International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

# A Systematic and Analytical Approach to Techniques and Tools in Topic Modeling

**Shanthi.S, Nithya R, Nagendraprabhu**

***Abstract:*** ***Topic modeling is one of the recent upcoming research areas of interest among the researchers. Topic Modeling is a straightforward way to examine the huge volumes of unstructured data. Each topic is a collection of words and these words usually bond together more frequently. When this technique is applied to a huge volume of data it can join words having same meanings and distinguish the uses of words with multiple meanings. The intention is to study and examine different topic modeling algorithms and to perform a brief literature review and analysis was performed and the obtained results are presented in this paper. Many techniques for topic modeling proposed by different researchers are put together and characteristics and drawbacks of various techniques have discussed. We present this paper with the intention that it will help few of the researchers in finding out the problems, present challenges and future scope of research in topic modeling.***

***Index Terms:*** ***Topic Modeling, Topic, Words, LDA, Supervised, Unsupervised***

## I. INTRODUCTION

In the present scenario, we are witnessing a huge volume of data generated such as plain text, audio, images, and video etc. Nowadays, the majority of the data are generated from online books and papers and from most of the social networking sites. Data generated through these sources are mostly unstructured data. So it is becoming very complicated to obtain preferred and relevant information .In this scenario, we need to have new tools for mining the data as well as fetching the information which we are currently looking for[1]. Topic modeling is one such widely used technique in the area of text mining. Topic modeling is defined as a process used to routinely recognize the topics available in a text also need to obtain hidden patterns demonstrated by text corpus. Thus, it's more useful in assisting in decision making. It is much different from the traditional rule based approach. It uses an unsupervised approach for finding the topics from a large cluster of texts [2].Objectives of topic model include learning the distribution of words, Learning distribution of topics and assigning every word in a particular document to a particular topic.

Each document consists of a bag of words. These objects need to be learned and not known in advance. The term learning includes first to define the model and second employ a learning algorithm.Topic Models can be employed in multiple purposes, including:

- Clustering of Documents
- Need to systematically arrange huge blocks of textual data
- Retrieving the Information from unstructured text format
- Feature selection
- Dimensionality reduction
- text summarization
- recommendation engine

Topic models are helpful in organizing, arranging and retrieving huge datasets of profiles from social media, emails, online customer reviews. Fig 1: Represents the functioning of the topic model.

## II. REVIEW METHODOLOGY

From the past few years, topic modeling has been widely used in many applications such as online reviews, biological and medical document mining, Document Mining, etc. Since the data generate is more and there is a need to identify the hidden themes and organize, summarize and search the documents based on these themes. The aim is to find out the various topic modeling techniques and analyze them and to discuss the various tools and challenges and issues involved in topic modeling. Therefore, in this section, based on a systematic literature review we discuss the most recent and related work on topic modeling.

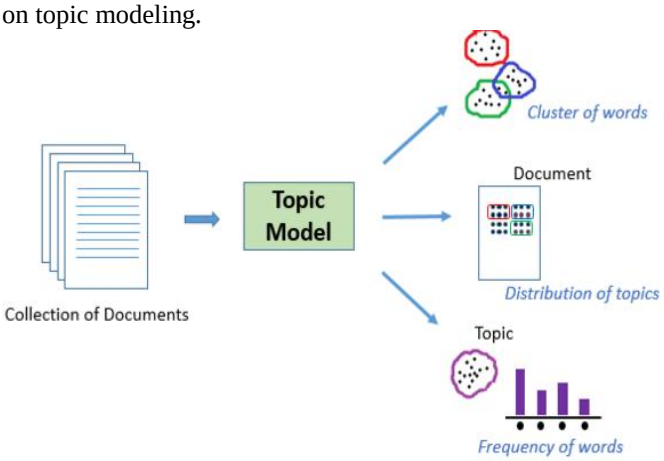


**Figure 1: Topic Model**

**Revised Manuscript Received on 30 May 2019.**

* Correspondence Author

**Shanthi.S*,** Research Professor with Malla Reddy College of Engineering and Technology, Hyderabad, Telangana, India.

**Nithya R,** Assistant Professor in the Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, affiliated to Jawaharlal Nehru Technical University, Anantapuram, India

**Nagendraprabhu**, Professor, Department of Computer Science and Engineering, Malla Reddy College of Engineering & Technlogy, Dhulapally, Secunderabad, India

### 2.1 Research Questions

To understand the importance of topic modeling, we need to find the right questions for research. The main work is to summarize the various techniques and methods used for analyzing the topic from unstructured data using topic modeling.

Following research questions were identified:

Q1: What is topic modeling and what are the different techniques used for topic modeling?

Q2: What are the tools used for Topic Modeling?

Q3: What are the challenges and Issues Faced in Topic Modeling?

## III. RELATED WORKS

| Author | Method | Analysis | Approach |
|---|---|---|---|
| Deerwester et al. [1990] | LSI - Latent semantic indexing | In order to make the semantic content it creates a Latent vector based representation for texts. | Unsupervised |
| Hofmann [2001] | PLSA - Probabilistic latent semantic analysis | Aim is to recognize and differentiate various contexts of word usage. | Unsupervised |
| Blei et al. [2003] | LDA - Latent Dirichlet allocation | Documents in LDA are a collection of a mixture of topics | Unsupervised |
| Griffiths and Tenenbaum [2004] | hLDA - Hierarchical latent Dirichlet allocation | HLDA models a tree of topics and represents the correlation of topics. | Unsupervised |
| Rosen-Zvi et al. [2004] | ATM –Author topic model | ATM uses metadata, which is associated with each document in corpus. | Unsupervised |
| Li and McCallum[ 2006] | PAM - Pachinko allocation model | PAM represents correlation between topics in the form of DAG directed acyclic graph | Unsupervised |
| Wang & McCallum [2006] | TOT - Topic over time | word co-occurrences, localization be modeled continuously | Supervised |
| Blei & Lafferty [2006] | DTM- Dynamic topic model | Considers arranging of the documents as per the order and provides a better posterior topical structure | Unsupervised |
| Teh et al. [2006a] | Hierarchical Dirichlet process (HDP) | No need of specifying the number of topics in advance and usually determined by using posterior inference. | Supervised |
| Blei and Lafferty [2007] | CTM - Correlated topic model | Logistic normal distribution is used to model pair wise correlations among various topics. | Unsupervised |
| Mcauliffe & Blei [2008] | sLDA - Supervised LDA | Each and every document is combined with continuous response variable and normal linear regression is used to model the response variables | Supervised |
| Lacoste-Julien et al.[2009] | discLDA - Discriminative variation on LDA | DiscLDA is a supervised form of LDA. Class labels are additionally introduced and unsupervised dimensionality reduction abilities are retained. | Supervised |
| Ramage et al. [2009] | LLDA - Labeled LDA | It introduces multiple labels for documents and establishes the relation between the labels. | Supervised |
| Chang & Blei [2010] | RTM - Relational topic model | Links between various documents are given by the distances among topic proportions of documents | Unsupervised |
| Ramage et al. [2011] | PLLDA - Partially labeled LDA | Extension of LLDA and introduces Latent Topics. | Supervised |
| Petinot et al.[2011] | HLLDA - Hierarchical labeled LDA | Improvement of hLDA and t introduced a label prior and establishes one-to-one communication between a label and topic | Supervised |
| Kataria et al. [2011] | WPAM -Wikipedia-based Pachinko allocation model | To learn the exact entity disambiguation models from Wikipedia | Semi supervised |
| Zhu et al. [2012] | medLDA -Maximum entropy discrimination LDA | To train and instruct supervised based topic models, max-margin principle is used and used for predicting tasks. | Supervised |
| Mao et al. [2012] | SSHLDA - semisupervised hierarchical topic model | It explores new topics in data space automatically | Semi supervised |
| Bakalov et al. [2012] | LPAM - Labeled Pachinko allocation model | keywords are automatically assigned to a particular taxonomy having multilabel documents | Unsupervised |
| Ma et al.[2012] | L-F-L-PAM -Labeled four-level Pachinko allocation model | Capture the correlations between multiple labels | Unsupervised |

| | | | |
|---|---|---|---|
| Mimno & McCallum[2012] | DMR - Dirichlet-multinomial regression topic model | It incorporates observed document features namely author, publication venue and for document topic distributions uses a log-linear prior. | Supervised |
| Nguyen et al.[2013] | SHLDA- Supervised hierarchical latent Dirichlet allocation | Multiple Paths are used for documents in the tree | Supervised |

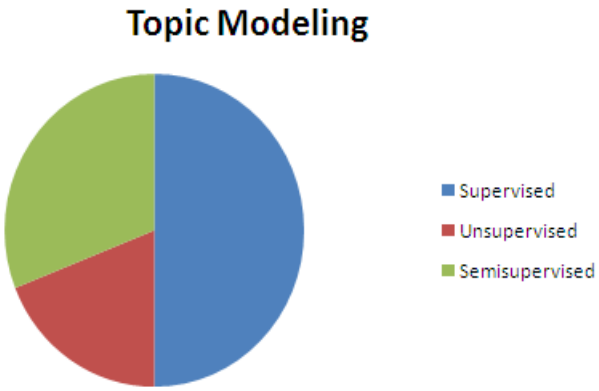


Fig 2 : Distribution of Approaches for Topic Modeling

## IV. TOOLKITS USED

Many toolkits are currently available and employed in many applications due to the development of topic models. Some of the toolkits are discussed and mostly utilized in the area of NLP natural language processing.

1. Gensim

Gensim introduced by Rehurek [2008] uses free Python library used for the automatic retrieval of various semantic topics from various documents. A plain text is the input and semantic topics are retrieved. Then topical similarity against documents can be checked. Some algorithms available in Gensim includes LSI, LDA, and Random Projections.

2. Topic Modeling Box

Stanford TMT proposed by Ramage and Rosen [2009] used Scala and created by Stanford NLP group. It is mainly intended to aid scientists and various researchers to examine huge volume of textual material. Many algorithms are used in TMT includes LDA, PLDA and LLDA.

3. MALLET

MALLET proposed by McCallum [2002] is a toolkit based on a Java-based package used for NLP natural language processing such as classification of document, biomedical applications, online reviews, clustering, topic modeling, and text mining related applications. MALLET supports algorithms such as LDA, PAM, HLDA..

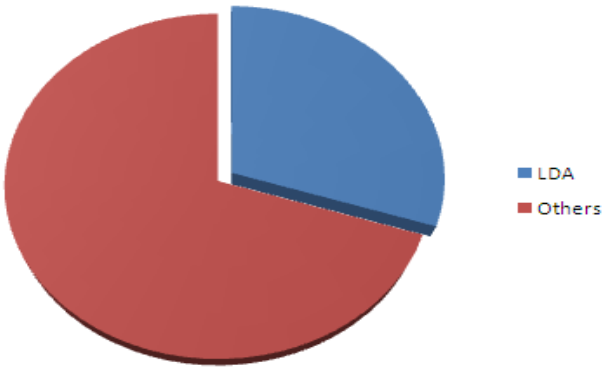


Fig 3 : Comparision of Topic Models

## V. RESEARCH CHALLENGES AND ISSUES

While implementing topic modeling many research challenges are being faced. Some of them include:

- Accuracy in Retrieving
- Identifying and retrieving hidden topic structure.
- Multinomial topic model Labeling
- Scalability
- Representing the mismatch between topic model and a label.
- Accurately topic interpretation
- Handling voluminous documents
- Creating a huge number of topics

## VI. CONCLUSION

In this paper, we have systematically reviewed the various techniques in topic modeling. From the review, effectiveness of each technique has been analyzed. A lot of approaches have been proposed for topic modeling, but most of them have used LDA based techniques for organizing, summarizing and retrieving the concerned topic or word from the unstructured data. Some of the questions raised have been solved, and it might be useful to researchers for carrying out their work.

## REFERENCES

1. Bicego M, Lovato P, Perina A, Fasoli M, Delledonne M et al (2012) Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM Trans Comput Biol Bioinform 9(6):1831–1836
2. Blei DM (2012) Probabilistic topic models. Commun ACM 55 (4):77–84
3. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391
4. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1–2):177–196.
5. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
6. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(Suppl 1):5228–5235.
7. Griffiths D, Tenenbaum M (2004) Hierarchical topic models and the nested chinese restaurant process. Adv Neural Inf Process Syst 16:17.

8. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on uncertainty in artificial intelligence, pp 487–494
9. Li W, McCallum A (2006) Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on Machine learning, pp 577–584
10. Wang X, McCallum A (2006) 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, pp 424–433
11. Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning, pp 113–120
12. Teh YW, Jordan MI, Beal MJ, Blei DM (2006a) Hierarchical dirichlet processes. J Am Stat Assoc 101(476):1566–1581
13. Blei DM, Lafferty JD (2007) A correlated topic model of science. Statistics 1(1):17–35
14. Mcauliffe JD, Blei DM (2008) Supervised topic models. In: Advances in neural information processing systems, pp 121–128.
15. Lacoste-Julien S, Sha F, Jordan M I (2009) DiscLDA: Discriminative learning for dimensionality reduction and classification.In: Advances in neural information processing systems, pp 897–904
16. Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: a supervised topic model for credit attribution in multi- labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing, pp 248–256
17. Chang J, Blei DM (2010) Hierarchical relational models for document networks. Ann Appl Stat 4(1):124–150
18. Ramage D, Manning CD, Dumais S (2011) Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 457–465
19. Petinot Y, McKeown K, Thadani K (2011) A hierarchical model of web summaries. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, vol 2, pp 670–675
20. Kataria SS, Kumar KS, Rastogi RR, Sen P, Sengamedu SH (2011) Entity disambiguation with hierarchical topic models. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1037–1045
21. Zhu J, Ahmed A, Xing EP (2012) MedLDA: maximum margin supervised topic models. J Mach Learn Res 13:2237–2278
22. Mao X-L, Ming Z-Y, Chua T-S, Li S, Yan H et al (2012) SSHLDA: a semi-supervised hierarchical topic model. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp 800–809
23. Bakalov A, McCallum A, Wallach H, Mimno D (2012) Topic models for taxonomies. In: Proceedings of the 12th ACM/IEEE- CS joint conference on digital libraries, pp 237–240
24. Ma H, Chen E, Xu L, Xiong H (2012) Capturing correlations of multiple labels: a generative probabilistic model for multi- label learning. Neurocomputing 92:116–123
25. Mimno D, McCallum A (2012) Topic models conditioned on arbitrary features with Dirichlet-multinomial regression.University of Massachusetts, Amherst 2008, pp 411–418
26. Nguyen V-A, Boyd-Graber JL, Resnik P (2013) Lexical and hierarchical topic regression. In: Advances in neural information processing systems, pp 1106–1114
27. LiuL, Tang L, Dong W, Yao S, Zhou W(4) An overview of topic modeling and its current applications in bioinformatics. SpringerPlus (2016) 5:1608 DOI 10.1186/s40064-016-3252-8
28. Rehurek R (2008) Gensim. http://radimrehurek.com/gensim/
29. Ramage D, Rosen E (2009) Stanford TMT. http://nlp.stanford.edu/software/tmt/tmt-0.4/
30. McCallum AK (2002) MALLET. http://mallet.cs.umass.edu/


## AUTHORS PROFILE

**Dr.S.Shanthi** received her Ph.D. degree from University Of Mysore, Mysore, India, in 2016, and M.E Degree from Sathyabama University, India in 2008. She is currently working as a Research Professor with Malla Reddy College of Engineering and Technology, Hyderabad, Telangana, India, an autonomous Institution under the affiliation of Jawaharlal Nehru Technological University, Hyderabad. She is an author and co-author of more than 20 papers in Technical Journals and Conference Proceedings, and she has contributed to Book Chapters in her areas of interest and to her credit she has 2 patents. Her research interests include image processing, wireless Adhoc and sensor networks, machine learning, Network Security.

Dr. R.Nidhya is presently working as Assistant Professor in the Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, affiliated to Jawaharlal Nehru Technical University, Anantapuram, India. She received the M.Tech and Ph.D degree from Anna University, Chennai. Her research interests include wireless body area network, network security and data mining. She published 12 papers in refereed international journals and 13 papers in conferences. She is an active member of ISTE, IAENG, SAISE and ISRD.

**Dr. S. Nagendra Prabhu** currently working as Professor, Department of Computer Science and Engineering, Malla Reddy College of Engineering & Technlogy, Dhulapally, Secunderabad, India. He completed his PhD in Network security in cloud computing from Anna University, Chennai, India. He completed his Master of Engineering in Network Engineering from Anna University and his Bachelor of Engineering in Computer Science and Engineering from K.C.G College of Engineering and Technology, Chennai. His research interest includes Cloud computing, Botnet attack, Web based network Security. Currently the author is doing research related security issues in Cloud Computing.

.