International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Latent Feature Word Representations to Enhance Topic Models for Text Mining Algorithms

Thayyaba Khatoon Mohammed, M. Gayatri, M. Sandeep, V. S. K. Reddy

***Abstract:** Dealing with large number of textual documents needs proven models that leverage the efficiency in processing. Text mining needs such models to have meaningful approaches to extract latent features from document collection. Latent Dirichlet allocation (LDA) is one such probabilistic generative process model that helps in representing document collections in a systematic approach. In many text mining applications LDA is useful as it supports many models. One such model is known as Topic Model. However, topic models LDA needs to be improved in order to exploit latent feature vector representations of words trained on large corpora to improve word-topic mapping learnt on smaller corpus. With respect to document clustering and document classification, it is essential to have a novel topic models to improve performance. In this paper, an improved topic model is proposed and implemented using LDA which exploits the benefits of Word2Vec tool to have pre-trained word vectors so as to achieve the desired enhancement. A prototype application is built to demonstrate the proof of the concept with text mining operations like document clustering.*

***Keywords:** Text mining, document clustering, LDA, topic modeling, Word2Vec*

I. INTRODUCTION

Modeling biomedical or other documents need a systematic approach. LDA [2] is one such proven approach that is widely used. Moreover, it supports different models like topic model, author model and author-topic model. There are many variants of LDA that are used for customized modeling and processing. Generative process models thus became popular and useful to text mining purposes. Conventional topic modeling made with LDA and its variants can inter distributions like topic-to-word and document-to-topic. It is based on the co-occurrence of words within given documents. More information on probabilistic topic models can be found in [3] while modeling hidden topics is studied in [5]. Topic models have got supervised and unsupervised extensions as investigated in [6].

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Dr Thayyaba Khatoon Mohammed, Professor, Malla Reddy College of Engineering & Technology, Hyderabad, India. Email: thayyaba.khatoon16@gmail.com
M. Gayatri, Associate Professor, Malla Reddy College of Engineering & Technology, Hyderabad, India. Email: gayatricse312@gmail.com
M. Sandeep, Associate Professor, Malla Reddy College of Engineering & Technology, Hyderabad, India. Email: Sandeep.nitc@yahoo.com
Dr.VSK Reddy, Principal &Professor, Malla Reddy College of Engineering & Technology, Hyderabad, India. Email:vskreddy2003@gmail.com

Though topic models have been around with many LDA variants, of late, the notion of latent features is introduced. Latent feature (LF) vectors are widely being used to process NLP tasks. Latent features permit a range of values that become a part of high-dimensional space which has proved to be efficient for modeling large corpus. Two latent feature models based on LDA and Dirichlet Multinomial Mixture Model (DMM) are explored in this paper. Based on these baseline process models, Word2Vec based variants are introduced and used for effective modeling of latent feature word representations. Our contributions in this paper are as follows.

1. We proposed two generative process models considering latent features that are based on LDA and DMM respectively.
2. We exploited the latent feature topic models for better representation or modelling to leverage performance of text mining operations like document clustering.
3. We built a prototype application to show the effectiveness of the proposed generative process models with latent feature vectors.

The remainder of the paper is structured as follows. Section 2 presents review of literature based on generative process models for systematic modeling of document corpora. Section 3 presents the LDA for modeling. Section 4 covers derivation of latent feature models that are used for improving text mining operations by using Word2Vec toolkit. Section 5 presents experimental results while section 6 provides conclusions besides directions for future work.

II. RELATED WORK

This section reviews literature on the LDA [2], [8] and its variants for topic modeling. Generative process models like LDA became instrumental in processing text documents. Rosen-Zvi et al. [1] derived author model from LDA to give importance to author based processing of documents. Shen et al. [2] on the other hand proposed a latent topic model that is meant for processing documents to obtain latent friends. An author-topic model focuses on both authors and topics at the same time. This model is proposed by Rosen-Zvi et al. [3] for text mining algorithms. Similarly, to represent topic and author community, Liu et al. [4] proposed a model known as Topic-Link LDA. While all the models can be used for different mining purposes, Melnykov and Maitra [5] focused on clustering applications that are based on generative process models. Fatema et al. [6] used micro blogs data in order to extract topics based on authors and other attributes like recipients and contents. Bishop [7] explored these models for machine learning as part of Information Retrieval (IR).

From the LDA many variants of topic models came into existence. One such variant is proposed by Blei [9] for generating probabilistic topic models. With respect to word co-occurrence statistics, Bullinaria and Levy [10] extracted semantic representations for better accuracy of processing textual content. Cai et al. [11] proposed a generative model for modelling hidden topics towards enhancing performance of text clustering. Cao et al. [12] on the other hand introduced a neural network model for getting artificial intelligence (AI) from textual documents. More on topic models can be found in [13] for improvements in single-label text categorization. Natural Language Processing (NLP) is crucial for text mining. Towards this end, Collobert and Weston [14] proposed a unified framework for NLP. Deerwester et al. [15] focused on latent semantic analysis and its usage of indexing in order to improve the performance of mining operations. When the text documents reflect sparsity, it is essential to deal with them as well. Eisenstein et al. [16] investigated on sparse additive generative models for solving the problem with such documents. Glorot et al. [17] on the other hand used deep learning to improve domain adaptation for sentiment analysis. Topic analysis for finding topics related to scientific studies is explored by Giffits and Steyvers [18]. Hingmire et al. [19] studied document classification and topic labelling with process models. Topic modeling with Twitter dataset for generating recommendations and realizing different applications is made in [20]. From the literature, it is understood that there is need for further improvement in the performance. Towards this end, in this paper, we proposed novel models with Word2Vec usage for better performance as it provides vectors that are rich in coverage and useful for text mining algorithms.

III. LATENT DIRICHLET ALLOCATION

It is a generative process model which is widely used for document clustering or text mining. It has provision to graphically present the model that will help in implementing different models such as author models, topic models and author-topic models. Figure 1 shows the LDA graphical model which provides a process model to work with documents.

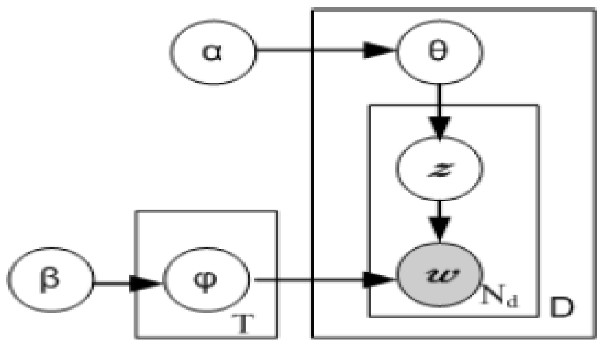

Figure 1: Graphical view of Latent Dirichlet allocation (LDA) [2]

Notation	Description
D	Set of documents
T	Set of topics
	Matrix reflecting distribution of topics
Theta	Matrix reflecting document-specific mixture weights for T
Alpha	Symmetric dirichlet prior
w	Indicates a single word in given document d
d	Indicates a single document from D
z	Topic that contains given word w
N_d	Number of words in the given document
Beta	Another symmetric dirichlet prior
Gama	Topics that participate in given document corpus

Table 1: Notations used in LDA

Different algorithms came into existence based on LDA model. They include variation inference, expectation propagation, and Gibbs sampling. The generative process for LDA as illustrated in Figure 1 is as follows.

$\theta_d \sim Dir(\propto)$ $\quad z_{d_i} \sim Cat(\theta_d)$

$\emptyset_z \sim Dir(\propto)$ $\quad w_{d_i} \sim Cat(\emptyset_{z_{d_i}})$

Dir stands for Dirichlet distribution while Cat denotes categorical distribution. The topic indicator for ith word denoted as wdi in given document d. The wdi is generated by a dirichlet multinomial component which topic-to-word in nature. For a given topic zdi categorical distribution is denoted as $Cat(\emptyset_{z_{d_i}})$.

IV. DERIVATION OF LATENT FEATURE TOPIC MODELS FROM LDA

Two novel models are derived from the LDA. Two probabilistic models are known as LF-LDA and LF-DMM. These two models help in achieving latent feature topic models. Figure 2 shows the models used in this paper along with along with Word2Vec which generates a set of vectors or feature vectors for words in the given corpus. The usage of Word2Vec provides higher level of accuracy in text mining applications.

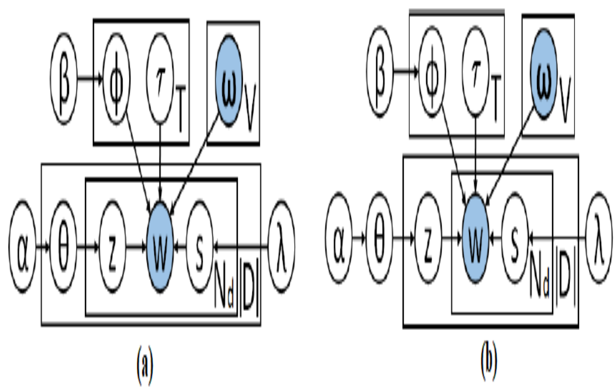

Figure 2: The latent feature topic models LF-LDA (a) and LF-DMM (b)

These two models are formed by original LDA and DMM models respectively. The traditional models are replaced by two component mixture of a topic to word dirichlet multinomial component and latent feature component. Both have resemblance with the original LDA or generative process models.

The difference is that the latent feature models define probability of a word given the topic with respect to categorical distribution. This relation is represented as in Eq. 1.

$$\mathrm{CatE}(w|T_t w^T) = \frac{\exp(T_t w_w)}{\sum_{w^1 \in W} \exp(T_t w_{w^1})} \quad (1)$$

The generative process model for the LF-LDA is as shown below.

$$\theta_d \sim Dir(\propto) \quad {}_z d_i \sim Cat(\theta_d)$$

$$\emptyset_z \sim Dir(\beta) \quad {}_w d_i \sim Ber(\lambda)$$

$$wd_i \sim (1 - sd_i)Cat(\emptyset_{d_i}) + sd_i\ CatE(Tz_{d_i}\omega^T)$$

Similarly, for LF-DMM, the generative process model is given as follows.

$$\theta \sim Dir(\propto) \quad {}_{zd} \sim Cat(\theta)$$

$$\emptyset_z \sim Dir(\beta) \quad {}_s d_i \sim Ber(\lambda)$$

$$wd_i \sim (1 - sd_i)Cat(\emptyset_{z_d}) + sd_i\ CatE(Tz_d\omega^T)$$

The word – topic assignment probability is explored for each world. Thus inference models are created for LF-LDA and LF-DMM. Then these models are used along with W2V in order to have observations on different datasets. The datasets used for empirical study are TMN and TMNTitle and F1 measure is used for performance analysis. With respect to document clustering, the W2V model has shown better performance as presented in Section 4.

V. EXPERIMENTAL RESULTS

An application is built with Java Swing API with Graphical User Interface (GUI). This intuitive application is meant for loading datasets and then perform required text mining operations on them as per the generative process models proposed in this paper. The prototype is used to provide good understanding about the process models that act on the given document corpus. Experimental results are presented in this section. A prototype application is built to validate the proposed latent feature topic models. The functional requirements of the proposed system are logically divided into modules namely IO Module, NLP Module, LF-TOPIC-LDA Module and Word2Vec Module. The IO module is responsible to provide functions related to inputs and outputs. The inputs are document corpus (collection of documents) and output is the result of clustering operation and performance. NSP module is meant for performing NLP operations while performing text mining (like clustering). It has provision for programmatically understanding words and sentences for similarity and other purposes. LF-TOPIC-LDA moduleis the generative process model that is derived from the LDA topic model. It is responsible to have latent feature weight computations and processing documents in the way LF-TOPIC-LDA (diagram) is graphically modelled.Word2Vec module when integrated with the system produces set of vectors from the pre-trained models for improving text mining process. In other words, the clustering performance will be improved with the usage of this module.

The application built for experiments is able to support different operations that are required by the proposed process models known as LF-LDA and LF-DMM. The interface reveals operations like Get Features, Evaluations, Get Models and Get Utility. The process models help in taking various kinds of input. Different input parameters are presented in Listing 1.

```
-alpha N     : Specify alpha
-beta N      : Specify beta
-corpus VAL  : Specify path to topic modeling corpus
-initers N   : Specify number of initial sampling iterations
-lambda N    : Specify mixture weight lambda
-model VAL   : Specify model
-name VAL    : Specify a name to a topic modeling
experiment
-niters N    : Specify number of EM-style sampling
iterations
-ntopics N   : Specify number of topics
-paras VAL   : Specify path to hyper-parameter file
-twords N    : Specify number of top topical words
-vectors VAL : Specify path to the file containing word
vectors
```

Listing 1: Shows the possible parameters that can be specified to fully exploit models

As presented in Listing 1, there are different input parameters that can be given to process models. They are known as alpha, beta, corpus, initers, lambda, model, name, niters, topics, paras, towards and vectors. By using the parameters as input, it helps user to exercise certain control over the models to deal with text mining activities.

```
Corpus size: 400 docs, 1857 words
Vocabulary size: 713
Number of topics: 4
alpha: 0.1
beta: 0.01
lambda: 0.6
Number of initial sampling iterations: 2000
Number of EM-style sampling iterations for the LF-LDA
model: 200
Number of top topical words: 20
Randomly initializing topic assignments...
Running Gibbs sampling inference:
Initial sampling iteration: 1
Initial sampling iteration: 2
Initial sampling iteration: 3
Initial sampling iteration: 4
...
Estimating topic vectors...
LFLDA sampling iteration: 133
Estimating topic vectors...
LFLDA sampling iteration: 134
Estimating topic vectors...
LFLDA sampling iteration: 135
Estimating topic vectors...
LFLDA sampling iteration: 136
Estimating topic vectors...
```

LISTING 2: Shows the iterative process pertaining to LF-LDA sampling

As presented in Listing 2, the proposed LF-LDA sampling is carried out. It is the iterative process that keeps estimating topic vectors. Once the iterative process is completed, it proceeds to write output to secondary storage.

```
0.9493914177249865 0.04887935639208968
1.0280283841054505E-6 0.0017281978545395662
4.332771063834005E-6 0.9999686401749415
2.3544280406781566E-8 2.7003509714359444E-5
4.161662502606706E-10 0.9997813822537462
3.326683469053791E-9 2.1861400340411567E-4
0.9999983278965086 1.150568375217415E-6
1.547133290367342E-10 5.213804027518439E-7
0.9895510669484859 0.0032073448580799734
8.382321769008487E-5 0.007157764975743864
0.9999997717495226 2.2796621946480572E-7
4.562120725580652E-16 2.8425758609934113E-10
6.809610528887646E-16 0.9999999999999698
1.1345015488965578E-16 2.921118266466978E-14
0.9999914984595634 1.681008789730536E-8
8.473930916133864E-6 1.0799432490031264E-8
0.9999897052854877 9.546693305837297E-6
3.4796269922924406E-7 4.000585072492128E-7
4.51295940479208E-6 2.6034936144083928E-6
0.9999927540690459 1.2947793494782057E-7
4.979350849728944E-11 0.9999999713892141
2.2007278921077675E-9 2.636026434184707E-8
4.918678935109998E-5 1.913847496070965E-6
0.9999450346247731 3.864738379680408E-6
1.6744305605390845E-10 2.1354083314551216E-13
8.95145501042084E-12 0.9999999998233919
1.2865760151187227E-11 0.999999999985571
4.840449050832555E-14 1.5148574350028663E-12
2.30710409018923E-8 0.9999999769289538
2.2525514932636024E-16 5.192512525354093E-15
0.011917696577342915 2.7812675849208594E-4
0.9878031488420475 1.0278221175978467E-6
0.999999806511151 1.3284825224816092E-7
6.054539370857234E-8 9.520297573731164E-11
0.002749329264201537 0.002529010644155105
4.78432944304802E-4 0.9942432271473386
1.6089160495763697E-4 0.9964075333714736
2.728914595752327E-6 0.0034288461089729437
…
Purity accuracy: 0.4925
NMI score: 0.17969378355038212
```

Listing 3: Presents an excerpt of results of the proposed models

As presented in Listing 3, the proposed models are providing results. The usage of Word2Vec model has resulted in more fine grained and accurate vectors that are used in making well informed decisions in text mining like document clustering. The models are employed for document classification. Topic coherence and latent feature extraction were improved in order to have better performance. The models proved that document – topic assignments are improved in terms of performance when compared with state of the art. Section 6 presents evaluation of results.

VI. EVALUATION OF RESULTS

Results are evaluated with F1 measure which is widely used to know the performance of text mining algorithms. TMN dataset and TMNtitle dataset are used with different number of topics like 7, 20, 40 and 80 for empirical study. However, the value of λ is set to 0.6 for empirical study. The results are as follows.

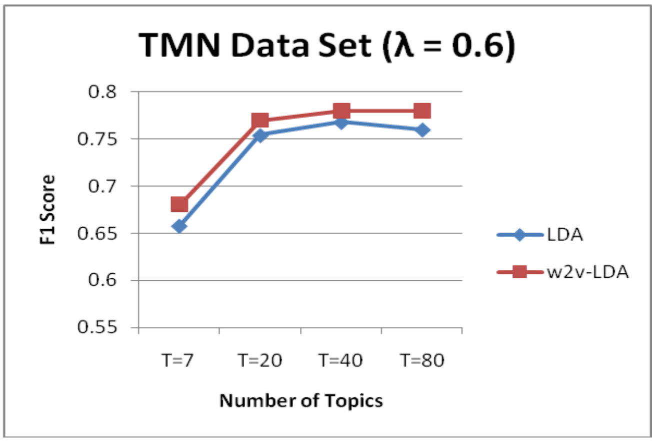

Figure 3: Number of topics vs. F1 score

Table 2: Performance comparison with TMN Dataset (λ = 0.6)

Method | No. of Titles	F1 Score Against Different No. of Titles			
	T=7	T=20	T=40	T=80
LDA	0.658	0.754	0.768	0.76
w2v-LDA	0.68	0.77	0.78	0.78

As presented in Figure 3, the number of topics is shown in horizontal axis and the vertical axis provides F1 score values against number of topics. The experimental results revealed that the number of topics has its influence on the F1 score. Another important observation found is that the proposed method has shown improved performance over the baseline LDA method.

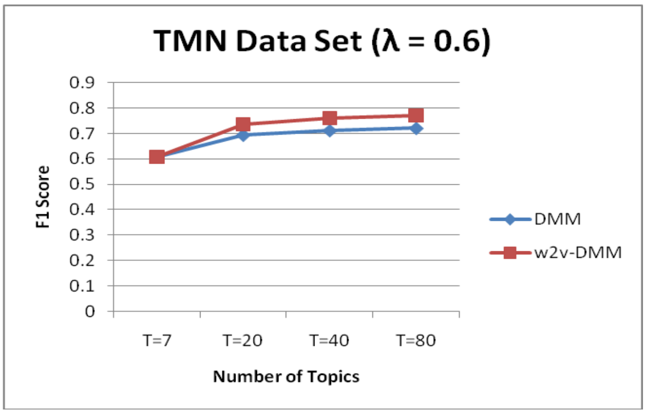

Figure 4: Number of topics vs. F1 score for baseline and proposed DMM

Retrieval Number: B2503129219/2019©BEIESP
DOI: 10.35940/ijeat.B2503.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 3: Performance comparison with TMN Dataset (λ = 0.6)

Method | No. of Titles	F1 Score Against Different No. of Titles			
	T=7	T=20	T=40	T=80
DMM	0.607	0.694	0.712	0.721
w2v-DMM	0.607	0.736	0.76	0.771

As presented in Figure 4, the number of topics is shown in horizontal axis and the vertical axis provides F1 score values against number of topics. The experimental results revealed that the number of topics has its influence on the F1 score. Another important observation found is that the proposed method has shown improved performance over the baseline DMM method.

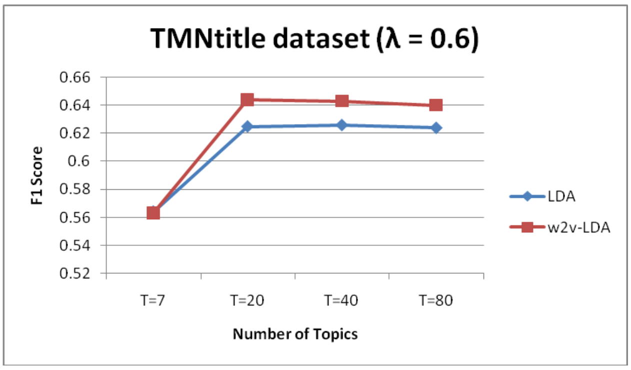

Figure 5: Number of topics vs. F1 score for baseline and proposed LDA (TMNtitle dataset is used)

Table 4: Performance comparison with TMNtitle Dataset (λ = 0.6)

Method | No. of Titles	F1 Score Against Different No. of Titles			
	T=7	T=20	T=40	T=80
LDA	0.564	0.625	0.626	0.624
w2v-LDA	0.563	0.644	0.643	0.64

As presented in Figure 5, the number of topics is shown in horizontal axis and the vertical axis provides F1 score values against number of topics. The experimental results revealed that the number of topics has its influence on the F1 score. Another important observation found is that the proposed method has shown improved performance over the baseline LDA method.

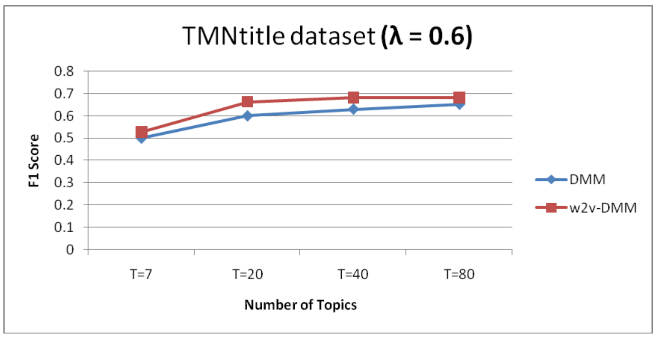

Figure 6: Number of topics vs. F1 score for baseline and proposed DMM (TMN title dataset is used).

Table 5: Performance comparison with TMNtitle Dataset (λ = 0.6)

Method | No. of Titles	F1 Score Against Different No. of Titles			
	T=7	T=20	T=40	T=80
DMM	0.5	0.6	0.63	0.652
w2v-DMM	0.528	0.663	0.682	0.681

As presented in Figure 6, the number of topics is shown in horizontal axis and the vertical axis provides F1 score values against number of topics. The experimental results revealed that the number of topics has its influence on the F1 score. Another important observation found is that the proposed method has shown improved performance over the baseline DMM method.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed novel topic models that are directly or indirectly derived from LDA generative process models. The new models are probabilistic and used to discover latent topics in document collections. The latent feature model is made up of LF-LDA and LF-DMM. Besides the models are integrated with the Word2Vec for having vectors with high quality suitable for discovering latent topics. The proposed model is capable of producing enhanced performance in text mining algorithms like document clustering. Different real world datasets are used for empirical study. Topic to word mapping is improved with the proposed model and the latent features are exploited in order to achieve better performance. A prototype application is built to evaluate the proposed model and record performance indicators. The empirical results revealed that the proposed model is performs better than the state of the art. As of now, the proposed system is slow when a large corpus is used. Therefore, the system can be extended further to incorporate more efficient solution and evaluated with large corpora. Another direction for future work is to apply the system for medical document and provide privacy preserving text mining.

REFERENCES

1. Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers and Padhraic Smyth. (2003). The Author-Topic Model for Authors and Documents. IEEE, p1-8.
2. Dou Shen, Jian-Tao Sun, Qiang Yang and Zheng Chen. (2006). Latent Friend Mining from Blog Data. IEEE, p1-10.
3. Michal Rosen-Zvi, ChaitanyaChemudugunta, Thomas Griffiths and Padhraic Smyth and Mark Steyvers. (2010). Learning Author-Topic Models from Text Corpora. ACM. 28 (1), p1-38.
4. Yan Liu, Alexandru Niculescu-Mizil and Wojciech Gryc. (2009). Topic-Link LDA: Joint Models of Topic and Author Community. Machine Learning, p1-8.
5. Melnykov and Maitra. (2010). Model-based clustering. IEEE, p1-15.
6. Nazneen Fatema, Rajani Kate McArdle and Jason Baldridge. (2014). Extracting Topics Based on Authors, Recipients and Content in Microblogs. ACM, p1-4.
7. Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc.
8. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

9. David M. Blei. 2012. Probabilistic Topic Models. Communications of the ACM, 55(4):77–84.
10. John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39(3):510–526.
11. Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. 2008. Modeling Hidden Topics on Document Manifold. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, pages 911–920.
12. Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A Novel Neural Topic Model and Its Supervised Extension. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pages 2210–2216.
13. Ana Cardoso-Cachopo. 2007. Improving Methods for Single-label Text Categorization. PhD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa. Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Advances in Neural Information Processing Systems 22, pages 288–296.
14. Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In Proceedings of the 25th International Conference on Machine Learning, pages 160–167.
15. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6):391– 407.
16. Jacob Eisenstein, Amr Ahmed, and Eric Xing. 2011. Sparse Additive Generative Models of Text. In Proceedings of the 28th International Conference on Machine Learning, pages 1041–1048.
17. Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In Proceedings of the 28th International Conference on Machine Learning, pages 513–520.
18. Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1):5228–5235.
19. Swapnil Hingmire, Sandeep Chougule, Girish K. Palshikar, and Sutanu Chakraborti. 2013. Document Classification by Topic Labeling. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pages 877–880.
20. Liangjie Hong and Brian D. Davison. 2010. Empirical Study of Topic Modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics, pages 80–88.

AUTHORS PROFILE

Dr. Thayyaba Khatoon Mohammed, has received B.Tech in CSE from KU in 2005, M.Tech and Ph.D in CSE from JNTU Hyderabad. She has published 15 papers in international Scopus indexed journals and 4papers in conferences at international level. She have life time membership in ISTE,CSI,IAENG.

M Gayatri has received the BTech in CSIT from JNTUH in 2005, MTech degree in Computer Science Engineering from JNVZ, JNTUK University, Vizianagaram, India, in 2011. She is pursuing PhD in JNTUK,AP. Her Research interests include Data Mining, Machine Learning, Web Mining

Sandeep Malle received the M.Tech degree in Computer Science Engineering and Information Security from National Institute of Technology, Calicut, Kerela, India, in 2010. He is currently pursuing the Ph.D. degree in data mining, Osmania University, Hyderabad, India. His research interests include Data Mining, Deep Learning, such as clustering and semi supervised classification and optimization techniques.

Dr. VSK Reddy, has received B.Tech in electronics and communication engineering from S.V university, Tirupati, India, M.Tech ECE from JNT University, Hyderabad, India. PhD in ECE from IIT, Kharagpur, India.He is Fellow of IETE & life time member of IEEE ISTE, CSI.