# Deep LDA : A new way to topic model

**Article** · June 2019

DOI: 10.1080/02522667.2019.1616911



**4 authors**, including:

Muzafar Bhat
Islamic University of Science and Technology
**5** PUBLICATIONS **0** CITATIONS

SEE PROFILE

Tanveer Ahmad Tarray
Islamic University of Science and Technology
**61** PUBLICATIONS **209** CITATIONS

SEE PROFILE

Journal of Information and Optimization Sciences

ISSN: 0252-2667 (Print) 2169-0103 (Online) Journal homepage: https://www.tandfonline.com/loi/tios20

# Deep LDA : A new way to topic model

Muzafar Rasool Bhat, Majid A Kundroo, Tanveer A Tarray & Basant Agarwal

**To cite this article:** Muzafar Rasool Bhat, Majid A Kundroo, Tanveer A Tarray & Basant Agarwal (2019): Deep LDA : A new way to topic model, Journal of Information and Optimization Sciences, DOI: 10.1080/02522667.2019.1616911

**To link to this article:** https://doi.org/10.1080/02522667.2019.1616911

Published online: 30 Jun 2019.

*Journal of Information & Optimization Sciences*
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
DOI : 10.1080/02522667.2019.1616911

Taylor & Francis
Taylor & Francis Group

# Deep LDA : A new way to topic model

Muzafar Rasool Bhat [†]
Majid A Kundroo
Tanveer A Tarray
*Department of Computer Sciences*
*Islamic University of Science & Technology*
*Awantipora 192122*
*Jammu & Kashmir*
*India*

Basant Agarwal *
*Department of Computer Science and Engineering*
*Indian Institute of Information Technology Kota (IIIT Kota)*
*MNIT Campus*
*Jaipur 302017*
*Rajasthan*
*India*

**Abstract**

Probabilistic topic models like Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA) and Biterm Topic Model (BTM) have been successfully implemented and used in many areas like movie reviews, recommender systems and text summarization etc. These models however become computationally heavy if tested on humongous corpus. Keeping in view the vide acceptability of Deep Neural network based machine learning, this research proposes two deep neural network variants (2NN DeepLDA and 3NN DeepLDA) of existing topic modeling technique Latent Dirichlet Allocation (LDA) with specific aim to handle large corpuses with less computational efforts. Two proposed models (2NN DeepLDA and 3NN DeepLDA) are used to mimic the statistical process of Latent Dirichlet Allocation. Reuters-21578 dataset has been used in the study. Results computed from LDA are compared with the proposed models (2NNDeepLDA and 3NNDeepLDA) using Support Vector Machine (SVM) classifier. Proposed models have shown significant accuracy besides computational effectiveness in comparison to traditional LDA.

***Keywords:*** *Deep Learning, LDA, SVM Classifier, Topic Modelling, Keras, Tensorflow.*

[†]*E-mail:* muzafarrasool@gmail.com
**E-mail:* basant@skit.ac.in (Corresponding Author)

## 1. Introduction

Topic modeling is essentially defined as a statistical way of text mining to identify a latent (hidden) pattern in a corpus of data. In simpler terms, it is a statistical technique which groups 'few' words across the corpus into topics. This method is alternatively defined as a scientific way to trace clusters of words (called topics) in large collections of text. Topic in statistical parlance is a multinomial distribution of words that co-occur in statistically significant ways. Topics are learned from collections of documents by iteratively running a topic model over a given collection. Few sampling techniques like Gibbs Sampling and Collapsed Gibbs sampling are generally used to understand the random process behind document generation [1-3]. Topic modeling has application in many natural language applications such as sentiment analysis, question answering, summarization etc. [10, 11]

This process of topic modeling, as illustrated Figure 1, can be precisely explained by an imaginatively working through a text article with different set of color highlighters. As article is read through, if supposedly we use different colors to highlight keywords that refer to different themes as we come across them while reading the article. After going through each keyword (words other than stop words) of the article and coloring each keyword with a highlighter, if we collect keywords colored with a common color, then each group of keywords colored with a common color

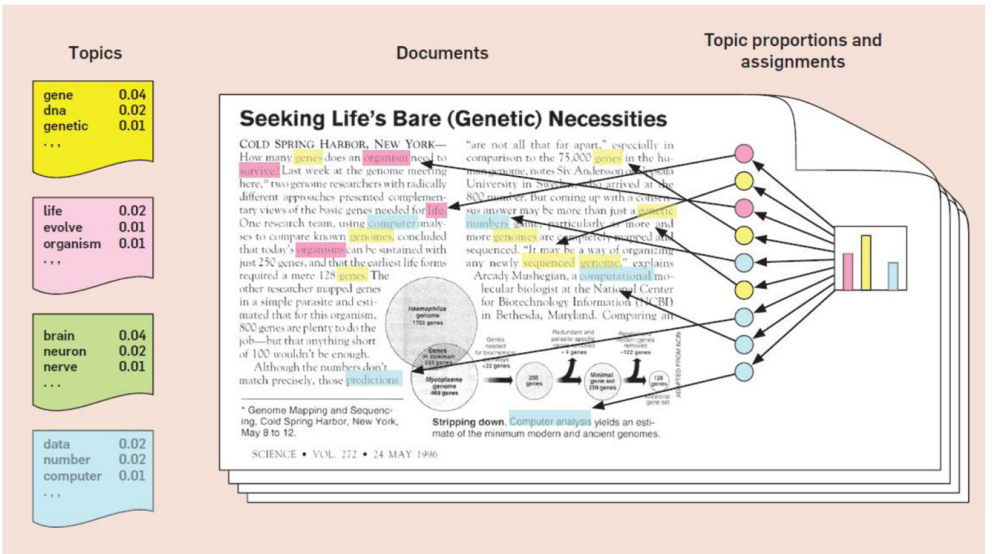


**Figure 1**

**Illustration of Probabilistic Topic Models. (FromBlei, D. 2012)**

can be presumed to represent one theme (topic). Total number of different colors gives us total number of topics discussed in the article.

A good topic model is believed to identify a meaningful cluster of words. Cluster of words that can be unambiguously labeled as a meaningful representation of some common topic. Common topic modeling techniques include PLSI, LDA and Biterm Topic Model (BTM).

This research explores possibility of modelling the statistical process of LDA using Deep Neural Network to mimic its working [4-6]. The proposed models are believed to reduce the computational cost required in LDA to extract topics (themes) of a huge corpus. Remaining sections of this paper are organized in a following sequence. Section 2 discusses LDA and its generative process. Section 3 discusses experimental setup of our study with detailed specification of our proposed models. Section 4 visualizes the results obtained from experimentation of proposed models (2NN DeepLDA and 3NN DeepLDA) on Reuters-21578 [7]. Few interim results are also listed in this section as well. Future direction of the proposed research is mentioned in the last section besides conclusion of the proposed study.

## 2. Latent Dirichlet Allocation (LDA) and its Generative Process.

Latent Dirichlet Allocation (LDA) [1], proposed by Bleiet.al., is a generative probabilistic model for collection of discrete data such as text corpora, genome sequences, collection of images etc. Though LDA can be used to model discrete data of any domain, however to explain its generative process, we will specifically focus on text collections. Understanding generative process of LDA vis-à-vis text collections can however be generalized to any other domain. In text collections, LDA explores possibility of representing documents as random mixtures over latent topics, where each topic is a multinomial distribution over words. In implementation of LDA, we need to have a collection $D$ of documents $d_i$, where each $d_i = (w_{i1,} w_{i2,} w_{i3}\ldots.w_{iN})$.

Latent Dirichlet Allocation has a pure probabilistic perspective about document generation process. Two multinomial distributions generally represented as $\phi_t$ and $\theta_d$ are pivotal in LDA. Fundamental assumption about LDA is that each word $w_i$ extracted from collection of documents is probabilistically associated with latent topic(s) $t_k$. This probability is approximated using a multinomial distribution $\phi_t$ over vocabulary V of the corpus such that

$$P(w_i \mid z_k = t) = \phi_{t(zk,wi)}, \tag{2.1}$$

$\phi_t$ is computed from a Dirichlet distribution with $\beta$ as a prior. It is also assumed that each document of a collection is probabilistically associated with a latent topic $Z_i$. This probability too is approximated using a multinomial distribution $\theta_d$ over d documents such that

$$P(Z_i = t \mid d) = \theta_{(d,z)}, \tag{2.2}$$

$\theta_{(d,\,z)}$ is computed from a Dirichlet distribution with $\alpha$ as a prior. To approximate latent topics from text collection, LDA precisely focuses on following estimates (computations/ approximations)

1. To find an estimate of $\phi_{t,}$ LDA focuses on multinomial distribution for terms with respect to each topic *t*. i.e. list of terms (words) in decreasing order of probability for each topic t.
2. To find an estimate of $\theta_{d,}$ LDA traces a multinomial distribution for each document with respect to topic *t*. i.e. list of documents in decreasing order of probability for a topic *t*.

LDA being generative model in nature, its generative process assumes that each document is generated by following two steps in sequence. Initially for topic (theme) composition of a document, instance from a multinomial distribution over *K* topics is drawn with parameter $\theta_d$, generated from Dirichlet distribution with $\alpha$ as a prior. Words in a document are subsequently selected by drawing a topic $Z_i = t$ from this distribution. To choose a word to suit a topic, words are drawn from a multinomial distribution with parameter $\phi_i$, generated from a Dirichlet distribution with $\beta$ as prior.

This process as explained by [8] can be probabilistically represented using a joint probability distribution over random variables (*D*, *Z*, $\phi$, $\theta$). That is

$$P(D,Z,\phi,\theta)\alpha\left(\prod_{t}^{k} P\phi_i \mid \beta\right)\left(\prod_{d}^{D} \theta_d \mid \alpha\right)\left(\prod_{ti}^{N} \phi_{zi,wi}\theta_{di,zi}\right) \tag{2.3}$$

Where $\phi_{zi,\,wi}$ is the $w_i^{th}$ element in $\phi$, $\phi_{d,\,wi}$ is $z_d$th element in .From Figure 2, it is apparent that only observed variables in LDA are the words within the document. The other variables which include latent topic assignments Z, document distribution over topics and topic distribution over words $\phi$ are all unobserved. The priors or hyper parameters to Dirichlet distributions associating documents to topics and topics to words are taken as input from the user. Better choice of priors can lead to better results. Estimation of multinomial distributions $\phi$ and which are empirically generated

from Dirichlet distributions with priors $\beta$ and $\alpha$ respectively requires computation of latent topic assignments $Z$, $P(z \mid D, \alpha, \beta)$ which can't be computed directly from the joint distribution given the fact that posterior distribution is intractable.

To approximate posterior distribution, few inference algorithms have been suggested in literature. These algorithms include Variational Approximation, Expectation Propagation, Laplace approximation and Gibbs sampling. These algorithms can differ in speed and accuracy

The generative process, as explained above, is therefore summarized in following few steps :

1. Choose $\theta_l$ ~Dir ($\alpha$) where $l \in \{1, 2, ..., L\}$ andDir ($\alpha$) is a Dirichlet distribution with $\alpha$ as a prior.
2. Choose $\phi_l$ ~ Dir ($\beta$) where $p \in \{1, 2, ..., p\}$ and Dir ($\beta$) is a Dirichlet distribution with $\beta$ as a prior, p represents number of topics used to topic model corpus D.
3. For each word position ($m$, $n$), where $n \in \{1, 2, ..., N\}$ and $m \in \{1, 2, ..., L\}$
   a. Choose a topic $z_{m,n}$ ~ *multinomial* ($\theta$)
   b. Choose a word $w_{m,n}$ ~ *multinomial* ($\phi_{m,n}$)

Parameters to topic model a given corpus essentially depend on corpus. A proper selection of optimal parameters to topic model using Latent Dirichlet Allocation needs to be thoroughly experimented.

This generative process can be explained precisely using following plate notation:

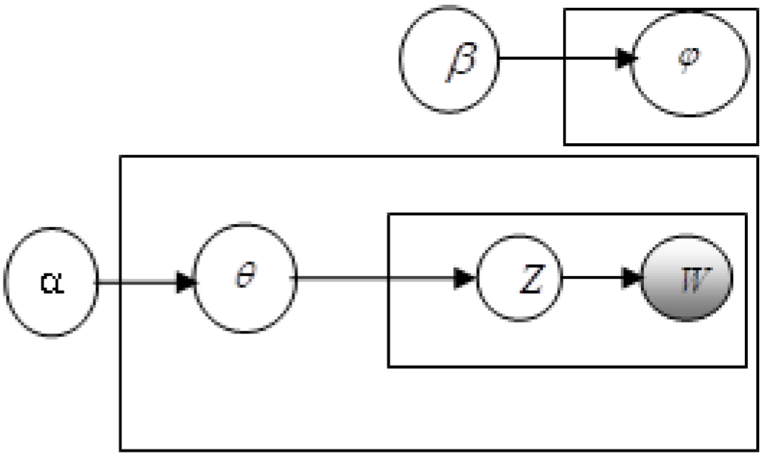


**Figure 2**

**Plate Notation of LDA**

## 3. Proposed Approach

In this paper we attempt to mimic generative process of LDA model, illustrated in Figure 2, using deep neural network variant with an objective to reduce the computational time required in LDA. Deep neural networks approach was adapted for its easy implementation using current state of art besides its possibility of generalization as well as its ability to handle large collection of data. The proposed models (2NN DeepLDA and 3NN DeepLDA) was implemented in Python. Few standard libraries like genism, NLTK and Keras etc. were also used during experimentation. Following stepwise implementation was followed during the experimentation.

i. To learn topics from the given dataset, gensim's multicore LDA model was used to perform topic extraction in addition to computation of Topic Document Distribution as well $\theta$ as Topic Word Distribution $\phi$.

ii. As a standard practice, optimal number of topics to topic model the given dataset was selected by iteratively running the model for number of topics ranging from 10 to 50.Trained model with fine-tuned parameters (hyper-parameters $\alpha$ and $\beta$) and suitable number of topics $k$ is saved for further comparisons.

iii. Accuracy of the Saved model trained on the given corpus "Reuters-21758" was computed using Support Vector Machine (SVM) Classifier.

Two sequential models (2NN DeepLDA and 3NN DeepLDA), as proposed in this paper, are built using Keras, an open source neural network library, with two and three hidden layers. These proposed models are aimed to mimic working of traditional Topic Modeling technique LDA. The input layer size of the proposed deep models (2NN and 3NN DeepLDA) is equal to the size of the vocabulary extracted from the given Dataset remained after prerequisite pre-processing. Vocabulary extracted from the data set was stemmed using NLTK's WorldNet Stemmer.

Previously trained LDA Model is used as a supervisor for the training of neural networks.

Input documents were first converted into bag of words(BOW) representation and then BOW representation of documents is fed into the neural network. Output of LDA is used as a label for supervised training of neural network so that it can learn Topic Document Distribution as well as Topic Word Distribution.

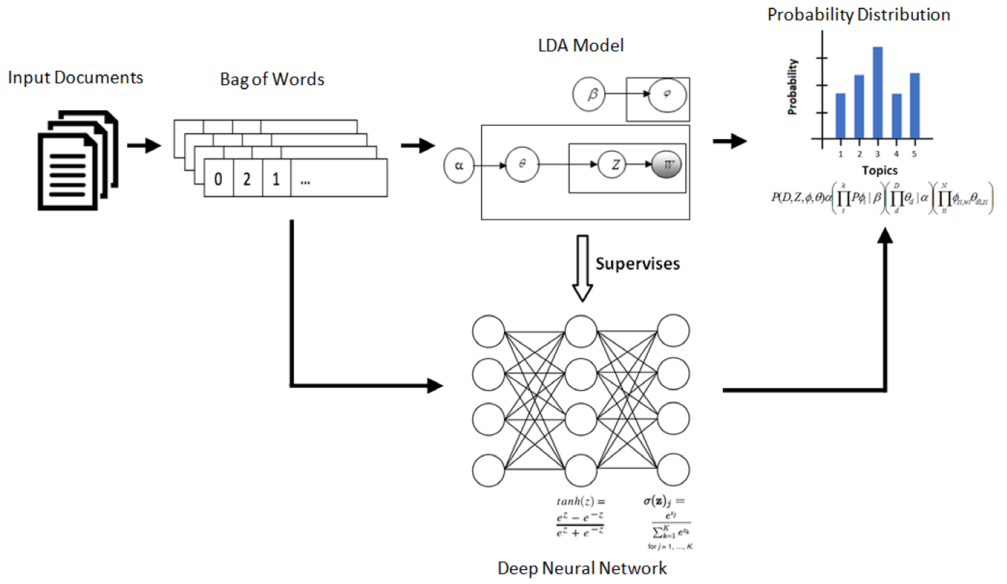


**Figure 3**

**Flow Diagram of proposed approach**

We have used *tanh* activation function in our neural networks for hidden layers and *softmax* activation function is used at the output layer of neural networks. We have used stochastic gradient descent optimizer and categorical cross entropy as loss function.

## 4. Experimental Setup

### 4.1 *Dataset used and necessary Preprocessing*

Reuters-21758 [7] dataset, a benchmark dataset for contemporary research in various fields like document classification, clustering and topic modeling, is used in our study. This dataset was imported using NLTK library in python. Necessary pre-processing which include stop word removal, removal of punctuation marks and removal of insignificant symbols like numbers and digits was done prior its use in the experimentation. For stemming of the extracted vocabulary, NLTK's WordNet stemmer was used. As a prerequisite, the whole collection of 9,160 documents was divided into two exclusive sets of training and testing with respective sizes of 6577 and 2583 documents.

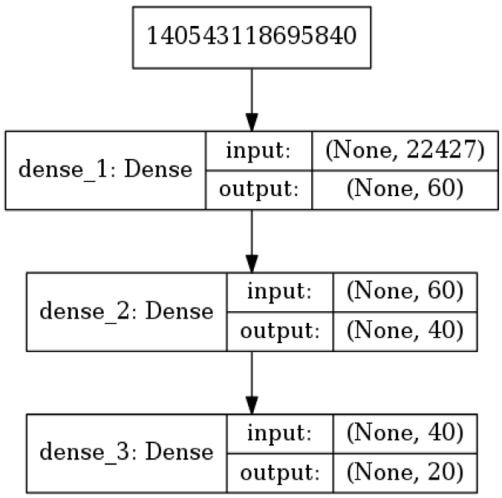


**Figure 4**

**Model Summary of the proposed 3NN DeepLDA**

4.2 *Implementation in Python*

i. Gensim's multicore LDA is initially used totrain the LDA model on selected training dataset for 10-50 number of topics for100iterations. Model is saved for subsequent use.

ii. Support Vector Classifier (SVC) is used to calculate the accuracy. Accuracy of LDA is also recorded in this step.

iii. Keras sequential model with two hidden layers and three hidden, as mentioned in Figure 3 are created with following specifications.

   a. Input layer with number of nodes equal to the size of vocabulary (length of dictionary) extracted from Reuters-21758.

   b. Activation function, loss function, and optimizer are respectively chosen as 'tanh', 'categorical_crossentropy' and 'sgd'.

   c. Both models (2NN DeepLDA and 3NN DeepLDA) are trained in 100 epochs. Models are saved the model for subsequent use.

   d. Accuracy of the models is recorded using Support Vector Classifier (SVC).

## 5. Results

Figure 5 shows the plot of accuracy vs number of topics. For our experiment we have chosen variable number of topics ranging from 10 to 40 with an interval of 5 (i.e., 10, 15,20,…, 40 number of topics) and plotted

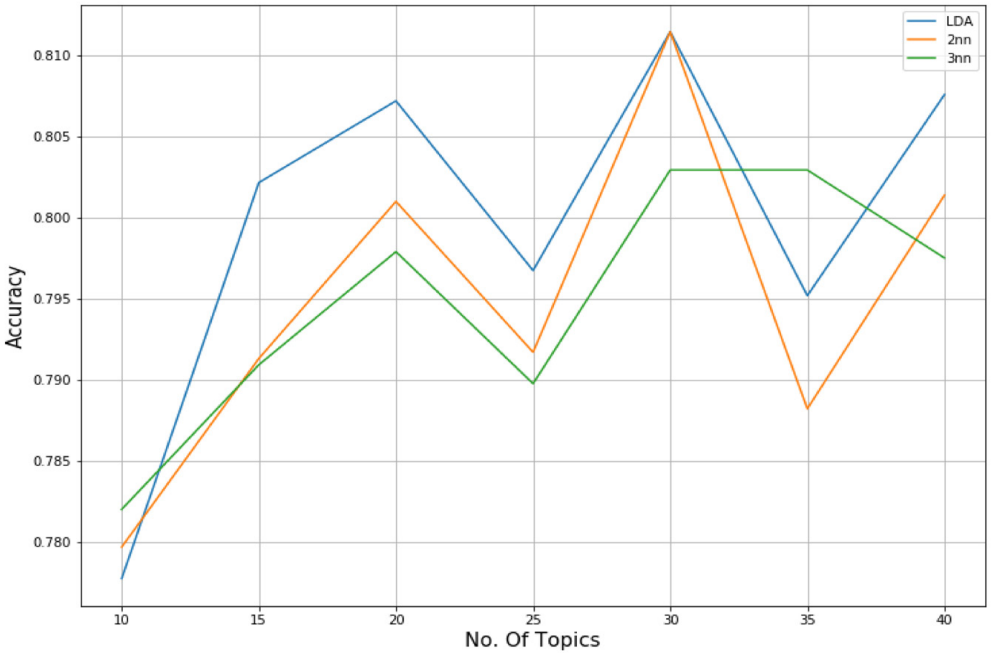


**Figure 5**

**Accuracy of LDA, 2NN DeepLDA and 3NN DeepLDAusing Varying number of Topics**

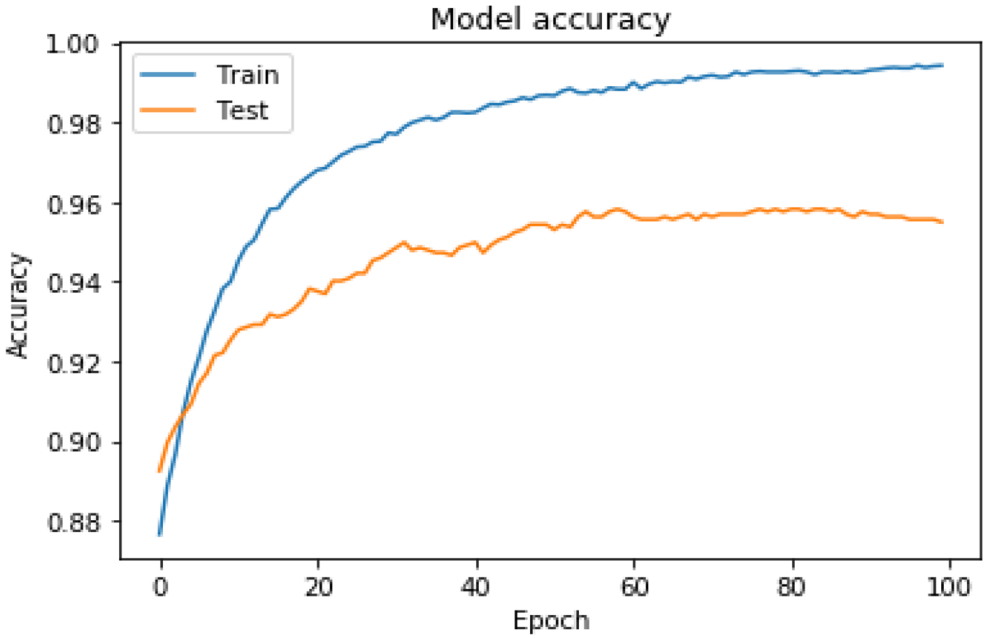


**Figure 6**

**Accuracy of Proposed Model (3NN DeepLDA)**

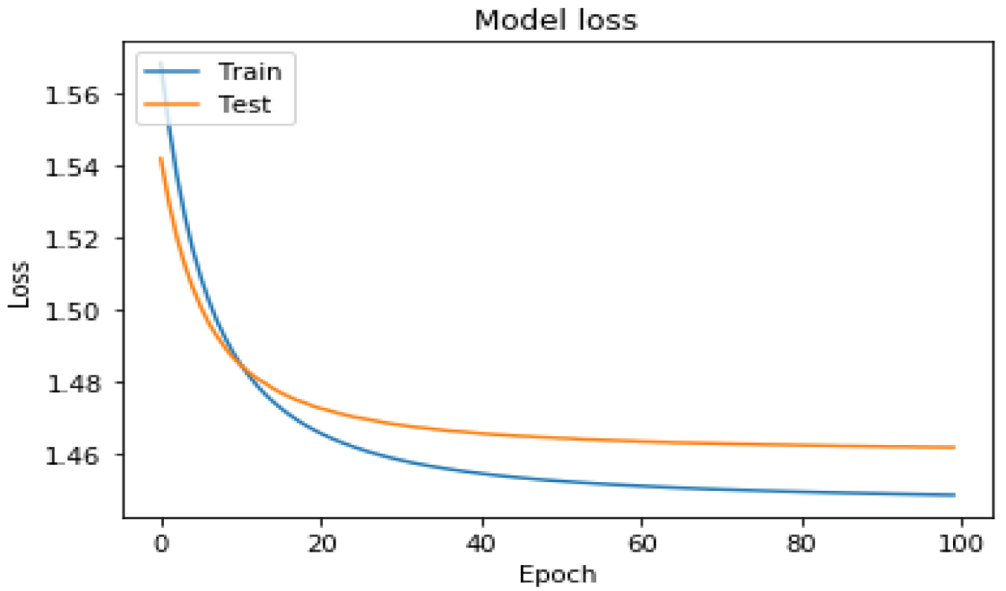


**Figure 7**

**Loss of Proposed Model (3NN DeepLDAModel)**

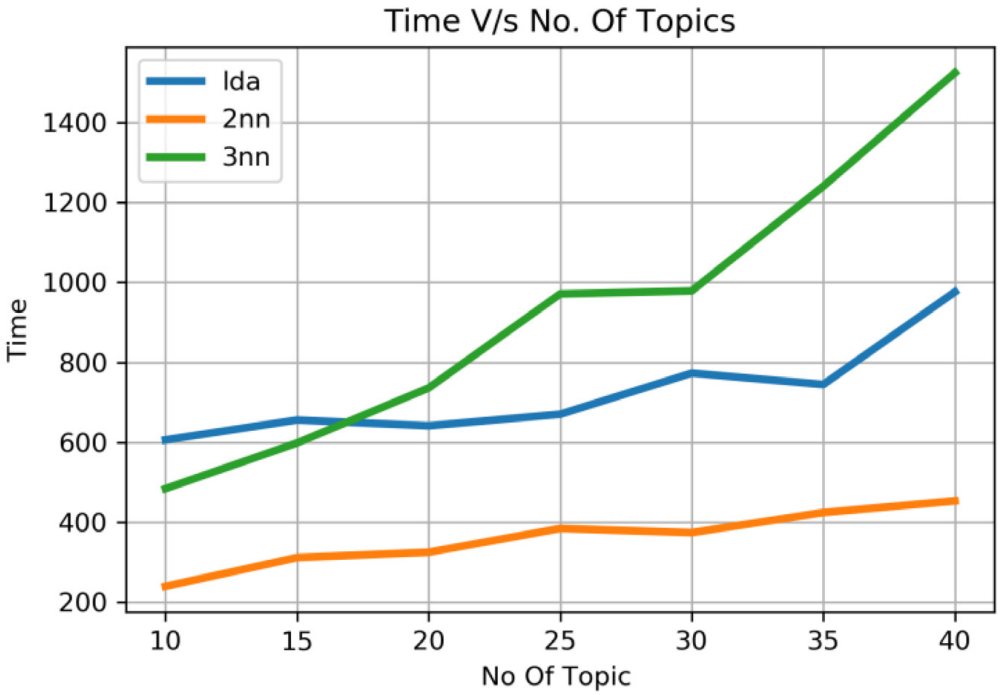


**Figure 8**

**Comparison of existing LDA with two Proposed Models (2NN DeepLDA and 3NN DeepLDA)**

their accuracy as obtained from LDA,2NN,3NN. The graph represents the accuracies of all three algorithms with varying number of topics chosen.

The graph (shown in fig 5) signifies that among 2NN and 3NN, 2NN model showed almost same accuracy as that of LDA. This proves that 2NN model can be used as an alternative to traditional LDA. Figure 6 and 7 shows the accuracy and loss for the proposed model.

Figure 8 is the plot of number of topics vs time computed by LDA, 2NN, 3NN. The figure demonstrates the time taken by LDA,2NN, 3NN for variable number of topics ranging from 10 to 40 with an interval of 5 (i.e., 10, 15, 20,…, 40 number of topics). The plot signifies that 2NN outperforms 3NN and LDA with respect to time taken for computing topics and thus proves our claim that 2NN takes lesser time than LDA and 3NN and can be used for topic modelling of large corpuses.

## 6. Conclusion and Future Scope

In this paper, we investigate the possibility of mimicking the well-known Topic modelling technique Latent Dirichlet Allocation with a Deep neural networking variant DeepLDA using two possible arrangements. Two models 2NN DeepLDA and 3NN DeepLDA that have been proposed in this study have shown significant accuracy. Initially topics were learned from the dataset Reuters-21578 using Latent Dirchlet allocation and were subsequently used as labels to the documents within the corpus. Proposed models have outperformed the existing topic model and are capable of handling humungous data besides showing significant improvement in slashing requisite computational effort as its evident from figure 9.Proposed 2NN DeepLDA showed accuracy almost same as LDA, however 3NN DeepLDA is outperforming both (Figure 5).The proposed models (2NN DeepLDA has reduced the requisite computational time (Figure 9).

Though this study specifically focused on LDA, however it can provide basis for developing deep neural networking based variant of Biterm Topic Model (BTM), existing topic modeling technique for short texts messages.

## 7. Acknowledgment

This research is funded by Islamic University of Science and Technology, Awantipora, J&K, India.

**References**

[1] Blei, David M, Ng Andrew "Latent Dirchlet Allocation", *Journal of machine Learning research*, 993-1022: (2003).

[2] Dahl, B David, "Model-based clustering for expression data via a Dirichlet process mixture model", Bayesian inference for gene expression and proteomics, 201-218, (2006).

[3] Ishwaran, Hemant, and Lancelot F. James. "Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information."*Journal of Computational and Graphical Statistics* :(2000).

[4] Gauri Jain, Manisha Sharma, and Basant Agarwal. "Spam detection on social media using semantic convolutional neural network." *International Journal of Knowledge Discovery in Bioinformatics* (IJKDB) 8.1,12-26: (2018).

[5] Sandeep Bhargava, and Seema Choudhary. "Behavioral Analysis of Depressed Sentimental Over Twitter: Based on Supervised Machine Learning Approach.", 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), (2018).

[6] Basant Agarwal, Namita Mittal. "Semantic feature clustering for sentiment analysis of English reviews." *IETE Journal of Research* 60.6 ,414-422: (2014).

[7] Lewis, David. "Reuters-21578." *Test Collections* 1 : (1987).

[8] Andrzejewski, David Michael, Mark Craven, and Xiaojin Zhu. Incorporating domain knowledge in latent topic models. Diss. University of Wisconsin–Madison: (2010).

[9] Shafi, Kh Muhammad, Muzafar Rasool Bhat, and Tariq Ahmad Lone. "Sentiment analysis of print media coverage using deep neural networking." *Journal of Statistics and Management Systems* 21.4, 519-527 : (2018).

[10] Shrawan Ram, Shloak Gupta, Basant Agarwal, "Devanagri Character Recognition Model Using Deep Convolution Neural Network", *In Journal of Statistics and Management Systems*, 21 (4), pp:593–599, (2018).

[11] Shikhar Seth, Basant Agarwal, "A hybrid deep learning model for detecting diabetic retinopathy", *In Journal of Statistics and Management Systems*, Taylor Francis, 21 (4), pages: 569–574(2018).

*Received December, 2018*

*Revised March, 2019*