Design and Development of Topic Identification Using Latent Dirichlet Allocation

P. Lakshmi Prasanna, S. Sandeep, V. Kantha Rao, and B. Sekhar Babu

Abstract Data storing and retrieving are the most important task in the present condition. Storing can be ended based on the topic that the document describes. Text mining generates documents from the collection of topics. To identify the topics, we have to categorize the documents; to classify, we are using topic modeling. Text mining technique is used for discovering latent semantic structure which is a fragment of topic modeling. Various research areas that make use of probabilistic modeling includes software engineering, political science, and medical science. A topic model is a probability-based model that discovers the major themes which are a group of documents. The main idea is to treat the documents as mixtures of topics in the topic model, and every topic is viewed as a probability distribution of the words. This research work aims to propose a model called topic modeling using LDA, and this model has been experimented on two datasets, where one is two news group dataset, and other is twenty news group dataset, and finally, all the results are tabulated.

1 Introduction

LDA is an unconventional probably based model for compilations of detached information and consequently more suitable for document data. An unsubstantiated accession was supposedly to be utilized for identifying and penetrating the bunch of terms in huge clusters of documents. This method imagines that each text is a combination of concepts; each and every word is accredited to each of the concepts that has limited prospect. This model also uncovers various topics that the texts correspond to and to what extent every individual concept is at hand in a document. LDA bears to discover the credibility dispersions over terms; then, it finds arrays of words that form together with definite possibility. Aforementioned arrays are marked as "topic."

LDA is a Bayesian inference model designed by David et al. [1]. Every text is linked to a likelihood allocation more than topics, and further, topics are probability

P. Lakshmi Prasanna (✉) · S. Sandeep · V. Kantha Rao · B. Sekhar Babu
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
e-mail: pprasanna@kluniversity.in

V. Bindhu et al. (eds.), *International Conference on Communication, Computing and Electronics Systems*, Lecture Notes in Electrical Engineering 733,
https://doi.org/10.1007/978-981-33-4909-4_63

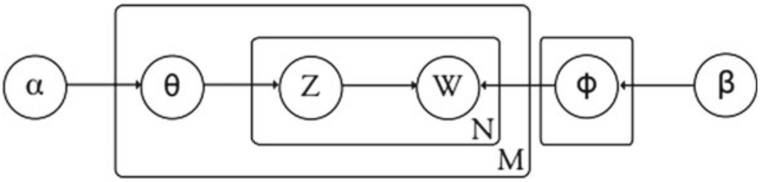


Fig. 1 Block diagram of LDA

segregation son terms. Every expected result of an arbitrary is changeable with the possibility that the happening will occur and come out of the distribution as an equation that links them. For example, in a sport coin flips for decision of who will go first and second as a deciding factor, there are two probable results out of it: heads or tails. Heads is personified as 1 and tails as 0. Here is an indiscriminate variable which is marked X. And dualistic expected reactions is delineated by X as 0 or X as 1. The probability allotment of X or $P(X = x)$ is: $P(X = 0) = 0.5$ $P(X = 1) = 0.5$. Figure 1 represents the block diagram of LDA.

2 LDA Criterion

To attain rest results, the below mentioned few parameters can be helpful.

1. Topics Number—Quantity of topics which have to be acquired through the body.
2. α and β are hyper parameters—α symbolizes document-topic density, and β symbolizes topic and token density. Depending upon the upper worth of α, texts are possessed of added topics and greater the β; themes are made up of bigger pack of terms [2].
3. Iterations no—The duration for algorithm to come together (or how do we sense that the algorithm has converged)?

1. The hidden topics digit in the corpus (K) needs to be explained previous to initiating the practice of the model.
2. Since the perplexity reduces very slowly, algorithm convergence needs a large number of iterations.
3. If the training information set is large, then the algorithm requires a long time for preparation.
4. Being a probabilistic model, LDA requires additional clarifications in preparing statistical inference. Hence, it does not toil on same small texts like sentence categorization or tweets.
5. For providing meaningful knowledge to advance excellence of topics, LDA does not take into account connection of concepts or words.

3 Topic

To identify what is a topic and how it can be interpreted has been a continuous point of discuss in the literature. Commonly, themes provide an outline to the subject matter present in the texts being considered as well as brief explanation of the contents of the dataset. Whether the statement "topics" is even the appropriate word to use for the groups of words generated by the topic model evaluation which has been under radar for decades. "Topic" might not constantly be apt word to be use for the divisions which are generated through topic modeling [3]. Some have recommended that "discourse" may be sophisticated, as subjects are not constantly amalgamated linguistically. For instance, mathematics, lineage, reasoning, science, theory, century, and epistemology shall not be well thought-out semantically coherent as they are seen as discourse of subjects found contained by philosophy. When every word has a different meaning, it would be intricate to give them an explicit topic [4].

3.1 Topic Models

Topic models are a class of algorithms which exploit co-occurrences of words in documents in order to discover hidden sets of words which explain the co-occurrence patterns and are referred to as topics. Probabilistic topic models explain pragmatic documents with an underlying, hidden probabilistic model. The observed documents are assumed to be random samples from this model. In a probabilistic topic model, each document is associated in the midst of a probability allocation over a set of ideas, and concepts get associated with a prospective disposal upon the set of words. Similar documents share a similar topic distribution. Topic models are often employed in regard to text mining errands, e.g., to have understanding and visualizing the content of large document corpora or for detecting relations between topics and other variables of interest [5]. Additionally, topic models can be employed as a mean for dimensionality reduction (documents are mapped to a lower dimensional topic space), as input for prediction tasks, in recommender systems (e.g., for predicting semantically related tags) or in information retrieval (e.g., to understand and disambiguate the topic of query terms).

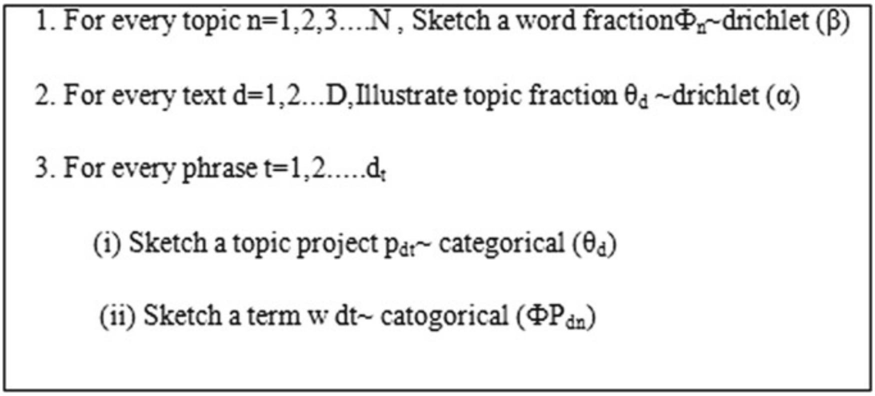
3.2 Topic Modeling Using LDA

For finding topics, LDA algorithm was used by applying Bayesian belief networks theorem to calculate probability and topic identification of each term and each document. For feature reductions, LDA Algorithm was considered to reduce features and filtered top terms of each topic and store it in the filtered document term matrix.

Latent Dirichlet allocation is being primarily utilized for evaluating texts [2]. It presumes that here is N no. of topics on how texts will be produced, and each topic is corresponding to multinomial circulation on words in the terminology. A document $w_d = \{w_{dt}\} d_{tt} = 1$ is accomplished by variety a concoction and these topics and fragmenting words from the mixture [1]. Figure 2 indicates the procedure of LDA Algorithm, and Table 1 represents the notations of LDA.

Applied the Bayesian probability to identify the model.

- Represent β as the total terms of the documents.
- Represent θ as specifying all the topics.
- Display all the terms in the documents with a minimum range of 10.
- Dividing the terms into topics.
- All terms are displayed in the topic wise.
- Display all the probability values with words.
- Based on these probabilities to create a word cloud.



1. For every topic n=1,2,3....N , Sketch a word fractionΦ_n~drichlet (β)

2. For every text d=1,2...D,Illustrate topic fraction θ_d ~drichlet (α)

3. For every phrase t=1,2.....d_t

(i) Sketch a topic project p_{dt}~ categorical (θ_d)

(ii) Sketch a term w dt~ catogorical (ΦP_{dn})

Fig. 2 Procedure of LDA Algorithm

Table 1 Notations of LDA

Symbol	Description
N	No. of topics
Y	No. of unique words in the vocabulary
D	No. of documents
DT	No. of words in the documents DT
θ_d	Proportion of topics specific to documents
On	Proportion of words specific to topic N
P_{dn}	Identify the topics of *n*th word in document d
W_{dt}	Identify the word in document d
αβ	Parameters of Dirichlet distribution

4 Procedure of LDA

See Fig. 2.

5 Notations of LDA Algorithm

Here, the first step depicts the number of topics and after that represents all word momentarily apportion toward the topics, and the procedure is completed absurdly, and from time to time, similar terms can be functional to various topics. The final step shows the updated adaptation of the topic assignment depended on their credibility as per the above criteria:

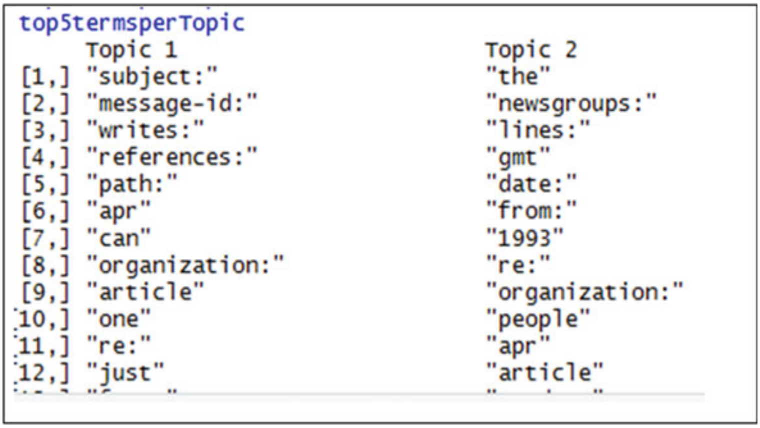
1. Primary criteria is about the length of prevalence is that tokens across the topics—it can be called as $P(w/t)$.
2. The another criteria is about the topic of the issues in the document $P(t/d)$.

As per the Bayesian belief network theorem $P(t/w) = P(t/d) * P(w/t)$. To calculate probability of each term. Figure 3, 4, 5, 6, and 7 represent the outcome of topic modeling using LDA.

Top terms per topics for 2 groups data (Figs. 3 and 4).

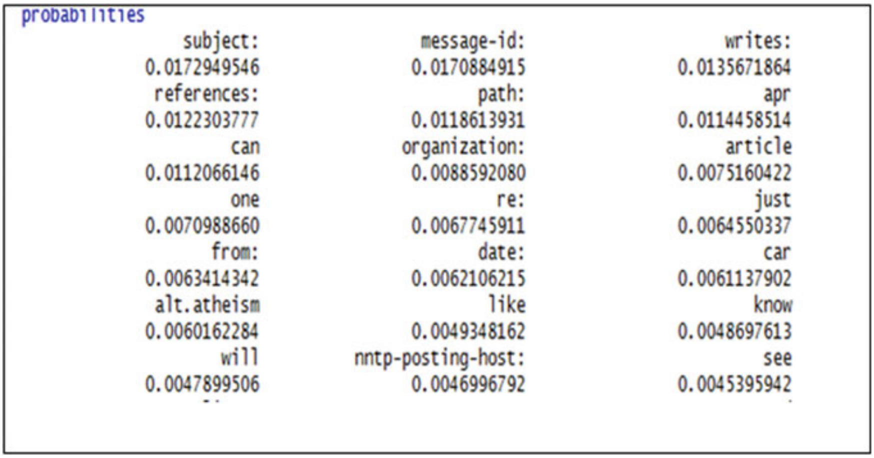
Top terms with probabilities for 20 news groups data (Figs. 5 and 6).

Filtered document term matrix (FDTM) (Fig. 7).



```
topStermsperTopic
      Topic 1                    Topic 2
[1,]  "subject:"                 "the"
[2,]  "message-id:"              "newsgroups:"
[3,]  "writes:"                  "lines:"
[4,]  "references:"              "gmt"
[5,]  "path:"                    "date:"
[6,]  "apr"                      "from:"
[7,]  "can"                      "1993"
[8,]  "organization:"            "re:"
[9,]  "article"                  "organization:"
[10,] "one"                      "people"
[11,] "re:"                      "apr"
[12,] "just"                     "article"
```

Fig. 3 Top terms per topics for 2 groups data

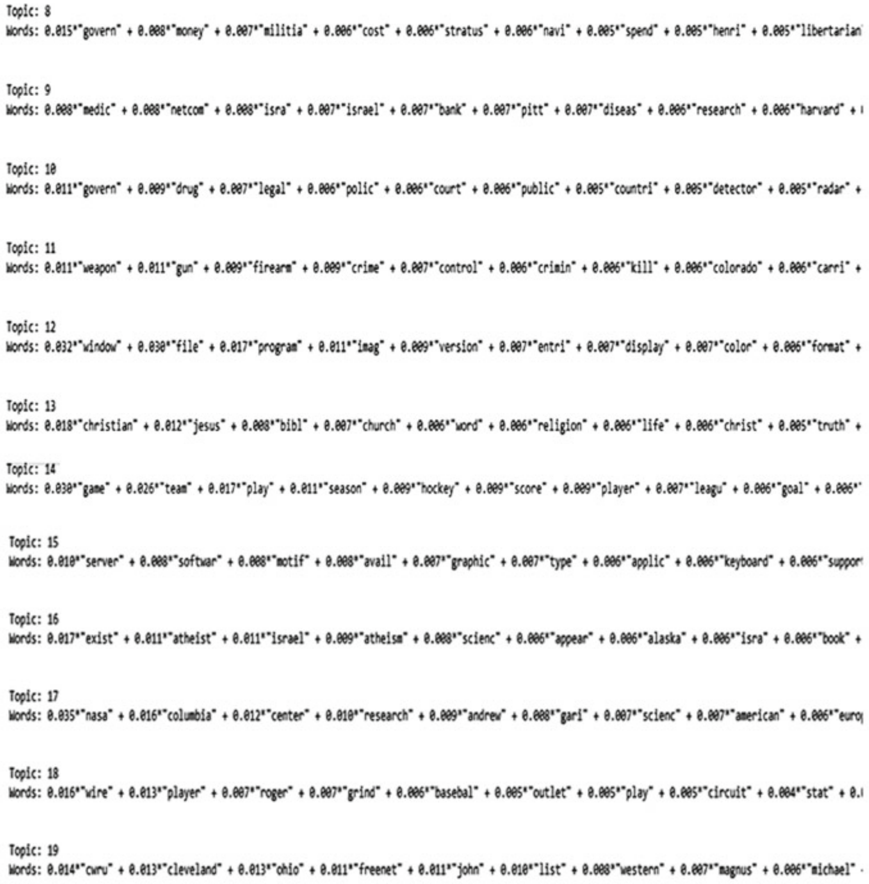


```
probabilities
        subject:          message-id:          writes:
    0.0172949546         0.0170884915     0.0135671864
     references:                path:              apr
    0.0122303777         0.0118613931     0.0114458514
             can        organization:          article
    0.0112066146         0.0088592080     0.0075160422
             one                  re:             just
    0.0070988660         0.0067745911     0.0064550337
           from:                date:              car
    0.0063414342         0.0062106215     0.0061137902
     alt.atheism                like              know
    0.0060162284         0.0049348162     0.0048697613
            will  nntp-posting-host:               see
    0.0047899506         0.0046996792     0.0045395942
```

Fig. 4 Top terms with probabilities for 2 groups data

6 Conclusion

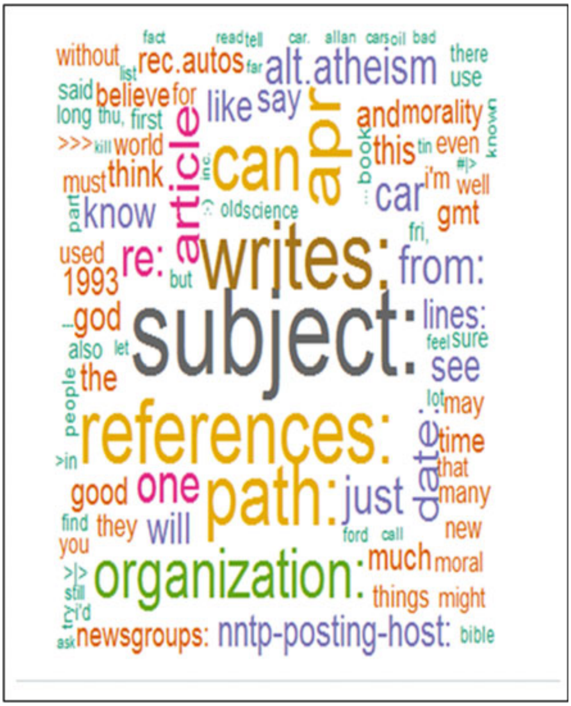
Topic modeling began as of text mining method designed for disclosing latent semantic structure within a compilation of texts. In text mining, each archive is produced from anthology of topics. It relies upon probabilistic modeling that has a huge assortment of relevance such as image detection, semantic understanding, and automatic music improvisation recognition. In this chapter to proposed topic modeling employing latent Dirichlet allocation [LDA], the LDA works backward to learn the topic illustration in all texts and the word allotment to every topic. The main focus of this paper is on LDA algorithms, and the outcomes will be displayed in 20 news group dataset and 2 groups dataset.

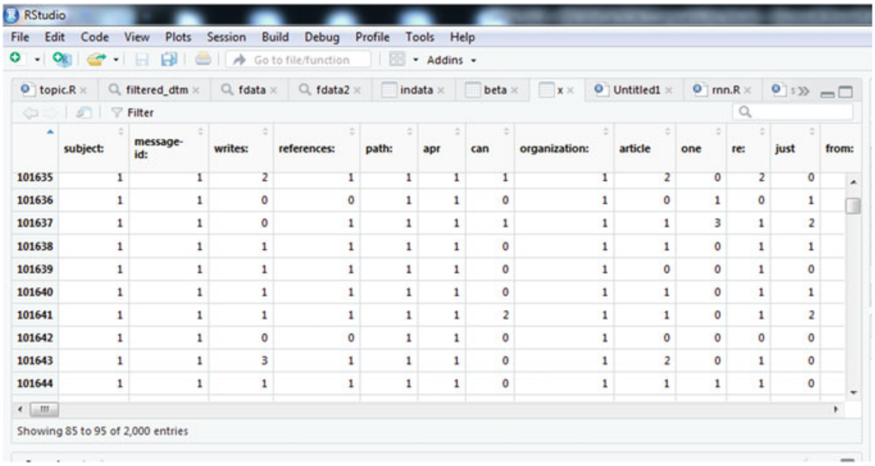


```
Topic: 8
Words: 0.015*"govern" + 0.008*"money" + 0.007*"militia" + 0.006*"cost" + 0.006*"stratus" + 0.006*"navi" + 0.005*"spend" + 0.005*"henri" + 0.005*"libertarian

Topic: 9
Words: 0.008*"medic" + 0.008*"netcom" + 0.008*"isra" + 0.007*"israel" + 0.007*"bank" + 0.007*"pitt" + 0.007*"diseas" + 0.006*"research" + 0.006*"harvard" +

Topic: 10
Words: 0.011*"govern" + 0.009*"drug" + 0.007*"legal" + 0.006*"polic" + 0.006*"court" + 0.006*"public" + 0.005*"countri" + 0.005*"detector" + 0.005*"radar" +

Topic: 11
Words: 0.011*"weapon" + 0.011*"gun" + 0.009*"firearm" + 0.009*"crime" + 0.007*"control" + 0.006*"crimin" + 0.006*"kill" + 0.006*"colorado" + 0.006*"carri" +

Topic: 12
Words: 0.032*"window" + 0.030*"file" + 0.017*"program" + 0.011*"imag" + 0.009*"version" + 0.007*"entri" + 0.007*"display" + 0.007*"color" + 0.006*"format" +

Topic: 13
Words: 0.018*"christian" + 0.012*"jesus" + 0.008*"bibl" + 0.007*"church" + 0.006*"word" + 0.006*"religion" + 0.006*"life" + 0.006*"christ" + 0.005*"truth" +

Topic: 14
Words: 0.030*"game" + 0.026*"team" + 0.017*"play" + 0.011*"season" + 0.009*"hockey" + 0.009*"score" + 0.009*"player" + 0.007*"leagu" + 0.006*"goal" + 0.006*"

Topic: 15
Words: 0.010*"server" + 0.008*"softwar" + 0.008*"motif" + 0.008*"avail" + 0.007*"graphic" + 0.007*"type" + 0.006*"applic" + 0.006*"keyboard" + 0.006*"suppor

Topic: 16
Words: 0.017*"exist" + 0.011*"atheist" + 0.011*"israel" + 0.009*"atheism" + 0.008*"scienc" + 0.006*"appear" + 0.006*"alaska" + 0.006*"isra" + 0.006*"book" +

Topic: 17
Words: 0.035*"nasa" + 0.016*"columbia" + 0.012*"center" + 0.010*"research" + 0.009*"andrew" + 0.008*"gari" + 0.007*"scienc" + 0.007*"american" + 0.006*"euro

Topic: 18
Words: 0.016*"wire" + 0.013*"player" + 0.007*"roger" + 0.007*"grind" + 0.006*"basebal" + 0.005*"outlet" + 0.005*"play" + 0.005*"circuit" + 0.004*"stat" + 0.

Topic: 19
Words: 0.014*"cwru" + 0.013*"cleveland" + 0.013*"ohio" + 0.011*"freenet" + 0.011*"john" + 0.010*"list" + 0.008*"western" + 0.007*"magnus" + 0.006*"michael" -
```

Fig. 5 Top terms with probabilities for 20 news groups data

Fig. 6 News groups dataset of word cloud





	subject:	message-id:	writes:	references:	path:	apr	can	organization:	article	one	re:	just	from:
101635	1	1	2	1	1	1	1	1	2	0	2	0	
101636	1	1	0	0	1	1	0	1	0	1	0	1	
101637	1	1	0	1	1	1	1	1	1	3	1	2	
101638	1	1	1	1	1	1	0	1	1	0	1	1	
101639	1	1	1	1	1	1	0	1	0	0	1	0	
101640	1	1	1	1	1	1	0	1	1	0	1	1	
101641	1	1	1	1	1	1	2	1	1	0	1	2	
101642	1	1	0	0	1	1	0	1	0	0	0	0	
101643	1	1	3	1	1	1	0	1	2	0	1	0	
101644	1	1	1	1	1	1	0	1	1	1	1	0	

Showing 85 to 95 of 2,000 entries

Fig. 7 Filtered document term m0atrix for filtered features

References

1. BleiDM, Andrew Y (2003) Latent Drichlent allocation. J Mach Learn Res
2. Tong Z, Zhang H (2016) A text mining research based on Lda topic modelling. In: The sixth international conference on computer science, engineering and information technology
3. DayaSagar KV, Shyam Krishna C, Lalith Kumar G, Surya Teja P, Charless Babu G (2018) A method for finding threated web sites through crime data mining and sentiment analysis. Int J EngTechnol (UAE) 7(2):62–65
4. Wallach HM (2008) Structured topic models for language, PhD thesis
5. Roose H, Roose W, Daenekindt S (2018) Trends in contemporary art discourse: using topic models to analyze 25 years of professional art criticism. CulturSociol 12:303–324
6. Kousar A, Subrahmanyam K (2019) Feature selection, optimization and clustering strategies of text documents. Int J Electr Comput Eng 9(2):1313–1320
7. BleiDM (2012) Surveying a suite of algorithms that offer a solution to managing large document archives. Commun ACM
8. Bastani1 K, Namavari1 H, Shaffer J (2016) Latent Dirichlet Allocation (LDA) for topic modeling of the CFPB consumer complaints. IEEE
9. Kaur PC, Ghorpade T, RamraoAdik V (2017) Extraction of unigram and bigram topic list by using Latent Dirichlet Markov allocation and sentiment classification. In: 2017 international conference on energy, communication, data analytics and soft computing
10. PotharajuSP, Sreedevi M, AndeVK, TirandasuRK (2019) Data mining approach for accelerating the classification accuracy of cardiotocography. ClinEpidemiol Global Health
11. Poornima BK, Deenadayalan D, Kangaiammal A (2017) Text preprocessing on extracted text from audio/video using R. Int J Comput Intell Inform 6(4)
12. Sajid A, Jan S et al (2017) Automatic topic modeling for single document short texts. In: International conference on frontiers of information technology (FIT).
13. Sleeman J, Halem M, Finin T (2017) Discovering scientific influence using cross-domain dynamic topic modeling. In: 2017 IEEE international conference on big data (big data)
14. Sharma N, Yalla P (2017) Classifying natural language text as controlled and uncontrolled for UML diagrams. Int J Adv Comput Sci Appl
15. Sapul MSC, Aung TH, Jiamthapthaksin R (2017) Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms. In: 2017 14th international joint conference on computer science and software engineering (JCSSE)
16. Lakshmi Prasanna P, Rajeswara Rao D (2017) Literature survey on text classification: a review.J Adv Res Dyn Control Syst 9(12):2270–2280