SN Computer Science (2021) 2:469
https://doi.org/10.1007/s42979-021-00876-2

ORIGINAL RESEARCH

A Trend Analysis of Significant Topics Over Time in Machine Learning Research

Deepak Sharma[1] · Bijendra Kumar[1] · Satish Chand[2] · Rajiv Ratn Shah[3]

Received: 26 July 2020 / Accepted: 13 September 2021 / Published online: 22 September 2021

Abstract
A vast number of research papers on numerous topics publish every year in different conferences and journals. Thus, it is difficult for new researchers to identify research problems and topics manually, which research community is currently focusing on. Since such research problems and topics help researchers to be updated with new topics in research, it is essential to know trends in research based on topic significance over time. Therefore, in this paper, we propose a method to identify the trends in machine learning research based on significant topics over time automatically. Specifically, we apply a topic coherence model with latent Dirichlet allocation (LDA) to evaluate the optimal number of topics and significant topics for a dataset. The LDA model results in topic proportion over documents where each topic has its probability (i.e., topic weight) related to each document. Subsequently, the topic weights are processed to compute average topic weights per year, trend analysis using rolling mean, topic prevalence per year, and topic proportion per journal title. To evaluate our method, we prepare a new dataset comprising of 21,906 scientific research articles from top six journals in the area of machine learning published from 1988 to 2017. Extensive experimental results on the dataset demonstrate that our technique is efficient, and can help upcoming researchers to explore the research trends and topics in different research areas, say machine learning.

Keywords Research trend analysis · Information retrieval · Machine learning · Latent Dirichlet allocation

Introduction

Recently, the vast number of scientific papers are published very rapidly and it is tiresome for the researchers to become streamlined with the state-of-the-art research area [27]. As a result of increasing scientific papers, there would be growing opportunity of algorithms and tools are essential to match the consistently increasing rate of the scientific output [8]. Algorithms and tools can support in examining huge collections of document in structured and alternative advanced techniques in as compared to traditional techniques. Because the conventional keyword searches cannot always detect the themes and the main concept within the articles which can be shared among similar articles [37]. The themes (a.k.a. topics) in the articles uncovered by applying the unsupervised algorithms are called as topic models [5, 6, 15, 25]. The themes are also known as thematic or latent structures from the vast collection of documents. These themes are naturally arising from the probabilistic characteristic of the collection of documents, and per-se no earlier annotation or labeling is necessary. As a consequence, the thematic structures can be used to systematically classify or summarize documents up to an extent that would be inconceivable to do manually. In [20, 21, 34], topic modeling algorithms have confirmed to be very beneficial in clarifying the major concepts within a set of documents and the algorithms are fast as compared to conventional review methodology.

Blei et al. [6] proposed latent Dirichlet allocation (LDA) as one of the most known and highly researched topic models. LDA is a generative probabilistic topic model that reduces the limitations of other well-known topic model algorithms such as latent semantic indexing (LSI) proposed by [15] and probabilistic latent semantic indexing (pLSI)

✉ Deepak Sharma
deepak.btg@gmail.com

1 Department of Computer Engineering, Netaji Subash Institute of Technology, University of Delhi, Sector-3, Dwarka, New Delhi 110078, India

2 School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi 110067, India

3 Multimodal Digital Media Analysis Lab, Indraprastha Institute of Information Technology, New Delhi 110020, India

proposed by [25]. In LDA, documents as multinomial distributions over *k* latent topics and each topic is modeled as a multinomial distribution over the fixed vocabulary. As such, LDA captures the heterogeneity of research topics or ideas within scientific publications and can be viewed as a mixed membership model in [17].

In this work, the LDA topic model is employed on machine learning articles published in well-respected mainstream journals in the past three decades, i.e., 1988–2017. Machine learning is an interdisciplinary area of research in various research areas such as statistics, artificial intelligence, and databases and has been studied steadily since the word 'machine learning' was coined by Arthur Samuel. Actually, since the last three decades, existing machine learning techniques have been applied to large-scale environments for data processing or have been extended to numerous application areas such as stock market, fraud detection, weather forecasting, etc. Furthermore, the algorithm is reformed according to newly emerged technology. So understanding the machine learning research themes of the past three decades will help upcoming researchers in studying current machine learning trends and applying it to practical applications. The machine learning field has revived many times and is acknowledged for its existence for many decades. The progress of machine learning has been presented very well in [13, 28]. For many years, the researchers in artificial intelligence are facing challenges in building systems that can imitate the intelligence like humans. The researchers are inspired to apply machine learning algorithms to enable a computer to communicate with human beings, write and publish sport match reports, locate the suspected terrorist, and autonomously drive cars. These machine learning algorithms are used typically to acquire information 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 as discussed in [16]. Notably, the admiration of machine learning research inspires us to understand the research trends in this field since the existing machine learning techniques have applied to various application areas such as fraud detection, the stock market, weather forecasting, etc. Additionally, the algorithms changed according to newly emerged technology. So understanding the machine learning research themes from 1988 to 2017 will help to study the machine learning trends and to apply it in practical applications.

Latent Dirichlet allocation is a generative probabilistic topic model that intends to reveal latent or hidden thematic representations from a text corpus. The latent structure represented as topics with topic proportions per document expressed by hidden variables that LDA postulate within the dataset. As referred from related work, it understood that the topic weight of topic proportion per document was not explored in uncovering the research trend in machine learning. 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 emerging technology. So, understanding the machine learning research themes of the past three decades will help to study the current machine learning trends and apply it to practical applications. The primary motivation of this work was to intellectualize the evolution of research topics in machine learning over a period of three decades, i.e., 1988–2017. This work allowed us to visualize and examine the development of research topics over time. The motivation of the study in analyzing the trends of significant topics over time in machine learning research are as follows:

(i) There is a need to prepare or collect the dataset related to machine learning research.
(ii) There is a need to identify the significant topics in machine learning research that is not covered by other state of the arts previously.
(iii) It is required to compute the average topic weights of significant topics per year.
(iv) Trend analysis of significant topics is required to show their growth using rolling mean.
(v) It is required to compute topic prevalence of significant topics per year and compute the proportions of the topic weight of significant topic per journal title.

The contribution of this work is as follows:

(i) We have prepared a dataset of machine learning research for the period of 1988–2017 to uncover topics.
(ii) We have explored the topic coherence in evaluating the optimal number of topics in the dataset.
(iii) We have identified the significant topics in the dataset by ranking the topic coherence score over an optimal number of topics in the dataset.
(iv) Also, we have found the average topic weights and topic prevalence of significant topics per year.
(v) Finally, we have found the trend of significant topics growth using rolling mean and topic weight proportion per journal title.

The rest of the paper is organized as follows. The next section provides the related work on trend analysis. The following section introduces the methodology and its explanation of each step and the next section discusses the evaluation process of methodology. Finally, the last section concludes the paper.

Related Work

The thematic structure can utilize in finding trends in research. The trends in research can be examined and determined manually or analytically. The manual process specifies an intuition into the articles, but it is not at all free from partiality as researchers are inclined towards more cited papers in [43]. In contrary manual tagging is very thorough and requires proficiency in the documents of subject-matter expert, whereas the algorithmical analysis based on an automatic process in [11, 12, 35] by using topic modeling.

The topic model inputs a corpus, uncovering the topics and improves the semantic meaning of the vocabulary. Both clustering methods and topic analysis can employ topic modeling. Nonetheless, the topic analysis is more suitable as compared to clustering for detection of trends in research articles of the dataset in [18]. 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. The uncovering of latent topics from textual data has been successfully applied in several research area by utilizing LDA topic model. In [20] performed LDA on the collection of abstracts (i.e. 28,154) of the journal Proceedings of the National Academy of Sciences(PNAS) to identify topics and to depict their relationship to the PNAS classification scheme. Gatti et al. [19] used LDA on abstracts (i.e. 80,757) from 37 primary journals from the fields of operations research and management science (OR/MS) to attain intuitiveness into the current and historical publication trends. Similarly, [39] followed the same approach within the field of transportation research on 17,163 abstracts from 22 leading transportation journals and by [42] within the area of conservation science on 9834 abstracts. Apart from being executed on abstract data, LDA was also applied to 12,500 full-text research articles with-in the field of computational linguistics by [22], 2326 articles from neural information processing systems papers (NIPS) by [41], and 1060 articles within agricultural and resource economics by [3]. In [36] employed LDA to understand the research trends and topics in software effort estimation. In the work proposed by [26], LDA was performed to find trends in 3962 ITU-T recommendations. The authors extracted the representative topics for each 4-year period and the trend graphs of each topic using ITU-T recommendations.

Topic coherence applied with LDA model for identifying the optimal number of topics solutions and significant topics in the dataset. The significant topics determined by ranking the topic coherence score over an optimal number of topics in the dataset. Each topic contains a set of topic words and word weights. The word_weight is the probability of each topic_word in the topic. The topic distribution over documents resulted in the probability of each topic,i.e., topic_weight for each document in the dataset. A list of topics with topic weight was generated for each article. The topic weight of each topic in articles were ordered as per its year of publication. This arrangement determines the behavior of topic over time w.r.t topic weights. We believe that topic weight gave an intuition to understand the trends in research.

The inference drawn from related work is that there is need to apply LDA to identify the trends in machine learning research and process topic weight as a result of topic proportion over documents in finding the trends in the topic over time since 1988–2017.

Methodology

In this section as in Fig. 1, discusses the methodology or flow chart used for data preprocessing, followed by the LDA topic model. Moreover, discussing the dataframe created for analysis and approach for solving was the

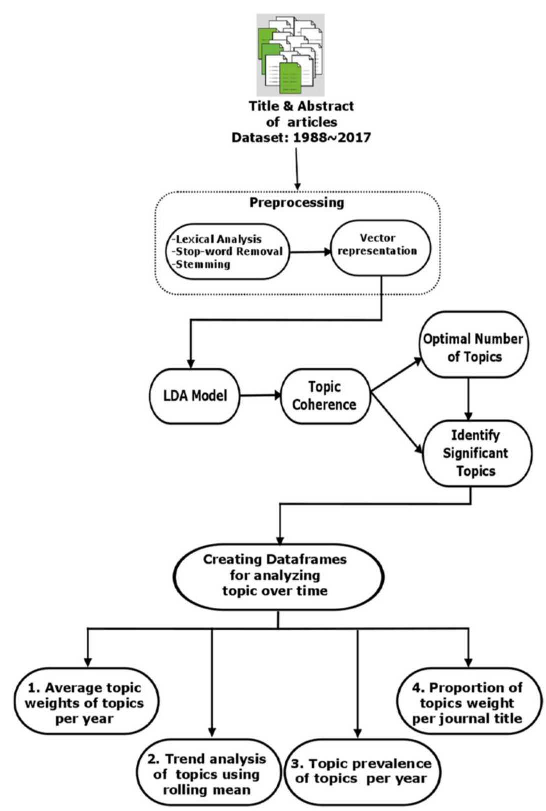

Fig. 1 Methodology or flowchart of the study

objective of our study. Algorithm 1 describes the steps of performing the trend analysis of significant topics over time.

Algorithm 1. Trend Analysis of Significant topics over time

#Preprocessing
1: Perform the lexical analysis, stop-word removal and stemming on the input dataset as discussed in section 3.1.
2: After this, transform the corpus into vector form to be feed into LDA model.
#LDA Model
3: The LDA model trained of the preprocessed data prepared in the preprocessing steps.
4: Create 99 different LDA models by varying the number of topics from 2 to 100.
5: For each number of topics, calculate the coherence score (C_v).
6: Choose the optimal number of topics in the dataset based on the highest value of coherence score.
7: Identifying the significant topics based on the coherence score of each topic from the optimal number of topic as selected in previous step.
#Creating dataframes for analyzing topic over time
8: Create the following dataframes for further analysis as shown in Table 1.
a: Create df.topicLabels dataframe consists of two columns as *topic_id* and *topic_words*.
b: Create df.wordWeights dataframe consists of three columns as *topic_id*, *topic_word*, and *word_weight*.
c: Create df.docTopicWeights dataframe consists of three columns as *doc_id*, *topic_id*, and *topic_weight*.
d: Create df.datasetDetail dataframe consists of three columns as *doc_id*, *year*, and *title*.
9: Used the following notations as shown in Table 2 for further calculations.
10: Compute the average topic weights of topics per year using Eq. 1 and Eq. 2.
11: Compute the rolling mean for trend analysis using Eq. 3.
12: Compute the topic prevalence per year using Eq. 4 to Eq. 6.
13: Compute the proportions of significant topic weights per journal using Eq. 7.

Text Processing

The preprocessing phase involves the elimination of noisy words/characters from the dataset, and it was performed by executing the following steps. Initially, the titles and abstracts of the articles were tokenized into tokens. The generated tokens were converted into lowercase letters in each document. The elimination of punctuation characters, apostrophe, commas, quotation marks, exclamation points, question marks, and hyphen was performed. Further, the numeric values are removed to get only the textual tokens. Then, the standard English words were as given in nltk python package [4] and were customized into stop-word list [10] with the phrases used to develop the literature dataset were removed. Afterward, for preparing a useful literature dataset, the word forms are stemmed from their original root form by using the Porter Stemmer algorithm [31]. It stems the tokens for each document and converts the inflected words to their base stem. Finally, we transformed documents into sparse vectors. The text files in a corpus contain titles and abstracts of articles. The bag-of-words document was a representation used for converting the documents into vectors. In this representation, each article was 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. The sparse vectors are created by counting merely the number of occurrences of each distinct word and convert each word to its integer word_id. The above steps are used to transform a corpus into vector representation for the LDA model.

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 [6]. Figure 2 shows the LDA graphical model. In LDA, a document (M) was considered as a mixture of latent topics (z), and each term (w) in the document was related with one of these topics. Using the latent clues, the topic model connects words having a similar meaning and differentiates the words having different meaning [38, 43]. 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 [2, 6, 29]. It produces topic models based on the three input parameters, namely, number of topics (k), 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 (θ). This parameter is considered as some "pseudo words", divided evenly between all topics present in every document, no matter how the other words were allocated to topics. The parameter β is per-word-weight of the Dirichlet prior over topic-word distributions (ϕ). The magnitude of the distribution (the sum over all words) 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.

Topic Coherence Measurement

After executing the LDA topic model, each topic includes words with a probability assigned to the words. The topic contains words with high probability are those words that likely to accompany more commonly in the topic distribution. The topics with words having high-probability, usually the top 10 words, are used to semantically label and interpret the topics. The evaluation of the quality of generated topics based on the measures such as the predictive likelihood of

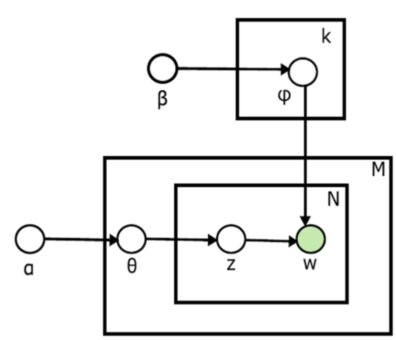

Fig. 2 LDA graphical model

held-out data proposed by [40]. Nevertheless, such a measure shows negative correlation with domain experts [14], by accomplishing the topics with high predictive likelihood less consistent from a domain expert perspective. The topic coherence measurement is specifically essential when generated topics are used for understanding the development and trends within a research area. Topic coherence measures proposed by researchers as a qualitative approach which automatically uncover the coherence of a topic [1, 30], and the underlying idea is rooted in the distributional hypothesis of linguistics [23]; words with similar meanings tend to occur in similar contexts. The topics are said to be coherent if most or all of the words, e.g., the topic's top N words, are related. The computational challenge is to obtain a metric that correlates remarkably with domain experts labeling or ranking data, such as topic ranking data obtained by word and topic intrusion tests [14].

Topic coherence is the metric which 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 interpret-ability [7]. The labeling or ranking of topics by domain experts are often considered to be the gold standard, and therefore, a method that correlates smoothly is a good sign of topic interpret-ability. The multitude of topic coherence measures and their correlation with domain experts are empirically and systematically explored by a recent study by [33]. Their systematic way uncovered a new unexplored coherence measure, labeled as C_v, to achieve the highest correlation with all available domain experts topic ranking data. As a result, this study adopts the C_v coherence measure for topic coherence calculations. C_v is based on four parts:

(i) segmentation of the data into word pairs,
(ii) calculation of word or word pair probabilities,
(iii) calculation of a confirmation measure that quantifies how strongly a word set supports another word set, and
(iv) finally, aggregation of individual confirmation measures into an overall coherence score.

Thus, this subsection discusses the topic coherence measurement for finding the optimal number of topics in dataset.

Creating the Dataframes for Analyzing Topic Over Time

After execution of LDA, the results are stored in dataframes for further analysis. Dataframes are two-dimensional data structure having unique columns of attributes for analysis. It helps us to manipulate the data to the topic change over time across different publication years. Table 1 shows the dataframes created for analyzing the research topic over time. Afterward, combining a series of dataframes to create a large composite dataframe. The composite dataframes contains seven columns such as {*index_pos*, *topic_id*, *topic_weight*, *topic_words*, *doc_id*, *year*, *journalTitle*}. Each row of this dataframe contains index_pos as the numeric index value of each *doc_id*, and the *topic_weight* of each *topic_id* belong to each *doc_id* with its *journalTitle* and *year* of publication. Additionally, the *topic_weight* of *topic_id* is inserted as zero in the dataframe if the *topic_id* didn't belong to *doc_id*. Finally, this composite data-frame ready for further analysis.

Table 2 represents the notations used in this work.

Computing the Average Topic Weights of Topics Per Year

The topics in the dataset evaluated by running the LDA model. Each topic contains a set of *topic_words* and *word_weight*. The *word_weight* is the probability of each *topic_word* in the topic. The topic distribution of each article is computed. A list of *topic_id* with *topic_weight* generated for each article. Initially, the topic weights are normalized of each topic belonging to each article in the dataset. The normalized topic weight is calculated as defined in Eq. (1):

Table 1 Dataframes used in this work

Dataframe	Columns	Description
df_topicLabels	*topic_id*, *topic_words*	Each row of this dataframe contains the *topic_id* and its corresponding *topic_words*. The *topic_words* contains the top ten words of each *topic_id*
df_wordWeights	*topic_id*, *topic_word*, *word_weight*	Each row of this dataframe contains the *word_weight* of each *topic_word* belong to *topic_id*
df_docTopicWeights	*doc_id*, *topic_id*, *topic_weight*	Each row of this dataframe contains the *topic_weight* of each *topic_id* concerning each *doc_id* in the dataset
df_datsetDetail	*doc_id*, *year*, *title*	Each row of this dataframe contains the *doc_id* and its publication *year* and the title of the journal, i.e., *title*

Table 2 Notations used in this work

Notation used	Notation meaning
D	The set of research articles in the dataset. For each $doc\_id \in \{d_0, d_1, \dots, d_{(\|D\|-1)}\}$, and $index\_pos \in \{0, 1, \dots, \|D\|-1\}$ where $\|D\|$ is the total number of articles in the dataset. Each $index\_pos$ corresponding to doc_id. The doc_id is a unique name for each research article in dataset
T	The set of topics in the dataset. For each $topic\_id \in \{t_0, t_1, \dots, t_{(\|T\|-1)}\}$, where $\|T\|$ is the total number of topics
Y	The set of a total number of years of publication. For each $year \in \{y_0, y_1, \dots, y_{(\|Y\|-1)}\}$, where $\|Y\|$ is the total number of years
W^i	The set of topic words for topic i. For each $topic\_words \in \{w^i_0, w^i_1, \dots, w^i_{(\|W^i\|-1)}\}$ where $i \in topic\_id$, and $\|W^i\|$ is the total number of topic words for each topic
WW^i	The set of word weights of corresponding topic words for topic i. For each $word\_weight \in \{ww^i_0, ww^i_1, \dots, ww^i_{(\|WW^i\|-1)}\}$ where $i \in topic\_id$, and $\|WW^i\|$ is the total number of word weights for each topic
TW^i_d	The set of topic weights of topic i corresponding to research article d. For each $topic\_weights_d \in \{tw^i_d \mid i \in topic\_id, d \in index\_pos\}$
TD_y	The set of total research articles published in year y. For each $total\_docs \in \{td_{Y_0}, td_{Y_1}, \dots, td_{Y_{(\|Y\|-1)}}\}$ where $yinY$.
J	The set of journal title in the dataset. For each $title \in \{j_0, j_1, \dots, j_{(\|J\|-1)}\}$ where $\|J\|$ is the total number of journal title in dataset.

$$norm\_topic\_weight^i_d = \frac{tw^i_d}{\sum_{i=0}^{i=|T_d|} tw^i_d} \tag{1}$$

where $d \in D$, $i \in T_d$, $tw^i_d \in TW^i_d$, and $T_d \subset T$.

Now, insert the *norm_topic_weight* to the composite data frame as discussed above. The average topic weight is computed by adding all of the weights for a given topic in a time period and dividing by the total number of documents in that time period as defined in Eq. (2):

$$avg\_tw^t_y = \frac{\sum_{t\in T_y, y\in Y} norm\_topic\_weight^t_y}{td_y} \tag{2}$$

where $y \in Y$,$td_y \in TD_y$, and $T_y \subset T$.

Finally, insert the *avg_tw* to the composite dataframe for further analysis.

Rolling Mean Method for Trend Analysis

Rolling mean (a.k.a. moving average) is one of the critical tools used to analyze the time series data. In a nutshell, moving average is simple weighted mean (sum) calculated over a selected historical time range. The text data is noisy, and the LDA model is applied to identify the topics from a dataset. The LDA topics contain topic-words with their topic-weight as a probability of each topic-word in the topic. The topic weight of topics for each year in the dataset is evaluated using the LDA model. Therefore, calculate the rolling for each topic t at year y is defined as in Eq. (3):

$$rm^t_y = p \times \sum_{i=1}^{w} tw^t_{y+i-1} \tag{3}$$

where $p = \frac{1}{w}$, $i = \{1, \dots, w\}$, $y \in Y$, and integer w determines the averaging window width. Thus, the rolling mean method was applied to the topics for finding the trends in dataset.

Computing Topic Prevalence of Topics Per Year

Another approach is used as a topic prevalence to calculate the topic significance over time. Topic prevalence is determining whether a topic is significantly present with the maximum topic weight for a document and then computing the percentage of documents in a given year where the topic is significantly present.

Topic prevalence can be computed by identifying the topic with the maximum topic weight per document, grouping the results by year, adding up the number of top occurrences of each topic per year and dividing them by the total number of documents per year. Initially, find the topic t with a maximum topic weight per document d using Eq. (4):

$$max\_topic\_weight^t_d = \max\{norm\_topic\_weight^t_d\} \tag{4}$$

where $d \in D$, $t \in T_d$, $T_d \subset T$.

Then, computing the occurrences of each topic t per year y using Eq. (5):

$$max\_count^t_y = \sum_{t\in T_y, y\in Y} [P] \tag{5}$$

where $[P] = [max\_topic\_weight^t_y = norm\_topic\_weight^t_y]$, $y \in Y$, $t \in T_y$, $T_y \subset T$. Here, $[\cdots]$ is the Inversion brackets. $[P]$ is defined to be 1 if P is true, and 0 if it is false. Finally, calculate the topic prevalence of each topic t for each year y using Eq. (6):

$$topic\_prevalence^t_y = \frac{max\_count^t_y}{td_y} \tag{6}$$

where $y \in Y$, $t \in T_y$, $T_y \subset T$, and $td_y \in TD_y$.

Computing the Proportions of Significant Topic Weights Per Journal

In this subsection, computing the proportions of significant topic weights for each journal title to see the overall distribution of topics within different subset of the dataset. Finally, calculate the proportion of significant topic t for each journal title j using Eq. (7):

$$prop\_topic\_per\_journal_j^t = \sum_{t \in T_j, j \in J} norm\_topic\_weight_j^t \quad (7)$$

where $j \in J, t \in T_j, T_j \subset T$. In the next section, this methodology is applied to understand the trend analysis of significant topic over time in machine learning research.

Evaluation

In this section, discusses the dataset, topic coherence as an evaluation metric, experimental setting, result, and discussion of our work.

Dataset

The research data were collected from various well-known journals published with high-quality research articles in machine learning. We include the established journals like Journal of Machine Learning Research (JMLR), IEEE Transactions on Neural Networks (IEEE-NN), IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE-PAMI), Science Direct Pattern Recognition (ScD-PR), Science Direct Neural Networks (ScD-NN), and Springer Machine Learning (Sp-ML). The titles and abstracts of research papers were considered from the electronic library of the mentioned journal articles. Recognizing significant contribution to research, we have included journal articles only for our work. The corpus has prepared by collecting articles in the order of its publication time, and results are drawn from the time span of 30 years, i.e., from 1988–2017. Table 3 lists the number of articles included in our work according to the journals. Each dataset has considered a separate corpus.

Table 3 The number of articles included in this study

S. no.	Journal name	Duration	#Years	#Articles published
1	JMLR	2000–2017	18	1755
2	IEEE-NN	1990–2017	28	4349
3	IEEE-PAMI	1988–2017	30	4630
4	ScD-PR	1988–2017	30	6567
5	ScD-NN	1988–2017	30	3294
6	Springer-ML	1988–2017	30	1311
Total				21,906

Creating the LDA Models

The LDA model trained on the preprocessed data prepared in the above sections. We created 99 different LDA models by varying the number of topics from 2 to 100. The Dirichlet parameters are set to be symmetrical for the smoothing of words within topics $\beta = \frac{1}{V}$, where V is the size of vocabulary and topics within the documents $\alpha = \frac{1}{|T|}$, where $|T|$ is number of topics. On choosing, $\alpha < 1$, the modes of the Dirichlet distribution are nearby to the corners, thus preferring merely a few topics for every document and leaving the larger part of topic proportions very close to zero. the Python Gensim [32] library for topic modeling is used for creating our LDA models. Approximation of the posterior distribution of our LDA models was performed through variation inference called online LDA by [24]. Gensim implemented variation inference as online LDA. In E-step, the convergence iteration parameter is set to 100 for the variational distributions where per document parameters are fit (see Algorithm 2 in [24]).

Topic Coherence

As explained in "Topic coherence measurement", we have created (99 in total) LDA model and calculated the C_v coherence score for each model. Segmentation of top pairs is gathered by combining every word from the top 10 words with every other word from the top 10 words. The below subsection discusses the evaluation of an optimal number of topics and significant topics after applying topic coherence.

Evaluating the Optimal Number of Topics

In an unstructured set of documents, where the numbers of appropriate trends are not known 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 [44]. There is no traditional measure to defend the optimal number of solutions. However, the topic coherence is run from topics 2 to 100 for the dataset to find the optimal range of topic solutions. The maximum coherence score leads to an optimal number of topics for the dataset. As in Fig. 3, shows the optimal numbers of the topic is 40 for our dataset. Based on these heuristics and findings of the study [9], the optimal number of topic solutions for identifying the trends chosen as 40 for the LDA model.

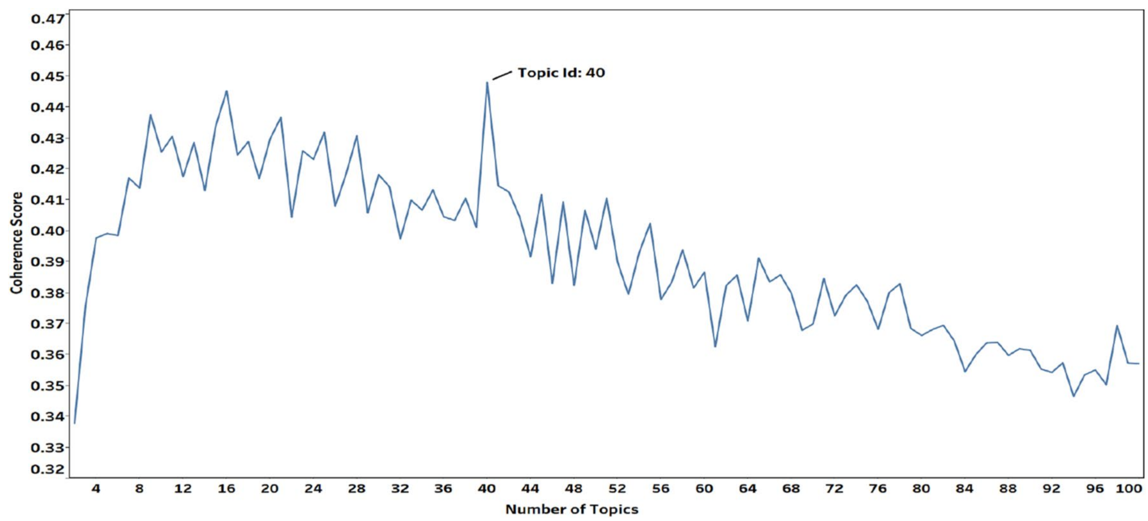

Fig. 3 Evaluating the optimal number of topics for the dataset (1988–2017)

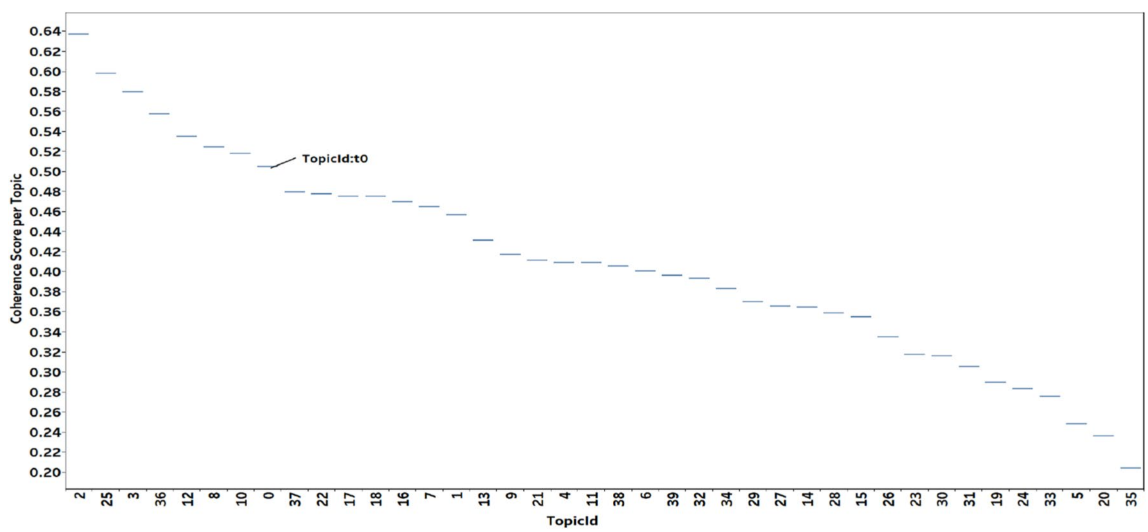

Fig. 4 Choosing significant topics based on coherence score per topic

Identify the Significant Topics

By running LDA model on a dataset, we obtained 40 topics and assigned each topic an ID number range from t_0 to t_{39}. The significant topics are evaluated by running the topic coherence on each topic which results in a sequence of a similarity measure for each topic. As in Fig. 4 shows the similarity measure for each topic arranged in decreasing order and elbow method is used to identify the significant topics for our dataset. The top eight topics are identified as significant topics by domain experts. The top eight significant topics with topic IDs are: $\{t_2, t_{25}, t_3, t_{36}, t_{12}, t_8, t_{10}, t_0\}$. Table 4 list down the statistics summary of topics using topic-weight. The topics range from 93% of the tokens in a document to 1% (excluding the zero values), with an average at 9% and a median value of 5%. The most frequent value is near 1%, which indicates that the data predominately describes topics that have a minor presence in the documents. Table 5 shows

Table 4 Summary statistics of topics

S. no.	Statistics	Topic_weight value
1	Max	0.933098
2	Min	0.010054
3	Average	0.091008
4	Median	0.054765
5	Most frequent value	0.013000

Table 5 List of significant topics and topic labels with their topic words and weights

S. no.	Topic id	Topic label	Topic words with their weights			
1	t_2	Neural network input and output	Neural	0.0978	Neural_network	0.0254
			Network	0.0696	Input	0.0251
			Train	0.0456	Weight	0.0224
			Layer	0.0317	Output	0.0220
			Learn	0.0259	Recurr	0.0200
2	t_{25}	A phase response model for brain behavior	Model	0.0514	Respons	0.0109
			Activ	0.0337	Studi	0.0096
			Brain	0.0137	Dynam	0.0095
			Behavior	0.0133	Input	0.0089
			Mechan	0.0117	Phase	0.0087
3	t_3	Real-time illumination estimation of scene images	Imag	0.0553	Scene	0.0239
			Camera	0.0370	Illumin	0.0217
			Reconstru	0.0360	Depth	0.0192
			Motion	0.0207	Light	0.0172
			Estim	0.0272	Align	0.0167
4	t_{36}	Implementing neural architecture using analog circuits	Neuron	0.0599	Process	0.0191
			Spike	0.0325	Neural	0.0163
			Implement	0.0305	Analog	0.0148
			Coupl	0.0250	Comput	0.0146
			Circuit	0.0202	Simul	0.0144
5	t_8	Synchronization of neural network with time delay	Neural	0.0733	Synchron	0.0232
			Network	0.0685	Global	0.0203
			Stabil	0.0475	Result	0.0200
			Delay	0.0441	Vari	0.0169
			Paper	0.0161	Neural_network	0.0345
6	t_8	Human motion detection and event tracking	Track	0.0765	Human	0.0365
			Sequenc	0.0667	Frame	0.0254
			Video	0.0635	Trajectori	0.0242
			Tempor	0.0382	Event	0.0240
			Motion	0.0370	Person	0.0207
7	t_{10}	Subspace method for linear discriminant analysis using local features	Dimension	0.0487	Propos	0.0252
			Space	0.0421	Linear	0.0249
			Discrimin	0.0398	Project	0.0221
			Method	0.0352	Subspac	0.0221
			Local	0.0265	Featur	0.0207
8	t_0	Measuring performance of classification accuracy of nearest neighbor	Classif	0.1609	Perform	0.0245
			Classifi	0.1265	Method	0.0226
			Class	0.0842	Accuraci	0.0206
			Decis	0.0316	Nearest	0.0203
			Train	0.0248	Neighbor	0.0191

the significant topics with topic_id and top 10 topic words for each topic with the corresponding word_weight of the period 1988–2017.

Results and Discussion

This section describes the result and discussion on the average topic weights of significant topics per year, trend analysis of significant topics using rolling mean, the topic prevalence of significant topic per year, and proportions of significant topics per journal title.

Average Topic Weights of Significant Topic Per Year

In this subsection, using "Computing the average topic weights of topics per year", we are aggregating the topic weights to evaluate the average of topic weights for each year. The average topic weight is computed by adding all of the weights for a given topic in a period and dividing by the total number of documents in that period. This gives us the average weight of the topic over all documents in the corpus. As in Fig. 5 showed the average topic weights of significant topics. The topic t_2 based on neural network shows the steady increase in average topic weight from 1988 till 1995, later on the scope of neural network decrease due to lack of computational resources. The average topic weight of topics t_{25}, t_3, t_{36}, and t_0 shows neutral during the time period. Furthermore, the average topic weight of topics t_{12}, t_8, and t_{10} shows a steady increase in research area.

Trend Analysis of Significant Topics Using Rolling Mean

As discussed in "Rolling mean method for trend analysis", the rolling mean method is used to highlight trends in the data and to compute the overall trajectory of a topic and to visualize the average on a rolling time window. The rolling mean strategy developed particularly for time series data, or data that is produced on regular intervals by some recording instrument. It is used for minimizing the dips and spikes of a particular year to find the research trends in data. Here, the articles of mentioned journals are collected for every month of a given period. The rolling time window of 3 years is considered for our experimental work. By computing, the trajectory of a particular topic using rolling mean provides a more abstracted depiction of the topic weights than average topic weight. As in Fig. 6 showed the rolling mean topic weights of significant topics. The topic t_2 described

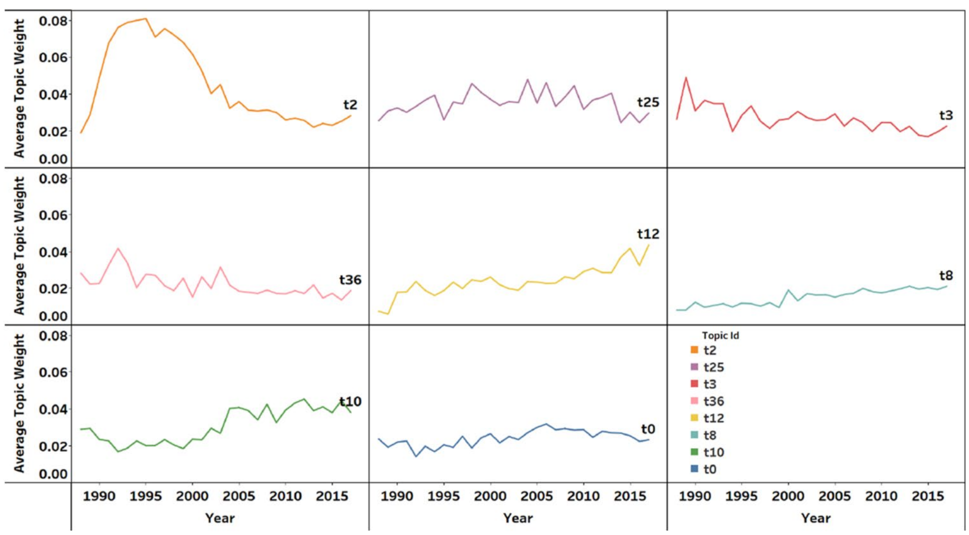

Fig. 5 Average topic weights of significant topics

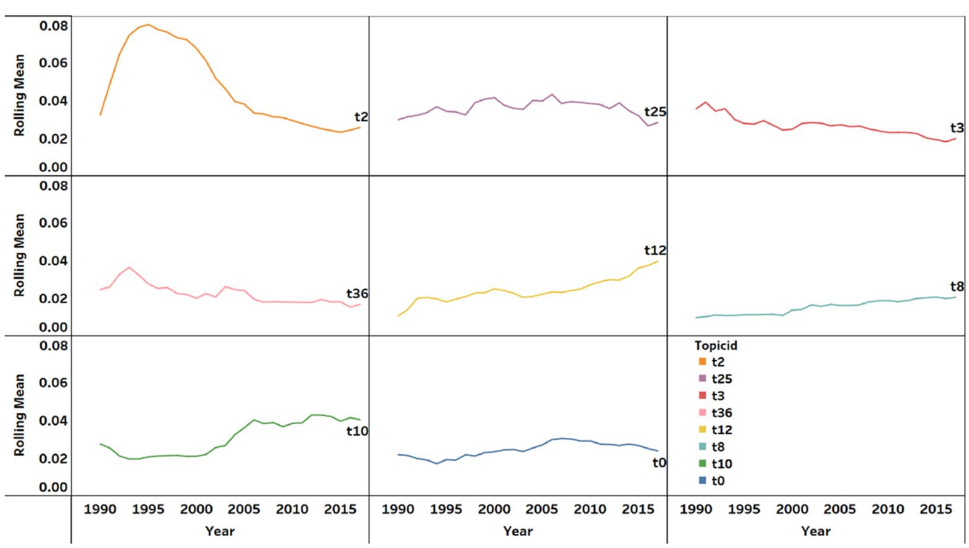

Fig. 6 Trend analysis of significant topics using rolling mean

the steady increase in trend till 1995, then trend steadily falling down. The trajectory of topics t_{25}, t_3, and t_{36} showed an increase in early 5 years and later on it decreasing during the rest of time period. The rolling mean of topics t_{12}, and t_{10} showed a steady increase in the trajectory of the topic throughout the time period. Finally, trend of the topics t_8, and t_0 showed a slow increase in their trajectory since 1988.

Topic Prevalence of Significant Topics Per Year

As discussed in "Computing topic prevalence of topics per year", topic prevalence is determining whether a topic is significantly present with the highest topic weight for an article and then computing the percentage of articles in a given year where the topic is significantly present. If we observe the figures for the average topic weights per year, the two sets of lines look very similar but not same. As in Fig. 7 showed the topic prevalence of significant topics. The topic t_2 shows the highest prevalence in the years 1993 and 1995, the prevalence gradually decreasing in the later years. The topics t_{25}, t_3, and t_{36} shows the highest topic prevalence in the years 2009, 1989 and 2003 respectively. Similarly, the highest topic prevalence of topics t_{12}, t_8, t_{10}, and t_0 in the year 2017, 2015, 2012 and 2006 respectively.

Proportions of Significant Topics Per Journal Title

As discussed in "Computing the proportions of significant topic weights per journal", Fig. 8 showed the proportions of significant topic weights in each journal title. The significant topics contributed for each journal titles mentioned as IEEENN, IEEEPAMI, JMLR, SDNN, SDPR, and SPRINGER as 31.30%, 18.28%, 12.24%, 38.17%, 16.13%, and 12.41% respectively than rest of the topics. The topics proportion of topics t_2, t_{25}, t_{36}, and t_{12} prominently belongs to two journal titles i.e., IEEENN and SDNN. Moreover,

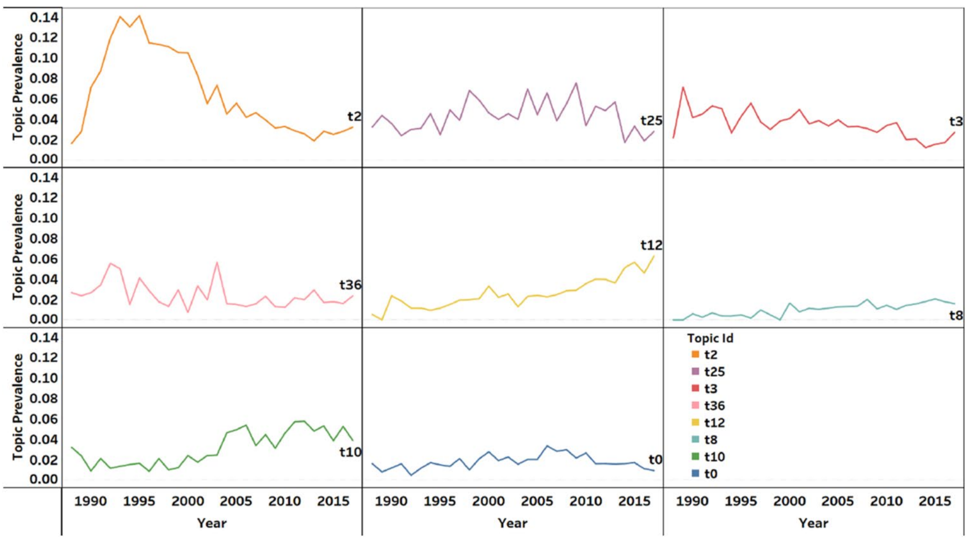

Fig. 7 Topic prevalence of significant topics

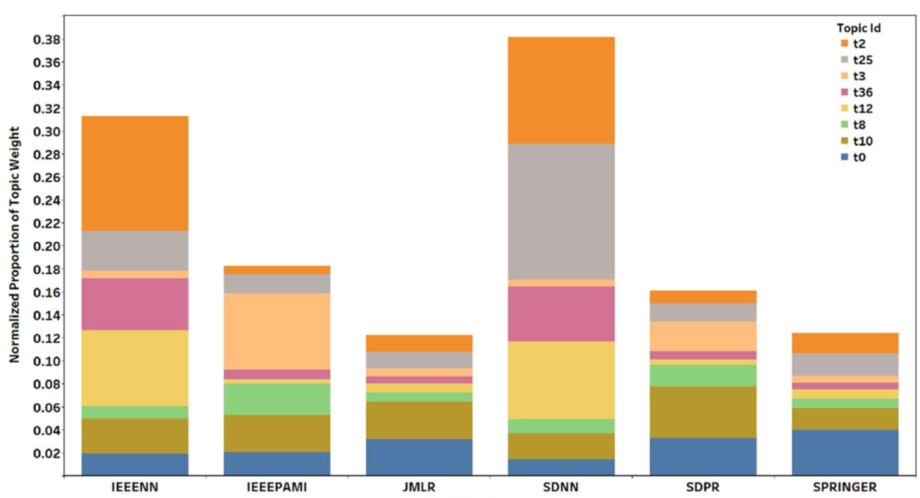

Fig. 8 Normalized proportion of significant topic weights in each journal

the topic proportion of topics t_0, and t_{10} uniformly belong to all journal titles. The topic proportion of topics t_3, and t_8 prominently belong two journal titles i.e., IEEEPAMI and SDPR. The topics generated by the LDA model has widely spread across different journals. Thus, the trend analysis of significant topics in machine learning research for the period was analyzed. This analysis can motivate the future researchers to understand the trends of the machine learning topics and give them the opportunity to explore further.

Conclusions

In this work, we have carried out a trend analysis of research topics over time in the machine learning research done over the last three decades. The LDA topic model is applied for evaluating the trends using the topic weight of significant topics. The dataset of machine learning research is prepared to uncover the topics and understand the trends in this area. 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.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Aletras N, Stevenson M. Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th international conference on computational semantics (IWCS 2013)–long papers. 2013. p. 13–22.
2. Alghamdi R, Alfalqi K. A survey of topic modeling in text mining. Int J Adv Comput Sci Appl (IJACSA). 2015;6(1), Citeseer.
3. Alston J, Pardey P. Six decades of agricultural and resource economics in Australia: an analysis of trends in topics, authorship and collaboration. Aust J Agric Resour Econ. 2016;60(4):554–68.
4. Loper E, Bird S. Nltk: the natural language toolkit. 2002. arXiv:cs/0205028.
5. Blei DM, Lafferty JD. Topic models. In: Text mining. Chapman and Hall/CRC; 2009. p. 101–124.
6. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
7. Blei DM, Jordan MI, et al. Variational inference for Dirichlet process mixtures. Bayesian Anal. 2006;1(1):121–43.
8. Boyack KW, Klavans R. Creation of a highly detailed, dynamic, global model and map of science. J Am Soc Inf Sci. 2014;65(4):670–85.
9. Bradford RB. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In: Proceedings of the 17th ACM conference on information and knowledge management. ACM; 2008. p. 153–62.
10. Buckley C, Salton G. Stopword list 2. 1995. http://www.lextek.com/manuals/onix/stopwords2.html.
11. Campbell JC, Hindle A, Stroulia E. Latent Dirichlet allocation: extracting topics from software engineering data. In: The art and science of analyzing software data. Elsevier; 2016. p. 139–59.
12. Canini K, Shi L, Griffiths T. Online inference of topics with latent Dirichlet allocation. In: Artificial intelligence and statistics. 2009. p. 65–72.
13. Carbonell JG, Michalski RS, Mitchell TM. Machine learning: a historical and methodological analysis. AI Mag. 1983;4(3):69.
14. Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. Reading tea leaves: how humans interpret topic models. In: Advances in neural information processing systems. 2009. p. 288–96.
15. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41(6):391–407.
16. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012;55(10):78–87.
17. Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications. Proc Natl Acad Sci. 2004;101(suppl 1):5220–7.
18. Evangelopoulos N, Zhang X, Prybutok VR. Latent semantic analysis: five methodological recommendations. Eur J Inf Syst. 2012;21(1):70–86.
19. Gatti CJ, Brooks JD, Nurre SG. A historical analysis of the field of or/ms using topic models. 2015. arXiv:1510.05154
20. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci. 2004;101(suppl 1):5228–35.
21. Grimmer J, Stewart BM. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit Anal. 2013;21(3):267–97.
22. Hall D, Jurafsky D, Manning CD. Studying the history of ideas using topic models. In: Proceedings of the 2008 conference on empirical methods in natural language processing. 2008. p. 363–71.
23. Harris ZS. Distributional structure. Word. 1954;10(2–3):146–62.
24. Hoffman M, Bach FR, Blei DM. Online learning for latent Dirichlet allocation. In: Advances in neural information processing systems. 2010. p. 856–64.
25. Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. 1999. p. 50–7.
26. Hwang MH, Ha S, In M, Lee K. A method of trend analysis using latent Dirichlet allocation. Int J Control Autom. 2018;11(5):173–82.
27. Larsen P, Von Ins M. The rate of growth in scientific publication and the decline in coverage provided by science citation index. Scientometrics. 2010;84(3):575–603.
28. Marr B. A short history of machine learning—every manager should read. 2016. https://www.forbescom/sites/bernardmarr/2016/02/19/a-shorthistory-of-machine-learning-every-managershould-read.
29. Mavridis T, Symeonidis AL. Semantic analysis of web documents for the generation of optimal content. Eng Appl Artif Intell. 2014;35:114–30.
30. Newman D, Lau JH, Grieser K, Baldwin T. Automatic evaluation of topic coherence. In: Human language technologies: The 2010

annual conference of the North American chapter of the association for computational linguistics. 2010. p. 100–8.
31. Porter MF. An algorithm for suffix stripping. Program. 1980;14(3):130–7.
32. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, Citeseer. 2010.
33. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. 2015. p. 399–408.
34. Rusch T, Hofmarcher P, Hatzinger R, Hornik K, et al. Model trees with topic model preprocessing: an approach for data journalism illustrated with the wikileaks Afghanistan war logs. Ann Appl Stat. 2013;7(2):613–39.
35. Saini S, Kasliwal B, Bhatia S, et al. Language identification using g-lda. Int J Res Eng Technol. 2013;2(11):42–5. Citeseer.
36. Sehra SK, Brar YS, Kaur N, Sehra SS. Research patterns and trends in software effort estimation. Inf Softw Technol. 2017;91:1–21.
37. Srivastava AN, Sahami M. Text mining: classification, clustering, and applications. Boca Raton: Chapman and Hall/CRC; 2009.
38. Steyvers M, Griffiths T. Probabilistic topic models. In: Handbook of latent semantic analysis, vol 427, no 7. 2007. p. 424–40.
39. Sun L, Yin Y. Discovering themes and trends in transportation research using topic modeling. Transport Res Part C Emerg Technol. 2017;77:49–66.
40. Wallach HM, Murray I, Salakhutdinov R, Mimno D. Evaluation methods for topic models. In: Proceedings of the 26th annual international conference on machine learning. ACM; 2009. p. 1105–12.
41. Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. p. 424–33.
42. Westgate MJ, Barton PS, Pierson JC, Lindenmayer DB. Text analysis tools for identification of emerging topics and research gaps in conservation science. Conserv Biol. 2015;29(6):1606–14.
43. Yalcinkaya M, Singh V. Patterns and trends in building information modeling (bim) research: a latent semantic analysis. Autom Constr. 2015;59:68–80.
44. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinform BioMed Central. 2015;16:S8.