al., "Trend analysis of streamflow drought events in Nebraska," Water Resources Management 22.2 (2008): 145-164.

[39] Zhao, Weizhong, et al., "A heuristic approach to determine an appropriate number of topics in topic modeling," BMC bioinformatics. Vol. 16. No. 13. BioMed Central, 2015.

[40] Bradford, Roger B., "An empirical study of required dimensionality for large-scale latent semantic indexing applications," Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

[41] Sidorova, Anna, et al., "Uncovering the intellectual core of the information systems discipline," Mis Quarterly (2008): 467-482.

[42] Salton, Gerard, Anita Wong, and Chung-Shu Yang, "A vector space model for automatic indexing," Communications of the ACM 18.11 (1975): 613-620.

[43] Rehurek, Radim, and Petr Sojka, "Software framework for topic modeling with large corpora," In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010.

[44] Delp, P., et al., "Delphi: System tools for project planning," Columbus, OH: National Center for Research in Vocational Education, Ohio State University (1977): 45-56.

Authors' Profiles

Deepak Sharma has received his B.E. (in Computer Engineering) and M.Tech. (in Information Technology) from Bharati Vidyapeeth College of Engineering, Bharati Vidyapeeth University, Pune respectively. Currently, he is pursuing Ph.D. in Topic Modeling and Trend Analysis from Department of Computer Engineering, Netaji Subhas Institute of Technology, New Delhi, India. His research interest in data mining, natural language processing, text mining, topic modeling.

Bijendra Kumar did his Bachelor of Engineering from H.B.T.I. Kanpur, India. He has done his Ph.D. from Delhi University, Delhi, India in 2011. Presently he is working as a Professor in Computer Engineering Division, Netaji Subhas Institute of Technology, Delhi, India. His areas of research interests are text mining video applications, watermarking, the design of algorithms, cloud computing.

Satish Chand did his M.Sc. in Mathematics from Indian Institute of Technology, Kanpur, India, and M.Tech. In Computer Science from Indian Institute of Technology, Kharagpur, India, and Ph.D. from Jawaharlal Nehru University, New Delhi, India. Presently he is working as a Professor in School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi, India. Areas of his research interest are text mining, trend analysis, multimedia broadcasting, networking, video-on-Demand, cryptography, and image processing.

How to cite this paper: Deepak Sharma, Bijendra Kumar, Satish Chand, "A Trend Analysis of Machine Learning Research with Topic Models and Mann-Kendall Test", International Journal of Intelligent Systems and Applications(IJISA), Vol.11, No.2, pp.70-82, 2019. DOI: 10.5815/ijisa.2019.02.08